UNIVERSIDADE ESTADUAL DE MARINGÁ
CENTRO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE MATEMÁTICA
MESTRADO PROFISSIONAL EM MATEMÁTICA EM REDE NACIONAL

Juliana Ferreira da Silva

ANÁLISE DE CORRELAÇÃO E REGRESSÃO LINEAR E SUA IMPORTÂNCIA PARA O ENSINO MÉDIO

Maringá, PR

2024

UNIVERSIDADE ESTADUAL DE MARINGÁ
CENTRO DE CIÊNCIAS EXATAS
DEPARTAMENTO DE MATEMÁTICA
MESTRADO PROFISSIONAL EM MATEMÁTICA EM REDE NACIONAL

ANÁLISE DE CORRELAÇÃO E REGRESSÃO LINEAR E SUA IMPORTÂNCIA PARA O ENSINO MÉDIO

Juliana Ferreira da Silva

Dissertação de Mestrado submetida ao Programa de Mestrado Profissional em Matemática em Rede Nacional (PROFMAT), da Universidade Estadual de Maringá para a obtenção do título de Mestre em Matemática.
Orientador: Rodrigo Martins

Maringá/PR

2024

Dados Internacionais de Catalogação na Publicação (CIP)
(Biblioteca Setorial BSE-DMA-UEM, Maringá, PR, Brasil)

Silva, Juliana Ferreira da
S586a Análise de correlação e regressão linear e sua importância para o Ensino Médio / Juliana Ferreira da Silva. -- Maringá, 2024.
[62] f. : il., color.

Orientador: Prof. Dr. Rodrigo Martins.
Dissertação (mestrado) - Universidade Estadual de Maringá, Centro de Ciências Exatas, Departamento de Matemática, 2024.

1. Correlação linear. 2. Regressão linear. 3. Análise de microdados. 4. ENEM. I. Martins, Rodrigo, orient. II. Universidade Estadual de Maringá. Centro de Ciências Exatas. Programa de Mestrado Profissional em Matemática em Rede Nacional - PROFMAT. III. Título.

CDD 22.ed. 519.537

Edilson Damasio CRB9-1.123

JULIANA FERREIRA DA SILVA

ANÁLISE DE CORRELAÇÃO E REGRESSÃO LINEAR E SUA IMPORTÂNCIA PARA O ENSINO MÉDIO

Dissertação apresentada ao Programa de Mestrado Profissional em Matemática em Rede Nacional do Departamento de Matemática, Centro de Ciências Exatas da Universidade Estadual de Maringá, como parte dos requisitos necessários para a obtenção do título de Mestre em Matemática tendo a Comissão Julgadora composta pelos membros:

COMISSÃO JULGADORA:

Prof. Dr. Rodrigo Martins
UEM - Universidade Estadual de Maringá (Orientador)

Prof. Dr. Vinicius Araujo Peralta
UTFPR - Universidade Tecnológica Federal do Paraná / Cornelio Procópio

Profa. Dra. Patricia Hilário Tacuri Córdova
UEM - Universidade Estadual de Maringá

Aprovada em: 22 de março de 2024
Local de defesa: Auditório do Departamento de Matemática – Bloco F67 (sala 217)

À Laura de Paula Alves *(in memoriam)*, minha querida avó.

Agradecimentos

Agradeço a Deus pela minha vida e por me dar a capacidade e oportunidade de realizar um curso de mestrado logo após o término da graduação.

Também agradeço imensamente ao meu Orientador, Rodrigo Martins, pela enorme paciência, incentivos, dedicação e orientação neste árduo trabalho.

Aos meus pais, Antonio Luiz e Nilva, pelo incentivo, apoio e ajuda nesse período de aulas online, presencial, e no período de elaboração da dissertação, bem como a minha avó Laura que nos deixou em 2022.

Além disso, gostaria de expressar minha gratidão às minhas colegas Fernanda e Renata, que foram excelentes companhias de estudo e incentivo.

Aos professores que ministraram as disciplinas, pelas correções e ensinamentos que me permitiram aprimorar minha formação profissional ao longo do curso.

"Viva como se fosse morrer amanhã. Aprenda como se fosse viver para sempre."

Mahatma Gandhi (1869-1948)

Resumo

A correlação e a regressão linear são estudadas na parte da estatística chamada inferencial pois permite a análise de dados coletados e fazer predições acerca da influência de uma variável em outra. Essa correlação pode ser avaliada por meio do coeficiente de Pearson, teste de hipóteses, gráficos de resíduos, coeficiente de determinação, intervalos de precisão. Parte desses assuntos estão entre aqueles previstos a serem estudados no Ensino Médio, de acordo com a Base Nacional Comum Curricular. Uma aplicação da correlação e regressão é apresentada na análise de Microdados do Exame Nacional do Ensino Médio (ENEM) de 2022, dos estudantes que realizaram a prova em Maringá - PR, em que, com a utilização do Software estatístico R, pode-se encontrar relações interessantes e relevantes para estudo.

Palavras-chave: Correlação linear. Regressão linear. Análise microdados ENEM.

Abstract

Correlation and linear regression are studied in the part of statistics called inferential as they allow the analysis of collected data and make predictions about the influence of one variable on another. This correlation can be evaluated using the Pearson coefficient, hypothesis testing, residual graphs, coefficient of determination, precision intervals. Some of these subjects are among those planned to be studied in High School, according to the National Common Curricular Base. An application of correlation and regression is presented in the Microdata analysis of the 2022 National High School Exam (ENEM), of students who took the test in Maringá - PR, in which, with the use of R statistical software, one can find interesting and relevant relationships for study.

Keywords: Linear correlation. Linear regression. ENEM microdata analysis.

Lista de ilustrações

Sumário

INTRODUÇÃO

A estatística, ciência que estuda dados, sejam eles observações, contagens, medições ou respostas, pode ser aplicada em diversas áreas do conhecimento humano, como áreas econômicas, sociais, culturais, políticas, educacionais, de saúde, meio ambiente, e proporciona a todos as ferramentas necessárias para entender aspectos e problemas que necessitam de uma melhor compreensão, desde um nível mais abrangente a um nível mais pontual, quando por exemplo o Estado faz o censo da população para poder tomar decisões em relação a políticas públicas, ou quando avaliamos nossas chances de ganhar numa loteria.

O estudo da estatística inferencial, mais especificamente da correlação e regressão permite identificar relações entre variáveis que, a princípio, não possuem uma relação direta, como estudado em álgebra, nas funções, em que temos a associação de cada elemento de um conjunto com apenas outro elemento do outro conjunto.

A análise da correlação e regressão tratada neste trabalho tem fundamento nos dados coletados pela Comissão Federal de Comércio dos Estados Unidos sobre 29 marcas de cigarros e a quantidade de alcatrão, nicotina e CO_2 presentes em cigarros de 100 mm de comprimento. Esses dados servirão como base para aplicação dos conceitos teóricos.

Dessa forma, o trabalho está organizado da seguinte maneira: No capítulo 1 trazemos conteúdos básicos da estatística, apresentando os conceitos de gráfico de dispersão, desvio padrão, variância, covariância, escore padrão e teste de hipóteses. Usamos como texto base deste capítulo os livros [6], [7] e [9].

Após essa pequena explanação, o Capítulo 2 apresenta os dados dos cigarros, por marca, quantidade de alcatrão, nicotina e CO_2 presentes em cada um, conceitos de correlação e regressão, bem como suas representações por meio de gráficos de dispersões, cálculo do coeficiente de correlação de Pearson, e suas fórmulas devidamente explicadas, exposição esta que não se encontra nos clássicos livros de estatística, testes de hipóteses, reta de regressão, gráfico de resíduos, coeficiente de determinação, erro padrão e intervalo de previsão.

No Capítulo 3 trataremos da importância do estudo da correlação e regressão no Ensino Médio, de acordo com disposições da Base Nacional Comum Curricular (BNCC),

bem como o uso de softwares computacionais gratuitos e acessíveis em que basta ter acesso a internet para usá-los, sem necessidade de fazer instalações de programas, e são excelentes ferramentas para estudo da estatística e matemática, podendo ajudar ainda mais no estudo e entendimento do assunto aqui proposto, através de comandos básicos.

Finalmente, no Capítulo 4, faremos a análise dos microdados do Exame Nacional do Ensino Médio (ENEM) aplicado no ano de 2022, em relação aos alunos da cidade de Maringá - PR, com o auxílio do software R, em que é usada uma linguagem e um ambiente de desenvolvimento integrado, livre e gratuito, bastante útil em análises estatísticas pois permite a manipulação de grandes bancos de dados, como é o caso dos microdados do ENEM.

Nessa análise, elaboraremos os gráficos de dispersão relacionando a faixa etária com nota em matemática; ano de conclusão no ensino médio e nota em matemática; grau de instrução do pai/homem responsável e nota em matemática; grau de instrução da mãe/mulher responsável e nota em matemática; e renda mensal e nota em matemática. Utilizando os conceitos trabalhados nos capítulos anteriores poderemos identificar correlações, encontrar a reta de regressão que melhor explica o modelo e determinar o coeficiente de determinação, que expressa a porcentagem de quanto uma variável é explicada pela outra.

1 Preliminares

Neste capítulo, apresentaremos conceitos básicos da estatística que proporcionarão uma conhecimento elementar de assuntos que aparecerão nos próximos capítulos desse trabalho.

Trataremos dos gráficos de dispersão, indispensáveis para o estudo da correlação e regressão linear, já que nos fornecem uma noção intuitiva a respeito da correlação das variáveis. Relembraremos conceitos das medidas de variação da estatística descritiva (desvio padrão, variância e covariância), apresentaremos o conceito de escore padrão e do teste de hipóteses, que também se mostrará uma ferramenta útil para avaliação da significação do Coeficiente de Pearson, no Capítulo 2.

1 Gráfico de Dispersão

Um diagrama de dispersão é um gráfico que relaciona pares de dados (x, y) com um eixo horizontal x e um eixo vertical y.

Os dados coletados relacionam um conjunto de valores x com seu conjunto correspondente y.

Para construir o gráfico de dispersão, faz-se um eixo horizontal onde serão atribuídos os valores da primeira variável e um eixo vertical para os valores da segunda variável, e após marca-se os pontos.

O gráfico de dispersão é muito usado para inferir sobre a existência ou não de correlação entre as variáveis, assunto a ser abordado no Capítulo 2.

A seguir um exemplo de um gráfico de dispersão:

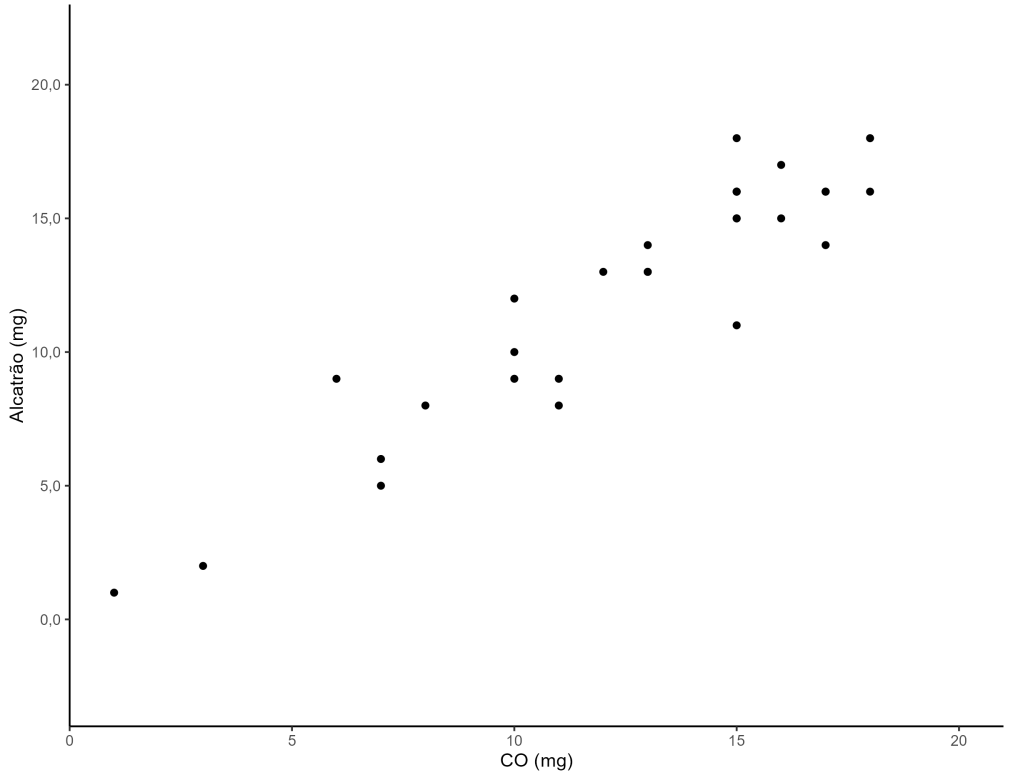


Figura 1 – Gráfico de Dispersão: Alcatrão x CO

O gráfico acima relaciona a quantidade de Alcatrão (em miligrama) encontrado em 29 marcas de cigarros de acordo com dados da Federal Trade Commission, no eixo vertical e a quantidade de Monóxido do Carbono (em miligrama) no eixo vertical, em cigarros de 100 mm (10 cm) de comprimento, que tem filtro e não são do tipo mentolado ou *light*.

2 Desvio Padrão, Variância e Covariância

O desvio padrão (s), variância (s^2) e covariância são medidas de variação da estatística descritiva, cujo objetivo é resumir ou descrever as características mais importantes de um conjunto de dados.

2.1 Desvio Padrão

Segundo [9], o desvio padrão de um conjunto de valores amostrais é uma medida da variação dos valores em torno da média. É uma espécie de desvio médio dos valores em relação a média, que é calculado através das seguintes fórmulas:

$$s = \sqrt{\frac{\sum(x-\overline{x})^2}{n-1}} \qquad \text{ou} \qquad s = \sqrt{\frac{n.\sum(x)^2 - (\sum x)^2}{n(n-1)}}.$$

em que $\overline{x}$ representa a média dos valores de x e n é o número de amostras.

Semelhante fórmula pode ser aplicada para o desvio padrão de uma população, diferindo apenas no denominador, que é o tamanho da população:

$$\sigma = \sqrt{\frac{\sum(x-\mu)^2}{N}}$$

em que μ é a média populacional e N é o número de valores de uma população finita.

A explicação para essa mudança no denominador se dá em razão da subestimação do valor obtido pelo desvio padrão amostral em relação ao desvio padrão da população. Para compensar, fazemos a divisão por um número menor, ou seja, $n-1$, aumentando seu valor geral. [1]

Para melhor compreensão do desvio padrão, vamos continuar a utilizar os dados do gráfico de dispersão acima que compara a quantidade de alcatrão e monóxido de carbono e vamos calcular o desvio padrão da variável quantidade de alcatrão, em que os dados são: 16, 16, 16, 9, 1, 8, 10, 16, 14, 13, 13, 15, 16, 9, 11, 2, 18, 15, 13, 15, 17, 9, 12, 14, 5, 6, 8, 18, 18 para a variável x, a média é $12,2$ e $n = 29$. Assim:

$$s = \sqrt{\frac{(16-12,2)^2 + (16-12,2)^2 + ... + (18-12,2)^2}{29-1}} = 4,71.$$

2.2 Variância

A variância é uma medida de variação/dispersão igual ao quadrado do desvio padrão.

$$s^2 = \frac{\sum(x-\overline{x})^2}{n-1} \qquad \text{ou} \qquad \sigma^2 = \frac{\sum(x-\mu)^2}{N}.$$

A variância possui a vantagem de ser um estimador não enviesado, ou seja, não subestima nem superestima a variância populacional, porém, a unidade do valor resultante de seu cálculo não é a mesma utilizada nos valores originais.

[1] Para melhor visualização e interpretação dessa mudança, veja: https://pt.khanacademy.org/computer-programming/challenge-unbiased-estimate-of-population-variance/1169428428. https://pt.khanacademy.org/computer-programming/will-it-converge-towards-1/1167579097.

Ainda usando os dados coletados no exemplo do gráfico de dispersão, vamos calcular a variância da quantidade de alcatrão encontrado nos cigarros.

$$s^2 = 4,71^2 = 22,1.$$

2.3 Covariância

De acordo com [7], a covariância é uma medida entre duas variáveis, que mede o total da concentração dos pontos em torno da média. Portanto, dados n pares de valores $(x_1, y_1), ..., (x_n, y_n)$, chamaremos de covariância entre duas variáveis X e Y

$$cov(X, Y) = \frac{\sum(x - \overline{x})(y - \overline{y})}{n}.$$

Isto é, a média dos produtos dos valores centrados das variáveis. Porém, a covariância possui um defeito: seu valor calculado depende diretamente das unidades de medida.

Tome o exemplo do cigarro, em que comparamos a quantidade de alcatrão com o monóxido de carbono.

Temos:

$$cov(X, Y) = \frac{(16 - 12,2).(15 - 12,3) + ... + (18 - 12,2).(15 - 12,3)}{29} = 18,43.$$

Ambos estão na mesma unidade, porém haverá casos em que a comparação poderá se dar entre idade (anos) e massa corporal (kg), o que divergirá bastante da relação entre idade e altura, por exemplo. Assim, ficamos reféns das unidades. Entretanto, veremos mais adiante uma maneira de contornar esse problema.

3 Z - score ou Escore Padrão

O Z-score ou Escore Padronizado, conforme [6], nos fornece uma forma muito útil para comparar valores de dados dentro do mesmo conjunto ou em conjuntos diferentes, sem correr o risco de que diferentes unidades utilizadas pelas variáveis interfiram no resultado final.

Ele também representa a quantidade de desvios padrões que um determinado valor da variável se encontra distante da média, sendo que nas distribuições normais (em forma

de sino), por preceitos impíricos, é comum que 95% de todos os valores se encontrem a 2 desvios padrões da média.

Possui as seguintes fórmulas para a amostra e população, respectivamente:

$$z = \frac{x - \overline{x}}{s} \qquad \text{ou} \qquad z = \frac{x - \mu}{\sigma}.$$

Isto é, quando temos a média e o desvio padrão do conjunto de dados, é possível mensurar a "distância"a que se encontra um determinado dado em relação a média.

Essa transformação de x (chamada de pontuação bruta) em um z-score possibilita a representação de valores na distribuição normal padrão, onde a média tem valor zero e um desvio padrão equivale a uma unidade. Além disso, outro benefício é que agora podemos calcular áreas abaixo da curva normal padronizada com base na Tabela Normal Padrão, que representa a probabilidade de determinado evento ocorrer.

4 Teste de Hipóteses

O teste de hipóteses, também chamado de teste de significância, é um procedimento que permite testar uma afirmativa sobre uma propriedade da população.

Ele ocorre da seguinte forma: existe uma afirmação que deve ser traduzida em uma sentença matemática bem como seu complemento, e veremos que quando uma for falsa a outra deve ser verdadeira, ou seja, rejeitamos ou deixamos de rejeitar uma hipótese e pode-se chegar a conclusão do teste de hipóteses, com um certo grau de confiança.

A hipótese H_0, (ou hipótese nula), deve conter a afirmativa expressa por $\geq, \leq$ ou $=$, ou seja, que contenha a igualdade, deixando para H_1 (ou hipótese alternativa) $>, <$ ou $\neq$. Isso se dá pelo fato de que o pesquisador, ao fazer seu teste de hipóteses deve colocar sua afirmativa na hipótese alternativa, pois este teste não apoia uma afirmativa de um parâmetro que seja igual a um valor específico. Isto é, admite-se que H_0 seja igual a um valor, mas deseja-se que H_0 seja rejeitada de modo que H_1 seja aceita.

A seguir, toma-se o nível de significância α e calcula-se a estatística de teste, que pode se dar com base no score Z, t ou χ, conforme o diagrama abaixo:

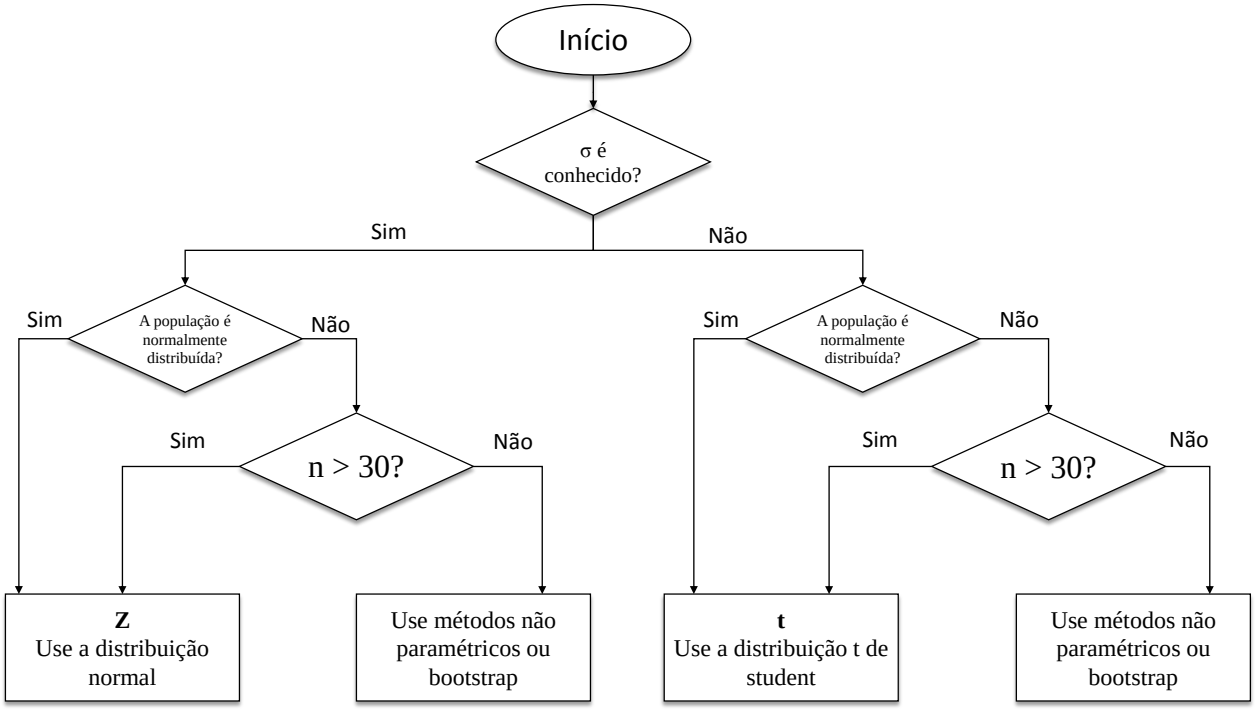


Figura 2 – Diagrama de escolha entre Z e t, disponível em [9]

.

Em posse do valor da estatística de teste e α podemos usar o gráfico de distribuição normal padrão ou t de student para definir o valor crítico e a região crítica, que representam o ponto e o intevalo de valores que possibilitam a rejeição da hipótese nula e a evidência suficiente para garantir que a afirmação inicial pode ser aceita.

Para definição do valor crítico, de acordo com α, consulte a tabela da distribuição t de Student, disponível no Apêndice A e a tabela de distribuição normal padrão para (Z) em [9].

Após essa breve retomada de assuntos relevantes para este trabalho, podemos de fato adentrar no conteúdo principal, que é a correlação e regressão linear.

2 CORRELAÇÃO E REGRESSÃO

O estudo da correlação e regressão proporciona a aplicação de métodos de estatística inferencial para detectar relações entre dados que não são descritas por funções, mas possuem uma relação que pode ser representada também por uma equação.

Para essa verificação, vamos analisar o conjunto de dados que apresenta a concentração de alcatrão, nicotina e monóxido de carbono (CO) em diversas marcas de cigarros, que possuem 100 milímetros (10 centímetros) de comprimento, tem filtro e não são do tipo mentolado ou light.

Para melhor compreensão desses componentes do cigarro, temos uma explanação em [2] sobre a nicotina, alcatrão e o monóxido de carbono.

A Nicotina está presente nas folhas de tabaco e é considerada um estimulante, uma vez que excita muitas células cerebrais e excita a atenção. Não chega a ser tão danosa quanto o alcatrão e o monóxido de carbono, mas seu papel é mais traiçoeiro – quando, no esforço para obtê-la, as pessoas acabam inalando monóxido e os sub-produtos do alcatrão. A nicotina provoca contração e acúmulo de gordura (colesterol) nas paredes das artérias, diminuindo a passagem do sangue e, consequentemente, propiciando ao derrame cerebral e ao infarto do miocárdio.

Alcatrão é uma das maiores ameaças à saúde contidas no cigarro, podendo resultar vários tipos de câncer. Além disso, suas pequenas partículas destroem os alvéolos, o qual apresenta uma grande quantidade de vasos sanguíneos de calibre reduzido, causando sérios problemas respiratórios, como enfisema, por exemplo, doença respiratória crônica, caracterizada pelo acúmulo de ar ou gás em alguma parte do corpo.

O Monóxido de Carbono (CO) é um gás que passa facilmente dos alvéolos pulmonares para a corrente sanguínea, onde se combina com a hemoglobina (substância do sangue que transporta O_2 para os tecidos), forma-se, então, a carboxiemoglobina (COHb), gerando carência de O_2 no organismo pela dificuldade da hemoglobina em transportar o oxigênio.

A seguir são apresentados os dados coletados pela Federal Trade Commision, a agência federal de proteção ao consumidor dos Estados Unidos da América, na tabela e

em gráficos de dispersão.

MARCA	**ALCATRÃO** (mg)	**NICOTINA** (mg)	**CO** (mg)
American Filter	16	1,2	15
Benson & Hedges	16	1,2	15
Camel	16	1,0	17
Capri	9	0,8	6
Carlton	1	0,1	1
Cartier Vendome	8	0,8	8
Chelsea	10	0,8	10
GPC Approved	16	1,0	17
Hi-Lite	14	1,0	13
Kent	13	1,0	13
Lucky Strike	13	1,1	13
Malibu	15	1,2	15
Marlboro	16	1,2	15
Merit	9	0,7	11
Newport Stripe	11	0,9	15
Now	2	0,2	3
Old Gold	18	1,4	18
Pall Mall	15	1,2	15
Players	13	1,1	12
Raleigh	15	1,0	16
Richland	17	1,3	16
Rite	9	0,8	10
Silva Thins	12	1,0	10
Tareyton	14	1,0	17
Triumph	5	0,5	7
True	6	0,6	7
Vantage	8	0,7	11
Viceroy	18	1,4	15

MARCA	**ALCATRÃO** (mg)	**NICOTINA** (mg)	**CO** (mg)
Winston	16	1,1	18

Tabela 1 – Conjunto de Dados: Alcatrão, Nicotina e CO em cigarros

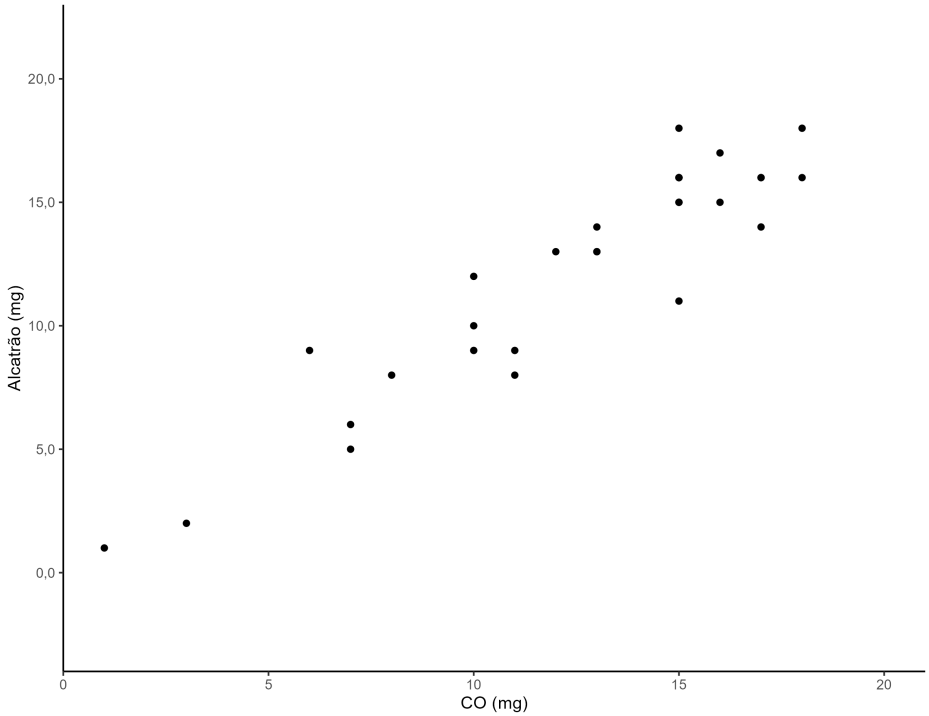


Figura 3 – Gráfico de Dispersão: Alcatrão x CO

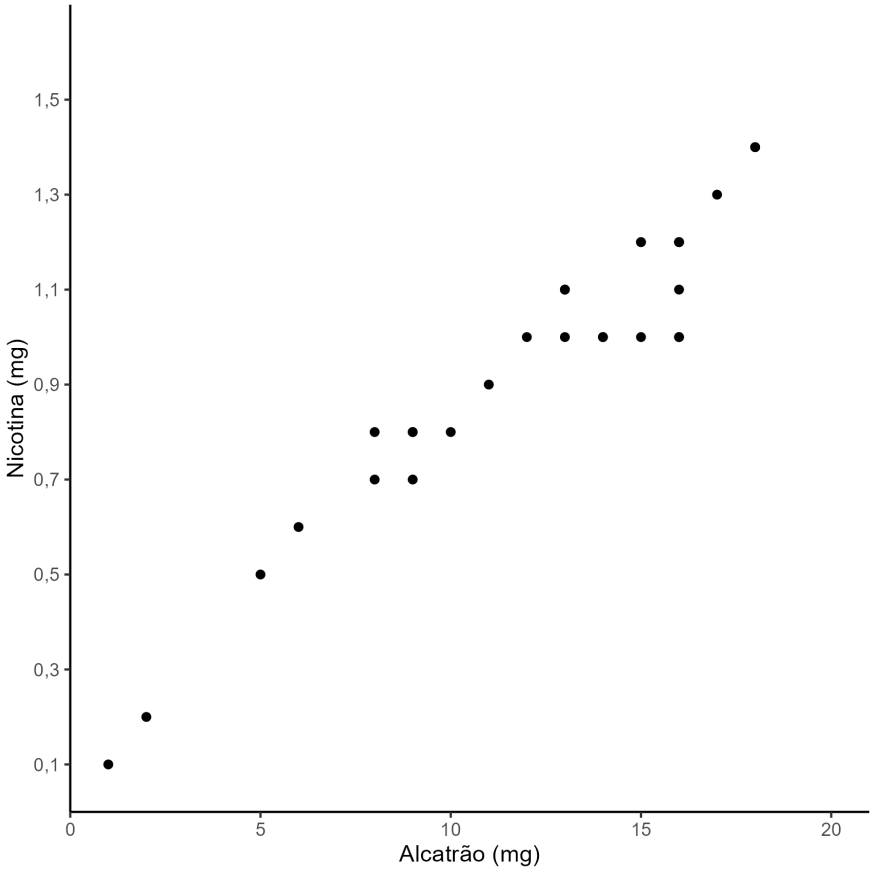


Figura 4 – Gráfico de Dispersão: Alcatrão x Nicotina

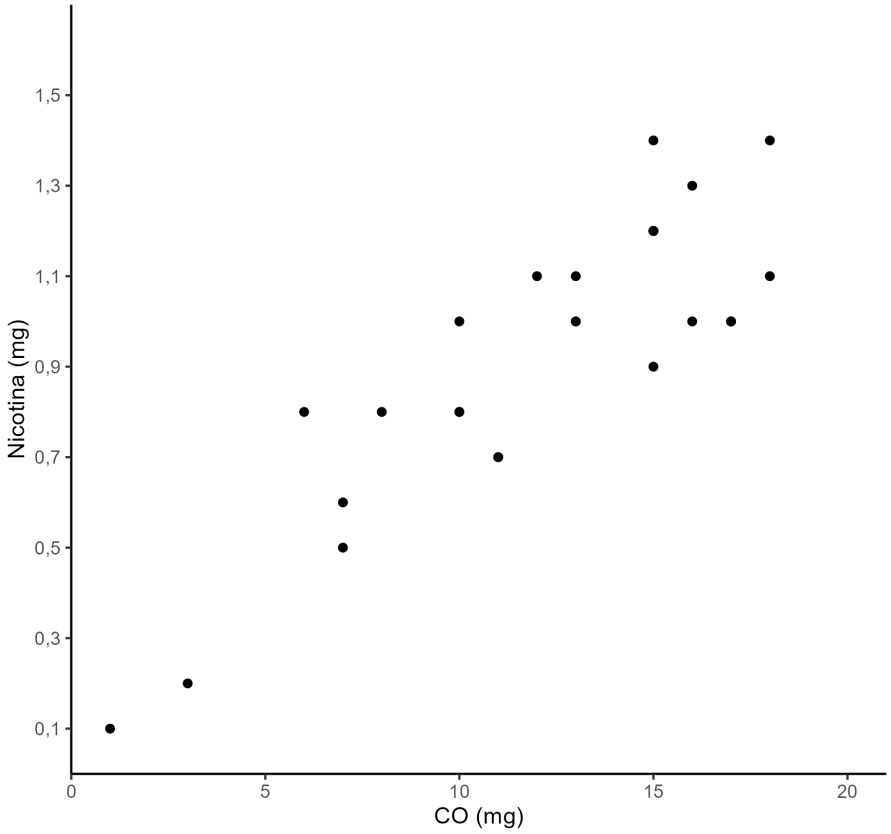


Figura 5 – Gráfico de Dispersão: Nicotina x CO

A primeira vista, olhando os gráficos percebe-se uma relação entre a nicotina e o alcatrão, pois apresenta um certo padrão de crescimento, apresentando uma relação linear já nos demais gráficos, não se visualiza padrões nítidos.

Note-se que essas análises feitas acima são extremamente subjetivas, mas com o auxílio de ferramentas estatísticas estaremos aptos a inferir se realmente há uma correlação entre as variáveis, e caso haja, prever um resultado usando uma equação e analisar quão preciso é esse resultado.

5 CORRELAÇÃO

A verificação da correlação entre duas variáveis se dá por meio do Coeficiente de Correlação Linear r, também conhecido como Coeficiente de Correlação de Produto de Momentos de Pearson, em homenagem ao estatístico inglês Karl Pearson (1857-1936), que o desenvolveu originariamente.

Alguns requisitos devem ser observados para sua aplicação:

1. A amostra de dados emparelhados *(x,y)* é uma amostra *aleatória* de dados quantitativos independentes;
2. O exame visual do diagrama de dispersão deve se aproximar do padrão de uma reta;
3. Quaisquer *outliers* devem ser removidos caso se saiba que se trata de erro.

O coeficiente r mede a intensidade da relação linear entre os valores quantitativos emparelhados x e y em uma amostra através da seguinte fórmula:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}} \tag{1}$$

em que:

- r representa o coeficiente de correlação linear para uma amostra;
- n representa o número de pares de dados presentes;
- $\sum$ representa a soma dos itens indicados, de 1 até n.

A fórmula apresentada acima facilita os cálculos manuais, tornando prático seu uso em uma planilha, por exemplo, mas podemos encontrar outras fórmulas equivalentes, que nos ajudam a entender o significado e compreensão de como r funciona. São elas:

$$r = \frac{\sum \left[\frac{(x - \overline{x})}{s_x} \frac{(y - \overline{y})}{s_y} \right]}{n - 1}, \tag{2}$$

$$r = \frac{\sum (x - \overline{x})(y - \overline{y})}{(n - 1)s_x s_y} \tag{3}$$

e

$$r = \frac{s_{xy}}{\sqrt{s_{xx}}\sqrt{s_{yy}}} \tag{4}$$

A fórmula apresentada em (2) é a que melhor explica o significado do coeficiente de correlação de Pearson. Tomando as retas perpendiculares aos eixos x e y que passam por $(\overline{x}, \overline{y})$, podemos dividir o gráfico de dispersão em 4 quadrantes, em que estão distribuídos os pontos que representam os pares dos dados coletados, conforme figura abaixo:

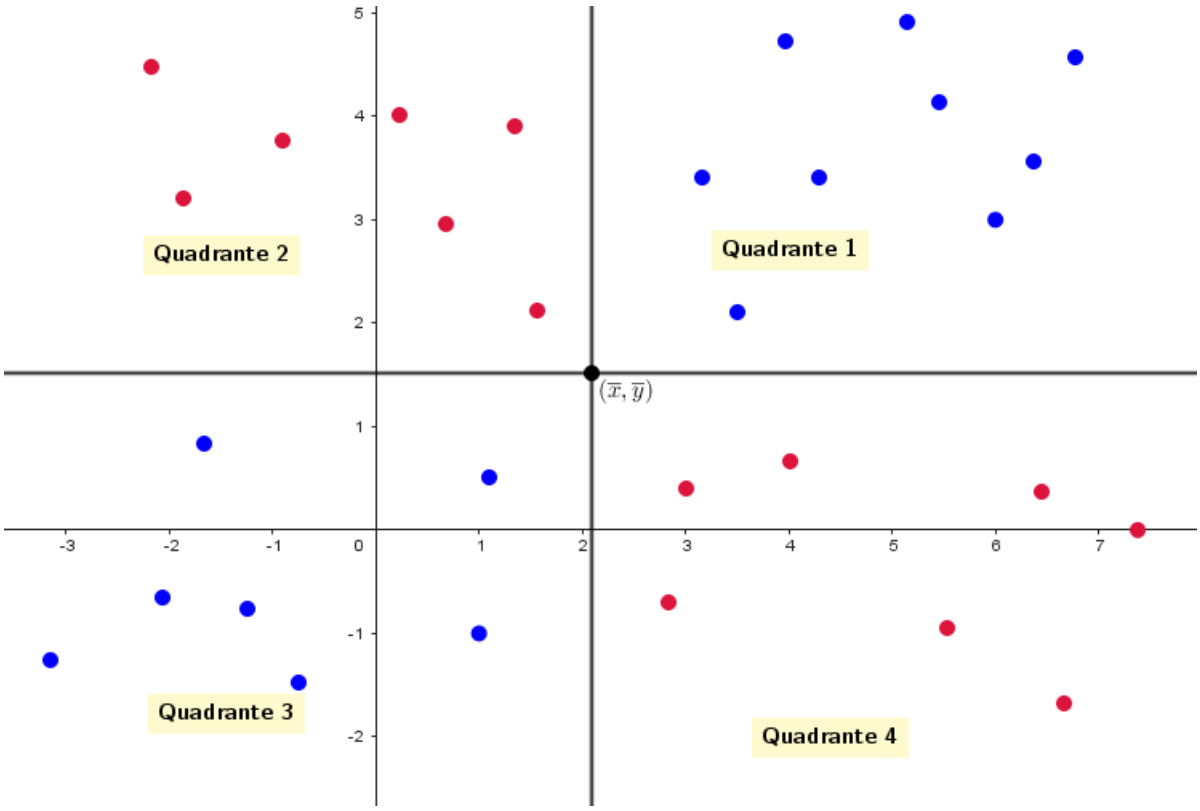


Figura 6 – Gráfico dividido em novos quadradantes

Note que, na Figura 6, os pontos em azul estão dispostos em um padrão de crescimento, e estão localizados nos quadrantes 1 e 3, de forma que, no quadrante 1 $(x > \overline{x})$ e $(y > \overline{y})$ e no quadrante 3 $(x < \overline{x})$ e $(y < \overline{y})$, e portanto em (2), no seu numerador, temos que $\sum (x - \overline{x})(y - \overline{y})$ nos retornará um valor positivo, ou seja, a somatória dos desvios dos valores observados em relação a média, sugerindo uma correlação positiva.

Analogamente, podemos fazer essa análise com os pontos vermelhos da Figura 6, em que os pontos estão distribuídos entre os quadrantes 2 e 4, e no quadrante 2 obtemos $(x < \overline{x})$ e $(y > \overline{y})$ e no quadrante 4 $(x > \overline{x})$ e $(y < \overline{y})$, obtendo uma somatória dos desvios negativa, sugerindo uma correlação negativa.

E por fim, se os os pontos estão espalhados nos quatro quadrantes, $\sum (x - \overline{x})(y - \overline{y})$ apresenta um valor próximo de zero, não indicando correlação linear.

Ainda na fórmula (2), temos que os desvios das médias são divididos pelo desvio padrão. Isso ocorre pois os transformamos em z-score com o objetivo de padronizarmos as variáveis.

Finalmente, a divisão do somatório por $n - 1$ é utilizado pelo motivo que é utilizado

no desvio padrão das amostras, visando obter um valor não enviesado para a população.

Agora, partimos para um desenvolvimento da fórmula (2) onde será possível obter as demais, usando expedientes matemáticos, que apesar de mencionados nos livros de Estatística, como [9], não são demonstrados.

Assim, começando de (2) temos, por uma simples manipulação matemática, que ela é correspondente a (3):

$$r = \frac{\sum \left[\frac{(x-\overline{x})}{s_x} \frac{(y-\overline{y})}{s_y} \right]}{n-1} = \frac{\sum (x-\overline{x})(y-\overline{y})}{(n-1)s_x s_y}.$$

Agora, partindo de (2) para (1), temos:

$$\begin{aligned} r &= \frac{\sum \left[\frac{(x-\overline{x})}{s_x} \frac{(y-\overline{y})}{s_y} \right]}{n-1} \\ &= \frac{\sum \left[\frac{(x-\overline{x})}{\sqrt{\frac{\sum (x-\overline{x})^2}{n-1}}} \frac{(y-\overline{y})}{\sqrt{\frac{\sum (y-\overline{y})^2}{n-1}}} \right]}{n-1} \\ &= \frac{\sum (x-\overline{x})(y-\overline{y})}{\sqrt{\sum (x-\overline{x})^2} \sqrt{\sum (y-\overline{y})^2}}. \end{aligned} \tag{5}$$

Desenvolvendo

$$\begin{aligned} \sum (x-\overline{x})(y-\overline{y}) &= \sum (xy - x\overline{y} - y\overline{x} + \overline{x}.\overline{y}) \\ &= \sum xy - \sum x\overline{y} - \sum y\overline{x} + \sum \overline{x}.\overline{y} \\ &= \sum xy - \overline{y} \sum x - \overline{x} \sum y + \overline{x}.\overline{y} \sum 1 \\ &= \sum xy - \overline{y} \sum x \frac{1}{n}.n - \overline{x} \sum y \frac{1}{n}.n + \overline{x}.\overline{y}.n \\ &= \sum xy - \overline{y}.\overline{x}.n - \overline{x}.\overline{y}.n + \overline{x}.\overline{y}.n \\ &= \sum xy - 2.\overline{y}.\overline{x}.n + \overline{x}.\overline{y}.n \\ &= \sum xy - \overline{y}.\overline{x}.n; \end{aligned} \tag{6}$$

$$\begin{aligned} \sum (x-\overline{x})^2 &= \sum (x^2 - 2.x.\overline{x} + \overline{x}^2) \\ &= \sum x^2 - 2.\overline{x} \sum x + \overline{x} \sum 1 \\ &= \sum x^2 - 2.n.\overline{x}^2 + n.\overline{x}^2 \\ &= \sum x^2 - n.\overline{x}^2; \end{aligned} \tag{7}$$

$$\begin{aligned}\sum(y-\overline{y})^2 &= \sum(y^2-2.y.\overline{y}+\overline{y}^2)\\ &= \sum y^2-2.\overline{y}\sum y+\overline{y}\sum 1\\ &= \sum y^2-2.n.\overline{y}^2+n.\overline{y}^2\\ &= \sum y^2-n.\overline{y}^2. \qquad (8)\end{aligned}$$

Substituindo (6), (7) e (8) em (5), temos:

$$\begin{aligned}r &= \frac{\sum xy-n.\overline{y}.\overline{x}}{\sqrt{\sum x^2-n.\overline{x}^2}\sqrt{\sum y^2-n.\overline{y}^2}}\\ r &= \frac{\sum xy-n.\frac{\sum x}{n}.\frac{\sum y}{n}}{\sqrt{\sum x^2-n.\left(\frac{\sum x}{n}\right)^2}\sqrt{\sum y^2-n.\left(\frac{\sum y}{n}\right)^2}}.\end{aligned}$$

Multiplicando e dividindo por n

$$\begin{aligned}r &= \frac{\sum xy-n.\frac{\sum x}{n}.\frac{\sum y}{n}}{\sqrt{\sum x^2-n.\left(\frac{\sum x}{n}\right)^2}\sqrt{\sum y^2-n.\left(\frac{\sum y}{n}\right)^2}}.\left(\frac{n}{n}\right)\\ r &= \frac{n.\sum xy-\sum x\sum y}{\sqrt{\frac{n.\sum x^2-(\sum x)^2}{n}}\sqrt{\frac{n.\sum y^2-(\sum y)^2}{n}}}.\left(\frac{1}{n}\right)\\ r &= \frac{n.\sum xy-\sum x\sum y}{\sqrt{n.\sum x^2-(\sum x)^2}\sqrt{n.\sum y^2-(\sum y)^2}}. \qquad (9)\end{aligned}$$

Observe que (9) é a mesma fórmula de (1).

Vamos calcular o Coeficiente de Correlação r dos dados emparelhados de alcatrão (coluna 1) e nicotina (coluna 2) da Tabela 1 usando a fórmula (1):

$$r=\frac{n(\sum xy)-(\sum x)(\sum y)}{\sqrt{n(\sum x^2)-(\sum x)^2}\sqrt{n(\sum y^2)-(\sum y)^2}}$$

x	y	$x.y$	x^2	y^2
16,0	1,2	19,2	256	1,44
16,0	1,2	19,2	256	1,44
16,0	1,0	16	256	1
⋮	⋮	⋮	⋮	⋮
8,0	0,7	5,6	64	0,49
18,0	1,4	25,2	324	1,96
16,0	1,1	17,6	256	1,21
$\Sigma x = 351,0$	$\Sigma y = 27,3$	$\Sigma x.y = 369,5$	$\Sigma x^2 = 4849$	$\Sigma y^2 = 28,45$

Tabela 2 – Determinação das Estatísticas Usadas para Calcular r

Substituindo os valores encontrados na tabela acima na fórmula (1), obtemos:

$$r = \frac{29(369,5) - (351,0)(27,5)}{\sqrt{29(4849) - (351,0)^2}\sqrt{29(28,48) - (27,3)^2}}$$

$$r = 0,961368566.$$

Como interpretar *r=0,961* (considerando um arredondamento de 3 casas decimais)?

Inicialmente, devemos perceber que da forma como foi construída a fórmula (1) o valor de r sempre estará entre -1 e 1. Dessa forma, podemos dizer que quanto mais próximo de -1 ou 1, mais forte a relação entre as variáveis, e contrario sensu, quanto mais perto de 0 conclui-se que não há correlação linear significativa entre as variáveis x e y.

Mais uma vez, devemos impor critérios objetivos para analisar essa proximidade de -1 ou 1. Há dois métodos que podem ser utilizados para testar se existe ou não a correlação linear. O método 1 consiste em usar a distribuição t de Student e o método 2 leva em consideração a análise da Tabela de Valores Críticos do Coeficiente de Correlação de Pearson r.

Vamos analisar cada um individualmente.

5.1 Teste de Hipótese Formal para Correlação - Método 1

O método 1 toma como base a distribuição t de Student, com uma estatística de teste em que se utiliza $t = \frac{r}{s_r}$, em que s_r representa o valor do desvio padrão amostral dos valores de r, podendo ser escrito como $\sqrt{\frac{(1-r^2)}{n-2}}$, em que $n-2$ diz respeito ao grau de liberdade.

Tome as seguintes hipóteses:

$H_0 : \rho = 0$ Não há correlação linear.

$H_1 : \rho \neq 0$ Há correlação linear.

Nesse caso, se a estatística de teste t possuir um valor maior que o valor crítico, em módulo, a ser consultado na tabela do Apêndice A, podemos rejeitar a hipótese nula e concluir que existe correlação linear; do contrário, se t, em módulo, é menor que o valor crítico, deixamos de rejeitar a hipótese nula e concluimos que não há correlação linear.

Vamos aos cálculos: seja, $\alpha = 0,05$, $r = 0,961$ e $n = 29$, a estatística teste é:

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}} = \frac{0,961368566}{\sqrt{\frac{1-0,961368566^2}{29-2}}} = 18,14770788.$$

Agora, devemos confrontar esse valor com o valor crítico na tabela do Apêndice A, onde $\alpha = 0,05$ distribuido em duas caudas e temos $n - 2 = 29 - 2 = 27$ graus de liberdade. Portanto $t = \pm 2,052$, como mostra a figura abaixo:

Distribuição t: Valores Críticos t					
	Área em Uma Cauda				
	0.005	**0.01**	**0.025**	**0.05**	**0.10**
	Área em Duas Caudas				
Graus de Liberdade	**0.01**	**0.02**	**0.05**	**0.10**	**0.20**
24	2,797	2,492	2,064	1,711	1,318
25	2,787	2,485	2,060	1,708	1,316
26	2,779	2,479	2,056	1,706	1,315
27	2,771	2,473	2,052	1,703	1,314
28	2,763	2,467	2,048	1,701	1,313
29	2,756	2,462	2,045	1,699	1,311
30	2,750	2,457	2,042	1,697	1,310

Figura 7 – Fragmento da tabela t de Student

Assim, o valor crítico $\pm 2,052$ define as regiões que a sua esquerda e direita representam as regiões de rejeição da hipótese nula, conforme figura abaixo, e como $|18,148| > 2,052$ está nessa região, podemos rejeitar a hipótese nula e afirmar que há correlação linear significativa entre os dados.

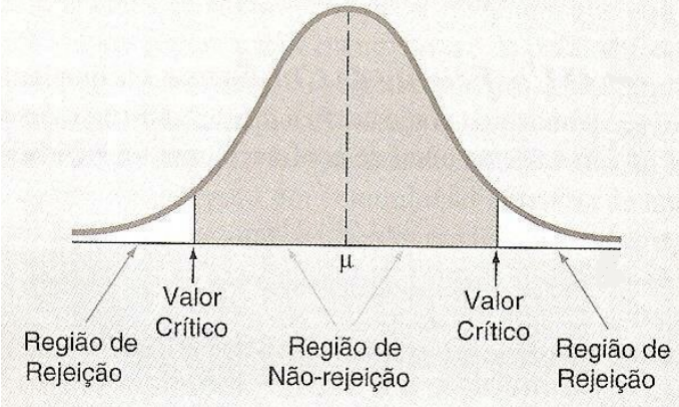


Figura 8 – Gráfico de distribuição, disponível em [4]

5.2 Teste de Hipótese Formal para Correlação - Método 2

O método 2 consiste em consultar na Tabela de Valores Críticos do Coeficiente de Correlação de Pearson r, na linha que representa o número de pares de dados presente, eleger um nível de confiança ou significância, que geralmente adota-se $\alpha = 0,05$ ou $\alpha = 0,01$ e comparar o valor de r obtido com o valor crítico estabelecido na tabela.

Se $\mid r \mid >$ valor crítico, rejeita-se H_0 e conclui-se que há uma correlação linear, mas se $\mid r \mid \leq$ valor crítico, então deixamos de rejeitar H_0, e conclui-se que não há correlação linear.

No exemplo dos componentes dos cigarros, temos $r = 0,961$, $n = 29$ e $\alpha = 0,05$. Conforme podemos ver na imagem abaixo, usando a tabela do Apêndice B, temos que os valores críticos de r são $r = \pm 0,361$. Portanto, como $|0,961| > 0,361$, podemos rejeitar a hipótese nula.

Valores Críticos do Coeficiente de Correlação de Pearson r		
n	$\alpha = 0,05$	$\alpha = 0,01$
20	0,444	0,561
25	0,396	0,505
30	0,361	0,463
35	0,335	0,430
40	0,312	0,402

Figura 9 – Fragmento da tabela dos valores críticos do coeficiente de correlação de Pearson

Conclusão: utilizando os dois métodos (teste t-student e teste de correlação de Pearson) em ambos chegamos na condição de rejeitar a hipótese nula. Isso significa que: "Há evidência suficiente para apoiar a afirmativa de que existe uma correlação linear entre a quantidade de nicotina e alcatrão presentes nos cigarros em análise".

6 REGRESSÃO

Uma vez estabelecida a correlação entre as variáveis, podemos determinar uma reta que melhor representa essa relação, chamada de *reta de regressão* ou *reta de melhor ajuste* ou *reta de mínimos quadrados* e sua respectiva *equação de regressão*.

Aqui estamos interessados em definir uma variável que não é completamente determinada pela outra, ao contrário do que acontece com funções, amplamente estudadas em Álgebra, em que duas variáveis de relacionam de forma determinística.

Usaremos, entretanto, a mesma equação da reta $y = mx + b$, que adaptada será $\hat{y} = b_0 + b_1 x$, que expressará a relação entre a variável x (chamada de **variável explanatória** ou **variável previsora** ou **variável independente**) e $\hat{y}$ (chamada de **variável resposta** ou **variável dependente**), onde b_0 é o coeficiente linear e b_1 é o coeficiente angular (inclinação da reta).

A inclinação da reta é dada por:

$$b_1 = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}. \tag{10}$$

E o coeficiente linear é dado por

$$b_0 = \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2} \qquad ou \qquad b_0 = \overline{y} - b_1\overline{x}. \tag{11}$$

Note que nessa equação de regressão, b_0 e b_1 representam as estatísticas amostrais, que são utilizados para estimarmos os parâmetros populacionais β_0 e β_1, respectivamente, em $y = \beta_0 + \beta_1 x$.

Dessa forma, vamos determinar, agora, a equação de regressão para o exemplo do alcatrão e nicotina.

Utilizando os dados constantes da Tabela 1 e a fórmula (10), temos:

$$b_1 = \frac{29(369,5) - (351)(27,3)}{29(4849) - (351)^2}$$

$$b_1 = 0,065.$$

E b_0 pode ser encontrado usando uma das versões da fórmula para b_0, constante em (11). Vamos utilizar a segunda forma, em que devemos calcular $\overline{y}$ e $\overline{x}$ e após substituir os seus valores.

$$\overline{y} = \frac{\sum y}{n} = \frac{27,3}{29} = 0,941 \qquad e \qquad \overline{x} = \frac{\sum x}{n} = \frac{351}{29} = 12,103.$$

Assim,

$$\begin{aligned} b_0 &= \overline{y} - b_1\overline{x} \\ &= 0,941 - (0,065).(25,103) \\ &= 0,154. \end{aligned}$$

Então a reta de regressão é:

$$\hat{y} = 0,154 + 0,065x.$$

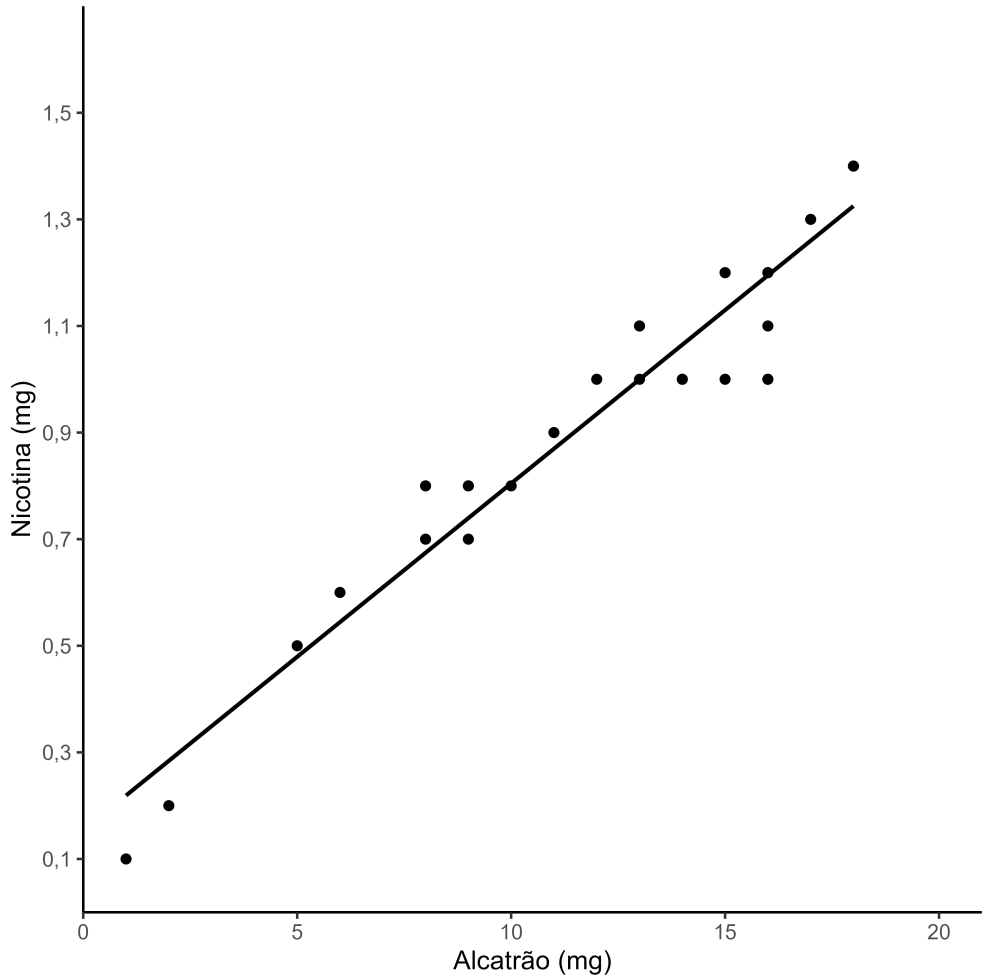


Figura 10 – Reta de Regressão: Alcatrão x Nicotina

A partir dessa equação de regressão podemos fazer previsões. Por exemplo, podemos prever qual será a quantidade de nicotina presente em um cigarro medindo apenas a variável alcatrão. Suponha que um cigarro contém 20 mg de alcatrão, então:

$$\hat{y} = 0,154 + 0,0065.(20)$$

$$\hat{y} = 1,454,$$

ou seja, se um cigarro contiver 20 mg de alcatrão, estima-se que terá 1,454 mg de nicotina.

6.1 Resíduos e a Propriedade dos Mínimos Quadrados

Para o melhor ajuste da reta de regressão, devemos observar a menor distância vertical entre os pontos dados e a reta. Essas distâncias são chamadas *resíduos* ou *desvio não explicado*. Dessa forma, podemos estabelecer a seguinte equação:

$$\text{resíduo} = (y \text{ observado}) \text{ - } (y \text{ previsto}) = y - \hat{y}.$$

Com o auxílio do Método dos Mínimos Quadrados pode-se compreender a origem da fórmulas (10) e (11) que estabelecem o valor do coeficiente angular e linear, respectivamente, da reta de regressão que objetiva minimizar a soma dos resíduos quadráticos de forma a obter a melhor aproximação dos valores reais.

Tome $M = {d_1}^2 + {d_2}^2 + {d_3}^2 + ... + {d_n}^2$ a soma dos n resíduos quadráticos. Sejam (x_k, y_k) os valores observados e tabelados, e $r(x) = \hat{y} = b_0 + b_1 x$ a reta que queremos obter através da regressão.

Então,

$$\begin{aligned} M(b_o, b_1) = \sum_{k=1}^{n} [d_k]^2 &= \sum_{k=1}^{n} [y_k - r(x_k)]^2 \\ &= \sum_{k=1}^{n} [y_k - (b_0 + b_1 x_k)]^2 \\ &= \sum_{k=1}^{n} [y_k - b_0 - b_1 x_k)]^2. \end{aligned} \tag{12}$$

Como queremos minimizar a diferença, fazemos a derivada parcial de (12) em relação a b_0 e b_1:

$$\begin{cases} \dfrac{\partial M}{\partial b_0} = 0 \\ \dfrac{\partial M}{\partial b_1} = 0 \end{cases} \Rightarrow \begin{cases} 2\displaystyle\sum_{k=1}^{n} [y_k - b_0 - b_1 x_k](-1) = 0 \\ 2\displaystyle\sum_{k=1}^{n} [y_k - b_0 - b_1 x_k](-x_k) = 0 \end{cases}$$

$$\Rightarrow \begin{cases} \displaystyle\sum_{k=1}^{n} [-y_k + b_0 + b_1 x_k] &= 0 \\ \displaystyle\sum_{k=1}^{n} [-x_k y_k + b_0 x_k + b_1 {x_k}^2] &= 0 \end{cases}$$

$$\Rightarrow \begin{cases} \displaystyle\sum_{k=1}^{n} -y_k + \sum_{k=1}^{n} b_0 + \sum_{k=1}^{n} b_1 x_k &= 0 \\ \displaystyle\sum_{k=1}^{n} -x_k y_k + \sum_{k=1}^{n} b_0 x_k + \sum_{k=1}^{n} b_1 {x_k}^2 &= 0 \end{cases}$$

$$\Rightarrow \begin{cases} \displaystyle\sum_{k=1}^{n} b_0 + \sum_{k=1}^{n} b_1 x_k &= \sum_{k=1}^{n} y_k \\ \displaystyle\sum_{k=1}^{n} b_0 x_k + \sum_{k=1}^{n} b_1 {x_k}^2 &= \sum_{k=1}^{n} x_k y_k \end{cases}$$

$$\Rightarrow \begin{cases} b_0 \sum_{k=1}^{n} 1 + b_1 \sum_{k=1}^{n} x_k &= \sum_{k=1}^{n} y_k \\ b_0 \sum_{k=1}^{n} x_k + b_1 \sum_{k=1}^{n} {x_k}^2 &= \sum_{k=1}^{n} x_k y_k \end{cases} .$$

Podemos então escrever a seguinte matriz:

$$\begin{bmatrix} \sum n & \sum x \\ \sum x & \sum x^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} \sum y \\ \sum xy \end{bmatrix} .$$

Utilizando o Método de Cramer para resolução desse produto de matrizes, obtemos:

$$D = \begin{vmatrix} n & \sum x \\ \sum x & \sum x^2 \end{vmatrix} = n(\sum x^2) - (\sum x)^2 \ ;$$

$$D_{b_0} = \begin{vmatrix} \sum y & \sum x \\ \sum xy & \sum x^2 \end{vmatrix} = (\sum y)(\sum x^2) - (\sum x)(\sum xy) \ ;$$

$$D_{b_1} = \begin{vmatrix} n & \sum y \\ \sum x & \sum xy \end{vmatrix} = n(\sum xy) - (\sum x)(\sum y) \ .$$

Então

$$\begin{aligned} \frac{D_{b_0}}{D} = b_0 &= \frac{(\sum y)(\sum x^2) - (\sum x)(\sum xy)}{n(\sum x^2) - (\sum x)^2}; \\ \frac{D_{b_1}}{D} = b_1 &= \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}. \end{aligned}$$

Dessa forma, fica demonstrado como as fórmulas(10) e (11) para os coeficientes linear e angular, respectivamente foram obtidas.

6.2 Gráfico de Resíduos

Ainda na análise dos resíduos, podemos elaborar um gráfico de resíduos, que se mostra uma forma relevante para avaliar os resultados obtidos pela reta de regressão. Para sua construção, usamos os mesmos valores do eixo x do gráfico de dispersão, ou seja, a quatidade de nicotina, e no eixo y serão aplicados os resíduos.

Para o exemplo utilizado neste Capítulo, temos:

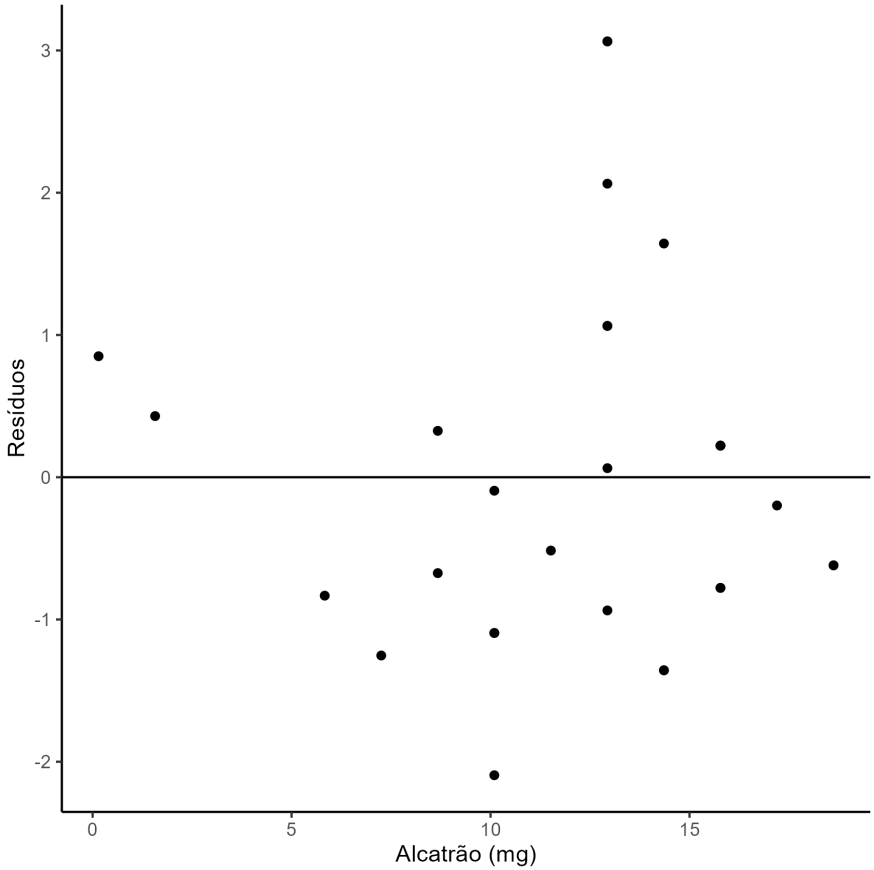


Figura 11 – Gráfico de Resíduos

A análise do gráfico acima, se dá da seguinte maneira:

- Se o gráfico de resíduos não mostra qualquer padrão entre seus pontos, isso quer dizer que a equação de regressão é uma boa representação da correlação entre as duas variáveis.

- Se o gráfico de resíduos mostra algum padrão sistemático entre seus pontos, isso quer dizer que a equação de regressão não é uma boa representação da correlação entre as duas variáveis.

No caso do gráfico da Figura 11, percebe-se que não temos um padrão na distribuição dos pontos, portanto, pode-se dizer que a reta de regressão $\hat{y} = 0,154 + 0,065x$ descreve bem a correlação entre as variáveis nicotina e alcatrão presentes nos cigarros em análise.

6.3 Variação Explicada e Não-Explicada

Os cálculos feitos até o momento nos proporcionaram encontrar uma reta de regressão que melhor se ajusta aos dados em análise. Porém essa reta não é exatamente precisa, e não poderia ser, pois não estamos trabalhando com funções. Entretanto, essa reta apresenta melhores resultados do que se considerarmos o simples cálculo da média dos dados. Assim, podemos perceber que há variações nos resultados das previsões em função de desvios que não são explicados pela variável dependente.

Retomando o exemplo dos componentes do cigarro, percebemos uma correlação significativa entre o Alcatrão e a Nicotina, ou seja, boa parte da Nicotina presente nos cigarros pode ser explicada pela quantidade de Alcatrão, porém há outros fatores, como por exemplo o CO, que a influenciam também.

Ou seja, essas diferenças entre a previsão e o valor real dos dados amostrais se dá por um *desvio* explicado e um desvio não explicado, que somados indicam o desvio total. Em outras palavras, temos:

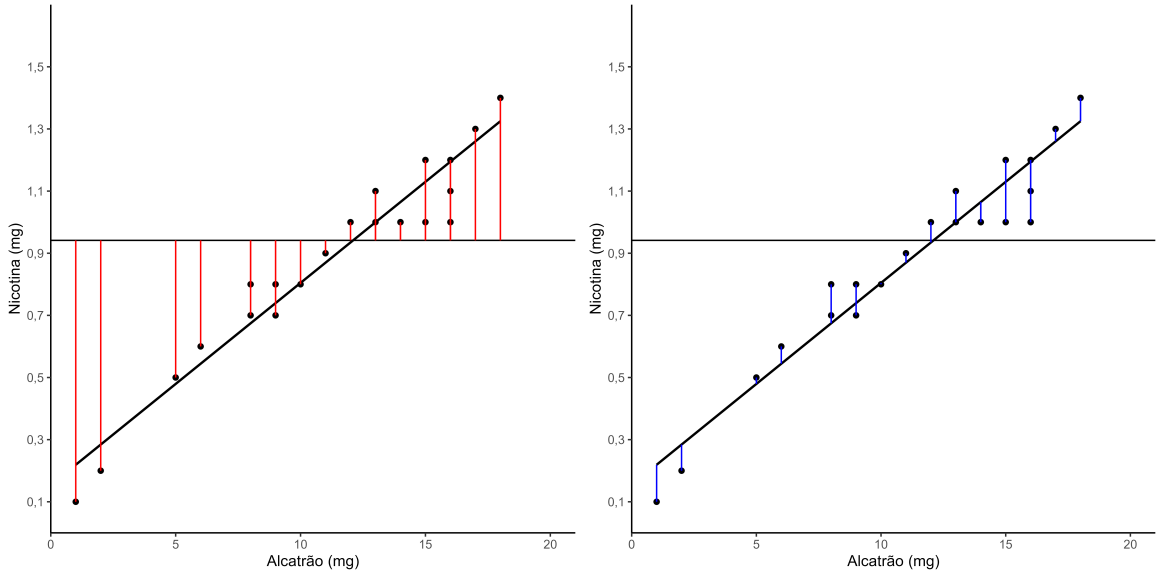
Desvio total: $y - \bar{y}$;

Desvio explicado: $y - \hat{y}$;

Desvio não-explicado: $\hat{y} - \bar{y}$;

em que $\bar{y}$ é a média dos valores de y; $\hat{y}$ é o valor estimado pela regressão e y é o valor do dado amostral.

Para melhor visualização dos desvios, considere a figura abaixo:



(a) Desvio Explicado (b) Desvio Não-Explicado

Figura 12 – Desvios Explicados e Não-Explicados

As figuras 12a e 12b permitem a visualização do desvio total e do desvio não-explicados, respectivamente. O desvio total se dá pela distância vertical (vermelha) entre o ponto do dado amostral e a reta horizontal que representa o valor médio da variável dependente (y); já o desvio não-explicado é evidenciado pela distância vertical (azul) entre o ponto do dado amostral e o valor estimado pela reta de regressão.

Assim, temos a seguinte relação:

$$\begin{array}{ccccc} \text{Desvio Total} & = & \text{Desvio Explicado} & + & \text{Desvio Não-Explicado} \\ (y-\bar{y}) & = & (\hat{y}-\bar{y}) & + & (y-\hat{y}). \end{array}$$

E se elevarmos todos os termos ao quadrado obteremos as *variações* totais, explicada e não-explicada, assim como ocorre com o desvio padrão e a variância. Isso é relevante pois permite determinar o Coeficiente de Determinação, que explica a quantidade de variação em y que é explicada pela reta de regressão.

$$\text{Coeficiente de Determinação} = \frac{\text{variação explicada}}{\text{variação total}}$$

Com o auxílio do software R, podemos calcular as variações:

- Variação total $= 2,750345$;
- Variação não explicada $= 0,2083949$;

- Variação explicada $= 2,750345 - 0.2083949 = 2,5419501$.

Assim, para determinar o coeficiente de determinação, temos:

$$\text{Coeficiente de Determinação} = \frac{\text{variação explicada}}{\text{variação total}} = \frac{2,5419501}{2,750345} = 0.9242295.$$

Porém, esse Coeficiente de Determinação também pode ser obtido de outra forma, bastante simples, que é calcular r^2, ou seja, elevar o Coeficiente de Pearson ao quadrado:

$$r^2 = (0,961368566)^2 = 0.9242.$$

Portanto, calculando o Coeficiente de Determinação das duas formas, obtivemos os mesmos resultados, 0,9242, ou seja, 92,42%. Esse valor indica que a reta de regressão é 92,42% melhor que a reta de média para expressar a relação entre Alcatrão e Nicotina presente nos cigarros.

6.4 Intervalos de Previsão

Após verificar a correlação entre dados, estabelecer a reta de regressão, que melhor representa essa correlação, fazer estimativas e entender as variações entre os valores preditos e os reais, podemos finalmente avaliar o *erro padrão da estimativa*, que é uma medida conjunta da dispersão dos pontos fornecidos pela amostra ao redor da reta de regressão, através das seguintes fórmulas:

$$s_e = \sqrt{\frac{\sum(y-\hat{y})^2}{n-2}} \qquad \text{ou} \qquad s_e = \sqrt{\frac{\sum y^2 - b_0 \sum y - b_1 \sum x.y}{n-2}}. \tag{13}$$

O raciocínio por trás dessas fórmulas é semelhante ao raciocínio para calcular o desvio padrão, ou seja, no primeiro medimos como os pontos amostrais se afastam da reta de regressão, enquanto que no segundo medimos como os pontos se afastam da sua média.

Portanto, o erro padrão da estimativa para o exemplo da Nicotina e Alcatrão presentes nos cigarros é, utilizando a segunda fórmula da Equação 13:

$$\begin{aligned} s_e &= \sqrt{\frac{28,45 - (0,154).(27,3) - (0,065).(369,5)}{29-2}} \\ &= 0,09195409483. \end{aligned}$$

Ainda com base nesse erro padrão de estimativa é possível calcular o *intervalo de previsão* para verificar o quão confiável é realmente a estimativa prevista, ou seja, se elas se encontram numa faixa com determinado grau de confiança. Usaremos a seguinte fórmula para determinar um intervalo de previsão para um dado individual:

$$E = t_{\alpha/2}.s_e\sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n(\sum x^2) - (\sum x)^2}}, \tag{14}$$

em que x_0 indica o valor dado de x, $t_{\alpha/2}$ é o valor encontrado na tabela de distribuição t: valores críticos, com $n-2$ graus de liberdade (Apêndice A) e s_e é o erro padrão de estimativa encontrado em (13).

O intervalo então fica dessa forma:

$$\hat{y} - E < y < \hat{y} + E$$

. Assim, para o exemplo em estudo nesse trabalho, vimos que uma estimativa para 20mg de Alcatrão presente no cigarro teríamos 1,454mg de Nicotina. Agora vamos construir um intervalo de previsão de 95% para $x_0 = 20$ e aferir se 1,454mg de Nicotina é preciso.

$$\begin{aligned} E &= (2,052).(0,09195409483)\sqrt{1 + \frac{1}{29} + \frac{29(20 - 12.10345)^2}{29(4849) - (351)^2}} \\ &= 0,201314359. \end{aligned}$$

O intervalo é:

$$\begin{aligned} 1,454 - 0,201314359 < &\quad y \quad < 1,454 + 0,201314359 \\ 1,252685641 < &\quad y \quad < 1,655314359. \end{aligned}$$

Portanto, sabemos que para 20mg de alcatrão estamos 95% certos de que a nicotina estará entre 1,25mg e 1,65mg.

3 Correlação e Regressão no Ensino Médio

A Base Nacional Comum Curricular (BNCC), disponível em [1], de 2017, que dispõe sobre a estrutura dos currículos escolares para o Ensino Médio (EM), indica as aprendizagens essenciais que todo estudante deve adquirir, no aspecto acadêmico, humano, social e propõe dez competências gerais que se traduzem em conhecimentos, habilidades atitudes e valores, que podem lhe guiar nas complexidades da vida cotidiana, no pleno exercício de sua cidadania e no mundo do trabalho.

Dessa forma, para atingir esses objetivos, a BNCC, seguindo as determinações da Lei de Diretrizes e Bases da Educação Nacional (LDB), estrutura o Ensino Médio em quatro áreas do conhecimento: Linguagens e suas Tecnologias, Matemática e suas Tecnologias, Ciências da Natureza e suas Tecnologias, Ciências e Humanas e Sociais Aplicadas (mantendo Língua Portuguesa e Matemática como componentes curriculares obrigatórios em todos os anos do EM).

Na área da Matemática e suas Tecnologias, temos as unidades de conhecimento: Números, Álgebra, Geometria, Grandezas e Medidas, Probabilidade e Estatística. No que diz respeito ao assunto estudado neste trabalho, pode-se focar na competência específica 5, que a BNCC descreve como

> (EM13MAT510) Investigar e estabelecer conjecturas a respeito de diferentes conceitos e propriedades matemáticas, empregando recursos e estratégias como observação de padrões, experimentações e tecnologias digitais, identificando a necessidade, ou não, de uma demonstração cada vez mais formal na validação das referidas conjecturas.

Essa competência vem ao encontro do que se propõe este trabalho ao fazer uma reflexão aos leitores, e principalmente aos professores, que atuantes em sala de aula, precisam levar aos estudantes um conteúdo previsto no currículo escolar, mas de forma a estimular um pensamento questionador e crítico, de maneira a proporcionar um aprendizado ativo.

A correlação e regressão linear pode ser de bastante útil nesse sentido, pois está

intimamente relacionado com o estudo algébrico das funções, mais precisamente das de 1° grau. Ao introduzir o conteúdo das funções, o professor em sala pode apresentar ao aluno situações em que há uma relação entre as variáveis e que podem ou não ser descritas por funções.

Alguns exemplos podem ser citados: quantidade comprada de uma certa mercadoria e o preço a ser pago; a quantidade de energia elétrica/água consumida em um mês e o cobrado na fatura; a remuneração de um trabalhador em razão das horas trabalhadas mensalmente; entre outras.

Em contraponto, podem ser apresentadas outras situações que intuitivamente podem parecer se tratar de funções, mas que não apresentam uma perfeita correspondência, como por exemplo: o peso de um carro e o consumo de combustível; temperatura do ambiente e desempenho de maratonistas; altura dos pais e altura dos filhos; preço de um apartamento e o número de cômodos; quantidade de lixo produzido por uma família e o número de habitantes daquela casa; o orçamento disponível para um filme e a sua correspondente bilheteria; entre outros.

Para o aluno, saber diferenciar o que pode ou não ser tratado como função, exige um pensamento muito além do superficial, e esse processo de raciocínio o leva a aprimorar suas capacidades e compreender, que apesar de não de tratar de uma função propriamente dita, podemos ainda estabelecer uma relação entre as variáveis, que serão objeto de estudo oportunamente, através da correlação e regressão linear.

Ainda nesse sentido, a BNCC traz a seguinte habilidade:

> (EM13MAT510) Investigar conjuntos de dados relativos ao comportamento de duas variáveis numéricas, usando tecnologias da informação, e, se apropriado, levar em conta a variação e utilizar uma reta para descrever a relação observada.

Essa habilidade se compatibiliza com o assunto da regressão linear, tratado na Seção 2.2, e ainda a descreve como um conteúdo compatível a ser estudados nos três anos do EM.

Um excelente auxiliar na análise e interpretação de dados disponíveis, que são

coletados de situações do dia a dia, sejam eles informacionais, numéricos, lógicos é o software Excel (ou seus equivalentes: Calc ou Google Planilhas), amplamente conhecido e de fácil acesso, em que a partir de simples comandos é possível analisar dados fornecidos ou coletados, identificar conclusões válidas, confirmando algo que estava apenas no campo da hipótese, que podem auxiliar na tomada de decisões.

Para melhor visualização do uso da tecnologia da informação, vamos analisar os dados informados em [9], dados estes que foram coletados da Motion Picture Association of America, a respeito do orçamento de filmes e sua respectiva bilheteria.

Primeiramente, devemos transportar os dados para a planilha:

	A	B
1	Orçamento (em milhões US$)	Receita (em milhões US$)
2	62	65
3	90	64
4	50	48
5	35	57
6	200	601
7	100	146
8	90	47

Figura 13 – Tabela de dados: Orçamento e Receita

A partir da tabela, podemos facilmente gerar uma gráfico de dispersão dos dados, selecionando a tabela, inserir, gráfico. É provável que o gráfico gerado já seja o gráfico de dispersão, porém se for mostrado outro, é possível alterar o tipo do gráfico, clicando sobre ele, aparecerá o editor de gráfico e basta ir em configurações.

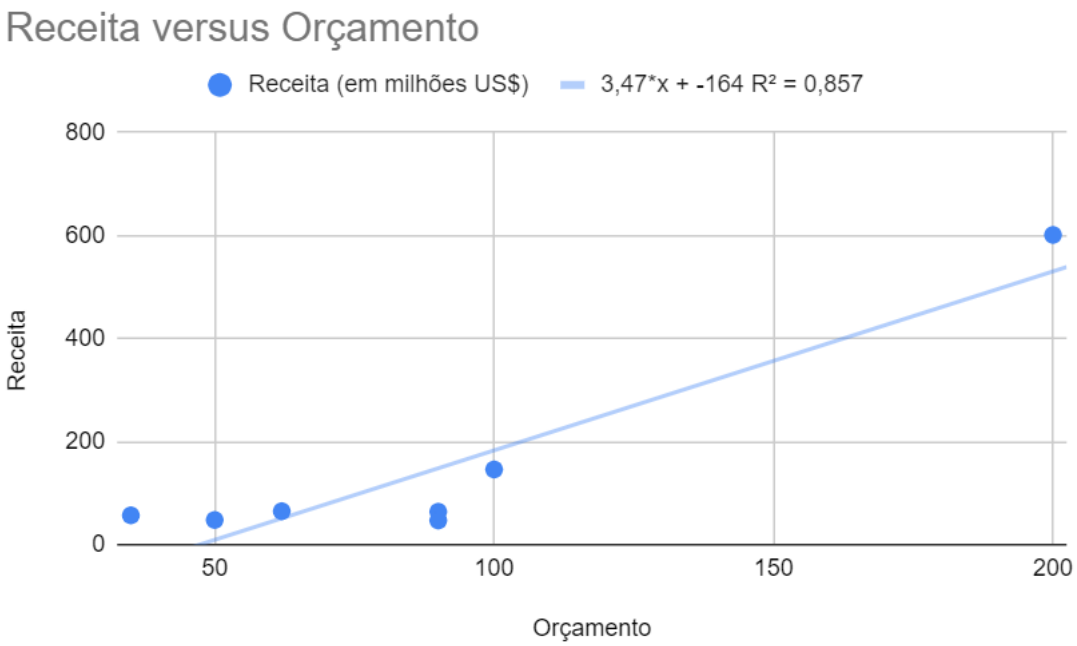


Figura 14 – Gráfico de dispersão: Orçamento e Receita

Note que na Figura 14 já consta a equação da reta de regressão e o r^2 (Coeficiente de Determinação). Esses dados são gerados automaticamente pelo Google Planilhas e não aparecendo automaticamente podem ser inseridos seguindo o seguinte caminho: duplo clique no gráfico, no "Editor de Gráfico", selecionar "Personalizar", opção "Série", em seguida, "Linha de tendência", em "Marcador", escolher a opção "Usar equação". Por fim, para apresentar r^2 basta marcar o quadro correspondente.

Para determinar o Coeficiênte de Pearson, podemos simplesmente fazer a raiz quadrada do r^2 ou ainda, usar um comando bastante simples, sem fazer contas, originário do próprio software, em que basta digitar a função =PEARSON(dados_y; dados_x), selecionando a coluna da variável independente, depois a coluna da variável dependente, separados por ponto e vírgula.

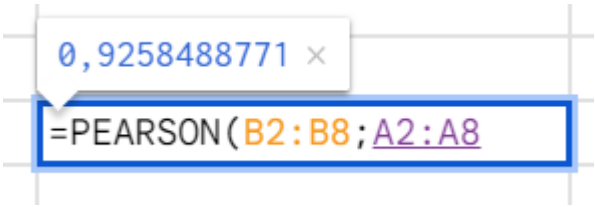


Figura 15 – Coeficiente de Correlação de Pearson para os dados dos filmes

Finalmente usando a tabela constante no Apêndice B, ou seja, o Método 2 apresentado no Capítulo 2, podemos avaliar o Coeficiente de Correlação e verificar se existe uma correlação, que nosso caso, usando essa pequena amostra de apenas 7 filmes, e um nível de significância de 5% é possível afirmar que há evidência suficiente para apoiar a informação de que existe uma correlação linear entre o orçamento de um filme e sua receita.

Ressalta-se que esse exemplo pode ser usado em sala de aula, os alunos podem pesquisar filmes aleatórios, agregar os dados e construir uma tabela semelhante a feita em 13, fazer o gráfico, a reta de regressão, fazer estimativas e comprovar a correlação.

4 Análise dos Microdados ENEM 2022

Utilizando-se dos conhecimentos adquiridos nesse trabalho acerca da correlação e regressão linear e com o auxílio do Software R, vamos analisar os microdados disponibilizados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira - INEP em relação ao Exame Nacional do Ensino Médio - ENEM, aplicado no ano de 2022 e verificar se podemos inferir se existem correlações interessantes e relevantes para estudo.

Os Microdados do ENEM estão disponíveis em (https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem) e fazendo o download é possível acessar os cadernos de provas aplicadas nos dois dias e os respectivos gabaritos, instruções, documentos técnicos, um dicionário que explica as siglas utilizadas nos microdados e os microdados em si, em formato .csv (um formato de arquivo que significa “comma-separated-values”, ou seja, valores separados por vírgulas).[3]

Através desse arquivo é possível acessar a faixa etária, sexo, estado civil, cor/raça, nacionalidade, situação de conclusão do ensino médio, ano de conclusão, tipo de escola do Ensino Médio, tipo de instituição que concluiu ou concluirá o Ensino Médio, se o inscrito fez a prova com intuito de apenas treinar seus conhecimentos, dados da escola, dados do local de aplicação da prova, dados da prova objetiva, dados da redação e ainda dados do questionário socioeconômico.

Esses dados foram devidamente anonimizados, por se tratarem de dados pessoais de pessoas físicas, que por determinação legal, recebem a proteção ao respeito a privacidade, conforme dispõe a Lei 13.709/2018, conhecida como Lei Geral de Proteção de Dados (LGPD). Dessa forma, o INEP reformulou os microdados, ocultando nome e números do CPF e do registro civil, bem como código da escola, a variável idade foi agrupada em "faixa etária", as variáveis “município de nascimento” e “residência do participante” foram retiradas, pois também permitem a identificação pessoal, caso sejam combinadas com outros dados cadastrais.

Por se tratar de um arquivo de um tamanho considerável, o Software utilizado para manipulação dos dados foi o R, ambiente bastante utilizado em análises estatísticas, confecção de gráficos, produção de documentos e relatórios, criação de sites interativos e

aplicativos.

Uma explicação de como fazer o download desse programa compatível com diversos sistemas operacionais pode ser obtido em [8].

Após a instalação deve-se fazer o procedimento de instalar pacotes e carregar o arquivo "MICRODADOS_ENEM_2022"o que pode ser um pouco demorado. Para obter resultados mais práticos, optou-se por restringir a análise dos dados para a cidade de Maringá - PR, visto que foram 3.576.105 inscritos no ENEM 2022 e os que realizaram a prova na cidade de Maringá foram 7.274.

É possível fazer essa limitação utilizando o código do município, a ser obtido no site do IBGE [5] e fazer esse filtro.

A partir disso, podemos escolher variáveis para relacionarmos duas a duas e verificar se possuem correlação.

A princípio foi feita a análise entre a faixa etária dos participantes do exame na cidade de Maringá a sua respectiva nota na prova de Matemática, e o resultado no gráfico de dispersão foi o seguinte:

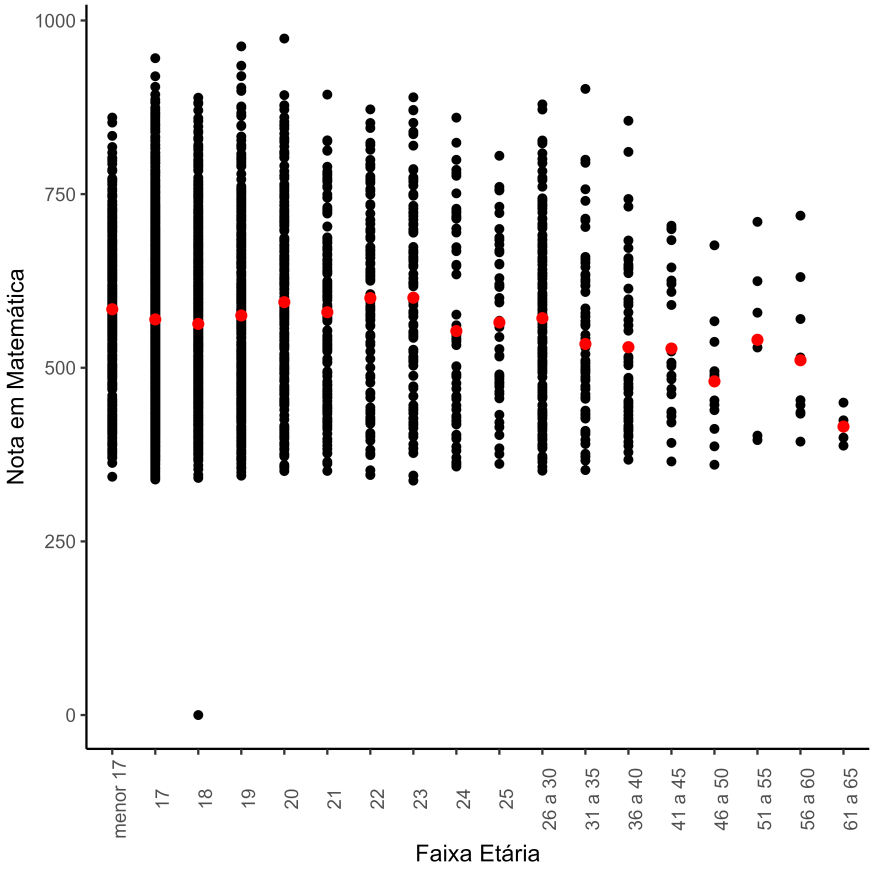


Figura 16 – Gráfico de Dispersão: Faixa Etária x Nota em Matemática

A Figura 16 apresenta os resultados dos estudantes de acordo com a faixa etária.

Para melhor análise utilizamos uma função estatística disponível no software R para resumir os pontos presentes no gráfico por faixa etária utilizando apenas um ponto que representa a média (que é o ponto vermelho).

Visualmente temos um padrão negativo, decrescente, o que indica uma correlação decrescente, em que quanto maior a idade, menor a nota na prova de matemática. Com o auxilio do software R, pode-se extrair algumas informações relevantes: a reta de regressão que melhor representa a relação é $\hat{y} = -1.4940\hat{x} + 576.3825$, o coeficiente de determinação $r^2 = 0.001423$, o que significa que apenas $0,1423\%$ da nota em matemática é explicada pela idade.

Também podemos analisar a relação entre o ano de conclusão do ensino médio e a nota em matemática. Obtivemos o seguinte gráfico:

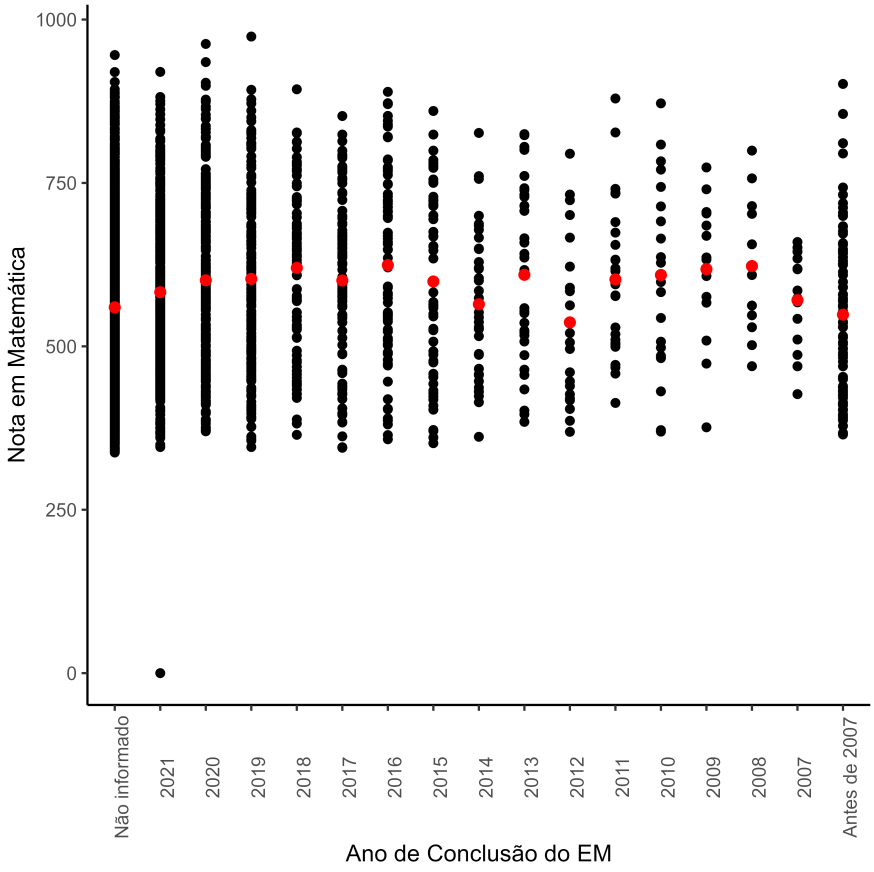


Figura 17 – Gráfico de Dispersão: Ano de Conclusão no EM x Nota em Matemática

A Figura 17 também apresenta as médias (pontos vermelhos) agora em relação ao ano de conclusão dos participantes do ENEM. Visualmente percebe-se uma noção de crescimento, indicando que aqueles que terminaram o Ensino Médio a mais tempo, obtém melhores notas. Vamos verificar essa conjectura.

Fazendo a análise pelo Software R, observou-se que: a reta de regressão que melhor representa a relação é $\hat{y} = 2.5833\hat{x} + 567.3078$, o coeficiente de determinação $r^2 = 0.004708$, o que significa que apenas $0,4708\%$ da nota em matemática é explicada pelo ano de conclusão do EM.

Agora, vamos analisar a correlação entre a resposta do participante ao questionário sobre o grau de instrução do pai ou homem responsável e a sua respectiva nota em matemática.

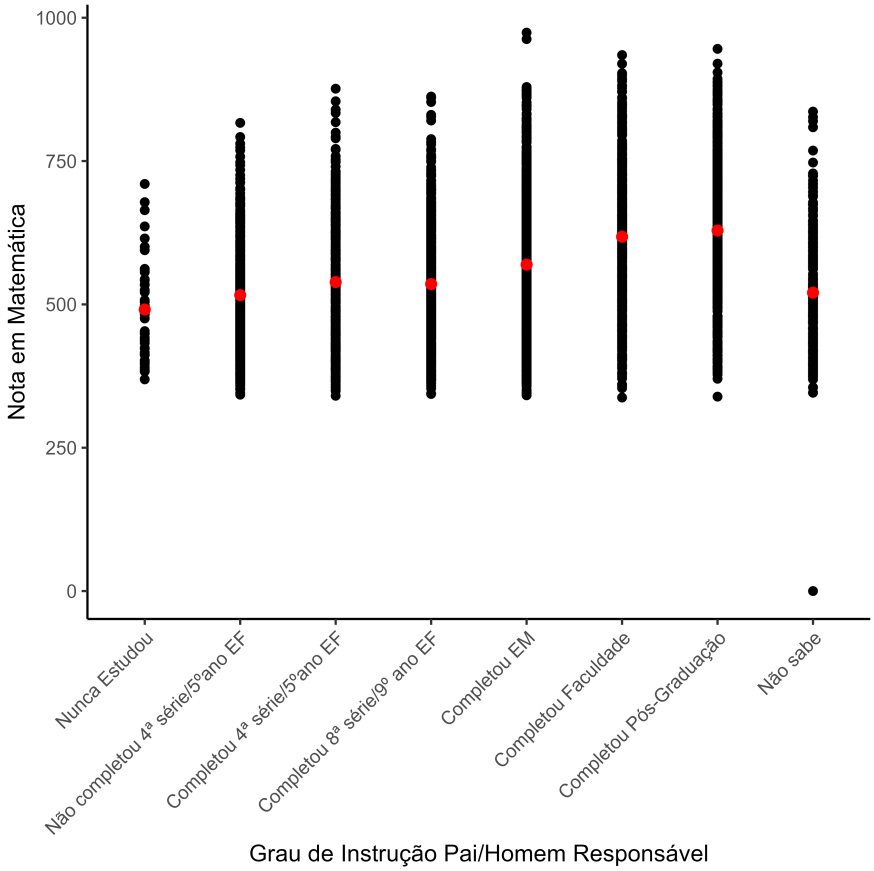


Figura 18 – Gráfico de Dispersão: Grau de Instrução do Pai/Homem Responsável x Nota em Matemática

Podemos visualizar um padrão de crescimento, ignorando o último resultado, em que a resposta foi "Não sabe".

De acordo com o Software R, temos um coeficiente de determinação de $0,112$, o que significa que $11,2\%$ da nota em matemática é explicada pelo nível de estudo do pai/homem responsável.

Outra análise diz respeito ao grau de instrução da mãe ou mulher responsável pelo participante. Temos o seguinte gráfico:

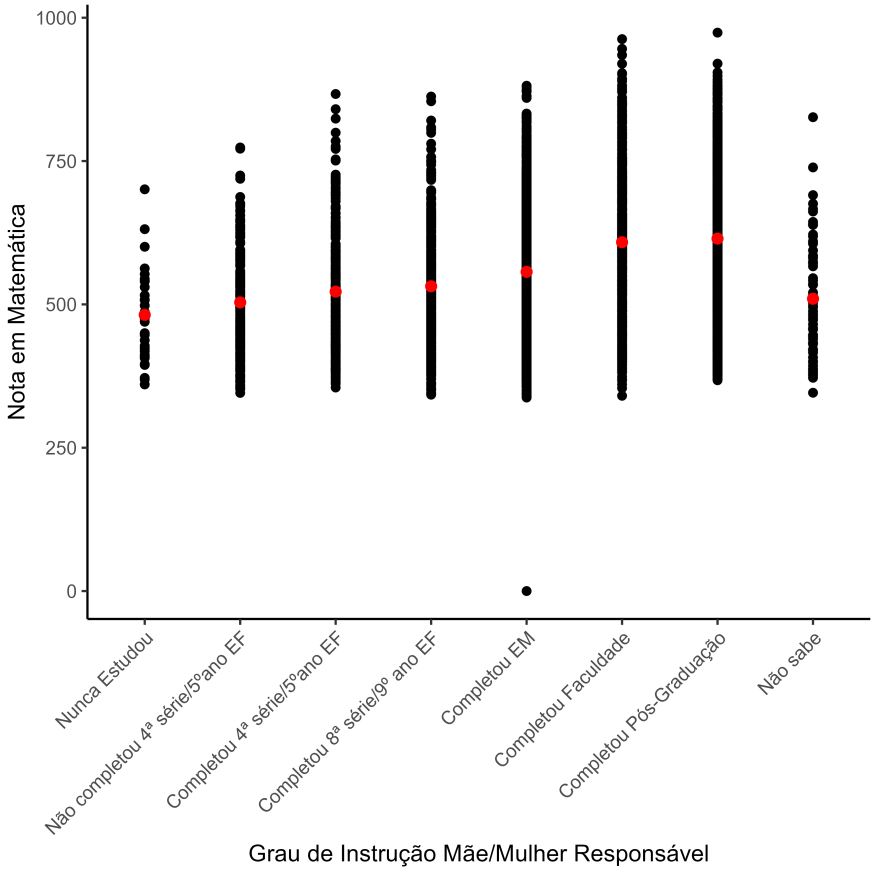


Figura 19 – Gráfico de Dispersão: Grau de Instrução da Mãe/Mulher Responsável x Nota em Matemática

Novamente temos um padrão de crescimento, desconsiderando a última coluna, em que a resposta foi "Não sabe". De acordo com o Software R, temos que o coeficiente de determinação $r^2 = 0,1035$, ou seja, $10,35\%$ do resultado da nota obtida em matemática se explica pelo nível de educação que a mãe ou mulher responsável completou.

Por último, fizemos a análise que relaciona a renda mensal da família e a nota em matemática, conforme a figura a seguir:

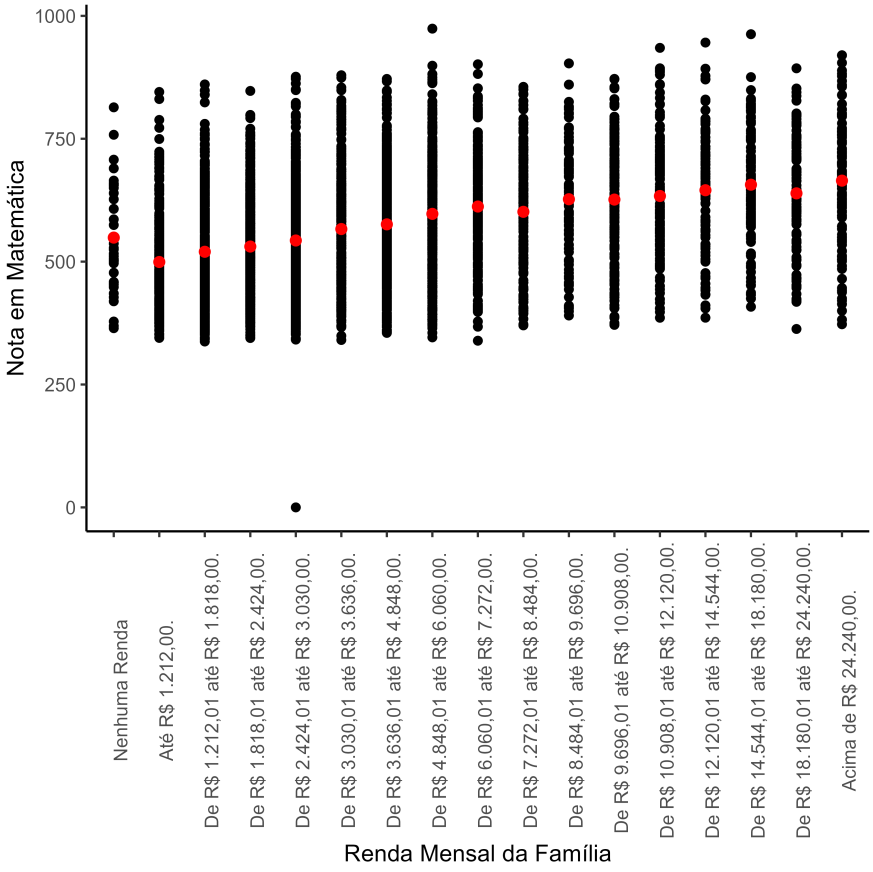


Figura 20 – Gráfico de Dispersão: Renda Mensal x Nota em Matemática

Nessa correlação, temos também visualmente um padrão crescente, o que indica que quanto maior a renda da família do participante, maior a nota obtida. O coeficiente de determinação, de acordo com o R, é $r^2 = 0,1492$, ou seja, $14,92\%$ da nota de matemática é explicada pela renda familiar.

5 CONCLUSÃO

Neste trabalho fizemos uma breve introdução a conceitos básicos da estatística, como gráfico de dispersão, desvio padrão, variância, covariância, z-score, teste de hipóteses, e que foram necessários para melhor entendimento dos assuntos que vieram a seguir, objeto do trabalho.

Em seguida, partimos para uma análise pormenorizada da correlação linear e regressão linear, tratamos dos gráficos de dispersão que nos dão uma indicação de que se trata de uma correlação e pudemos verificar isso adequadamente mediante a realização de testes, análise de coeficiente, tanto de Pearson como de determinação.

Ressalta-se que as fórmulas de Correlação Linear apesar de aparecem nos livros de Estatística, como [9], [7] ou [6], em quatro fórmulas equivalentes, não são realizados aprofundamentos sobre essas versões e aqui, neste trabalho, pudemos aprofundar e explicá-las, de forma mais algébrica, proporcionando um significado mais amplo aos seus objetivos.

Ademais, de acordo com a BNCC, verificamos que a correlação e regressão linear são conhecimentos que podem ser abordados no Ensino Médio através da investigação de dados quanto ao modo como duas variáveis se comportam, observando padrões e regularidades. Ainda, esse assunto tem grande relevância no aprendizado do aluno, que estimula a investigação e questionamento acerca de pensamentos que podem parecer intuitivos mas que podem ser cientificamente comprovados.

Trouxemos ainda, de maneira detalhada, um exemplo de dados que foram analisados por meio do Software Google Planilhas, que sem maiores dificuldades pode ser aplicado em sala de aula com alunos do Ensino Médio, relacionando esse conteúdo com o de Funções do 1° Grau.

Para finalizar, fizemos uma análise dos Microdados do ENEM, utilizando a correlação linear, por meio do Software R, que apresentou correlações significantes entre o grau de escolaridade do pai e da mãe no desempenho do aluno, bem como uma correlação entre a renda mensal familiar e a nota na prova de matemática, em relação os estudantes que realizaram o exame na cidade de Maringá - PR.

Todas essas correlações citadas dão indícios de uma correlação positiva, ou seja,

quanto maior o grau de instrução do pai ou da mãe e maior a renda familiar, melhor foi o desempenho do aluno na prova de matemática, representando um possível somatório de aproximadamente 50% da nota do aluno.

Acrescenta-se que para trabalhos futuros uma outra abordagem a ser feita utilizando esses microdados divulgados pelo INEP ano a ano, pode-se incluir mais cidades, regiões, ou compara-los, utilizar outros coeficientes ou outras regressões como a regressão múltipla, onde podemos incluir em uma mesma equação mais de uma variável independente, e analisar o grau de correlação de cada uma em relação a variável dependente.

Referências

[1] BRASIL. Base Nacional Comum Curricular. Disponível em:<http://portal.mec.gov.br/index.php?option=com_docman&view=download&alias=79601-anexo-texto-bncc-reexportado-pdf-2&category_slug=dezembro-2017-pdf&Itemid=30192>. Acesso em: 10 dez. 2023.

[2] **Drogas - Tabaco - Componentes.** Disponível em: <https://cenpre.furg.br/drogas?id=53>. Acesso em: 10/02/2024.

[3] Entenda o que é o formato CSV e saiba como importar e exportar esses arquivos. Disponível em:<https://rockcontent.com/br/blog/csv/>. Acesso em: 02 jan. 2024.

[4] Gráfico t de Student. Disponível em:<https://www.researchgate.net/figure/Figura-6-Grafico-teste-t-de-Student_fig2_364134380>. Acesso em: 10 jan. 2024.

[5] IBGE. Disponível em:<https://cidades.ibge.gov.br/brasil/pr/maringa/panorama>. Acesso em 09/02/2024.

[6] LARSON, Ron, FARBER, Betsy. **Estatística Aplicada**; tradução Luciane Ferreira Pauleti Vianna. 4. ed. São Paulo: Pearson Prentice Hall, 2010.

[7] MORETTIN, Pedro Alberto. BUSSAB, Wilton O. **Estatística Básica**. 9. ed. São Paulo: Saraiva, 2017.

[8] Projeto Kit Cálculo. Disponível em:<www.dma.uem.br/kit/software>. Acesso em: 09 fev. 2024.

[9] TRIOLA, Mário F. **Introdução à Estatística**. Rio de Janeiro: LTC, 2012.

APÊNDICE A – Tabela Distribuição t: Valores Críticos t

Distribuição t: Valores Críticos t					
	Área em Uma Cauda				
	0.005	**0.01**	**0.025**	**0.05**	**0.10**
	Área em Duas Caudas				
Graus de Liberdade	**0.01**	**0.02**	**0.05**	**0.10**	**0.20**
1	63,657	31,821	12,706	6,314	3,078
2	9,925	6,965	4,303	2,920	1,886
3	5,841	4,541	3,182	2,353	1,638
4	4,604	3,747	2,776	2,132	1,533
5	4,032	3,365	2,571	2,015	1,476
6	3,707	3,143	2,447	1,943	1,440
7	3,499	2,998	2,365	1,895	1,415
8	3,355	2,896	2,306	1,860	1,397
9	3,250	2,821	2,262	1,833	1,383
10	3,169	2,764	2,228	1,812	1,372
11	3,106	2,718	2,201	1,796	1,363
12	3,055	2,681	2,179	1,782	1,356
13	3,012	2,650	2,160	1,771	1,350
14	2,977	2,624	2,145	1,761	1,345
15	2,947	2,602	2,131	1,753	1,341
16	2,921	2,583	2,120	1,746	1,337
17	2,898	2,567	2,110	1,740	1,333
18	2,878	2,552	2,101	1,734	1,330
19	2,861	2,539	2,093	1,729	1,328
20	2,845	2,528	2,086	1,725	1,325
21	2,831	2,518	2,080	1,721	1,323
22	2,819	2,508	2,074	1,717	1,321
23	2,807	2,500	2,069	1,714	1,319

Distribuição t: Valores Críticos t					
	Área em Uma Cauda				
	0.005	**0.01**	**0.025**	**0.05**	**0.10**
	Área em Duas Caudas				
Graus de Liberdade	**0.01**	**0.02**	**0.05**	**0.10**	**0.20**
24	2,797	2,492	2,064	1,711	1,318
25	2,787	2,485	2,060	1,708	1,316
26	2,779	2,479	2,056	1,706	1,315
27	2,771	2,473	2,052	1,703	1,314
28	2,763	2,467	2,048	1,701	1,313
29	2,756	2,462	2,045	1,699	1,311
30	2,750	2,457	2,042	1,697	1,310
31	2,744	2,453	2,040	1,696	1,309
32	2,738	2,449	2,037	1,694	1,309
34	2,728	2,441	2,032	1,691	1,307
36	2,719	2,434	2,028	1,688	1,306
38	2,712	2,429	2,024	1,686	1,304
40	2,704	2,423	2,021	1,684	1,303
45	2,690	2,412	2,014	1,679	1,301
50	2,678	2,403	2,009	1,676	1,299
55	2,668	2,396	2,004	1,673	1,297
60	2,660	2,390	2,000	1,671	1,296
65	2,654	2,385	1,997	1,669	1,295
70	2,648	2,381	1,994	1,667	1,294
75	2,643	2,377	1,992	1,665	1,293
80	2,639	2,374	1,990	1,664	1,292
90	2,632	2,368	1,987	1,662	1,291
100	2,626	2,364	1,984	1,660	1,290
200	2,601	2,345	1,972	1,653	1,286
300	2,592	2,339	1,968	1,650	1,284
400	2,588	2,336	1,966	1,649	1,284
500	2,586	2,334	1,965	1,648	1,283

Distribuição t: Valores Críticos t					
	Área em Uma Cauda				
	0.005	**0.01**	**0.025**	**0.05**	**0.10**
	Área em Duas Caudas				
Graus de Liberdade	**0.01**	**0.02**	**0.05**	**0.10**	**0.20**
750	2,582	2,331	1,963	1,647	1,283
1000	2,581	2,330	1,962	1,646	1,282
2000	2,578	2,328	1,961	1,646	1,282
Grande	2,576	2,326	1,960	1,645	1,282

APÊNDICE B – Valores Críticos do Coeficiente de Correlação de Pearson r

Valores Críticos do Coeficiente de Correlação de Pearson r		
n	$\alpha = 0{,}05$	$\alpha = 0{,}01$
4	0,950	0,999
5	0,878	0,959
6	0,811	0,917
7	0,754	0,875
8	0,707	0,834
9	0,666	0,798
10	0,632	0,765
11	0,602	0,735
12	0,576	0,708
13	0,553	0,684
14	0,532	0,661
15	0,514	0,641
16	0,497	0,623
17	0,482	0,606
18	0,468	0,590
19	0,456	0,575
20	0,444	0,561
25	0,396	0,505
30	0,361	0,463
35	0,335	0,430
40	0,312	0,402
45	0,294	0,378
50	0,279	0,361
60	0,254	0,330
70	0,236	0,305

80	0,220	0,286
90	0,207	0,269
100	0,196	0,256