



UNIVERSIDADE ESTADUAL PAULISTA “JÚLIO DE MESQUITA FILHO”
Instituto de Geociências e Ciências Exatas
Campus de Rio Claro

Método dos mínimos quadrados aplicado ao lançamento de foguetes propulsionados a ar comprimido

Gilberto Caetano da Silva Junior

Dissertação apresentada ao Programa de Pós-Graduação – Mestrado Profissional em Matemática em Rede Nacional-PROFMAT como requisito parcial para a obtenção do grau de Mestre

Orientadora
Profa. Dra. Sidineia Barrozo

2017

111 Silva Junior, Gilberto Caetano da
X111x Método dos mínimos quadrados aplicado ao lançamento de fo-
 guetes propulsionados a ar comprimido/ Gilberto Caetano da Silva
 Junior- Rio Claro: [s.n.], 2017.
 107 f.: fig., tab.

 Dissertação (mestrado) - Universidade Estadual Paulista, Insti-
 tuto de Geociências e Ciências Exatas.
 Orientadora: Sidineia Barrozo

 1. Regressão linear. 2. Ensino de matemática. 3. Lançamento
 de foguetes. 4. Propulsão a ar comprimido. I. Título

TERMO DE APROVAÇÃO

Gilberto Caetano da Silva Junior

MÉTODO DOS MÍNIMOS QUADRADOS APLICADO AO LANÇAMENTO DE
FOGUETES PROPULSIONADOS A AR COMPRIMIDO

Dissertação APROVADA como requisito parcial para a obtenção do grau de Mestre no Curso de Pós-Graduação Mestrado Profissional em Matemática Universitária do Instituto de Geociências e Ciências Exatas da Universidade Estadual Paulista “Júlio de Mesquita Filho”, pela seguinte banca examinadora:

Profa. Dra. Sidineia Barrozo
Orientadora

Prof. Dra. Eliris Cristina Rizzioli
Departamento de Matemática, UNESP - Rio Claro

Prof. Dra. Marisa Veiga Capela
Departamento de Físico-Química, UNESP - Araraquara

Rio Claro, 18 de outubro de 2017

Aos meus familiares e amigos.

Agradecimentos

Agradeço inicialmente a Deus, pois nos momentos de dificuldades e dúvidas renovei minha perseverança na tua palavra.

À minha esposa Cláudia e minha filha Larissa pelo companherismo, compreensão e apoio.

À SBM - Sociedade Brasileira de Matemática por estruturar um programa de Pós-Graduação que viabiliza o acesso de muitos professores que anseiam por capacitação e melhoramento profissional.

À CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, por disponibilizar bolsa de estudo, sem a qual não seria possível diminuir a carga horária de trabalho para que pudesse ter uma maior dedicação aos estudos.

À minha orientadora Professora Dra. Sidneia Barrozo pelos ensinamentos, sugestões e correções que nortearam a realização deste trabalho.

A todos os professores do Departamento de Matemática da Unesp de Rio Claro, em especial à Professora Dra. Suzinei Aparecida Siqueira Marconato pelo apoio e incentivo, ao Professor Dr. Jamil Viana Pereira pela paciência e ensinamentos imprescindíveis a minha formação e à Professora Dra. Elíris Cristina Rizziolli pelo incentivo e generosidade.

Aos meus colegas de curso: Helba, Henrique, Ênio, Joyce e Irma que estiveram presentes e me ajudaram nessa caminhada que empreendi.

A minha colega de trabalho, professora Katia Lucas.

“Tenho a impressão de ter sido somente um garoto brincando e me divertindo na praia, encontrando, de vez em quando, um seixo mais liso ou uma concha mais bonita que a normal, enquanto que o grande oceano da verdade permanece todo desconhecido diante de mim.”

Isaac Newton.

Resumo

Neste trabalho, apresentamos um relato de experimento realizado junto aos alunos de ensino fundamental de uma escola pública municipal e efetuamos o ajuste de curva dos dados observados por meio do método dos mínimos quadrados. Para tanto, discutimos a concepção e aplicação desse método a partir de resultados oriundos do cálculo diferencial, da álgebra linear e alguns conceitos estatísticos. Do cálculo diferencial estudamos a minimização dos erros de aproximação por meio da investigação dos pontos de mínimo da função erro. Da álgebra linear determinamos os parâmetros da função ajustada através da discussão e solução de um sistema de equações lineares resultante do conjunto de derivadas parciais nulas que estabelecem o ponto crítico da função erro. Da estatística utilizamos alguns conceitos e formulações que tratam da intensidade da relação entre as variáveis, bem como, das incertezas na variável dependente e nos parâmetros da função ajustada.

Palavras-chave: Regressão linear, Ensino de matemática, Lançamento de foguetes, Propulsão a ar comprimido.

Abstract

In this work, we present a report of an experiment carried out with elementary school students of a municipal public school, and we performed the curve adjustment of the observed data through the least squares method. For this, we discuss the conception and application of this method from results derived from differential calculus, linear algebra and some statistical concepts. From the differential calculation, we study the minimization of approximation errors by investigating the minimum points of the error function. From linear algebra, we determine the parameters of the adjusted function through the discussion and solution of a system of linear equations resulting from the set of null partial derivatives that establish the critical point of the error function. From statistics, we use some concepts and formulations that deal with the intensity of the relationship between variables, as well as the uncertainties in the dependent variable and the parameters of the adjusted function.

Keywords: Linear regression, Mathematics teaching, Rocket launch, Compressed air propulsion.

Lista de Figuras

1.1	Ponto de mínimo	27
1.2	Ponto de mínimo e sela	33
1.3	Região convexa e não convexa	34
1.4	Função convexa e segmento secante	35
1.5	Plano tangente ao gráfico e vetor gradiente perpendicular	36
3.1	Dispersão das medidas observadas em relação à média	68
3.2	Gráfico de dispersão dos pontos experimentais	70
3.3	Gráficos de funções polinomiais ajustadas	71
3.4	Representação dos erros (desvios verticais) e_i	71
3.5	Distribuição de frequência dos valores listados na tabela 2.3	74
3.6	Tendência da distribuição de frequência	75
3.7	Aproximação do histograma à distribuição gaussiana	76
3.8	Distribuição gaussiana	78
3.9	Área sob a curva normal entre a média e z	80
3.10	Barras de incerteza	82
4.1	Lançamento do melhor foguete	94
4.2	Gráfico de dispersão dos pares ordenados (x,y)	96
4.3	Gráfico de dispersão e linha de tendência	97

Lista de Tabelas

3.1	Peso da saca de farinha em três medidas	67
3.2	Média, mediana, desvios absolutos e quadrado dos desvios dos pesos da saca de farinha em estudo.	69
3.3	Série de medições, intervalo de valores e frequências relativas.	73
4.1	Desempenho do foguete em função do número de bombeadas	95
4.2	Tabela de somatório dos dados do experimento	99
4.3	Tabela resumida de somatório dos dados do experimento	102
4.4	Tabela de alguns somatórios dos dados do experimento, dos desvios em relação a função e à média, do número de observações e do número de parâmetros da função ajustada	103

Sumário

1	Noções de cálculo	11
2	Noções de álgebra linear	47
3	Método dos mínimos quadrados	66
3.1	Breve histórico	66
3.2	Concepção do método	67
3.2.1	Abordagem não probabilística	68
3.2.2	Abordagem probabilística	72
3.3	Problema de mínimos quadrados	81
3.4	Avaliação do ajuste	90
4	O experimento	92
4.1	Motivação para realização do experimento	92
4.2	Relato do experimento	94
4.3	Ajuste de curva	101
5	Considerações finais	105
	Referências	106

Introdução

Ao realizarmos um experimento, as observações e medições resultantes da coleta de dados estão frequentemente sujeitas a erros aleatórios e incontrolláveis. Então, para que possamos obter uma estimativa confiável das grandezas que estamos analisando, devemos tentar estabelecer uma relação de causa e efeito entre as variáveis representativas dessas grandezas.

Nesse sentido, a formulação de um modelo matemático pode ser o primeiro passo na tentativa de explicar valores de uma variável em termos de outra. E o ajuste dos parâmetros do modelo aos dados experimentais possibilitam validá-lo, ou seja, dão a medida do quanto o modelo representa, de fato, o fenômeno que está sendo estudado.

Um dos métodos mais conhecidos para ajuste de parâmetros é o método dos mínimos quadrados, que em síntese, consiste em determinar os parâmetros que minimizam a diferença entre os dados experimentais e o modelo teórico.

Nessa perspectiva, e considerando um experimento realizado junto aos alunos do ensino fundamental de uma escola pública municipal, discutiremos nesse trabalho a concepção e aplicação desse método a partir de resultados oriundos do cálculo diferencial, da álgebra linear e de alguns conceitos estatísticos.

Assim, nos capítulos 1 e 2 apresentaremos noções de cálculo e álgebra linear cujo encadeamento das definições e teoremas fundamentam a discussão e aplicação do método dos mínimos quadrados. Iniciaremos o capítulo 3 com um breve histórico sobre o surgimento do método, a fim de entendermos os fatores que motivaram o seu desenvolvimento. Em seguida, discutiremos sua concepção sob dois pontos de vista: o não probabilístico, onde buscamos entender a métrica do método sem nos preocuparmos com o caráter aleatório dos dados observados; e o probabilístico, onde a aleatoriedade dos dados experimentais é imprescindível para determinarmos as melhores chances para as estimativas dadas pelo método. Descreveremos ainda, o problema de mínimos quadrados considerando as incertezas inerentes ao ajuste e um critério para avaliação de sua qualidade.

No capítulo 4, apresentaremos a motivação para a realização do experimento, o seu delineamento, o tratamento dos dados observados, as atividades realizadas pelos alunos como consequência do experimento e, finalizaremos com o ajuste de curva dos dados observados de acordo com os estudos realizados nos capítulos anteriores.

1 Noções de cálculo

Como veremos no capítulo 3, o método dos mínimos quadrados trabalha com minimização de erros na predição de dados experimentais. Do ponto de vista do cálculo, a minimização do erro implica em determinar uma função cujos pontos críticos sejam pontos de mínimo. Assim, este capítulo tem por objetivo descrever os resultados do cálculo que são relevantes para a determinação dos pontos críticos de tal função.

Estamos interessados em uma função polinomial de duas variáveis, então observemos as seguintes definições:

Definição 1.1. *Uma bola aberta de raio r centrada em um ponto $A = (x_0, y_0) \in \mathbb{R}^2$, é o conjunto de todos os pontos $P = (x, y) \in \mathbb{R}^2$ tal que $\sqrt{(x - x_0)^2 + (y - y_0)^2} < r$.*

Exemplo 1.2. O conjunto $\{(x, y) \in \mathbb{R}^2; x^2 + y^2 < r^2\}$ é uma bola aberta em \mathbb{R}^2 pertencente ao interior de uma circunferência.

Definição 1.3. *Seja D um conjunto de pares ordenados $(x, y) \in \mathbb{R}^2$. Uma função real f de duas variáveis em D é uma correspondência que associa um único número real $z = f(x, y)$ a cada par ordenado (x, y) em D .*

Definição 1.4. *O conjunto D , denominado domínio de f , é o maior subconjunto de \mathbb{R}^2 para o qual faz sentido a correspondência em questão, e o conjunto de valores de z assumidos por f é um subconjunto de \mathbb{R} denominado imagem de f , e dado por $Im(f) = \{z = f(x, y); (x, y) \in D\}$.*

Exemplo 1.5. Seja a função de duas variáveis $f(x, y) = \sqrt{25 - x^2 - y^2}$. O domínio de f é o conjunto de todos os pares ordenados (x, y) para os quais $25 - x^2 - y^2 \geq 0$, ou seja, $D = \{(x, y) \in \mathbb{R}^2; x^2 + y^2 \leq 25\}$. Observamos que $0 \leq f(x, y) \leq 5$ e, portanto, a imagem de f é o conjunto de todos os números reais no intervalo fechado $[0, 5]$, isto é, $Im(f) = \{z \in \mathbb{R}; 0 \leq z \leq 5\}$.

Definição 1.6. *Dizemos que $f : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$ é uma função polinomial de duas variáveis e grau n quando pode ser escrita na forma*

$$f(x, y) = a_0x^n + a_1x^{n-1}y + \cdots + a_{n-1}xy^{n-1} + a_ny^n = \sum_{i=0}^n a_i x^{n-i} y^i,$$

onde n é um número inteiro não negativo e a_0, a_1, \dots, a_n são números reais.

Tendo em vista as definições descritas anteriormente e tomando uma função polinomial $f : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$, dada por $z = f(x, y)$, queremos garantir que tal função seja diferenciável em todo ponto de \mathbb{R}^2 . Para tanto, vamos descrever e discutir as seguintes definições e teoremas:

Definição 1.7. (Limite) *Seja $f : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$ uma função definida sobre uma bola aberta e L um número real. Dizemos que existe o limite de $f(x, y)$ quando (x, y) tende (x_0, y_0) e este limite é igual a L , o que denotamos por $\lim_{(x,y) \rightarrow (x_0,y_0)} f(x, y) = L$, se e somente se, para todo número real $\varepsilon > 0$ for possível encontrar um número real $\delta > 0$, tal que $|f(x, y) - L| < \varepsilon$, sempre que $(x, y) \in D$ e $0 < \sqrt{(x - x_0)^2 + (y - y_0)^2} < \delta$.*

Teorema 1.8. *Seja a função f definida sobre uma bola aberta e na vizinhança do ponto (x_0, y_0) , exceto possivelmente no próprio ponto (x_0, y_0) , e uma constante real c . Assim:*

- a) se $f(x, y) = c$, então $\lim_{(x,y) \rightarrow (x_0,y_0)} f(x, y) = c$;
- b) se $f(x, y) = x$, então $\lim_{(x,y) \rightarrow (x_0,y_0)} f(x, y) = x_0$;
- c) se $f(x, y) = y$, então $\lim_{(x,y) \rightarrow (x_0,y_0)} f(x, y) = y_0$.

Demonstração: a) Dado $\varepsilon > 0$ arbitrário, queremos mostrar que existe $\delta > 0$ tal que se

$$0 < \sqrt{(x - x_0)^2 + (y - y_0)^2} < \delta,$$

então

$$|f(x, y) - c| < \varepsilon.$$

De fato, para $f(x, y) = c$, temos que

$$|f(x, y) - c| = |c - c| = 0 < \varepsilon, \quad \forall (x, y) \in \mathbb{R}^2.$$

Logo,

$$\lim_{(x,y) \rightarrow (x_0,y_0)} f(x, y) = c, \quad \forall (x_0, y_0) \in \mathbb{R}^2.$$

b) Dado $\varepsilon > 0$ qualquer, queremos mostrar que existe $\delta = \varepsilon$, tal que se

$$0 < \sqrt{(x - x_0)^2 + (y - y_0)^2} < \delta,$$

então

$$|f(x, y) - x| \leq \varepsilon.$$

De fato, tomando $f(x, y) = x$, temos que

$$|f(x, y) - x_0| = |x - x_0| = \sqrt{(x - x_0)^2} \leq \sqrt{(x - x_0)^2 + (y - y_0)^2} < \delta = \varepsilon.$$

Portanto,

$$\lim_{(x,y) \rightarrow (x_0,y_0)} f(x, y) = x_0.$$

c) Seja $f(x, y) = y$. De maneira análoga ao que foi feito no item (b), temos que

$$\lim_{(x,y) \rightarrow (x_0,y_0)} f(x, y) = y_0.$$

■

Teorema 1.9. (Propriedades do limite) *Sejam as funções f e g definidas sobre uma bola aberta e na vizinhança do ponto (x_0, y_0) , exceto possivelmente no próprio ponto (x_0, y_0) , e C uma constante. Se $\lim_{(x,y) \rightarrow (x_0,y_0)} f(x, y) = L$ e $\lim_{(x,y) \rightarrow (x_0,y_0)} g(x, y) = M$, então:*

$$1) \quad \lim_{(x,y) \rightarrow (x_0,y_0)} [f(x, y) + g(x, y)] = L + M;$$

$$2) \quad \lim_{(x,y) \rightarrow (x_0,y_0)} f(x, y) \cdot g(x, y) = L \cdot M;$$

$$3) \quad \lim_{(x,y) \rightarrow (x_0,y_0)} C \cdot f(x, y) = C \cdot L$$

Demonstração: (Lima[13])

1) Pela definição 1.7, demonstrar que

$$\lim_{(x,y) \rightarrow (x_0,y_0)} [f(x, y) + g(x, y)] = L + M,$$

implica em demonstrar que dado $\varepsilon > 0$, existe $\delta > 0$ tal que:

$$|[f(x, y) + g(x, y)] - (L + M)| < \varepsilon$$

sempre que

$$0 < \sqrt{(x - x_0)^2 + (y - y_0)^2} < \delta.$$

Como por hipótese, existem os limites

$$\lim_{(x,y) \rightarrow (x_0,y_0)} f(x, y) = L \quad e \quad \lim_{(x,y) \rightarrow (x_0,y_0)} g(x, y) = M,$$

os valores

$$|f(x, y) - L| \quad e \quad |g(x, y) - M|$$

podem tornar-se arbitrariamente pequenos quando escolhermos (x, y) suficientemente próximo de (x_0, y_0) . Em particular, podem tornar-se menores que $\frac{\varepsilon}{2}$. Desta forma temos que

$$|f(x, y) - L| < \frac{\varepsilon}{2}$$

sempre que

$$0 < \sqrt{(x - x_0)^2 + (y - y_0)^2} < \delta_1$$

e

$$|g(x, y) - M| < \frac{\varepsilon}{2}$$

sempre que

$$0 < \sqrt{(x - x_0)^2 + (y - y_0)^2} < \delta_2.$$

Tomando $\delta = \min\{\delta_1, \delta_2\}$, temos que:

$$|f(x, y) - L| < \frac{\varepsilon}{2}$$

e

$$|g(x, y) - M| < \frac{\varepsilon}{2}$$

sempre que

$$0 < \sqrt{(x - x_0)^2 + (y - y_0)^2} < \delta.$$

Usando as propriedades associativa e comutativa da adição, bem como a desigualdade triangular, temos:

$$\begin{aligned} |[f(x, y) + g(x, y) - (L + M)]| &= |[f(x, y) - L] + [g(x, y) - M]| \leq \\ &\leq |f(x, y) - L| + |g(x, y) - M| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon \end{aligned}$$

sempre que

$$0 < \sqrt{(x - x_0)^2 + (y - y_0)^2} < \delta.$$

Isso mostra que

$$\lim_{(x, y) \rightarrow (x_0, y_0)} [f(x, y) + g(x, y)] = L + M.$$

2) Para demonstrar que

$$\lim_{(x, y) \rightarrow (x_0, y_0)} f(x, y) \cdot g(x, y) = L \cdot M,$$

vamos inicialmente considerar o seguinte caso particular:

sejam as funções $f(x, y)$ e $h(x, y)$, tais que

$$\lim_{(x, y) \rightarrow (x_0, y_0)} f(x, y) = L$$

e

$$\lim_{(x, y) \rightarrow (x_0, y_0)} h(x, y) = 0.$$

Queremos demonstrar que

$$\lim_{(x, y) \rightarrow (x_0, y_0)} f(x, y) \cdot h(x, y) = 0,$$

ou seja, dado $\varepsilon > 0$, existe $\delta > 0$ tal que

$$|f(x, y) \cdot h(x, y)| < \varepsilon$$

sempre que

$$0 < \sqrt{(x - x_0)^2 + (y - y_0)^2} < \delta.$$

Assim, tomando $\varepsilon = 1$, temos que

$$|f(x, y) - L| < 1$$

sempre que

$$0 < \sqrt{(x - x_0)^2 + (y - y_0)^2} < \delta_1.$$

Mas,

$$|f(x, y)| = |f(x, y) - L + L| \leq |f(x, y) - L| + |L| < 1 + |L|.$$

Então,

$$|f(x, y)| < 1 + |L|$$

sempre que

$$0 < \sqrt{(x - x_0)^2 + (y - y_0)^2} < \delta_1.$$

Por outro lado, para

$$\lim_{(x, y) \rightarrow (x_0, y_0)} h(x, y) = 0,$$

temos que $|h(x, y)|$ pode tornar-se suficientemente pequeno quando aproximamos (x, y) de (x_0, y_0) , de tal forma que

$$|h(x, y)| < \frac{\varepsilon}{(1 + |L|)}$$

sempre que

$$0 < \sqrt{(x - x_0)^2 + (y - y_0)^2} < \delta_2.$$

Portanto, se tomarmos $\delta = \min\{\delta_1, \delta_2\}$, temos que

$$|f(x, y) \cdot h(x, y)| = |f(x, y)| \cdot |h(x, y)| < (1 + |L|) \cdot |h(x, y)| < (1 + |L|) \cdot \frac{\varepsilon}{(1 + |L|)} = \varepsilon$$

sempre que

$$0 < \sqrt{(x - x_0)^2 + (y - y_0)^2} < \delta.$$

Isso mostra que

$$\lim_{(x, y) \rightarrow (x_0, y_0)} f(x, y) \cdot h(x, y) = 0.$$

Vamos agora demonstrar que

$$\lim_{(x, y) \rightarrow (x_0, y_0)} f(x, y) \cdot g(x, y) = L \cdot M.$$

Isto é, dado $\varepsilon > 0$, existe $\delta > 0$ tal que

$$|[f(x, y) \cdot g(x, y)] - L \cdot M| < \varepsilon$$

sempre que

$$0 < \sqrt{(x - x_0)^2 + (y - y_0)^2} < \delta.$$

Temos que

$$\begin{aligned} [f(x, y) \cdot g(x, y) - L \cdot M] &= f(x, y) \cdot g(x, y) - f(x, y) \cdot M + f(x, y) \cdot M - L \cdot M = \\ &= f(x, y) \cdot [g(x, y) - M] + M \cdot [f(x, y) - L]. \end{aligned}$$

Como

$$\lim_{(x,y) \rightarrow (x_0, y_0)} f(x, y) = L$$

e

$$\lim_{(x,y) \rightarrow (x_0, y_0)} g(x, y) = M,$$

segue que

$$\lim_{(x,y) \rightarrow (x_0, y_0)} [f(x, y) - L] = 0 = \lim_{(x,y) \rightarrow (x_0, y_0)} [g(x, y) - M].$$

Portanto, pelos resultados do caso particular, temos que:

$$\lim_{(x,y) \rightarrow (x_0, y_0)} f(x, y) \cdot [g(x, y) - M] = 0 = \lim_{(x,y) \rightarrow (x_0, y_0)} M \cdot [f(x, y) - L].$$

Tomando os resultados obtidos na demonstração do item (1), temos que

$$\lim_{(x,y) \rightarrow (x_0, y_0)} [f(x, y) \cdot g(x, y) - L \cdot M] = 0,$$

ou seja,

$$\lim_{(x,y) \rightarrow (x_0, y_0)} [f(x, y) \cdot g(x, y)] = L \cdot M.$$

Como queríamos demonstrar.

- 3) Para demonstração de $\lim_{(x,y) \rightarrow (x_0, y_0)} C \cdot f(x, y) = C \cdot L$ usamos raciocínio análogo ao do item (2).

■

Definição 1.10. (Continuidade) *Seja f uma função definida sobre uma bola aberta contendo (x_0, y_0) . Dizemos que f é contínua em (x_0, y_0) se existe $\lim_{(x,y) \rightarrow (x_0, y_0)} f(x, y)$ e $\lim_{(x,y) \rightarrow (x_0, y_0)} f(x, y) = f(x_0, y_0)$. Dizemos que f é contínua num conjunto D , se ela for contínua em todos os pontos desse conjunto.*

Teorema 1.11. (Propriedades da continuidade) *Se as funções f e g são contínuas no ponto (x_0, y_0) e C é uma constante, então $(C \cdot f)$, $(f + g)$ e $(f \cdot g)$ também são contínuas em (x_0, y_0) .*

Demonstração: (Lima [13]) Tomando $f(x, y)$ e $g(x, y)$ contínuas em (x_0, y_0) , então, da definição 1.10 temos que

$$\lim_{(x,y) \rightarrow (x_0, y_0)} f(x, y) = f(x_0, y_0)$$

e

$$\lim_{(x,y) \rightarrow (x_0,y_0)} g(x,y) = g(x_0,y_0).$$

Dos itens (1), (2) e (3) do teorema 1.9, notamos que

$$\begin{aligned} \lim_{(x,y) \rightarrow (x_0,y_0)} [f(x,y) + g(x,y)] &= f(x_0,y_0) + g(x_0,y_0), \\ \lim_{(x,y) \rightarrow (x_0,y_0)} [f(x,y) \cdot g(x,y)] &= f(x_0,y_0) \cdot g(x_0,y_0), \end{aligned}$$

e

$$\lim_{(x,y) \rightarrow (x_0,y_0)} [C \cdot f(x,y)] = C \cdot f(x_0,y_0).$$

Logo:

$$[f(x,y) + g(x,y)], [f(x,y) \cdot g(x,y)] \quad e \quad [C \cdot f(x,y)]$$

são contínuas em (x_0, y_0) . ■

Teorema 1.12. *Polinômios nas variáveis x e y são funções contínuas em todo ponto de \mathbb{R}^2 .*

Demonstração: (Lima [13]): Dos itens (1) e (2) do teorema 1.9, segue por indução que se C_1, C_2, \dots, C_n são constantes e $f_1(x,y), f_2(x,y), \dots, f_n(x,y)$ são funções tais que

$\lim_{(x,y) \rightarrow (x_0,y_0)} f_i(x,y), \{i = 1, 2, \dots, n\}$ existem, então

$$\lim_{(x,y) \rightarrow (x_0,y_0)} \left[\sum_{i=1}^n C_i \cdot f_i(x,y) \right] = \sum_{i=1}^n C_i \cdot \left[\lim_{(x,y) \rightarrow (x_0,y_0)} f_i(x,y) \right]. \quad (1.1)$$

Do item (2) do teorema 1.9, segue por indução que:

$$\lim_{(x,y) \rightarrow (x_0,y_0)} [f_1(x,y) \cdots f_n(x,y)] = \left[\lim_{(x,y) \rightarrow (x_0,y_0)} f_1(x,y) \right] \cdots \left[\lim_{(x,y) \rightarrow (x_0,y_0)} f_n(x,y) \right]. \quad (1.2)$$

Do teorema 1.8, itens (b) e (c) e da equação (1.2), temos que se m e n são inteiros não negativos, então:

$$\lim_{(x,y) \rightarrow (x_0,y_0)} x^m = \left[\lim_{(x,y) \rightarrow (x_0,y_0)} x \right] \cdots \left[\lim_{(x,y) \rightarrow (x_0,y_0)} x \right] = x_0^m \quad (1.3)$$

e

$$\lim_{(x,y) \rightarrow (x_0,y_0)} y^n = \left[\lim_{(x,y) \rightarrow (x_0,y_0)} y \right] \cdots \left[\lim_{(x,y) \rightarrow (x_0,y_0)} y \right] = y_0^n. \quad (1.4)$$

Do item (2) do teorema 1.9, das equações (1.3) e (1.4), concluímos que se m, n são inteiros não negativos, então

$$\lim_{(x,y) \rightarrow (x_0,y_0)} x^m \cdot y^n = x_0^m \cdot y_0^n. \quad (1.5)$$

Das equações (1.1) e (1.5), concluímos que se $f(x,y)$ é um polinômio, então

$$\lim_{(x,y) \rightarrow (x_0,y_0)} f(x,y) = f(x_0,y_0). \quad (1.6)$$

Portanto, do teorema 1.11 e da equação (1.6) segue que polinômios nas variáveis x e y são funções contínuas em todo ponto de \mathbb{R}^2 . ■

Exemplo 1.13. Seja a função $f : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$, dada por

$$f(x, y) = x^4 + 5x^3y^2 + 6xy^4 + 6.$$

Temos que $f(x, y)$ é uma função polinomial, daí segue do teorema 1.12 e da equação (1.6) que

$$\lim_{(x,y) \rightarrow (x_0,y_0)} f(x, y) = f(x_0, y_0),$$

qualquer $(x_0, y_0) \in D$. Ou seja, $f(x, y)$ é contínua em todo ponto (x_0, y_0) em \mathbb{R}^2 .

Definição 1.14. Seja uma função $f : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$, dada por $z = f(x, y)$. Definimos a derivada parcial de f em relação a x por

$$\frac{\partial f}{\partial x}(x, y) = \lim_{\Delta x \rightarrow 0} \frac{f(x + \Delta x, y) - f(x, y)}{\Delta x}$$

se tal limite existir. Analogamente a derivada parcial de f em relação a y é definida por

$$\frac{\partial f}{\partial y}(x, y) = \lim_{\Delta y \rightarrow 0} \frac{f(x, y + \Delta y) - f(x, y)}{\Delta y}$$

se tal limite existir.

Exemplo 1.15. Seja a função $f(x, y) = 5x^2 + 2xy + 2y^2$. Queremos encontrar $\frac{\partial f}{\partial x}(x, y)$ e $\frac{\partial f}{\partial y}(x, y)$. Segue que:

$$\begin{aligned} \text{i. } \frac{\partial f}{\partial x}(x, y) &= \lim_{\Delta x \rightarrow 0} \frac{f(x + \Delta x, y) - f(x, y)}{\Delta x} = \\ &= \lim_{\Delta x \rightarrow 0} \frac{5(x + \Delta x)^2 + 2(x + \Delta x)y + 2y^2 - (5x^2 + 2xy + 2y^2)}{\Delta x} \\ &= \lim_{\Delta x \rightarrow 0} \frac{10x\Delta x + 5(\Delta x)^2 + 2y\Delta x}{\Delta x} \\ &= \lim_{\Delta x \rightarrow 0} (10x + 5\Delta x + 2y) = 10x + 2y. \end{aligned}$$

$$\begin{aligned} \text{ii. } \frac{\partial f}{\partial y}(x, y) &= \lim_{\Delta y \rightarrow 0} \frac{f(x, y + \Delta y) - f(x, y)}{\Delta y} = \\ &= \lim_{\Delta y \rightarrow 0} \frac{5x^2 + 2x(y + \Delta y) + 2(y + \Delta y)^2 - (5x^2 + 2xy + 2y^2)}{\Delta y} \\ &= \lim_{\Delta y \rightarrow 0} (2x + 4y + 2\Delta y) = 2x + 4y. \end{aligned}$$

Tomando

$$\frac{\partial f}{\partial x}(x, y) = 10x + 2y$$

e

$$\frac{\partial f}{\partial y}(x, y) = 2x + 4y,$$

temos que se $(x_0, y_0) = (1, 1)$ é um ponto particular do domínio de f , então

$$\frac{\partial f}{\partial x}(1, 1) = 10 + 2 = 12$$

e

$$\frac{\partial f}{\partial y}(1, 1) = 2 + 4 = 6.$$

Observação 1.16. Na prática, quando derivamos uma função parcialmente, tomamos uma variável e as outras consideramos como constante. Assim, a derivação é feita como se fosse para uma função ordinária, onde todas as regras de derivação para funções de uma variável continuam válidas. Portanto, usando a definição, é possível provar (*vide*[12]) que as derivadas parciais satisfazem as seguintes propriedades para cada variável:

- i. Se c é uma constante e se $f(x) = c$ para todo x , então $f'(x) = 0$;
- ii. Se f e g são funções, se h é a função definida por $h(x) = f(x) + g(x)$, e se $f'(x)$ e $g'(x)$ existem, então $h'(x) = f'(x) + g'(x)$;
- iii. Se f e g são funções, se h é a função definida por $h(x) = f(x) \cdot g(x)$, e se $f'(x)$ e $g'(x)$ existem, então $h'(x) = f'(x) \cdot g(x) + f(x) \cdot g'(x)$;
- iv. Se f e g são funções, se h é a função definida por

$$h(x) = \frac{f(x)}{g(x)}, \quad \text{onde } g(x) \neq 0,$$

e se $f'(x)$ e $g'(x)$ existem, então

$$h'(x) = \frac{g(x) \cdot f'(x) - f(x) \cdot g'(x)}{[g(x)]^2}.$$

Estamos interessados em funções polinomiais, onde as propriedades supramencionadas são necessárias para o cálculo das derivadas parciais dessas funções. Por atender particularmente nossos interesses, destacamos a seguinte propriedade:

Teorema 1.17. *A função $f(x) = x^n$, com n inteiro, é derivável para todo $x \in \mathbb{R}$ se $n \geq 0$ e derivável para $x \in \mathbb{R}^*$ se $n < 0$. Nos dois casos $f'(x) = nx^{n-1}$.*

Demonstração: Separamos a demonstração em duas partes: primeiro encontraremos a derivada de x^n para $n \geq 0$ usando a fórmula do binômio de Newton [12]. Em seguida, encontraremos a derivada de x^n para $n < 0$ usando a derivada do quociente [12].

1) Seja $f(x) = x^n$, com $n \geq 0$. Pela definição de derivada temos que

$$f'(x) = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} = \lim_{h \rightarrow 0} \frac{(x+h)^n - x^n}{h}.$$

Expandindo $(x+h)^n$, obtemos

$$(x+h)^n = \binom{n}{0}x^{n-0}h^0 + \binom{n}{1}x^{n-1}h^1 + \dots + \binom{n}{n-1}x^{n-(n-1)}h^{n-1} + \binom{n}{n}x^{n-n}h^n,$$

o que corresponde a

$$(x+h)^n = x^n + nx^{n-1}h + \dots + nxh^{n-1} + h^n.$$

Logo,

$$(x+h)^n - x^n = nx^{n-1}h + \dots + nxh^{n-1} + h^n.$$

Colocando h em evidência na igualdade anterior, obtemos

$$(x+h)^n - x^n = h(nx^{n-1} + \dots + nxh^{n-2} + h^{n-1}).$$

Dividindo ambos os lados da igualdade anterior por h , resulta em

$$\frac{(x+h)^n - x^n}{h} = (nx^{n-1} + \dots + nxh^{n-2} + h^{n-1}).$$

Portanto,

$$f'(x) = \lim_{h \rightarrow 0} \frac{(x+h)^n - x^n}{h} = \lim_{h \rightarrow 0} (nx^{n-1} + \dots + nxh^{n-2} + h^{n-1}) = nx^{n-1}.$$

2) Seja $f(x) = x^n$, com $n < 0$. Tomando $n = -m$, com $m > 0$ e $x^n = x^{-m} = \frac{1}{x^m}$, temos que $f(x) = \frac{1}{x^m}$. Se $x \neq 0$ então, pela derivada do quociente segue que

$$f'(x) = \frac{(1)'(x^m) - 1(x^m)'}{(x^m)^2} = \frac{-mx^{m-1}}{x^{2m}} = -mx^{-m-1} = nx^{n-1}.$$

■

Diante disto, e retomando o exemplo 1.15 podemos determinar $\frac{\partial f}{\partial x}(x, y)$ e $\frac{\partial f}{\partial y}(x, y)$ usando 1.17. Assim, temos que

$$\frac{\partial}{\partial x}(5x^2 + 2xy + 2y^2) = 2 \cdot 5x^{2-1} + 2x^{1-1}y = 10x + 2y$$

e

$$\frac{\partial}{\partial y}(5x^2 + 2xy + 2y^2) = 2xy^{1-1} + 2 \cdot 2y^{2-1} = 2x + 4y.$$

Definição 1.18. *Seja uma função $f : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$. Dizemos que o incremento de f no ponto $(x_0, y_0) \in D$, denotado por $\Delta f(x_0, y_0)$ é dado por:*

$$\Delta f(x_0, y_0) = f(x_0 + \Delta x, y_0 + \Delta y) - f(x_0, y_0).$$

Definição 1.19. *Uma função $f : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$ definida sobre uma bola aberta, é dita diferenciável em (x_0, y_0) se o incremento de f em (x_0, y_0) pode ser escrito por*

$$\Delta f(x_0, y_0) = \frac{\partial}{\partial x}(x_0, y_0)\Delta x + \frac{\partial}{\partial y}(x_0, y_0)\Delta y + \lambda_1\Delta x + \lambda_2\Delta y,$$

onde λ_1 e λ_2 são funções de Δx e Δy tal que $\lambda_1 \rightarrow 0$ e $\lambda_2 \rightarrow 0$ quando $(\Delta x, \Delta y) \rightarrow (0, 0)$.

Teorema 1.20. *Se a função $f : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$ definida sobre uma bola aberta é diferenciável no ponto $(x_0, y_0) \in D$, então ela é contínua neste ponto.*

Demonstração: (Leithold [12]) Das definições 1.18 e 1.19 segue que

$$f(x_0 + \Delta x, y_0 + \Delta y) - f(x_0, y_0) = \frac{\partial}{\partial x}(x_0, y_0)\Delta x + \frac{\partial}{\partial y}(x_0, y_0)\Delta y + \lambda_1\Delta x + \lambda_2\Delta y.$$

Logo,

$$f(x_0 + \Delta x, y_0 + \Delta y) = f(x_0, y_0) + \frac{\partial}{\partial x}(x_0, y_0)\Delta x + \frac{\partial}{\partial y}(x_0, y_0)\Delta y + \lambda_1\Delta x + \lambda_2\Delta y.$$

Assim, tomando o limite quando Δx e Δy tendem a zero temos

$$\begin{aligned} & \lim_{(\Delta x, \Delta y) \rightarrow (0,0)} f(x_0 + \Delta x, y_0 + \Delta y) = \\ & = \lim_{(\Delta x, \Delta y) \rightarrow (0,0)} \left[f(x_0, y_0) + \frac{\partial}{\partial x}(x_0, y_0)\Delta x + \frac{\partial}{\partial y}(x_0, y_0)\Delta y + \lambda_1\Delta x + \lambda_2\Delta y \right], \end{aligned}$$

podemos reescrever como:

$$\begin{aligned} \lim_{(\Delta x, \Delta y) \rightarrow (0,0)} f(x_0 + \Delta x, y_0 + \Delta y) &= \lim_{(\Delta x, \Delta y) \rightarrow (0,0)} f(x_0, y_0) + \lim_{(\Delta x, \Delta y) \rightarrow (0,0)} \frac{\partial}{\partial x}(x_0, y_0)\Delta x + \\ &+ \lim_{(\Delta x, \Delta y) \rightarrow (0,0)} \frac{\partial}{\partial y}(x_0, y_0)\Delta y + \lim_{(\Delta x, \Delta y) \rightarrow (0,0)} \lambda_1\Delta x + \lim_{(\Delta x, \Delta y) \rightarrow (0,0)} \lambda_2\Delta y. \end{aligned}$$

Mas

$$\lim_{\Delta x \rightarrow 0} \Delta x = \lim_{\Delta y \rightarrow 0} \Delta y = 0.$$

Então, do item (2) do teorema 1.9, segue que

$$\lim_{(\Delta x, \Delta y) \rightarrow (0,0)} \frac{\partial}{\partial x}(x_0, y_0)\Delta x = 0 = \lim_{(\Delta x, \Delta y) \rightarrow (0,0)} \frac{\partial}{\partial y}(x_0, y_0)\Delta y$$

e

$$\lim_{(\Delta x, \Delta y) \rightarrow (0,0)} \lambda_1\Delta x = 0 = \lim_{(\Delta x, \Delta y) \rightarrow (0,0)} \lambda_2\Delta y.$$

Desta forma, temos que

$$\lim_{(\Delta x, \Delta y) \rightarrow (0,0)} f(x_0 + \Delta x, y_0 + \Delta y) = \lim_{(\Delta x, \Delta y) \rightarrow (0,0)} f(x_0, y_0).$$

Como

$$\lim_{(\Delta x, \Delta y) \rightarrow (0,0)} f(x_0, y_0) = f(x_0, y_0),$$

segue que

$$\lim_{(\Delta x, \Delta y) \rightarrow (0,0)} f(x_0 + \Delta x, y_0 + \Delta y) = f(x_0, y_0). \quad (1.7)$$

Tomando $x_0 + \Delta x = x$ e $y_0 + \Delta y = y$, temos que $f(x_0 + \Delta x, y_0 + \Delta y) = f(x, y)$ e $(\Delta x, \Delta y) \rightarrow (0, 0)$ é equivalente a $(x, y) \rightarrow (x_0, y_0)$. Então, da equação (1.7) segue que

$$\lim_{(x,y) \rightarrow (x_0,y_0)} f(x, y) = f(x_0, y_0).$$

Portanto, da definição 1.10 concluímos que f é contínua em (x_0, y_0) . ■

Definição 1.21. Dada uma função $f : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$ definida sobre uma bola aberta e um ponto $(x_0, y_0) \in D$. Dizemos que f é diferenciável em (x_0, y_0) se existirem $a, b \in \mathbb{R}$, tais que:

$$\lim_{(\Delta x, \Delta y) \rightarrow (x_0, y_0)} \frac{f(x_0 + \Delta x, y_0 + \Delta y) - f(x_0, y_0) - a\Delta x - b\Delta y}{\sqrt{(\Delta x)^2 + (\Delta y)^2}} = 0.$$

Teorema 1.22. Se uma função $f : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$ definida sobre uma bola aberta é diferenciável em um ponto $(x_0, y_0) \in D$, então existem as derivadas parciais $\frac{\partial f}{\partial x}(x_0, y_0)$ e $\frac{\partial f}{\partial y}(x_0, y_0)$.

Demonstração: (Leithold [12]) Da definição 1.21 temos que

i. tomando $(\Delta x, 0)$ segue que

$$\begin{aligned} \lim_{(\Delta x, \Delta y) \rightarrow (x_0, y_0)} \frac{f(x_0 + \Delta x, y_0 + \Delta y) - f(x_0, y_0) - a\Delta x - b\Delta y}{\sqrt{(\Delta x)^2 + (\Delta y)^2}} &= 0 \\ \iff \lim_{\Delta x \rightarrow 0} \frac{f(x_0 + \Delta x, y_0) - f(x_0, y_0) - a\Delta x}{\sqrt{\Delta x^2}} &= 0 \\ \iff \lim_{\Delta x \rightarrow 0} \frac{f(x_0 + \Delta x, y_0) - f(x_0, y_0) - a\Delta x}{|\Delta x|} &= 0 \\ \iff \lim_{\Delta x \rightarrow 0} \frac{f(x_0 + \Delta x, y_0) - f(x_0, y_0)}{\Delta x} &= a. \end{aligned}$$

ii. tomando $(0, \Delta y)$ segue que

$$\lim_{(\Delta x, \Delta y) \rightarrow (x_0, y_0)} \frac{f(x_0 + \Delta x, y_0 + \Delta y) - f(x_0, y_0) - a\Delta x - b\Delta y}{\sqrt{(\Delta x)^2 + (\Delta y)^2}} = 0$$

$$\begin{aligned} &\iff \lim_{\Delta y \rightarrow 0} \frac{f(x_0, y_0 + \Delta y) - f(x_0, y_0) - b\Delta y}{\sqrt{\Delta y^2}} = 0 \\ &\iff \lim_{\Delta y \rightarrow 0} \frac{f(x_0, y_0 + \Delta y) - f(x_0, y_0) - b\Delta y}{|\Delta y|} = 0 \\ &\iff \lim_{\Delta y \rightarrow 0} \frac{f(x_0, y_0 + \Delta y) - f(x_0, y_0)}{\Delta y} = b. \end{aligned}$$

Assim, pela definição 1.14 temos que $\frac{\partial f}{\partial x}(x_0, y_0) = a$ e $\frac{\partial f}{\partial y}(x_0, y_0) = b$. Portanto, se f é diferenciável no ponto $(x_0, y_0) \in D$, então a e b da definição 1.21 são as derivadas parciais nesse mesmo ponto. ■

Teorema 1.23. *Seja $f : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$ uma função de duas variáveis x e y definida sobre uma bola aberta. Supondo que $\frac{\partial f}{\partial x}(x_0, y_0)$ e $\frac{\partial f}{\partial y}(x_0, y_0)$ existam. Se $\frac{\partial f}{\partial x}(x_0, y_0)$ e $\frac{\partial f}{\partial y}(x_0, y_0)$ forem contínuas em $(x_0, y_0) \in D$, então f é diferenciável nesse ponto.*

Demonstração: (Leithold [12]) Considerando o teorema do valor médio para uma função de uma variável aplicado para uma função de duas variáveis (vide demonstração em [12]) e tomando um ponto $(x_0 + \Delta x, y_0 + \Delta y) \in D$, temos que

$$\Delta f(x_0, y_0) = f(x_0 + \Delta x, y_0 + \Delta y) - f(x_0, y_0).$$

Somando e subtraindo $f(x_0 + \Delta x, y_0)$ no lado direito da equação acima, obtemos

$$\Delta f(x_0, y_0) = [f(x_0 + \Delta x, y_0 + \Delta y) - f(x_0 + \Delta x, y_0)] + [f(x_0 + \Delta x, y_0) - f(x_0, y_0)] \quad (1.8)$$

Por hipótese $\frac{\partial f}{\partial x}$ e $\frac{\partial f}{\partial y}$ existem em $(x_0 + \Delta x, y_0 + \Delta y)$, então se tomarmos λ_1 e λ_2 respectivamente entre $(x_0, x_0 + \Delta x)$ e $(y_0, y_0 + \Delta y)$, segue que

$$f(x_0 + \Delta x, y_0 + \Delta y) - f(x_0 + \Delta x, y_0) = \Delta y \frac{\partial f}{\partial y}(x_0 + \Delta x, \lambda_2) \quad (1.9)$$

$$f(x_0 + \Delta x, y_0) - f(x_0, y_0) = \Delta x \frac{\partial f}{\partial x}(\lambda_1, y_0) \quad (1.10)$$

Substituindo (1.9) e (1.10) em (1.8), temos

$$\Delta f(x_0, y_0) = \Delta y \frac{\partial f}{\partial y}(x_0 + \Delta x, \lambda_2) + \Delta x \frac{\partial f}{\partial x}(\lambda_1, y_0) \quad (1.11)$$

Como $(x_0 + \Delta x, y_0 + \Delta y) \in D$, λ_1 e λ_2 estão respectivamente entre $(x_0, x_0 + \Delta x)$ e $(y_0, y_0 + \Delta y)$, e $\frac{\partial f}{\partial x}$ e $\frac{\partial f}{\partial y}$ são contínuas, temos que

$$\lim_{\Delta(x,y) \rightarrow (0,0)} \frac{\partial f}{\partial x}(\lambda_1, y_0) = \frac{\partial f}{\partial x}(x_0, y_0) \quad (1.12)$$

$$\lim_{\Delta(x,y) \rightarrow (0,0)} \frac{\partial f}{\partial y}(x_0 + \Delta x, \lambda_2) = \frac{\partial f}{\partial y}(x_0, y_0) \quad (1.13)$$

Se tomarmos

$$\beta_1 = \frac{\partial f}{\partial x}(\lambda_1, y_0) - \frac{\partial f}{\partial x}(x_0, y_0), \quad (1.14)$$

segue da equação (1.12) que

$$\lim_{\Delta(x,y) \rightarrow (0,0)} \beta_1 = 0. \quad (1.15)$$

Supondo

$$\beta_2 = \frac{\partial f}{\partial y}(x_0 + \Delta x, \lambda_2) - \frac{\partial f}{\partial y}(x_0, y_0), \quad (1.16)$$

segue da equação (1.13) que

$$\lim_{\Delta(x,y) \rightarrow (0,0)} \beta_2 = 0. \quad (1.17)$$

Substituindo (1.14) e (1.16) em (1.11), obtemos

$$\Delta f(x_0, y_0) = \Delta y \left[\frac{\partial f}{\partial y}(x_0, y_0) + \beta_2 \right] + \Delta x \left[\frac{\partial f}{\partial x}(x_0, y_0) + \beta_1 \right], \quad (1.18)$$

donde vem

$$\Delta f(x_0, y_0) = \Delta x \frac{\partial f}{\partial x}(x_0, y_0) + \Delta y \frac{\partial f}{\partial y}(x_0, y_0) + \beta_1 \Delta x + \beta_2 \Delta y. \quad (1.19)$$

Daí segue que das equações (1.15), (1.17) e (1.19) a definição 1.18 é verificada e pela definição 1.19 podemos concluir que a função f é diferenciável em $(x_0, y_0) \in D$. ■

Considerando as definições e teoremas que foram apresentados até aqui, queremos verificar se uma função polinomial $f(x, y)$ é diferenciável em todo ponto de \mathbb{R}^2 .

Seja então uma função polinomial $f : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$. Tendo em vista o que foi descrito na observação 1.16, bem como a propriedade apresentada pelo teorema 1.17, podemos dizer que as derivadas parciais de f são polinômios de grau $(n - 1)$, e, sendo assim, pelo teorema 1.12 também são contínuas em todos os seus pontos. Isso implica que existem as derivadas parciais de f em todos os pontos de \mathbb{R}^2 e estas são contínuas. Então, pelo teorema 1.23 concluimos que f é diferenciável em todo ponto de \mathbb{R}^2 .

Desta forma, podemos investigar se em alguns desses pontos as derivadas parciais são nulas. As derivadas parciais nulas consistem em uma condição necessária para que $(x_0, y_0) \in D$ seja um ponto onde f tenha extremo local, isto é, um ponto de máximo ou mínimo local.

Estamos interessados em determinar os pontos de mínimos de uma certa função polinomial. Então, considerando os resultados discutidos acima, sabemos que tal função é diferenciável em todo \mathbb{R}^2 , logo, esta função admite derivadas parciais em todos os pontos de D . Nesse sentido, e de acordo com Guidorizzi [9], se f admite derivadas parciais em todos os pontos de D , então os pontos $(x_0, y_0) \in D$ nos quais as derivadas parciais se anulam são, entre os pontos interiores de D , os únicos candidatos a extremos locais de f .

Em vista disto, para que possamos identificar os pontos candidatos a extremos locais, ou mais especificamente os pontos de mínimos locais de f , temos que garantir a existência de pontos $(x_0, y_0) \in D$ tais que as derivadas parciais sejam nulas, ou seja,

$$\frac{\partial f}{\partial x}(x_0, y_0) = \frac{\partial f}{\partial y}(x_0, y_0) = 0.$$

Vejamos as seguintes definições e teoremas:

Definição 1.24. *Seja uma função $f : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$ definida sobre uma bola aberta, dizemos que $(x_0, y_0) \in D$ é ponto de mínimo local de f se existe um disco aberto $B((x_0, y_0), r)$, tal que $f(x, y) \geq f(x_0, y_0)$ para todo $(x, y) \in B((x_0, y_0), r)$.*

Definição 1.25. *Um ponto (x_0, y_0) para o qual $\frac{\partial f}{\partial x}(x_0, y_0) = \frac{\partial f}{\partial y}(x_0, y_0) = 0$, é denominado ponto crítico de f .*

Teorema 1.26. *Se uma função $f : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$ definida sobre uma bola aberta é diferenciável em um ponto $(x_0, y_0) \in D$, então uma condição necessária para que (x_0, y_0) seja um ponto onde f tem um extremo local é que $\frac{\partial f}{\partial x}(x_0, y_0) = \frac{\partial f}{\partial y}(x_0, y_0) = 0$.*

Demonstração: (Leithold [12]) Por hipótese temos que $\frac{\partial f}{\partial x}(x_0, y_0)$ e $\frac{\partial f}{\partial y}(x_0, y_0)$ existem, pois f é diferenciável neste ponto (Teorema 1.22). Então, pela definição 1.14 existem os limites

$$\lim_{\Delta x \rightarrow 0} \frac{f(x_0 + \Delta x, y_0) - f(x_0, y_0)}{\Delta x}$$

e

$$\lim_{\Delta y \rightarrow 0} \frac{f(x_0, y_0 + \Delta y) - f(x_0, y_0)}{\Delta y}.$$

Supondo que f tenha um ponto de mínimo local em $(x_0, y_0) \in D$. Então, segue da definição 1.24 que

$$f(x_0 + \Delta x, y_0) \geq f(x_0, y_0)$$

e

$$f(x_0, y_0 + \Delta y) \geq f(x_0, y_0),$$

ou seja

$$f(x_0 + \Delta x, y_0) - f(x_0, y_0) \geq 0$$

e

$$f(x_0, y_0 + \Delta y) - f(x_0, y_0) \geq 0,$$

sempre que Δx e Δy forem suficientemente pequenos, tais que

$(x_0 + \Delta x, y_0)$ e $(x_0, y_0 + \Delta y)$ estejam em um disco aberto B contendo (x_0, y_0) . Desta forma, observamos que:

i. Se Δx aproxima-se de zero pela esquerda, então $\Delta x < 0$. Assim,

$$\frac{f(x_0 + \Delta x_0, y_0) - f(x_0, y_0)}{\Delta x} \leq 0,$$

já que o numerador é positivo. Logo, $\frac{\partial f}{\partial x}(x_0, y_0) \leq 0$;

ii. Analogamente, se Δx aproxima-se de zero pela direita, então $\Delta x > 0$. Assim,

$$\frac{f(x_0 + \Delta x_0, y_0) - f(x_0, y_0)}{\Delta x} \geq 0.$$

Logo, $\frac{\partial f}{\partial x}(x_0, y_0) \geq 0$;

iii. Utilizando o mesmo raciocínio para Δy , temos que Δy aproxima-se de zero pela esquerda, então $\Delta y < 0$. Logo,

$$\frac{f(x_0, y_0 + \Delta y_0) - f(x_0, y_0)}{\Delta y} \leq 0,$$

ou seja, $\frac{\partial f}{\partial y}(x_0, y_0) \leq 0$;

iv. Se Δy aproxima-se de zero pela direita, então $\Delta y > 0$. Assim,

$$\frac{f(x_0, y_0 + \Delta y_0) - f(x_0, y_0)}{\Delta y} \geq 0.$$

Logo, $\frac{\partial f}{\partial y}(x_0, y_0) \geq 0$.

De acordo com o que foi exposto, concluímos que se $\frac{\partial f}{\partial x}(x_0, y_0)$ e $\frac{\partial f}{\partial y}(x_0, y_0)$ existem, então as desigualdades

$$\frac{\partial f}{\partial x}(x_0, y_0) \leq 0, \frac{\partial f}{\partial x}(x_0, y_0) \geq 0$$

e

$$\frac{\partial f}{\partial y}(x_0, y_0) \leq 0, \frac{\partial f}{\partial y}(x_0, y_0) \geq 0$$

devem valer simultaneamente, o que necessariamente implica em

$$\frac{\partial f}{\partial x}(x_0, y_0) = \frac{\partial f}{\partial y}(x_0, y_0) = 0.$$

■

Exemplo 1.27. Dada a função $f(x, y) = 3x^2 - 4xy + 3y^2 + 8x - 17y + 30$

e o ponto de mínimo $(x_0, y_0) = (1, \frac{7}{2})$, queremos verificar se $\frac{\partial f}{\partial x}(x_0, y_0) = \frac{\partial f}{\partial y}(x_0, y_0) = 0$.

De fato, como f é polinomial, então é diferenciável em todo ponto $(x_0, y_0) \in \mathbb{R}^2$. E isso implica que existem as derivadas parciais $\frac{\partial f}{\partial x}(x_0, y_0)$ e $\frac{\partial f}{\partial y}(x_0, y_0)$. Portanto, para o ponto $(x_0, y_0) = (1, \frac{7}{2})$ temos que:

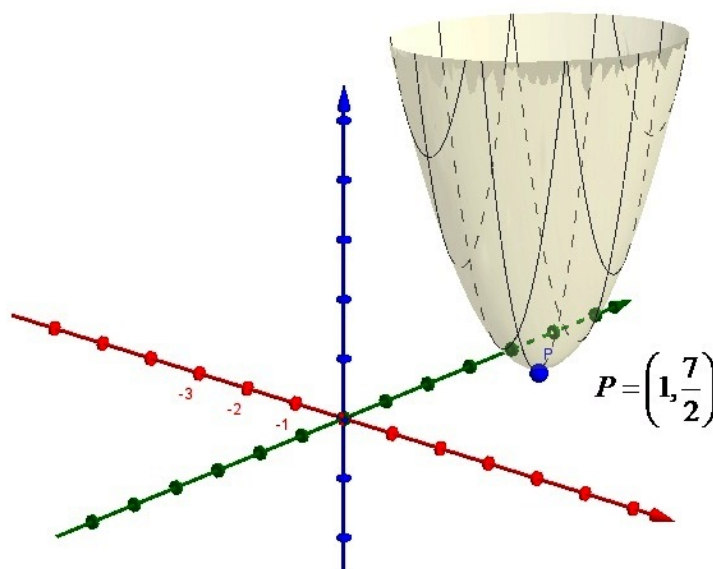
$$\frac{\partial f}{\partial x}(x_0, y_0) = 6x_0 - 4y_0 + 8 = 6 \cdot 1 - 4 \cdot \frac{7}{2} + 8 = 0$$

e

$$\frac{\partial f}{\partial y}(x_0, y_0) = -4x_0 + 6y_0 - 17 = -4 \cdot 1 + 6 \cdot \frac{7}{2} - 17 = 0.$$

A figura 1.1 ilustra o gráfico de f e a localização do ponto $P = (1, \frac{7}{2})$ sobre ele:

Figura 1.1: Ponto de mínimo



Fonte: Figura gerada pelo autor

O teorema 1.26 estabelece que os pontos críticos consistem em uma condição necessária para que $(x_0, y_0) \in D$ seja extremo local. Mas a anulação das derivadas parciais não é condição suficiente para que uma função tenha extremo local em $(x_0, y_0) \in D$. Um exemplo clássico desta situação pode ser constatado pela função $f(x, y) = x^2 - y^2$, em que $\frac{\partial f}{\partial x}(0, 0) = \frac{\partial f}{\partial y}(0, 0) = 0$, porém f não tem extremo local em $(0, 0)$. Neste caso, esses pontos são denominados pontos de *sela*. Desta forma, o teorema 1.26 nos fornece um critério para selecionar, entre os pontos de D , candidatos a extremos locais. No entanto, esse critério não nos garante que os pontos selecionados sejam de fato extremos locais.

Assim, precisamos de algum critério que estabeleça uma condição suficiente para que os pontos críticos de f sejam extremos locais. Segundo Guidorizzi [9], um cri-

tério usualmente denominado *teste da derivada segunda* nos garante essa condição. Observemos então, as definições e o teorema a seguir:

Observação 1.28. Seja uma função de duas variáveis $f : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$ definida sobre uma bola aberta. Em geral, $\frac{\partial f}{\partial x}$ e $\frac{\partial f}{\partial y}$ também são funções de duas variáveis. Denotamos estas funções por:

$$\frac{\partial f}{\partial x} = f_1 \quad e \quad \frac{\partial f}{\partial y} = f_2.$$

Definição 1.29. Se as derivadas parciais de f_1 e f_2 existem, podemos denominá-las *derivadas parciais segundas* de f . Assim, dizemos que:

i. A derivada parcial segunda de f , em relação a x , é a função dada por

$$\frac{\partial^2 f}{\partial x^2} = \frac{\partial f_1}{\partial x} = \lim_{\Delta x \rightarrow 0} \frac{f_1(x + \Delta x, y) - f_1(x, y)}{\Delta x};$$

ii. A derivada parcial segunda de f , em relação a y , é a função dada por

$$\frac{\partial^2 f}{\partial y^2} = \frac{\partial f_2}{\partial y} = \lim_{\Delta y \rightarrow 0} \frac{f_2(x, y + \Delta y) - f_2(x, y)}{\Delta y};$$

iii. A derivada parcial segunda de f , em relação a x e y , é a função dada por

$$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial f_2}{\partial x} = \lim_{\Delta x \rightarrow 0} \frac{f_2(x + \Delta x, y) - f_2(x, y)}{\Delta x};$$

iv. A derivada parcial segunda de f , em relação a y e x , é a função dada por

$$\frac{\partial^2 f}{\partial y \partial x} = \frac{\partial f_1}{\partial y} = \lim_{\Delta y \rightarrow 0} \frac{f_1(x, y + \Delta y) - f_1(x, y)}{\Delta y}.$$

Exemplo 1.30. Dada a função $f(x, y) = x^7 y^5 - 2x^3 y^2 + 5$. Queremos determinar todas as derivadas parciais segundas de f . Do teorema 1.17, da observação 1.28 e da definição 1.29 segue que:

i. $f_1 = \frac{\partial f}{\partial x} = 7x^6 y^5 - 6x^2 y^2;$

ii. $f_2 = \frac{\partial f}{\partial y} = 5x^7 y^4 - 4x^3 y;$

iii. $\frac{\partial^2 f}{\partial x^2} = \frac{\partial f_1}{\partial x} = 42x^5 y^5 - 12xy^2;$

iv. $\frac{\partial^2 f}{\partial y^2} = \frac{\partial f_2}{\partial y} = 20x^7 y^3 - 4x^3;$

v. $\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial f_2}{\partial x} = 35x^6 y^4 - 12x^2 y;$

vi. $\frac{\partial^2 f}{\partial y \partial x} = \frac{\partial f_1}{\partial y} = 35x^6 y^4 - 12x^2 y.$

Teorema 1.31. (Teste da derivada segunda) Dada uma função $f : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$, tal que suas derivadas parciais primeira e segunda sejam contínuas em uma bola aberta $B((x_0, y_0), r)$, com $(x_0, y_0) \in D$ ponto crítico de f . Então temos que:

i. f tem um valor de mínimo local em $(x_0, y_0) \in D$ se

$$\frac{\partial^2 f}{\partial x^2}(x_0, y_0) \cdot \frac{\partial^2 f}{\partial y^2}(x_0, y_0) - \left[\frac{\partial^2 f}{\partial x \partial y}(x_0, y_0) \right]^2 > 0 \quad e \quad \frac{\partial^2 f}{\partial x^2}(x_0, y_0) > 0;$$

ii. f tem um valor de máximo local em $(x_0, y_0) \in D$ se

$$\frac{\partial^2 f}{\partial x^2}(x_0, y_0) \cdot \frac{\partial^2 f}{\partial y^2}(x_0, y_0) - \left[\frac{\partial^2 f}{\partial x \partial y}(x_0, y_0) \right]^2 > 0 \quad e \quad \frac{\partial^2 f}{\partial x^2}(x_0, y_0) < 0;$$

iii. f não tem um valor de extremo local em $(x_0, y_0) \in D$ se

$$\frac{\partial^2 f}{\partial x^2}(x_0, y_0) \cdot \frac{\partial^2 f}{\partial y^2}(x_0, y_0) - \left[\frac{\partial^2 f}{\partial x \partial y}(x_0, y_0) \right]^2 < 0;$$

iv. Não podemos chegar a nenhuma conclusão se

$$\frac{\partial^2 f}{\partial x^2}(x_0, y_0) \cdot \frac{\partial^2 f}{\partial y^2}(x_0, y_0) - \left[\frac{\partial^2 f}{\partial x \partial y}(x_0, y_0) \right]^2 = 0.$$

Demonstração: (Leithold [12]) Faremos a demonstração apenas do item (i) do teorema 1.31, pois este é pertinente ao nosso interesse em determinar pontos de mínimo de uma certa função. Pois bem, seja

$$\Phi(x_0, y_0) = \frac{\partial^2 f}{\partial x^2}(x_0, y_0) \cdot \frac{\partial^2 f}{\partial y^2}(x_0, y_0) - \left[\frac{\partial^2 f}{\partial x \partial y}(x_0, y_0) \right]^2 > 0.$$

Dado que

$$\Phi(x_0, y_0) > 0 \quad e \quad \frac{\partial^2 f}{\partial x^2}(x_0, y_0) > 0,$$

queremos mostrar que $f(x_0, y_0)$ é um valor de mínimo local de f em D . Como as derivadas parciais segundas são contínuas em $B((x_0, y_0), r)$, segue que $\Phi(x_0, y_0)$ também é contínua em B . Logo, existe uma bola aberta $B_1((x_0, y_0), r_1)$, onde $r_1 \leq r$, tal que $\Phi(x_0, y_0) > 0$ e $\frac{\partial^2 f}{\partial x^2}(x_0, y_0) > 0$ para todo (x_0, y_0) em B_1 .

Sejam h e k constantes, ambas não nulas, tais que o ponto $(x_0 + h, y_0 + k)$ esteja em B_1 . Então as equações $x = x_0 + ht$ e $y = y_0 + kt$, com $0 \leq t \leq 1$ definem todos os pontos no segmento de reta de (x_0, y_0) a $(x_0 + h, y_0 + k)$ e todos esses pontos estão em B_1 .

Seja F a função de uma variável definida por

$$F(t) = f(x_0 + ht, y_0 + kt). \quad (1.20)$$

Pela fórmula de Taylor [12], temos que

$$F(t) = F(0) + F'(0)t + F''(\xi) \frac{t^2}{2!}, \quad (1.21)$$

onde F' e F'' são respectivamente as derivadas primeira e segunda de F e ξ está entre 0 e t .

Então se tomarmos $t = 1$ na equação (1.21), temos que

$$F(1) = F(0) + F'(0) + F'' \frac{(\xi)}{2}, \quad (1.22)$$

onde $0 < \xi < 1$.

Como $F(0) = f(x_0, y_0)$ e $F(1) = f(x_0 + h, y_0 + k)$, segue da equação (1.22) que

$$f(x_0 + h, y_0 + k) = f(x_0, y_0) + F'(0) + \frac{1}{2}F''(\xi). \quad (1.23)$$

Usando a regra da cadeia [12] para encontrarmos $F'(t)$ e $F''(t)$, obtemos

$$F'(t) = h \frac{\partial f}{\partial x}(x_0 + ht, y_0 + kt) + k \frac{\partial f}{\partial y}(x_0 + ht, y_0 + kt) \quad (1.24)$$

e

$$F''(t) = h^2 \frac{\partial^2 f}{\partial x^2}(x_0 + ht, y_0 + kt) + 2hk \frac{\partial^2 f}{\partial x \partial y}(x_0 + ht, y_0 + kt) + k^2 \frac{\partial^2 f}{\partial y^2}(x_0 + ht, y_0 + kt). \quad (1.25)$$

Substituindo t por 0 na equação (1.24) obtemos

$$F'(0) = h \frac{\partial f}{\partial x}(x_0, y_0) + k \frac{\partial f}{\partial y}(x_0, y_0),$$

pois pelo teorema 1.26 temos que

$$\frac{\partial f}{\partial x}(x_0, y_0) = \frac{\partial f}{\partial y}(x_0, y_0) = 0.$$

Assim segue que

$$F'(0) = 0 \quad (1.26)$$

e substituindo t por ξ na equação (1.25) obtemos

$$F''(t) = h^2 \frac{\partial^2 f}{\partial x^2}(x_0 + h\xi, y_0 + k\xi) + 2hk \frac{\partial^2 f}{\partial x \partial y}(x_0 + h\xi, y_0 + k\xi) + k^2 \frac{\partial^2 f}{\partial y^2}(x_0 + h\xi, y_0 + k\xi). \quad (1.27)$$

Substituindo agora as equações (1.26) e (1.27) na equação (1.23) obtemos

$$\begin{aligned} & f(x_0 + h, y_0 + k) - f(x_0, y_0) = \\ & = \frac{1}{2} \left[h^2 \frac{\partial^2 f}{\partial x^2}(x_0 + h\xi, y_0 + k\xi) + 2hk \frac{\partial^2 f}{\partial x \partial y}(x_0 + h\xi, y_0 + k\xi) + k^2 \frac{\partial^2 f}{\partial y^2}(x_0 + h\xi, y_0 + k\xi) \right]. \end{aligned} \quad (1.28)$$

Os termos entre colchetes no lado direito da equação 1.28 podem ser escritos por

$$\left[h^2 \frac{\partial^2 f}{\partial x^2}(x_0 + h\xi, y_0 + k\xi) + 2hk \frac{\partial^2 f}{\partial x \partial y}(x_0 + h\xi, y_0 + k\xi) + k^2 \frac{\partial^2 f}{\partial y^2}(x_0 + h\xi, y_0 + k\xi) \right] =$$

$$= \frac{\partial^2 f}{\partial x^2} \left[h^2 + 2hk \cdot \left(\frac{\frac{\partial^2 f}{\partial x \partial y}}{\frac{\partial^2 f}{\partial x^2}} \right) + k^2 \cdot \left(\frac{\frac{\partial^2 f}{\partial x \partial y}}{\frac{\partial^2 f}{\partial x^2}} \right)^2 - k^2 \cdot \left(\frac{\frac{\partial^2 f}{\partial x \partial y}}{\frac{\partial^2 f}{\partial x^2}} \right)^2 + k^2 \cdot \left(\frac{\frac{\partial^2 f}{\partial y^2}}{\frac{\partial^2 f}{\partial x^2}} \right) \right].$$

Assim, da equação (1.28) temos

$$f(x_0 + h, y_0 + k) - f(x_0, y_0) = \frac{\partial^2 f}{2\partial x^2} \cdot \left[\left(h + k \cdot \left(\frac{\frac{\partial^2 f}{\partial x \partial y}}{\frac{\partial^2 f}{\partial x^2}} \right) \right)^2 + \frac{\frac{\partial^2 f}{\partial x^2} \cdot \frac{\partial^2 f}{\partial y^2} - \left[\frac{\partial^2 f}{\partial x \partial y} \right]^2}{\left(\frac{\partial^2 f}{\partial x^2} \right)^2} \cdot k^2 \right]. \quad (1.29)$$

Como $\frac{\partial^2 f}{\partial x^2} \cdot \frac{\partial^2 f}{\partial y^2} - \left[\frac{\partial^2 f}{\partial x \partial y} \right]^2$ calculada em $(x_0 + h\xi, y_0 + k\xi)$ é igual a

$\Phi(x_0 + h\xi, y_0 + k\xi) > 0$. Então temos que

$$\frac{\partial^2 f}{2\partial x^2} \cdot \left[\left(h + k \cdot \left(\frac{\frac{\partial^2 f}{\partial x \partial y}}{\frac{\partial^2 f}{\partial x^2}} \right) \right)^2 + \frac{\frac{\partial^2 f}{\partial x^2} \cdot \frac{\partial^2 f}{\partial y^2} - \left[\frac{\partial^2 f}{\partial x \partial y} \right]^2}{\left(\frac{\partial^2 f}{\partial x^2} \right)^2} \cdot k^2 \right] > 0.$$

Daí segue que $f(x_0 + h, y_0 + k) - f(x_0, y_0) > 0$. Logo, $f(x_0 + h, y_0 + k) > f(x_0, y_0)$ para todo $(x_0 + h, y_0 + k) \neq (x_0, y_0)$ em B_1 . Portanto, pela definição 1.24, $f(x_0, y_0)$ é um valor mínimo local de f . ■

Exemplo 1.32. Seja a função $f : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$, definida por

$$f(x, y) = 4x^4 + 2y^2 - 2x^2 - 4y + 2.$$

Queremos determinar os extremos locais de f se existirem. Assim, temos que

$$\frac{\partial f}{\partial x}(x, y) = 16x^3 - 4x$$

e

$$\frac{\partial f}{\partial y}(x, y) = 4y - 4.$$

Tomando $\frac{\partial f}{\partial x}(x, y) = \frac{\partial f}{\partial y}(x, y) = 0$, temos $y = 1$ e três possibilidades para x , sendo: $x = -\frac{1}{2}$, $x = 0$ ou $x = \frac{1}{2}$. Logo, $\frac{\partial f}{\partial x}(x, y)$ e $\frac{\partial f}{\partial y}(x, y)$ se anulam nos pontos $\left(-\frac{1}{2}, 1\right)$, $(0, 1)$ e $\left(\frac{1}{2}, 1\right)$. Determinando as derivadas segundas de f obtemos

$$\frac{\partial^2 f}{\partial x^2} = 48x^2 - 4,$$

$$\frac{\partial^2 f}{\partial y^2} = 4$$

e

$$\frac{\partial^2 f}{\partial x \partial y} = 0.$$

Aplicando o teste da derivada segunda, segue que:

i. Para o ponto $\left(-\frac{1}{2}, 1\right)$ temos

$$\frac{\partial^2 f}{\partial x^2} \left(-\frac{1}{2}, 1\right) = 8 > 0$$

e

$$\frac{\partial^2 f}{\partial x^2} \left(-\frac{1}{2}, 1\right) \cdot \frac{\partial^2 f}{\partial y^2} \left(-\frac{1}{2}, 1\right) - \left[\frac{\partial^2 f}{\partial x \partial y} \left(-\frac{1}{2}, 1\right) \right]^2 = 8 \cdot 4 - 0 = 32 > 0.$$

Logo, pelo item (i) do teorema 1.31, f tem um valor mínimo local, o qual é dado por $f\left(-\frac{1}{2}, 1\right) = -\frac{1}{2}$;

ii. Para o ponto $(0, 1)$ temos

$$\frac{\partial^2 f}{\partial x^2}(0, 1) \cdot \frac{\partial^2 f}{\partial y^2}(0, 1) - \left[\frac{\partial^2 f}{\partial x \partial y}(0, 1) \right]^2 = (-4) \cdot 4 - 0 = -16 < 0.$$

Assim, pelo item (iii) do teorema 1.31, f não tem um valor de extremo local em $(0, 1)$, pois o ponto $(0, 1)$ é ponto de sela de f ;

iii. Para o ponto $\left(\frac{1}{2}, 1\right)$ temos

$$\frac{\partial^2 f}{\partial x^2} \left(\frac{1}{2}, 1\right) = 8 > 0$$

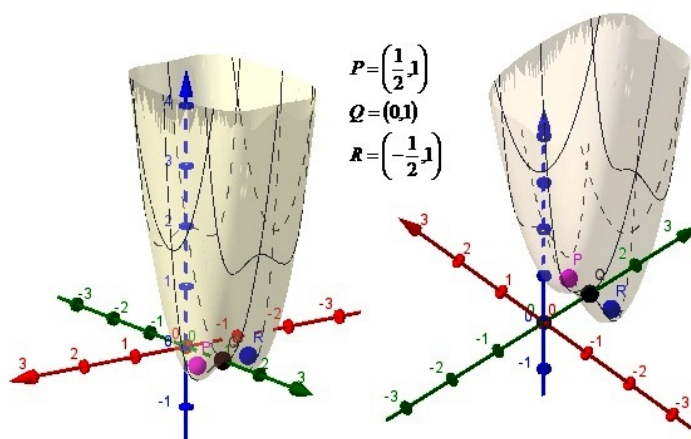
e

$$\frac{\partial^2 f}{\partial x^2} \left(\frac{1}{2}, 1\right) \cdot \frac{\partial^2 f}{\partial y^2} \left(\frac{1}{2}, 1\right) - \left[\frac{\partial^2 f}{\partial x \partial y} \left(\frac{1}{2}, 1\right) \right]^2 = 8 \cdot 4 - 0 = 32 > 0.$$

Logo, pelo item (i) do teorema 1.31, f tem um valor mínimo local em $\left(\frac{1}{2}, 1\right)$, o qual é dado por $f\left(\frac{1}{2}, 1\right) = -\frac{1}{2}$.

Portanto, f tem um valor mínimo local em cada um dos pontos $\left(-\frac{1}{2}, 1\right)$ e $\left(\frac{1}{2}, 1\right)$. A figura 1.2 ilustra o gráfico de f com os pontos de mínimo e de sela:

Figura 1.2: Ponto de mínimo e sela



Fonte: Figura gerada pelo autor

O teorema 1.31 testa os pontos candidatos a extremos locais selecionados de acordo com o critério estabelecido pelo teorema 1.26. Desta forma, se os candidatos se enquadram nas condições descritas pelo teorema 1.31, então são de fato pontos onde a função possui extremos locais. Em particular, o item (i) do mesmo teorema, nos dá condições de deduzir se os pontos que estamos investigando são efetivamente pontos de mínimos locais.

Observando o exemplo 1.32, percebemos que uma função pode apresentar vários pontos de mínimos locais. Mas para o estudo dos *mínimos quadrados*, estamos interessados em determinar que o erro seja o menor possível, isto é, do ponto de vista do cálculo, queremos identificar dentre os pontos de mínimos locais de uma certa função, aquele que seja um mínimo global. No entanto, o teorema 1.31 nada afirma sobre a globalidade do extremo local.

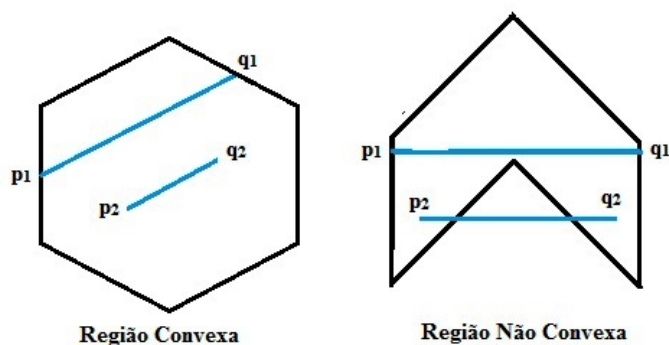
Em vista disso, precisamos estabelecer algum critério que nos garanta a globalidade do extremo local, ou mais especificamente, a globalidade do ponto de mínimo. A esse respeito, Bortolossi [4] afirma que existe uma classe específica de funções (funções convexas) para a qual é possível garantir a globalidade. Vejamos as seguintes definições e teoremas:

Definição 1.33. *Seja uma função $f : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$. Dizemos que $p_0 = (x_0, y_0) \in D$ é um ponto de mínimo global de f se para qualquer $p = (x, y) \in D$ temos que $f(p) \geq f(p_0)$. Neste caso, $f(p_0)$ é o valor mínimo global de f em D .*

Exemplo 1.34. *Sejam uma função $f : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$, dada por $f(x, y) = 1 + x^2 + y^2$ e o ponto $(x_0, y_0) = (0, 0)$. Note que $f(0, 0) = 1$ e como $x^2 + y^2 \geq 0$ para qualquer $(x, y) \in \mathbb{R}^2$, segue que $1 + x^2 + y^2 \geq 1 = f(0, 0)$. Logo, $f(x, y) \geq f(0, 0)$ para qualquer $(x, y) \in \mathbb{R}^2$, ou seja, $(x_0, y_0) = (0, 0)$ é um ponto de mínimo global de f .*

Definição 1.35. Dizemos que $D \subset \mathbb{R}^2$ é um conjunto convexo quando para quaisquer $p, q \in D$, verificamos que $[(1-t)p + tq] \in D$, para todo $t \in [0, 1]$. Isto é, se o segmento de reta que une dois pontos quaisquer de D está sempre contido em D . A figura 1.3 ilustra isto:

Figura 1.3: Região convexa e não convexa



Fonte: Figura gerada pelo autor

Definição 1.36. Seja uma função $f : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$, com $D \subset \mathbb{R}^2$ um conjunto convexo. Dizemos que f é uma função convexa quando

$$f[(1-t)p + tq] \leq (1-t)f(p) + tf(q),$$

para todo $p, q \in D$ e $t \in [0, 1]$. Caso a desigualdade seja verificada no sentido estrito, dizemos que a função é estritamente convexa.

Observação 1.37. (Bortolossi [4]) Uma vez escolhidos $p, q \in D$, para cada $t \in [0, 1]$ temos que:

- i. As expressões

$$x(t) = (1-t)p + tq \quad e \quad s(t) = (1-t)f(p) + tf(q)$$

compõem os pontos $(x(t), s(t))$ sobre a reta secante que passa pelos pontos $(p, f(p))$ e $(q, f(q))$;

- ii. As expressões

$$x(t) = (1-t)p + tq \quad e \quad f(x(t)) = f((1-t)p + tq)$$

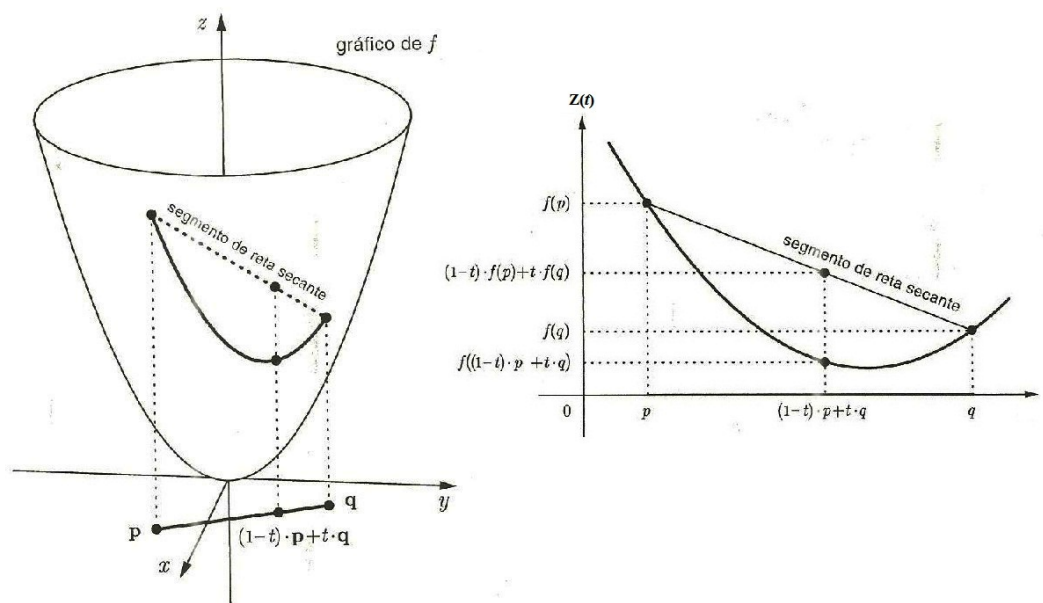
compõem os pontos $(x(t), f(x(t)))$ no gráfico de f .

Então, dizer que f satisfaz

$$f[(1-t)p + tq] \leq (1-t)f(p) + tf(q)$$

da definição 1.36, significa afirmar que o segmento de reta secante que passa pelos pontos $(p, f(p))$ e $(q, f(q))$ sempre está acima ou coincide com o gráfico de f sobre o segmento de reta que vai de p a q . A figura 1.4 ilustra essa afirmação.

Figura 1.4: Função convexa e segmento secante



Fonte: Adaptado de Bortolossi

Os teoremas a seguir recorrem à *derivada direcional* para estabelecer a convexidade de uma função diferenciável. Então, vamos efetuar alguns esclarecimentos para melhor entendimento desses teoremas. De acordo com Bortolossi [4], dada uma função $f : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$, onde $D \subset \mathbb{R}^2$ é um aberto, e um ponto $p = (a, b) \in D$, as derivadas direcionais representam, geometricamente, um deslocamento no domínio D de f por uma reta que passa pelo ponto $p = (a, b)$ e cuja direção é dada por um vetor unitário $\vec{v} = (v_1, v_2)$ denominado vetor diretor. Assim, a medida que efetuamos o deslocamento, analisamos o comportamento de f sobre a reta.

Mas, para que possamos realizar o deslocamento, devemos considerar o traço da curva parametrizada $\alpha(t) = p + t\vec{v}$ que descreve a mencionada reta, tal que esta é paralela ao vetor \vec{v} , $\alpha(0) = (a, b)$ e, para valores suficientemente pequenos de t , $\alpha(t) \in D$. Isto feito, analisamos o comportamento de f sobre a reta, estudando a função composta de uma única variável $f(\alpha(t)) = f(a + tv_1, b + tv_2)$. Ou seja, determinamos o limite

$$\lim_{t \rightarrow 0} \frac{f(\alpha(t)) - f(\alpha(0))}{t} = \lim_{t \rightarrow 0} \frac{f(a + tv_1, b + tv_2) - f(a, b)}{t} = \lim_{t \rightarrow 0} \frac{f(p + t\vec{v}) - f(p)}{t},$$

caso ele exista.

Assim, de acordo com a definição 1.14 podemos denotar

$$\lim_{t \rightarrow 0} \frac{f(p + t\vec{v}) - f(p)}{t} = \frac{\partial f}{\partial \vec{v}}(p) \quad (1.30)$$

e $\frac{\partial f}{\partial \vec{v}}(p)$ é o número real denominado derivada direcional de f no ponto $p = (a, b)$ e na direção do vetor $\vec{v} = (v_1, v_2)$.

Podemos determinar a derivada direcional $\frac{\partial f}{\partial \vec{v}}(p)$, sem fazer uso do cálculo do limite

$$\lim_{t \rightarrow 0} \frac{f(p + t\vec{v}) - f(p)}{t}.$$

Para tanto, basta que façamos o produto interno entre o vetor gradiente

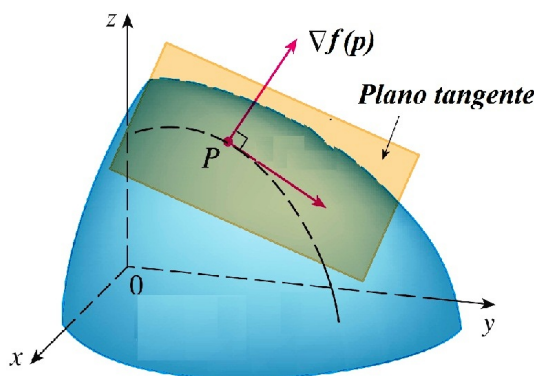
$$\nabla f(p) = \left(\frac{\partial f}{\partial x}(p), \frac{\partial f}{\partial y}(p) \right)$$

e o vetor diretor $\vec{v} = (v_1, v_2)$. Isto é [4]:

$$\frac{\partial f}{\partial \vec{v}}(p) = \nabla f(p) \cdot \vec{v} = \frac{\partial f}{\partial x}(p)v_1 + \frac{\partial f}{\partial y}(p)v_2. \quad (1.31)$$

Como a derivada direcional representa o deslocamento no domínio D de f , vale ressaltar que o vetor gradiente ∇f da função f indica o sentido e a direção na qual, por deslocamento a partir de um ponto $p = (a, b)$, obtemos a máxima variação dessa função. É perpendicular ao plano tangente ao gráfico de f no ponto $p = (a, b)$. Vejamos a figura 1.5:

Figura 1.5: Plano tangente ao gráfico e vetor gradiente perpendicular



Fonte: Figura gerada pelo autor

Teorema 1.38. *Seja um conjunto convexo $D \subset \mathbb{R}^2$ e $f : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$ uma função definida sobre uma bola aberta e diferenciável em todo ponto de D . Então, f é uma função convexa em D se, e somente se $f(q) \geq f(p) + (q - p) \cdot \nabla f(p)$, para quaisquer $p, q \in D$.*

Demonstração: (Amorim [1] e Hlenka [10])

(\Rightarrow) Seja f diferenciável e convexa em D . Tomando dois pontos quaisquer $p, q \in D$, definimos o vetor $\vec{v} = (q - p)$ e tomamos $t \in (0, 1]$. Assim, temos

$$f(p + t\vec{v}) = f(p + tq - tp) = f[tq + (1 - t)p].$$

Por hipótese

$$f[tq + (1 - t)p] \leq tf(q) + (1 - t)f(p) = tf(q) - tf(p) + f(p) = t[f(q) - f(p)] + f(p).$$

Portanto,

$$t[f(q) - f(p)] \geq t(p + t\vec{v}) - f(p).$$

Dividindo ambos os lados da desigualdade por $t > 0$, temos

$$f(q) - f(p) \geq \frac{t(p + t\vec{v}) - f(p)}{t}.$$

Tomando agora o limite quando $t \rightarrow 0^+$, obtemos

$$f(q) - f(p) \geq \lim_{t \rightarrow 0^+} \frac{t(p + t\vec{v}) - f(p)}{t}, \quad (1.32)$$

sendo que, de acordo com o visto em (1.30), o lado direito da desigualdade (1.32) é exatamente a derivada direcional de f na direção de \vec{v} , no ponto p . Observamos que embora \vec{v} não seja um vetor unitário, como t foi tomado no intervalo $(0, 1]$, o produto $t\vec{v}$ atende as condições da definição de derivada direcional.

Logo, a desigualdade (1.32) pode ser escrita como

$$f(q) - f(p) \geq \frac{\partial f}{\partial \vec{v}}(p). \quad (1.33)$$

Porém, conforme visto em (1.31)

$$\frac{\partial f}{\partial \vec{v}}(p) = \vec{v} \cdot \nabla f(p).$$

Assim, substituindo em (1.33), temos que

$$f(q) - f(p) \geq \vec{v} \cdot \nabla f(p),$$

o que é equivalente a

$$f(q) \geq f(p) + (q - p) \cdot \nabla f(p).$$

Portanto, f convexa para quaisquer $p, q \in D$, implica em $f(q) \geq f(p) + (q - p) \cdot \nabla f(p)$.

(\Leftarrow) Reciprocamente, tomando p, q quaisquer do domínio e $x = tq + (1 - t)p$ com $t \in [0, 1]$, por hipótese temos que

$$f(x) + (p - x) \cdot \nabla f(x) \leq f(p) \quad e \quad f(x) + (q - x) \cdot \nabla f(x) \leq f(q).$$

Multiplicando a primeira desigualdade por $(1 - t) > 0$, obtemos

$$(1 - t)f(x) + (1 - t)(p - x) \cdot \nabla f(x) \leq (1 - t)f(p). \quad (1.34)$$

Multiplicando a segunda desigualdade por $t > 0$, obtemos

$$tf(x) + t(q - x) \cdot \nabla f(x) \leq tf(q). \quad (1.35)$$

Somando as desigualdades (1.34) e (1.35), temos

$$tf(x) + (1 - t)f(x) + t(q - x) \cdot \nabla f(x) + (1 - t)(p - x) \cdot \nabla f(x) \leq tf(q) + (1 - t)f(p).$$

Efetuada algumas manipulações algébricas no lado esquerdo da desigualdade, obtemos

$$f(x) + [tq + (1 - t)p - x] \nabla f(x) \leq tf(q) + (1 - t)f(p).$$

Como $x = tq + (1 - t)p$, então

$$f(x) + [x - x] \nabla f(x) \leq tf(q) + (1 - t)f(p).$$

Daí segue que

$$f(x) \leq tf(q) + (1 - t)f(p).$$

Logo,

$$f[tq + (1 - t)p] \leq tf(q) + (1 - t)f(p).$$

Portanto, f é uma função convexa em D . ■

Teorema 1.39. *Seja um conjunto convexo $D \subset \mathbb{R}^2$ e $f : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$ uma função convexa em D . Então, todo ponto de mínimo local é mínimo global.*

Demonstração: (Amorim [1]) Supondo que $\bar{x} \in D$ seja um ponto de mínimo local que não seja global. Então, existe $y \in D$, tal que $f(y) < f(\bar{x})$. Tomando $z \in D$, com $z = ty + (1 - t)\bar{x}$ e $t \in (0, 1]$, uma vez que f é convexa, temos

$$f[ty + (1 - t)\bar{x}] \leq tf(y) + (1 - t)f(\bar{x}) = f(\bar{x}) + t[f(y) - f(\bar{x})],$$

ou seja,

$$f(z) \leq f(\bar{x}) + t[f(y) - f(\bar{x})].$$

Mas $f(y) - f(\bar{x}) < 0$, então para $t \in (0, 1]$ temos que $t[f(y) - f(\bar{x})] < 0$. Logo, $f(\bar{x}) + t[f(y) - f(\bar{x})] < f(\bar{x})$ e conseqüentemente $f(z) < f(\bar{x})$.

Considerando t suficientemente pequeno, podemos afirmar que z é arbitrariamente próximo de \bar{x} , e como $f(z) < f(\bar{x})$ e $z \in D$, temos uma contradição na hipótese de \bar{x} ser ponto de mínimo local que não seja global. Portanto, todo ponto de mínimo local deve ser mínimo global. ■

Definição 1.40. *Seja uma função $f : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$ e p um ponto crítico de f . Dizemos que a matriz quadrada $(n \times n)$ das derivadas parciais segundas de f é a matriz hessiana $Hf(p)$ de f .*

Observação 1.41. O teorema de Schwarz (vide [12]) afirma que se as derivadas parciais são contínuas até segunda ordem, então a ordem de derivação não importa, ou seja,

$$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x}.$$

Assim, como consequência desse teorema, temos que dada uma função $f : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$ e p um ponto crítico de f , a matriz hessiana $Hf(p)$ é simétrica em cada ponto de D quando as derivadas parciais segundas de f são contínuas em todo ponto de D .

Definição 1.42. *Seja H uma matriz quadrada de ordem $n \times n$. Um menor de ordem k é o determinante de uma submatriz $k \times k$ de H , obtida removendo-se $(n - k)$ linhas e $(n - k)$ colunas de H .*

Definição 1.43. *Um menor principal de ordem k é um menor de ordem k no qual as linhas e colunas de mesmo índice foram removidas. Mais especificamente, se as linhas i_1, i_2, \dots, i_{n-k} foram removidas, então as colunas removidas são as de índice j_1, j_2, \dots, j_{n-k} .*

Definição 1.44. *Seja H uma matriz¹ quadrada $n \times n$ formada por números reais. A forma quadrática associada à matriz H é a função escalar $Q : \mathbb{R}^n \rightarrow \mathbb{R}$, definida por $Q(h) = h^T \cdot H \cdot h$. Isto é,*

$$Q(h_1, h_2, \dots, h_n) = \begin{bmatrix} h_1 & h_2 & \dots & h_n \end{bmatrix} \cdot \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ \vdots & \dots & \dots & \vdots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \cdot \begin{bmatrix} h_1 \\ \vdots \\ h_n \end{bmatrix}$$

Definição 1.45. *Sejam H uma matriz $n \times n$ e $Q(h) = h^T \cdot H \cdot h$ a forma quadrática associada. Dizemos que:*

- i. a matriz H (ou a forma quadrática Q) é positiva definida quando $Q(h) > 0$ para todo $h \neq 0$ em \mathbb{R}^n ;*
- ii. a matriz H (ou a forma quadrática Q) é positiva semidefinida quando $Q(h) \geq 0$ para todo $h \neq 0$ em \mathbb{R}^n .*

Teorema 1.46. *Seja H uma matriz quadrada de ordem 2 simétrica. H é positiva e semidefinida se, e somente se, todos os seus menores principais são maiores ou igual a zero.*

¹As definições e teoremas pertinentes a teoria das matrizes serão descritas na seção 1.2

Demonstração: (Bortolossi [4]) Tomando a matriz

$$H = \begin{bmatrix} a & b \\ b & c \end{bmatrix}.$$

Da definição 1.44, temos que a forma quadrática associada à matriz H é dada por $Q(h) = h^T \cdot H \cdot h$. Em notação matricial, segue que

$$Q(h_1, h_2) = \begin{bmatrix} h_1 & h_2 \end{bmatrix} \cdot \begin{bmatrix} a & b \\ b & c \end{bmatrix} \cdot \begin{bmatrix} h_1 \\ h_2 \end{bmatrix}.$$

De onde obtemos

$$Q(h_1, h_2) = ah_1^2 + 2bh_1h_2 + ch_2^2.$$

Completando quadrado, segue que

$$\begin{aligned} Q(h_1, h_2) &= a \left(h_1^2 + \frac{2b}{a} h_1 h_2 + \frac{b^2}{a^2} h_2^2 \right) + ch_2^2 - \frac{ab^2 h_2^2}{a^2} \\ Q(h_1, h_2) &= a \left(h_1 + \frac{b}{a} h_2 \right)^2 + \frac{a^2 ch_2^2 - ab^2 h_2^2}{a^2} \\ Q(h_1, h_2) &= a \left(h_1 + \frac{b}{a} h_2 \right)^2 + \frac{ac - b^2}{a} h_2^2. \end{aligned} \quad (1.36)$$

Do item (ii) da definição 1.45, temos que a matriz H é positiva e semidefinida quando $Q(h_1, h_2) \geq 0$. Ou seja, quando

$$a \left(h_1 + \frac{b}{a} h_2 \right)^2 + \frac{ac - b^2}{a} h_2^2 \geq 0.$$

Das definições 1.42 e 1.43 temos que os menores principais da matriz H , são

$$|a| = a, \quad |c| = c \quad e \quad \begin{vmatrix} a & b \\ b & c \end{vmatrix} = ac - b^2.$$

Assim, queremos mostrar que $a \left(h_1 + \frac{b}{a} h_2 \right)^2 + \frac{ac - b^2}{a} h_2^2 \geq 0$ se, e somente se, $a \geq 0$, $c \geq 0$ e $ac - b^2 \geq 0$.

(\Rightarrow) Supondo que $Q(h_1, h_2) \geq 0$ implica em $a \geq 0$, $c \geq 0$ e $ac - b^2 \geq 0$. Então, tomando a equação (1.36) e fazendo

$$Q(1, 0) = a \cdot \left(1 + \frac{b}{a} \cdot 0 \right)^2 + \frac{ac - b^2}{a} \cdot 0,$$

resulta em

$$Q(1, 0) = a$$

Tomando a mesma equação e fazendo

$$Q(0, 1) = a \cdot \left(0 + \frac{b}{a} \cdot 1\right)^2 + \frac{ac - b^2}{a} \cdot 1$$

$$Q(0, 1) = a \cdot \left(\frac{b}{a}\right)^2 + \frac{ac - b^2}{a},$$

$$Q(0, 1) = \frac{b^2}{a} + \frac{ac - b^2}{a},$$

obtemos

$$Q(0, 1) = c.$$

Como por hipótese $Q(h_1, h_2) \geq 0$, temos respectivamente que $Q(1, 0) = a \geq 0$ e $Q(0, 1) = c \geq 0$. Ou seja, $Q(h_1, h_2) \geq 0$ implica em $a \geq 0$ e $c \geq 0$.

Vamos agora considerar três subcasos:

i. Tomando $Q\left(\frac{-b}{a}, 1\right)$ e $a > 0$, temos que

$$Q\left(\frac{-b}{a}, 1\right) = a \left(\frac{-b}{a} + \frac{b}{a}\right)^2 + \frac{ac - b^2}{a} \cdot 1$$

$$Q\left(\frac{-b}{a}, 1\right) = \frac{ac - b^2}{a}.$$

Como por hipótese $Q(h_1, h_2) \geq 0$, então $Q\left(\frac{-b}{a}, 1\right) = \frac{ac - b^2}{a} \geq 0$ e, portanto, $ac - b^2 \geq 0$;

ii. Tomando $Q\left(1, \frac{-b}{c}\right)$ e $c > 0$, temos que

$$Q\left(1, \frac{-b}{c}\right) = a \left(1 + \frac{b}{a} \cdot \left(\frac{-b}{c}\right)\right)^2 + \frac{ac - b^2}{a} \cdot \left(\frac{-b}{c}\right)^2$$

$$Q\left(1, \frac{-b}{c}\right) = a \left(1 - \frac{b^2}{ac}\right)^2 + \frac{ac - b^2}{a} \cdot \left(\frac{-b}{c}\right)^2$$

$$Q\left(1, \frac{-b}{c}\right) = \frac{(ac - b^2)^2}{ac^2} + \frac{(ac - b^2) \cdot (-b^2)}{ac^2}$$

$$Q\left(1, \frac{-b}{c}\right) = \frac{(ac - b^2) \cdot (ac - b^2 + b^2)}{ac^2}$$

$$Q\left(1, \frac{-b}{c}\right) = \frac{ac - b^2}{c}.$$

Como por hipótese $Q(h_1, h_2) \geq 0$, então $Q\left(1, \frac{-b}{c}\right) = \frac{ac - b^2}{c} \geq 0$ e, portanto, $ac - b^2 \geq 0$;

iii. Tomando $Q(1, 1)$ e $a = c = 0$, temos

$$Q(1, 1) = a\left(1 + \frac{b}{a}\right)^2 + \frac{ac - b^2}{a} \cdot 1$$

$$Q(1, 1) = a\left(\frac{a + b}{a}\right)^2 + \frac{ac - b^2}{a}$$

$$Q(1, 1) = \frac{(a + b)^2}{a} + \frac{ac - b^2}{a}$$

$$Q(1, 1) = \frac{a^2 + 2ab + b^2 + ac - b^2}{a}$$

$$Q(1, 1) = a + 2b + c$$

Tomando $Q(1, -1)$ e $a = c = 0$, temos

$$Q(1, -1) = a\left(1 - \frac{b}{a}\right)^2 + \frac{ac - b^2}{a} \cdot (-1)^2$$

$$Q(1, -1) = a\left(\frac{a - b}{a}\right)^2 + \frac{ac - b^2}{a}$$

$$Q(1, -1) = \frac{(a - b)^2}{a} + \frac{ac - b^2}{a}$$

$$Q(1, -1) = \frac{a^2 - 2ab + b^2 + ac - b^2}{a}$$

$$Q(1, -1) = a - 2b + c$$

Como por hipótese $Q(h_1, h_2) \geq 0$ e $a = c = 0$, então $Q(1, 1) = 2b \geq 0$ e $Q(1, -1) = -2b \geq 0$, portanto, $b \geq 0$ e $b \leq 0$. Logo, $b = 0$ e conseqüentemente $ac - b^2 = 0$. Assim, nos três subcasos observamos que $ac - b^2 \geq 0$, ou seja, $Q(h_1, h_2) \geq 0$ implica em $ac - b^2 \geq 0$. Desta forma, concluímos que $Q(h_1, h_2) \geq 0$ implica em $a \geq 0$, $c \geq 0$ e $ac - b^2 \geq 0$.

(\Leftarrow) Reciprocamente, supondo que $a \geq 0$, $c \geq 0$ e $ac - b^2 \geq 0$ implicam em $Q(h_1, h_2) \geq 0$. Segue que:

i. Se $a > 0$, então tomando $c \geq b^2$ e $b^2 \geq \frac{b^2}{a}$ temos que:

$$c \geq \frac{b^2}{a},$$

multiplicando ambos os lados da desigualdade por a , obtemos

$$ac \geq b^2,$$

subtraindo b^2 em ambos os lados da desigualdade anterior, obtemos

$$ac - b^2 \geq 0,$$

dividindo ambos os lados da desigualdade por a , temos

$$\frac{ac - b^2}{a} \geq 0,$$

multiplicando ambos os lados da desigualdade anterior por h_2^2 , resulta em

$$\frac{ac - b^2}{a} h_2^2 \geq 0. \quad (1.37)$$

Tomando agora $\frac{a}{b} \geq -h_2$, com $b > 0$ e $-h_2 \geq \frac{-h_2}{h_1}$, com $h_1 > 0$, temos que:

$$\frac{a}{b} \geq \frac{-h_2}{h_1},$$

multiplicando ambos os lados da desigualdade por b , obtemos

$$a \geq \frac{-bh_2}{h_1},$$

multiplicando ambos os lados da desigualdade anterior por h_1 , segue que

$$ah_1 \geq -bh_2,$$

somando bh_2 a ambos os lados da desigualdade anterior, temos

$$ah_1 + bh_2 \geq 0,$$

elevando ambos os lados da desigualdade ao quadrado, obtemos

$$(ah_1 + bh_2)^2 \geq 0,$$

daí segue que

$$a^2h_1^2 + 2ah_1bh_2 + b^2h_2^2 \geq 0,$$

dividindo ambos os lados da desigualdade anterior por a , temos

$$ah_1^2 + 2bh_1h_2 + \frac{b}{a}h_2^2,$$

colocando a em evidência no lado esquerdo da desigualdade, obtemos

$$a \left(h_1^2 + 2h_1 \frac{b}{a} h_2 + \frac{b^2}{a^2} h_2^2 \right) \geq 0,$$

que resulta em

$$a \left(h_1 + \frac{b}{a} h_2 \right)^2 \geq 0. \quad (1.38)$$

Somando (1.37) e (1.38), obtemos

$$a \left(h_1 + \frac{b}{a} h_2 \right)^2 + \frac{ac - b^2}{a} h_2^2 \geq 0. \quad (1.39)$$

- ii. Se $c > 0$, então tomando $a \geq b^2$ e $b^2 \geq \frac{b^2}{c}$, temos por processo análogo ao realizado no item anterior que:

$$\frac{ac - b^2}{c} h_1^2 + c \left(h_1 + \frac{b}{c} h_2 \right)^2 \geq 0. \quad (1.40)$$

- iii. Se $a = c = 0$, então $ac - b^2 = -b^2 \geq 0$, que equivale a $b^2 \leq 0$, logo $b = 0$. Tomando a equação (1.39) temos que

$$a \left(h_1 + \frac{b}{a} h_2 \right)^2 + \frac{ac - b^2}{a} h_2^2 = 0$$

Assim, concluímos que $a \geq 0$, $c \geq 0$ e $ac - b^2 \geq 0$ implicam em $Q(h_1, h_2) \geq 0$.

Portanto, $Q(h_1, h_2) \geq 0$ implica em $a \geq 0$, $c \geq 0$ e $ac - b^2 \geq 0$, e, $a \geq 0$, $c \geq 0$ e $ac - b^2 \geq 0$ implicam em $Q(h_1, h_2) \geq 0$, então a matriz H é positiva e semidefinida. ■

Teorema 1.47. *Seja um conjunto convexo e aberto $D \subset \mathbb{R}^2$ e $f : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$ uma função continuamente duas vezes diferenciável em todo ponto de D . Então f é uma função convexa em D se, e somente se, a matriz hessiana $Hf(p)$ é positiva e semidefinida para todo $p \in D$.*

Demonstração: (Hlenka [10]) (\Rightarrow) Supondo f convexa em D e tomando dois pontos r e s quaisquer do domínio, temos por hipótese

$$f(r) \geq f(s) + (r - s)\nabla f(s). \quad (1.41)$$

Como f é continuamente duas vezes diferenciável sobre \mathbb{R}^2 , temos pela fórmula de Taylor, que existe $t \in (0, 1)$ tal que

$$f(r) = f(s) + (r - s)\nabla f(s) + \frac{1}{2}(r - s)^T \nabla^2 f(p)(r - s), \quad (1.42)$$

onde $\nabla^2 f(p)$ é a matriz hessiana $Hf(p)$ e $p = tr + (1 - t)s$ é um ponto do segmento de reta que une r e s .

Então, substituindo (1.42) em (1.41), obtemos

$$f(s) + (r - s)\nabla f(s) + \frac{1}{2}(r - s)^T \nabla^2 f(p)(r - s) \geq f(s) + (r - s)\nabla f(s),$$

que corresponde a

$$(r - s)^T \nabla^2 f(p)(r - s) \geq 0.$$

Denotando $(r - s) = h$, segue que

$$h^T \nabla^2 f(p)h \geq 0,$$

de onde vem

$$h^T Hf(p)h \geq 0.$$

Assim, segundo as definições 1.44 e 1.45, a matriz hessiana $Hf(p)$ é positiva e semidefinida.

(\Leftarrow) Reciprocamente supondo que a matriz hessiana $Hf(p)$ é positiva e semidefinida, e tomando os pontos $r, s, h \in D$ tais que $p = rt + (1 - t)s$ com $t \in (0, 1)$ e $h = (r - s)$. Temos por hipótese que

$$h^T Hf(p)h \geq 0,$$

o que corresponde a

$$(r - s)^T \nabla^2 f(p)(r - s) \geq 0. \quad (1.43)$$

Multiplicando ambos os lados da desigualdade (1.43) por $\frac{1}{2}$, obtemos

$$\frac{1}{2}(r - s)^T \nabla^2 f(p)(r - s) \geq 0. \quad (1.44)$$

Tomando $f(s) + (r - s)\nabla f(s)$ e somando em ambos os lados da desigualdade (1.44), resulta em

$$f(s) + (r - s)\nabla f(s) + \frac{1}{2}(r - s)^T \nabla^2 f(p)(r - s) \geq f(s) + (r - s)\nabla f(s). \quad (1.45)$$

Mas, pela fórmula de Taylor, temos que

$$f(s) + (r - s)\nabla f(s) + \frac{1}{2}(r - s)^T \nabla^2 f(p)(r - s) = f(r). \quad (1.46)$$

Então, substituindo (1.46) no lado esquerdo da desigualdade (1.45), obtemos

$$f(r) \geq f(s) + (r - s)\nabla f(s).$$

Portanto, f é convexa em D . ■

O teorema 1.38 nos garante a convexidade de uma dada função diferenciável. Garantida a convexidade, o teorema 1.39 nos diz que todo ponto mínimo local é mínimo global. No entanto, segundo Bortolossi [4], o teorema 1.38 não é prático para testarmos a convexidade da função, uma vez que exige que verifiquemos uma desigualdade para todo par de pontos p e q pertencente ao domínio de f . Nesse sentido, o teorema 1.47 é uma alternativa mais prática para constatar tal convexidade, pois atrela esta à classificação da matriz hessiana.

Estes resultados são relevantes para o estudo do método dos mínimos quadrados, pois se mostrarmos que a função que minimiza os erros é convexa, garantiremos que possui ponto de mínimo global, ou seja, garantiremos que tal função nos dá o menor erro possível na predição dos dados experimentais.

2 Noções de álgebra linear

No capítulo 3, constataremos que a minimização do erro requer a determinação de parâmetros de uma função que melhor se ajusta ao comportamento dos dados experimentais, isto é, dos valores observados no experimento. Para determinarmos tais parâmetros, necessitamos solucionar um sistema de equações lineares. Nesse sentido, esse capítulo tem por objetivo apresentar os resultados da álgebra linear que nos fornecem embasamento teórico para discutirmos a solução desse sistema.

Como veremos mais adiante nessa seção, um sistema de equações lineares pode ser escrito na forma matricial (equação matricial). Assim sendo, ao manipularmos as matrizes geradas pelo sistema, estamos efetuando operações básicas da álgebra matricial. Particularmente, nosso interesse está voltado para *inversibilidade* da matriz, pois essa propriedade é relevante para resolução da equação matricial correspondente ao sistema de equações lineares objeto de nosso estudo.

Em vista disto, para que possamos entender o encadeamento das operações matriciais envolvidas na solução do sistema, vamos observar as seguintes definições e teoremas:

Definição 2.1. *Sejam $I = \{1, 2, \dots, m\}$ e $J = \{1, 2, \dots, n\}$ dois subconjuntos de \mathbb{N} . Denominamos produto cartesiano de I e J , ao conjunto $I \times J$ cujos elementos são os pares ordenados de números (x, y) em que $x \in I$ e $y \in J$, isto é: $I \times J = \{(x, y); x \in I \text{ e } y \in J\}$. Denominamos matriz $m \times n$ toda aplicação $f: I \times J \rightarrow \mathbb{R}$, isto é, uma correspondência na qual associamos ao elemento genérico $(i, j) \in I \times J$ um único elemento $a_{ij} \in \mathbb{R}$. O número a_{ij} é chamado imagem do par (i, j) . A imagem da aplicação $f: I \times J \rightarrow \mathbb{R}$ é o conjunto $\{a_{11}, a_{12}, \dots, a_{mn}\} \subset \mathbb{R}$ e os elementos desse conjunto são os elementos da matriz. Representamos uma matriz A , $m \times n$, por uma tabela na qual os números a_{ij} são distribuídos em m linhas e n colunas, e na qual o número a_{ij} é colocado na intersecção das linhas de índice i com as colunas de índice j . Tal tabela é representada da seguinte maneira:*

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}_{m \times n} \quad \text{ou} \quad A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}_{m \times n}$$

Definição 2.2. Duas matrizes $A = (a_{ij})$ e $B = (b_{ij})$, de mesmo tipo $m \times n$, são iguais se apresentam todos os elementos correspondentes iguais, ou seja, se $a_{ij} = b_{ij}$, qualquer que seja $i \in \{1, 2, \dots, m\}$ e $j \in \{1, 2, \dots, n\}$.

Definição 2.3. Denominamos matriz identidade de ordem n , a matriz quadrada (número de linhas igual ao número de colunas) indicada por $I_n = (\delta_{ij})$, onde

$$\delta_{ij} = \begin{cases} 1 & \text{se } i = j \\ 0 & \text{se } i \neq j \end{cases}, \quad \text{com } i, j \in \{1, 2, \dots, n\}.$$

A matriz identidade pode ser representada por

$$I_n = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}_{n \times n}$$

Definição 2.4. Dadas duas matrizes $A = (a_{ij})$, $m \times p$, e $B = (b_{jk})$, $p \times n$, denominamos produto de A por B , indicado por AB , a matriz $C = (c_{ik})$, $m \times n$, tal que

$$c_{ik} = a_{i1} \cdot b_{1k} + a_{i2} \cdot b_{2k} + \cdots + a_{ip} \cdot b_{pk} = \sum_{j=1}^p a_{ij} \cdot b_{jk},$$

para $\{i = 1, 2, 3, \dots, m\}$ e $\{j = 1, 2, 3, \dots, n\}$.

Exemplo 2.5. Sejam as matrizes

$$A = \begin{bmatrix} 1 & 1 & 2 \\ 2 & 3 & 1 \end{bmatrix}_{2 \times 3} \quad \text{e} \quad B = \begin{bmatrix} 4 \\ 0 \\ 5 \end{bmatrix}_{3 \times 1},$$

temos que

$$AB = \begin{bmatrix} 1 \cdot 4 + 1 \cdot 0 + 2 \cdot 5 \\ 2 \cdot 4 + 3 \cdot 0 + 1 \cdot 5 \end{bmatrix}_{2 \times 1} = \begin{bmatrix} 14 \\ 13 \end{bmatrix}_{2 \times 1}.$$

Dentre as propriedades do produto de matrizes, destacamos duas que são pertinentes aos nossos interesses. São elas:

Teorema 2.6. (Propriedade associativa do produto de matrizes) Dadas as matrizes A, B e C , de tipos $m \times n$, $n \times p$ e $p \times r$ respectivamente, temos que $(AB)C = A(BC)$.

Demonstração: (Caroli [5]) Sejam $A = (a_{ij})$, $B = (b_{jk})$ e $C = (c_{ks})$, com $1 \leq i \leq m$, $1 \leq j \leq n$, $1 \leq k \leq p$ e $1 \leq s \leq r$.

Denotando $AB = (d_{ik})$, $BC = (e_{js})$, $(AB)C = (f_{is})$ e $A(BC) = (g_{is})$. Queremos mostrar que $f_{is} = g_{is}$. Da definição 2.4, temos que

$$f_{is} = \sum_{k=1}^p d_{ik} c_{ks},$$

mas

$$d_{ik} = \sum_{j=1}^n a_{ij}b_{jk}.$$

Então,

$$f_{is} = \sum_{k=1}^p \left(\sum_{j=1}^n a_{ij}b_{jk} \right) c_{ks} = \sum_{k=1}^p \left(\sum_{j=1}^n a_{ij}b_{jk}c_{ks} \right) = \sum_{j=1}^n \left(\sum_{k=1}^p a_{ij}b_{jk}c_{ks} \right) = \sum_{j=1}^n a_{ij} \left(\sum_{k=1}^p b_{jk}c_{ks} \right).$$

Pela mesma definição, temos

$$\sum_{k=1}^p b_{jk}c_{ks} = e_{js}.$$

Daí, segue que

$$\sum_{j=1}^n a_{ij} \left(\sum_{k=1}^p b_{jk}c_{ks} \right) = \sum_{j=1}^n a_{ij}e_{js}.$$

Portanto, $f_{is} = g_{is}$, o que implica em $(AB)C = A(BC)$. ■

Teorema 2.7. *Se A é uma matriz do tipo $m \times p$, então $AI_n = A = I_nA$.*

Demonstração: (Caroli [5]) Sejam as matrizes

$$A = (a_{ij}), \quad \text{com} \quad \begin{cases} 1 \leq i \leq m \\ 1 \leq j \leq p \end{cases} \quad \text{e} \quad I_n = (\delta_{ik}), \quad \text{com} \quad \begin{cases} 1 \leq j \leq p \\ 1 \leq k \leq n \end{cases}.$$

Denotando $AI_n = (b_{ik})$ e $I_nA = (c_{ik})$, e de acordo com a definição 2.4, podemos escrever

$$b_{ik} = \sum_{j=1}^p a_{ij}\delta_{jk} = a_{i1}\delta_{1k} + a_{i2}\delta_{2k} + \cdots + a_{ip}\delta_{pk}$$

e

$$c_{ik} = \sum_{j=1}^p \delta_{jk}a_{ij} = a_{i1}\delta_{1k} + a_{i2}\delta_{2k} + \cdots + a_{ip}\delta_{pk}$$

Da definição 2.3, temos

$$\delta_{ij} = \begin{cases} 1 & \text{se } i = j \\ 0 & \text{se } i \neq j \end{cases},$$

então segue que

$$b_{ik} = a_{i1} \cdot 0 + \cdots + a_{ik} \cdot 1 + \cdots + a_{ip} \cdot 0 = a_{ik},$$

e

$$c_{ik} = a_{i1} \cdot 0 + \cdots + a_{ik} \cdot 1 + \cdots + a_{ip} \cdot 0 = a_{ik}.$$

Assim, como $b_{ik} = a_{ik} = c_{ik}$, resulta que $AI_n = A = I_nA$. ■

Definição 2.8. *Denominamos matriz transposta da matriz $A = (a_{ij}), m \times n$, a matriz $A^t = (b_{ji}), n \times m$, onde $b_{ji} = a_{ij}$, com $\begin{cases} 1 \leq i \leq m \\ 1 \leq j \leq n \end{cases}$. Isto é, as linhas da matriz A coincidem ordenadamente com as colunas da matriz A^t .*

Exemplo 2.9. Seja a matriz

$$A = \begin{bmatrix} 1 & 5 & 7 \\ 4 & 3 & 2 \\ 9 & 6 & 8 \end{bmatrix}_{3 \times 3},$$

então pela definição 2.8, a matriz transposta de A é dada por:

$$A^t = \begin{bmatrix} 1 & 4 & 9 \\ 5 & 3 & 6 \\ 7 & 2 & 8 \end{bmatrix}_{3 \times 3}.$$

Definição 2.10. Dada uma matriz A de ordem n , denominamos de inversa de A a matriz B , tal que $AB = BA = I_n$. Denotamos a matriz inversa por A^{-1} .

Teorema 2.11. Se uma matriz quadrada $A = (a_{ij})$, de ordem n , é inversível, então a sua inversa é única.

Demonstração: (Caroli [5]) Supondo que existam duas matrizes B e C inversas da matriz A . Assim, de acordo com a definição 2.10, temos que:

- i. $AB = BA = I_n$,
- ii. $AC = CA = I_n$.

Do teorema 2.7, segue que

$$B = BI_n.$$

Como por hipótese

$$I_n = AC,$$

então podemos escrever

$$B = B(AC).$$

Do teorema 2.6, temos que

$$B(AC) = (BA)C.$$

Desta forma, segue que

$$B = (BA)C.$$

Mas, por hipótese

$$(BA) = I_n,$$

então

$$B = I_n C.$$

Do teorema 2.7, temos que

$$I_n C = C,$$

o que implica em $B = C$. Portanto, $B = C = A^{-1}$ é a única inversa da matriz A . ■

Exemplo 2.12. Sejam as matrizes

$$A = \begin{bmatrix} 3 & 1 \\ 5 & 2 \end{bmatrix}_{2 \times 2} \quad e \quad B = \begin{bmatrix} 2 & -1 \\ -5 & 3 \end{bmatrix}_{2 \times 2}.$$

Então, pela definição 2.4, temos que

$$AB = \begin{bmatrix} 3 & 1 \\ 5 & 2 \end{bmatrix}_{2 \times 2} \cdot \begin{bmatrix} 2 & -1 \\ -5 & 3 \end{bmatrix}_{2 \times 2} = \begin{bmatrix} 3 \cdot 2 + 1 \cdot (-5) & 3 \cdot (-1) + 1 \cdot 3 \\ 5 \cdot 2 + 2 \cdot (-5) & 5 \cdot (-1) + 2 \cdot 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

e

$$BA = \begin{bmatrix} 2 & -1 \\ -5 & 3 \end{bmatrix}_{2 \times 2} \cdot \begin{bmatrix} 3 & 1 \\ 5 & 2 \end{bmatrix}_{2 \times 2} = \begin{bmatrix} 2 \cdot 3 + (-1) \cdot 5 & 2 \cdot 1 + (-1) \cdot 2 \\ -5 \cdot 3 + 3 \cdot 5 & -5 \cdot 1 + 3 \cdot 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Portanto, temos que $AB = BA = I_n$ e de acordo com o teorema 2.11, B é a única matriz inversa de A . Assim, podemos denotar $B = A^{-1}$.

Observamos que a definição 2.10 nos permite verificar se duas matrizes são inversas ou não, assim como, o teorema 2.11 nos garante a unicidade da matriz inversa. No entanto, estes não nos fornecem informações para que possamos verificar se uma certa matriz admite inversa ou não, muito menos um critério para determiná-la.

Nessa perspectiva, as definições e teoremas a seguir nos indicarão que a admissibilidade da inversa de uma matriz é inerente ao cálculo do determinante. Por outro lado, se a matriz admite inversa, uma maneira de encontrá-la é fazendo uso da matriz adjunta. Vejamos então, tais definições e teoremas:

Definição 2.13. Seja o conjunto $J = \{1, 2, \dots, n\}$ um subconjunto de \mathbb{N} . Denominamos uma permutação de J a correspondência bijetiva $P : J \rightarrow J$. Denotamos a permutação por

$$P = \begin{pmatrix} 1 & 2 & \dots & n \\ P(1) & P(2) & \dots & P(n) \end{pmatrix},$$

onde a primeira fileira representa a ordem dos elementos a serem permutados e a segunda fileira a nova ordem dada pela permutação.

Definição 2.14. Dado o conjunto $J = \{1, 2, \dots, n\}$ com n elementos distintos, a quantidade de permutações possíveis desses elementos é dada por $n! = n \cdot (n-1) \cdot (n-2) \cdot \dots \cdot 3 \cdot 2 \cdot 1$. Denotamos o conjunto de todas as permutações de J por

$$S = \left\{ \begin{pmatrix} 1 & 2 & \dots & n \\ P_1(1) & P_1(2) & \dots & P_1(n) \end{pmatrix}, \dots, \begin{pmatrix} 1 & 2 & \dots & n \\ P_k(1) & P_k(2) & \dots & P_k(n) \end{pmatrix} \right\}.$$

Exemplo 2.15. Seja o conjunto $J = \{1, 2, 3\}$. A quantidade de permutações possíveis dos elementos de J é $3! = 6$, são elas:

$$\begin{pmatrix} 1 & 2 & 3 \\ P_1(1) & P_1(2) & P_1(3) \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix};$$

$$\begin{aligned} \begin{pmatrix} 1 & 2 & 3 \\ P_2(1) & P_2(2) & P_2(3) \end{pmatrix} &= \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}; \\ \begin{pmatrix} 1 & 2 & 3 \\ P_3(1) & P_3(2) & P_3(3) \end{pmatrix} &= \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}; \\ \begin{pmatrix} 1 & 2 & 3 \\ P_4(1) & P_4(2) & P_4(3) \end{pmatrix} &= \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}; \\ \begin{pmatrix} 1 & 2 & 3 \\ P_5(1) & P_5(2) & P_5(3) \end{pmatrix} &= \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix}; \\ \begin{pmatrix} 1 & 2 & 3 \\ P_6(1) & P_6(2) & P_6(3) \end{pmatrix} &= \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}. \end{aligned}$$

E o conjunto de todas as permutações de J é

$$S = \left\{ \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix}, \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix} \right\}.$$

Definição 2.16. *Seja uma permutação $P \in S$. Denotamos por Z o conjunto dos pares (i, j) com $1 \leq i < j \leq n$, tais que $P(i) < P(j)$ e por W o conjunto dos pares (i, j) com $1 \leq i < j \leq n$, tais que $P(i) > P(j)$. Seja t o número de pares (i, j) pertencentes a W , denominamos t por número de inversões de P .*

Exemplo 2.17. Sejam

$$P_1 = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 3 & 2 & 5 & 4 \end{pmatrix} \quad e \quad P_2 = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 \\ 3 & 1 & 2 & 5 & 4 \end{pmatrix}.$$

Temos que

- i. $P_1(2) > P_1(3)$ e $P_1(4) > P_1(5)$,
- ii. $P_2(1) > P_2(2)$, $P_2(1) > P_2(3)$ e $P_2(4) > P_2(5)$.

Então, o número de inversões de P_1 e P_2 são respectivamente $t = 2$ e $t = 3$.

Definição 2.18. *Seja uma permutação $P \in S$. Denotamos o sinal de P por*

$$\varepsilon(P) = (-1)^t,$$

onde t é o número de inversões de P .

Exemplo 2.19. Do exemplo 2.17 temos que $\varepsilon(P_1) = (-1)^2 = 1$ e $\varepsilon(P_2) = (-1)^3 = -1$.

Definição 2.20. *Sejam a matriz $A = (a_{ij})$ de ordem n , o conjunto $J = \{1, 2, \dots, n\}$ de índices j da matriz A , a permutação P dos elementos de J e $\varepsilon(P)$ o sinal da permutação P . Denominamos determinante da matriz A o número real dado por*

$$\sum_{k=1}^{n!} \varepsilon(P_k) a_{1p_k(1)} a_{2p_k(2)} \cdots a_{np_k(n)}$$

e denotamos por $\det(A)$ ou $|A|$.

Exemplo 2.21. Seja a matriz

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}_{3 \times 3}.$$

O conjunto $J = \{1, 2, 3\}$ de índices j da matriz A possui 6 permutações possíveis de seus elementos, são elas:

$$P_1 = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{pmatrix}, P_2 = \begin{pmatrix} 1 & 2 & 3 \\ 1 & 3 & 2 \end{pmatrix}, P_3 = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix},$$

$$P_4 = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 3 & 1 \end{pmatrix}, P_5 = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 1 & 2 \end{pmatrix}, P_6 = \begin{pmatrix} 1 & 2 & 3 \\ 3 & 2 & 1 \end{pmatrix}.$$

Observamos que o número de inversões em P_1, P_2, P_3, P_4, P_5 e P_6 é respectivamente $t = 0, t = 1, t = 1, t = 2, t = 2$ e $t = 3$. Assim, calculando o sinal $\varepsilon(P)$ de cada permutação temos: $\varepsilon(P_1) = (-1)^0 = 1$; $\varepsilon(P_2) = (-1)^1 = -1$; $\varepsilon(P_3) = (-1)^1 = -1$; $\varepsilon(P_4) = (-1)^2 = 1$; $\varepsilon(P_5) = (-1)^2 = 1$ e $\varepsilon(P_6) = (-1)^3 = -1$.

Tomando os sinais das permutações e calculando o determinante da matriz A pela definição, temos que

$$\det(A) = \sum_{k=1}^6 \varepsilon(P_k) a_{1p_k(1)} a_{2p_k(2)} a_{3p_k(3)} =$$

$$\varepsilon(P_1) a_{1p_1(1)} a_{2p_1(2)} a_{3p_1(3)} + \varepsilon(P_2) a_{1p_2(1)} a_{2p_2(2)} a_{3p_2(3)} + \varepsilon(P_3) a_{1p_3(1)} a_{2p_3(2)} a_{3p_3(3)} +$$

$$+ \varepsilon(P_4) a_{1p_4(1)} a_{2p_4(2)} a_{3p_4(3)} + \varepsilon(P_5) a_{1p_5(1)} a_{2p_5(2)} a_{3p_5(3)} + \varepsilon(P_6) a_{1p_6(1)} a_{2p_6(2)} a_{3p_6(3)},$$

o que corresponde a

$$\det(A) = a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} - a_{12}a_{21}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31},$$

ou

$$\det(A) = (a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32}) - (a_{11}a_{23}a_{32} + a_{12}a_{21}a_{33} + a_{13}a_{22}a_{31}).$$

Calcular o determinante de uma matriz de ordem $n > 3$ diretamente pela definição é um processo muito dispendioso, uma vez que o número de parcelas

$$\varepsilon(P_k) a_{1p_k(1)} a_{2p_k(2)} \cdots a_{np_k(n)}$$

que o compõe cresce rapidamente a medida que aumentamos a ordem da matriz. Nesse sentido, as definições e teoremas a seguir facilitarão o cálculo do determinante.

Observação 2.22. Denominamos o menor da definição 1.42 como menor complementar e o denotamos por D_{ij} do elemento a_{ij} da matriz quadrada A de ordem n .

Definição 2.23. Denominamos complemento algébrico ou cofator do elemento a_{ij} de uma matriz quadrada A de ordem n , ao produto do menor complementar D_{ij} por $(-1)^{i+j}$, isto é: $A_{ij} = (-1)^{i+j}D_{ij}$.

Definição 2.24. Fixando a coluna de índice 1 da matriz quadrada $A = (a_{ij})$ de ordem n , dizemos que o determinante de A é a soma dos produtos dos elementos da coluna de índice 1 pelos respectivos cofatores, ou seja: $\det(A) = \sum_{i=1}^n a_{i1}A_{i1}$.

Teorema 2.25. Seja a matriz quadrada $A = (a_{ij})$ de ordem n , se a matriz transposta de A é $A^t = (b_{ji})$, então $\det(A) = \det(A^t)$.

Demonstração: (Guelli [8]) Sejam as matrizes B_1 e B_2 , de ordem $(n-1)$, cujos elementos são os respectivos cofatores B_{ji} e A_{ij} , dos respectivos elementos $b_{ji} \in A^t$ e $a_{ij} \in A$. Assim, de acordo com as definições 2.23 e 2.24, desenvolvemos $\det(A^t)$ e $\det(A)$, obtendo:

$$\det(A^t) = b_{11}B_{11} - b_{21}B_{21} + b_{31}B_{31} - \dots + (-1)^{n+1}b_{n1}B_{n1} = \sum_{j=1}^n b_{j1}B_{j1}$$

e

$$\det(A) = a_{11}A_{11} - a_{21}A_{21} + a_{31}A_{31} - \dots + (-1)^{n+1}a_{n1}A_{n1} = \sum_{i=1}^n a_{i1}A_{i1}.$$

Pela definição 2.8 temos que $b_{ji} = a_{ij}$, então $B_{ji} = A_{ij}$. Daí segue que

$$b_{11}B_{11} - b_{21}B_{21} + b_{31}B_{31} - \dots + (-1)^{n+1}b_{n1}B_{n1} = a_{11}A_{11} - a_{21}A_{21} + a_{31}A_{31} - \dots + (-1)^{n+1}a_{n1}A_{n1},$$

o que resulta em

$$\sum_{j=1}^n b_{j1}B_{j1} = \sum_{i=1}^n a_{i1}A_{i1}.$$

Portanto, concluímos que

$$\det(A^t) = \det(A).$$

■

Exemplo 2.26. Seja a matriz $A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 7 & 5 \\ 9 & 8 & 6 \end{bmatrix}$ e sua transposta $A^t = \begin{bmatrix} 1 & 4 & 9 \\ 2 & 7 & 8 \\ 3 & 5 & 6 \end{bmatrix}$. O

desenvolvimento dos determinantes de A e A^t pela primeira coluna são dados por:

i. $\det(A) = a_{11}A_{11} - a_{21}A_{21} + a_{31}A_{31} = 2 + 48 - 99 = -49$;

ii. $\det(A^t) = b_{11}B_{11} - b_{21}B_{21} + b_{31}B_{31} = 2 + 42 - 93 = -49$.

Portanto, temos $\det(A) = \det(A^t) = -49$.

Teorema 2.27. (Teorema de Laplace) Se $A = (a_{ij})$ é uma matriz quadrada de ordem n , então, o $\det(A)$ é a soma dos produtos dos elementos de uma fila (linha ou coluna) de A pelos respectivos cofatores. Isto é: $\det(A) = \sum_{i=1}^n a_{ij}A_{ij}$, para cada índice j fixado ou $\det(A) = \sum_{j=1}^n a_{ij}A_{ij}$, para cada índice i fixado.

Demonstração: (SÁ [16]) Vamos demonstrar por indução.

Para $n = 2$, ou seja, para uma matriz $A = (a_{ij})$ de ordem $n = 2$, temos

$$\det(A) = a_{12}A_{12} + a_{22}A_{22} = a_{12}(-1)^{1+2}D_{12} + a_{22}(-1)^{2+2}D_{22} = a_{22}a_{11} - a_{12}a_{21}, \quad (2.1)$$

quando desenvolvemos o determinante pela segunda coluna. E

$$\det(A) = a_{11}A_{11} + a_{21}A_{21} = a_{11}(-1)^{1+1}D_{11} + a_{21}(-1)^{2+1}D_{21} = a_{22}a_{11} - a_{12}a_{21}, \quad (2.2)$$

quando desenvolvemos o determinante pela primeira coluna.

Assim, observamos que as equações (2.1) e (2.2) são coincidentes, portanto a proposição é verdadeira para $n = 2$.

Da observação 2.22 temos os menores complementares D_{ij} das entradas da matriz A . Supondo que a proposição seja verdadeira para menores complementares de uma matriz de ordem $(n-1)$, queremos mostrar que a proposição é verdadeira para menores complementares de uma matriz de ordem n .

Fixando a j -ésima coluna da matriz $A = (a_{ij})$, com $1 < j \leq n$ e desenvolvendo o determinante, obtemos

$$\det(A) = a_{1j}A_{1j} + a_{2j}A_{2j} + a_{3j}A_{3j} + \cdots + a_{nj}A_{nj}.$$

Pela definição 2.23 temos que

$$\det(A) = a_{1j}(-1)^{1+j}D_{1j} + a_{2j}(-1)^{2+j}D_{2j} + a_{3j}(-1)^{3+j}D_{3j} + \cdots + a_{nj}(-1)^{n+j}D_{nj}.$$

Mas $D_{1j}, D_{2j}, D_{3j}, \dots, D_{nj}$, são determinantes de matrizes de ordem $(n-1)$. Portanto, por hipótese de indução podemos calcular esses determinantes pelo teorema de Laplace. Fazendo o desenvolvimento pela primeira coluna da matriz gerada e denotando seus menores complementares por d_{ij} , segue que:

$$\begin{aligned} D_{1j} &= a_{21}(-1)^{2+1}d_{21} + a_{31}(-1)^{3+1}d_{31} + \cdots + a_{n1}(-1)^{n+1}d_{n1} = \sum_{i=2}^n a_{i1}(-1)^{i+1}d_{i1} \\ D_{2j} &= a_{11}(-1)^{1+1}d_{11} + a_{31}(-1)^{3+1}d_{31} + \cdots + a_{n1}(-1)^{n+1}d_{n1} = a_{11}d_{11} + \sum_{i=3}^n a_{i1}(-1)^{i+1}d_{i1} \\ &\vdots \\ D_{nj} &= a_{11}(-1)^{1+1}d_{11} + a_{21}(-1)^{2+1}d_{21} + \cdots + a_{(n-1)1}(-1)^{n-1+1}d_{(n-1)1} = \sum_{i=1}^{n-1} a_{i1}(-1)^{i+1}d_{i1}. \end{aligned}$$

Substituindo $D_{1j}, D_{2j}, D_{3j}, \dots, D_{nj}$ em $\det(A)$, obtemos:

$$\det(A) = a_{1j}(-1)^{1+j} \left\{ \sum_{i=2}^n a_{i1}(-1)^{i+1} d_{i1} \right\} + a_{2j}(-1)^{2+j} \left\{ a_{11}d_{11} + \sum_{i=3}^n a_{i1}(-1)^{i+1} d_{i1} \right\} + \dots \\ \dots + a_{nj}(-1)^{n+j} \left\{ \sum_{i=1}^{n-1} a_{i1}(-1)^{i+1} d_{i1} \right\}.$$

Tomando em $\det(A)$ todas parcelas onde a_{11} aparece, obtemos

$$a_{2j}(-1)^{2+j} a_{11} d_{11} + a_{3j}(-1)^{3+j} a_{11} d_{11} + \dots + a_{nj}(-1)^{n+j} a_{11} d_{11}.$$

Colocando em evidência o elemento a_{11} , temos

$$a_{11} \left\{ a_{2j}(-1)^{2+j} d_{11} + a_{3j}(-1)^{3+j} d_{11} + \dots + a_{nj}(-1)^{n+j} d_{11} \right\}. \quad (2.3)$$

Tomando o fator $\{a_{2j}(-1)^{2+j} d_{11} + a_{3j}(-1)^{3+j} d_{11} + \dots + a_{nj}(-1)^{n+j} d_{11}\}$ da expressão (2.3), observamos que este fator é o desenvolvimento, pela j -ésima coluna, de um determinante D da matriz $A = (a_{ij})$. Então, podemos escrevê-lo como

$$D = a_{2j}(-1)^{2+j} k_{2j} + a_{3j}(-1)^{3+j} k_{3j} + \dots + a_{nj}(-1)^{n+j} k_{nj}, \quad (2.4)$$

onde k_{ij} são os menores complementares da matriz gerada.

Mas o determinante desenvolvido na expressão (2.4) é o determinante D_{11} da matriz $A = (a_{ij})$. Logo,

$$a_{11} \left\{ a_{2j}(-1)^{2+j} d_{11} + a_{3j}(-1)^{3+j} d_{11} + \dots + a_{nj}(-1)^{n+j} d_{11} \right\} = a_{11} D_{11}.$$

Analogamente, se tomarmos em $\det(A)$ todas as parcelas onde a_{21} aparece, obtemos

$$a_{21} \left\{ a_{1j}(-1)^j d_{21} - a_{3j}(-1)^{1+j} d_{21} - \dots - a_{nj}(-1)^{n-2+j} d_{21} \right\} = -a_{21} D_{21}.$$

Assim, por hipótese de indução, se tomarmos somente as parcelas onde a_{n1} aparece, temos

$$\pm a_{n1} \left\{ a_{1j}(-1)^{1+j} d_{1j} + a_{2j}(-1)^{2+j} d_{2j} + \dots + a_{(n-1)j}(-1)^{n-1+j} d_{(n-1)j} \right\} = \pm a_{n1} D_{n1}.$$

Portanto,

$$\det(A) = a_{11} D_{11} - a_{21} D_{21} + \dots \pm a_{n1} D_{n1},$$

que é equivalente a

$$\det(A) = a_{11} A_{11} + a_{21} A_{21} + \dots \pm a_{n1} A_{n1}.$$

Deste modo, pela arbitrariedade da escolha de j , concluímos que a proposição é verdadeira para todas as colunas da matriz $A = (a_{ij})$.

Queremos agora, mostrar que a proposição é válida para todas as linhas da matriz $A = (a_{ij})$. Fixando sua i -ésima linha, com $1 < i \leq n$, tomando a matriz transposta

$A_t = b_{ji}$ da matriz A , denotando o menor complementar do elemento b_{ji} por d_{ji} e usando A_{ji}^t para indicar os cofatores de b_{ji} . Podemos de acordo com a hipótese do teorema de Laplace, calcular o determinante de A^t , fazendo seu desenvolvimento pela i -ésima coluna de sua matriz.

Assim sendo, temos que

$$\det(A^t) = b_{1i}A_{1i}^t + b_{2i}A_{2i}^t + \dots + b_{ni}A_{ni}^t,$$

isto é:

$$\det(A^t) = b_{1i}(-1)^{1+i}d_{1i} + b_{2i}(-1)^{2+i}d_{2i} + \dots + b_{ni}(-1)^{n+i}d_{ni}.$$

Mas pela definição 2.8 e pelo teorema 2.25 observamos que

$$\begin{aligned} \det(A^t) &= b_{1i}(-1)^{1+i}d_{1i} + b_{2i}(-1)^{2+i}d_{2i} + \dots + b_{ni}(-1)^{n+i}d_{ni} = \\ &= a_{1j}(-1)^{1+j}D_{1j} + a_{2j}(-1)^{2+j}D_{2j} + \dots + a_{nj}(-1)^{n+j}D_{nj} = \det(A). \end{aligned}$$

Deste modo, pela arbitrariedade da escolha de j , concluímos que a proposição é verdadeira para todas as linhas da matriz $A = (a_{ij})$.

Portanto, pelo princípio da indução, temos que a proposição é verdadeira para qualquer coluna ou linha da matriz $A = (a_{ij})$. ■

Exemplo 2.28. Seja a matriz $A = \begin{bmatrix} 1 & -1 & 2 \\ 2 & 1 & 3 \\ -4 & 5 & -2 \end{bmatrix}$, o desenvolvimento do determi-

nante de A pela segunda linha é coincidente com o desenvolvimento pela terceira coluna. Isto é:

$$\begin{aligned} \det(A) &= a_{21}A_{21} + a_{22}A_{22} + a_{23}A_{23} = a_{13}A_{13} + a_{23}A_{23} + a_{33}A_{33} = \\ &= 2 \begin{vmatrix} -1 & 2 \\ 5 & -2 \end{vmatrix} (-1)^{2+1} + 1 \begin{vmatrix} 1 & 2 \\ -4 & -2 \end{vmatrix} (-1)^{2+2} + 3 \begin{vmatrix} 1 & -1 \\ -4 & 5 \end{vmatrix} (-1)^{2+3} = \\ &= 2 \begin{vmatrix} 2 & 1 \\ -4 & 5 \end{vmatrix} (-1)^{1+3} + 3 \begin{vmatrix} 1 & -1 \\ -4 & 5 \end{vmatrix} (-1)^{2+3} + (-2) \begin{vmatrix} 1 & -1 \\ 2 & 1 \end{vmatrix} (-1)^{3+3} = 19 \end{aligned}$$

Teorema 2.29. Se A_1 é a matriz quadrada que se obtém da matriz quadrada $A = (a_{ij})$, de ordem n , trocando entre si as posições de duas linhas (ou colunas) distintas, então $\det(A_1) = -\det(A)$.

Demonstração: (Guelli [8]) Vamos demonstrar por indução.

Para a matriz quadrada $A = (a_{ij})$, de ordem 2, temos:

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \quad e \quad A_1 = \begin{bmatrix} a_{21} & a_{22} \\ a_{11} & a_{12} \end{bmatrix},$$

então segue que

$$\det(A) = a_{11}a_{22} - a_{21}a_{12} \quad e \quad \det(A_1) = a_{21}a_{12} - a_{11}a_{22}.$$

Assim, temos que $\det(A_1) = -\det(A)$. Portanto, para matrizes de ordem $n = 2$, a propriedade é verdadeira.

Supondo que a propriedade seja verdadeira para matrizes de ordem $(n - 1)$, queremos mostrar que é verdadeira para matrizes de ordem n . Então, desenvolvendo $\det(A_1)$ e $\det(A)$ pelos elementos da linha de índice p , admitindo que esta linha não seja nenhuma das duas que foram trocadas de lugar e sabendo que os menores complementares indicados por d_{ij} e e_{ij} das respectivas matrizes A e A_1 , satisfazem $d_{ij} = -e_{ij}$, de acordo com a hipótese. Observamos que

$$\det(A) = a_{p1}A_{p1} + a_{p2}A_{p2} + \cdots + a_{pn}A_{pn} = \sum_{j=1}^n a_{pj}A_{pj}$$

e

$$\det(A_1) = a_{p1}(-A_{p1}) + a_{p2}(-A_{p2}) + \cdots + a_{pn}(-A_{pn}) = -\sum_{j=1}^n a_{pj}A_{pj}.$$

Isto mostra que $\det(A_1) = -\sum_{j=1}^n a_{pj}A_{pj} = -\det(A)$. Portanto, pelo princípio de indução, a proposição é verdadeira para qualquer matriz quadrada de ordem n . ■

Exemplo 2.30. Sejam as matrizes $A = \begin{bmatrix} 2 & 1 & 8 \\ 3 & 7 & 3 \\ 4 & 2 & 2 \end{bmatrix}$ e $A_1 = \begin{bmatrix} 3 & 7 & 3 \\ 2 & 1 & 8 \\ 4 & 2 & 2 \end{bmatrix}$. Desenvolvendo

os $\det(A)$ e $\det(A_1)$ pela terceira linha, pois esta não é nenhuma das linhas permutadas, temos que

$$\det(A) = a_{31}A_{31} + a_{32}A_{32} + a_{33}A_{33} = 4 \cdot (-53) + 2 \cdot 18 + 2 \cdot 11 = -154$$

e

$$\det(A_1) = a_{31}(A_1)_{31} + a_{32}(A_1)_{32} + a_{33}(A_1)_{33} = 4 \cdot 53 + 2 \cdot (-18) + 2 \cdot (-11) = 154$$

Assim, concluímos que $\det(A_1) = -\det(A)$.

Teorema 2.31. Se uma matriz quadrada $A = (a_{ij})$, de ordem n , tem duas linhas (ou colunas) formadas por elementos respectivamente iguais, então $\det(A) = 0$.

Demonstração: (Guelli [8]) Seja uma matriz quadrada $A = (a_{ij})$, de ordem n . Supondo que as linhas de índices p e q dessa matriz sejam formadas por elementos respectivamente iguais, ou seja: $a_{pj} = a_{qj}$ para todo $j \in \{1, 2, \dots, n\}$, e que trocando entre si estas linhas, obtemos uma nova matriz A_1 , tal que $A_1 = A$. Então, de acordo com o teorema 2.29 e pela hipótese do teorema 2.31, temos respectivamente

$$-\det(A_1) = \det(A) \quad e \quad \det(A_1) = \det(A).$$

Daí segue que

$$-\det(A_1) = \det(A_1),$$

mas

$$-\det(A_1) = \det(A_1) \iff \det(A_1) = 0.$$

Portanto, concluímos que $\det(A) = 0$. ■

Exemplo 2.32. Seja a matriz $A = \begin{bmatrix} 3 & 1 & 2 \\ 4 & 1 & 5 \\ 3 & 1 & 2 \end{bmatrix}$, desenvolvendo o $\det(A)$ pela segunda

linha, obtemos

$$\det(A) = a_{21}A_{21} + a_{22}A_{22} + a_{23}A_{23} = 4 \cdot 0 + 1 \cdot 0 + 5 \cdot 0 = 0.$$

Portanto, temos que $\det(A) = 0$.

Teorema 2.33. (*Teorema de Cauchy*) Em toda matriz quadrada $A = (a_{ij})$, de ordem n , a soma dos produtos dos elementos de uma linha (ou coluna), ordenadamente, pelos cofatores dos elementos de outra linha (ou coluna) é igual a zero.

Demonstração: (Guelli [8]) Seja a matriz quadrada $A = (a_{ij})$, de ordem n . Substituindo em A a linha de índice s pela linha de índice r , obtemos a matriz A_1 com duas linhas iguais, então conforme o teorema 2.31, o $\det(A_1) = 0$. Assim sendo, se desenvolvermos o $\det(A_1)$ pelos elementos da linha de índice s , obtemos $\det(A_1) = \sum_{j=1}^n a_{sj}A_{sj}$,

mas $\det(A_1) = 0$, então $\sum_{j=1}^n a_{sj}A_{sj} = 0$. ■

Exemplo 2.34. Calculando a soma dos produtos dos elementos da primeira linha pelos

cofatores dos elementos da segunda linha da matriz $A = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$, temos

$$\begin{aligned} a(-1)^{2+1}(bi - ch) + b(-1)^{2+2}(ai - gc) + c(-1)^{2+3}(ah - gb) &= \\ = -a(bi - ch) + b(ai - gc) - c(ah - gb) &= \\ = -abi + ach + abi - bcg - ach + bcg &= 0 \end{aligned}$$

Definição 2.35. Dada uma matriz quadrada $A = (a_{ij})$, de ordem n , dizemos que a matriz dos cofatores, denotada por C , é a matriz que se obtém de A , substituindo cada elemento de A por seu cofator. Ou seja:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}_{n \times n} \rightarrow C = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \cdots & A_{nn} \end{bmatrix}_{n \times n}$$

Definição 2.36. Dada a matriz $A = (a_{ij})$, de ordem n , denominamos matriz adjunta e denotamos por \bar{A} a matriz transposta da matriz dos cofatores, isto é: $\bar{A} = C^t$.

Exemplo 2.37. Seja a matriz $A = \begin{bmatrix} 4 & 7 \\ 2 & 3 \end{bmatrix}$, da definição 2.23 temos que a matriz dos cofatores dos elementos de A é dada por $C = \begin{bmatrix} 3 & -2 \\ -7 & 4 \end{bmatrix}$, então $\bar{A} = \begin{bmatrix} 3 & -7 \\ -2 & 4 \end{bmatrix}$.

Definição 2.38. Dado um número real α e uma matriz quadrada $A = (a_{ij})$, de ordem n . Obtemos a matriz $\alpha A = (b_{ij})$, com $i, j \in \{1, 2, \dots, n\}$ quando multiplicamos por α todos os elementos de A .

Exemplo 2.39. Seja o número $\alpha = 3$ e a matriz $A = \begin{bmatrix} 4 & -1 \\ 0 & 2 \end{bmatrix}$, temos que o produto de α por A é dado da seguinte maneira:

$$\alpha A = \begin{bmatrix} 3 \cdot 4 & 3 \cdot (-1) \\ 3 \cdot 0 & 3 \cdot 2 \end{bmatrix} = \begin{bmatrix} 12 & -3 \\ 0 & 6 \end{bmatrix}.$$

Teorema 2.40. Sejam A , $\det(A)$, \bar{A} e I_n , respectivamente uma matriz quadrada de ordem n , seu determinante, sua matriz adjunta e a matriz identidade de ordem n . Então $A\bar{A} = \bar{A}A = \det(A)I_n$.

Demonstração: (Guelli [8]) Seja $A\bar{A} = (b_{ik})$, então pela definição 2.4 temos que

$$b_{ik} = \sum_{j=1}^n a_{ij} A_{jk},$$

onde A_{jk} , de acordo com as definições 2.35 e 2.36 são os elementos da matriz transposta da matriz de cofatores dos elementos da matriz A . Daí segue que

- i. Pelo teorema 2.27, $b_{ik} = \det(A)$ para $i = k$;
- ii. Pelo teorema 2.33, $b_{ik} = 0$ para $i \neq k$.

Então, a matriz $A\bar{A}$ é dada por

$$A\bar{A} = \begin{bmatrix} \det(A) & 0 & \dots & 0 \\ 0 & \det(A) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \det(A) \end{bmatrix}_{n \times n}.$$

Da definição 2.38 temos que

$$\det(A) \cdot \begin{bmatrix} 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1 \end{bmatrix}_{n \times n} = \begin{bmatrix} \det(A) & 0 & \dots & 0 \\ 0 & \det(A) & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \det(A) \end{bmatrix}_{n \times n}.$$

Logo,

$$A\bar{A} = \begin{bmatrix} \det(A) & 0 & \cdots & 0 \\ 0 & \det(A) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \det(A) \end{bmatrix}_{n \times n} = \det(A) \cdot \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}_{n \times n}.$$

Da definição 2.3 temos que

$$I_n = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}_{n \times n}.$$

Portanto, $A\bar{A} = \det(A)I_n$. ■

Teorema 2.41. *Se $A = (a_{ij})$ é uma matriz quadrada de ordem n tal que $\det(A) \neq 0$, então a matriz inversa da matriz A é $A^{-1} = \frac{1}{\det(A)}\bar{A}$.*

Demonstração: (Guelli [8]) Seja a matriz quadrada $A = (a_{ij})$, de ordem n . Supondo que $\det(A) \neq 0$ e considerando o teorema 2.40, temos que

$$\bar{A}A = \det(A)I_n.$$

Daí segue que

$$\frac{1}{\det(A)}\bar{A}A = I_n. \quad (2.5)$$

Pelo teorema 2.6 podemos escrever a equação (2.5) como

$$\left(\frac{1}{\det(A)}\bar{A}\right)A = I_n. \quad (2.6)$$

Conforme a definição 2.10, temos que

$$A^{-1}A = I_n \quad (2.7)$$

Então, comparando as equações (2.6) e (2.7), observamos

$$\left(\frac{1}{\det(A)}\bar{A}\right)A = A^{-1}A.$$

Desta forma, concluímos que

$$A^{-1} = \frac{1}{\det(A)}\bar{A}.$$

■

Exemplo 2.42. Sejam a matriz

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 3 & 2 \\ 4 & 7 & 5 \end{bmatrix}$$

e sua matriz adjunta

$$\bar{A} = \begin{bmatrix} 1 & 2 & -1 \\ -2 & 1 & 0 \\ 2 & -3 & 1 \end{bmatrix}.$$

Temos que $\det(A) = 1$, então pelo teorema 2.42 segue que

$$A^{-1} = \frac{1}{\det(A)} \bar{A} = \begin{bmatrix} 1 & 2 & -1 \\ -2 & 1 & 0 \\ 2 & -3 & 1 \end{bmatrix}.$$

Corolário 2.43. *A condição necessária e suficiente para que uma matriz quadrada $A = (a_{ij})$ de ordem n seja inversível, é que $\det(A) \neq 0$.*

Pelo teorema 2.41 e corolário 2.43 temos estabelecida a condição de admissibilidade da inversa de uma matriz. Esse resultado será utilizado mais adiante nesta seção, para que possamos discutir a solução de sistemas de equações lineares.

Estamos interessados em discutir um sistema em função de parâmetros. Nesse sentido, e de acordo com Iezzi et.al. [11], essa discussão implica em classificar o sistema quanto ao número de suas soluções, ou seja, em compatível, compatível indeterminado ou incompatível para cada valor do parâmetro. Para que possamos entender essa classificação, observemos as seguintes definições:

Definição 2.44. *Uma equação linear em n incógnitas x_1, x_2, \dots, x_n é uma equação da forma $a_1x_1 + a_2x_2 + \dots + a_nx_n = b$, onde os números $a_1, a_2, \dots, a_n \in \mathbb{R}$ são denominados coeficientes e $b \in \mathbb{R}$ é o termo independente da equação.*

Definição 2.45. *Denominamos solução de uma equação linear a n incógnitas à ênupla ou conjunto ordenado de n números $\{\alpha_1, \alpha_2, \dots, \alpha_n\}$ que, substituídos na equação em lugar de x_1, x_2, \dots, x_n , respectivamente, a transforma em uma sentença verdadeira, isto é: $\{\alpha_1, \alpha_2, \dots, \alpha_n\}$ é solução se, e somente se, $a_1\alpha_1 + a_2\alpha_2 + \dots + a_n\alpha_n = b$ é verdadeira.*

Exemplo 2.46. Considerando a equação $3x_1 + 2x_2 + x_3 - 2x_4 = -2$, temos que o conjunto ordenado $(1, -1, 1, 2)$ é solução da equação pois a sentença $3 \cdot 1 + 2 \cdot (-1) + 1 - 2 \cdot 2 = -2$ é verdadeira.

Definição 2.47. *Sejam a_{ij} e b_i números reais, com $1 \leq i \leq m$ e $1 \leq j \leq n$. Então, um sistema linear de m equações e n incógnitas é um sistema da seguinte forma:*

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m \end{aligned}$$

em que x_1, x_2, \dots, x_n são as incógnitas, a_{ij} os coeficientes e b_i os termos independentes do sistema de equações lineares.

Definição 2.48. *Uma n -upla ou conjunto ordenado de números $\{\alpha_1, \alpha_2, \dots, \alpha_n\}$ é solução do sistema de equações lineares se for, simultaneamente, solução de todas as equações do sistema, ou seja: $\{\alpha_1, \alpha_2, \dots, \alpha_n\}$ é solução se, e somente se, $a_{i1}\alpha_1 + a_{i2}\alpha_2 + \cdots + a_{in}\alpha_n = b_i$, para todo $i \in \{1, 2, \dots, n\}$.*

Exemplo 2.49. Seja o sistema de equações lineares

$$\begin{aligned} x + y + z &= 0 \\ 2x + 3y + z &= 0 \\ x + y + 2z &= 1 \end{aligned}$$

O sistema admite como solução o conjunto $(-2, 1, 1)$, pois:

$$\begin{aligned} (-2) + 1 + 1 &= 0 \quad (\text{verdadeira}) \\ 2 \cdot (-2) + 3 \cdot 1 + 1 &= 0 \quad (\text{verdadeira}) \\ (-2) + 1 + 2 \cdot 1 &= 1 \quad (\text{verdadeira}) \end{aligned}$$

Mas não admite $(2, -1, -1)$ como solução pois:

$$\begin{aligned} 2 + (-1) + (-1) &= 0 \quad (\text{verdadeira}) \\ 2 \cdot 2 + 3 \cdot (-1) + (-1) &= 0 \quad (\text{verdadeira}) \\ (-2) + (-1) + 2 \cdot (-1) &= 1 \quad (\text{falsa}) \end{aligned}$$

Definição 2.50. *Seja um sistema linear de m equações e n incógnitas. Dizemos que o sistema é:*

- i. Compatível, quando possui uma única solução;*
- ii. Compatível indeterminado, quando possui infinitas soluções;*
- iii. Incompatível, quando não possui solução.*

Assim, a definição 2.50 nos dá a classificação do sistema de equações quanto ao número de soluções. Pois bem, estamos interessados num sistema compatível em que o número de equações é igual ao número de incógnitas. A esse respeito, Iezzi et. al. [11] afirma que para discutirmos sistemas de n equações e n incógnitas devemos calcular o $\det(A)$, onde A é a matriz gerada pelo sistema. Bem como, analisarmos os seguintes casos:

- i. $\det(A) \neq 0$, então o sistema é compatível;
- ii. $\det(A) = 0$, então o sistema é compatível indeterminado ou incompatível.

As definições e o teorema a seguir nos garantem as condições para que tenhamos um sistema compatível. São eles:

Definição 2.51. *Um sistema linear de m equações e n incógnitas pode ser escrito em notação matricial, ou seja, na forma $AX = B$, onde:*

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}_{m \times n} ; \quad X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}_{n \times 1} \quad e \quad B = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}_{n \times 1} .$$

Temos então, que

$$AX = B \iff \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix} .$$

Definição 2.52. *Se o número de equações de um sistema linear for igual ao número de incógnitas, a matriz A é quadrada, portanto, existe $\det(A)$ que denominamos determinante do sistema.*

Definição 2.53. *Um sistema de n equações a n incógnitas é chamado sistema normal quando o determinante do sistema é diferente de zero.*

Teorema 2.54. *Seja um sistema linear de n equações e n incógnitas. Se a matriz A dos coeficientes, de tamanho $n \times n$, for inversível, então o sistema é classificado como compatível, ou seja, possui uma única solução indicada por $X = A^{-1}B$.*

Demonstração: Se a matriz $A = (a_{ij})$ dos coeficientes é quadrada de ordem n , então pela definição 2.52 existe o $\det(A)$. Da definição 2.53, temos que, se $\det(A) \neq 0$ então o sistema é denominado normal. Em vista disso e supondo que o sistema seja normal, segue do corolário 2.43 e do teorema 2.11 que existe A^{-1} e é única.

De fato, se tomarmos a equação matricial $AX = B$ correspondente ao sistema e multiplicarmos ambos os lados da igualdade por A^{-1} , obtemos:

$$A^{-1}AX = A^{-1}B.$$

Mas,

$$A^{-1}A = I_n.$$

Então, segue que

$$I_n X = A^{-1}B.$$

O que resulta em

$$X = A^{-1}B.$$

Portanto, concluímos que se o sistema for normal, a solução existe e é única. Em outras palavras, o sistema é compatível. ■

Exemplo 2.55. Seja o sistema

$$x + 2y + 4z = 0$$

$$3x + y + 2z = 0$$

$$x - y - z = 5$$

Escrevendo o sistema em notação matricial, obtemos:

$$AX = B \iff \begin{bmatrix} 1 & 2 & 4 \\ 3 & 1 & 2 \\ 1 & -1 & -1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 5 \end{bmatrix}.$$

Calculando o determinante da matriz A , a matriz de cofatores dos elementos (a_{ij}) e a matriz adjunta de A , temos:

i. $\det(A) = -5$

ii. $C = \begin{bmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{bmatrix} = \begin{bmatrix} 1 & 5 & -4 \\ -2 & -5 & 3 \\ 0 & 10 & -5 \end{bmatrix}$

iii. $\bar{A} = C^t = \begin{bmatrix} 1 & -2 & 0 \\ 5 & -5 & 10 \\ -4 & 3 & -5 \end{bmatrix}$

Daí segue que a solução do sistema $X = A^{-1}B$ é dada por:

$$X = \frac{1}{\det(A)} \cdot \bar{A} \cdot B = \frac{1}{(-5)} \cdot \begin{bmatrix} 1 & -2 & 0 \\ 5 & -5 & 10 \\ -4 & 3 & -5 \end{bmatrix} \cdot \begin{bmatrix} 0 \\ 0 \\ 5 \end{bmatrix}.$$

O que resulta em

$$X = \begin{bmatrix} 0 \\ -10 \\ 5 \end{bmatrix}.$$

Dos resultados apresentados nesta seção, em síntese, temos que dado um sistema de equações lineares, podemos denotá-lo na forma de equação matricial, ou seja $AX = B$, onde A , X e B são as matrizes geradas. Se a matriz A dos coeficientes for inversível, então a equação matricial possui solução e esta é única. Em outras palavras, a existência e unicidade da solução da equação matricial implica na existência e unicidade da solução do sistema de equações lineares correspondente.

3 Método dos mínimos quadrados

3.1 Breve histórico

De acordo com Crato [7], um dos maiores problemas com que sempre se debateram os astrônomos foi o da combinação de observações, feitas necessariamente com erros, para estimação de parâmetros de posição dos corpos celestes. Astrônomos da Grécia antiga, tais como Hiparco (180 – 125 *a.C.*), Eratóstenes (276 – 194 *a.C.*) e Aristarco (310 – 230 *a.C.*), aceitavam aproximações em suas medições e não se preocupavam com erros em suas mensurações, ou pelo menos, não escreveram sobre esses problemas.

Ainda de acordo com Crato [7], Tycho Brahe (1546 – 1601) astrônomo da era pré – telescópica foi um dos primeiros a se preocupar com as medidas e o rigor das observações. Fazia várias medições de um mesmo parâmetro, juntava essas observações, eliminava erros grosseiros e obtinha médias que utilizava para suas estimativas. Esse controle da precisão da medida era uma novidade para época.

O estudo matemático da combinação de observações, realizado de forma sistemática, iniciou-se com Roger Joseph Boscovich (1711 – 1787), Pierre Simon Laplace (1749 – 1827), Adrien - Marie Legendre (1752 – 1833) e Carl Friedrich Gauss (1777 – 1855). Tentava-se encontrar um método ideal de combinação de observações, em particular, tentava-se determinar os parâmetros das órbitas dos cometas a partir de medições pontuais obtidas em diferentes momentos.

A solução encontrada que teve maior desenvolvimento teórico e maior aplicação prática, foi a do método dos mínimos quadrados, publicada por Legendre em 1805 na sua obra intitulada *Nouvelles Méthodes pour la détermination des orbites des comètes* (Novos métodos para determinação das órbitas dos cometas), e por Gauss em 1809 na *Theoria Motus Corporum Coelestium* (Teoria do movimento dos corpos celestes). Embora Legendre tenha antecipado a apresentação de seus resultados, sabe-se que Gauss os tinha obtido muito antes, entre 1794 e 1795, por isso se atribui a este último a prioridade da criação do método.

3.2 Concepção do método

Queremos nesta seção, discorrer brevemente sobre os pressupostos que levaram ao desenvolvimento do método dos mínimos quadrados. Para tanto, vamos iniciar com o seguinte problema:

“Como estimar um parâmetro com base em medidas repetidas, com valores ligeiramente diferentes?”

Para melhor compreensão do problema apresentado, vejamos o exemplo a seguir:

Exemplo 3.1. *Uma balança mecânica foi utilizada para medir o peso de uma saca de farinha. Foram efetuadas três medições consecutivas, e observaram-se pequenas diferenças nos resultados. Veja tabela 3.1:*

Tabela 3.1: Peso da saca de farinha em três medidas

Série de medições	1 ^a	2 ^a	3 ^a
Peso em Kg	21	18	21

Fonte: Tabela gerada pelo autor

Como combinar essas observações de modo a obter uma estimativa plausível do peso da saca? Precisamos de uma medida que tende a tipificar, ou, representar melhor o conjunto de medidas. Uma maneira historicamente usual para determinar tal medida é a média aritmética. Nesse sentido, o próprio Gauss disse que:

“tem sido costume encarar como um axioma a hipótese de que, se uma quantidade foi determinada por várias observações diretas, feitas nas mesmas circunstâncias e com igual cuidado, então a média aritmética dos valores observados fornece o valor mais provável, se não rigorosamente, pelo menos com grande aproximação”. (citado por Crato [7])

De acordo com Stevenson [17], a média possui certas propriedades interessantes e úteis que explicam por que é a mais usada. São elas:

- i. Pode sempre ser calculada
- ii. Para um dado conjunto de valores, a média é única;
- iii. É sensível a todos os valores do conjunto. Assim, se um valor se modifica, a média também se modifica;
- iv. Somando-se, subtraindo-se, multiplicando-se ou dividindo-se uma constante a cada valor do conjunto, a média ficará respectivamente aumentada, reduzida, multiplicada ou dividida do valor dessa constante;

v. A soma dos desvios dos valores de um conjunto a contar da média é zero.

Desta forma, considerando as características da média e tomando-a como método de estimativa, é possível generalizar os princípios desse método a um problema mais complexo, como a determinação simultânea de vários parâmetros? Crato [7] afirma que a partir deste questionamento Gauss desenvolveu o método dos mínimos quadrados. Para tanto, ele atacou o problema de duas maneiras:

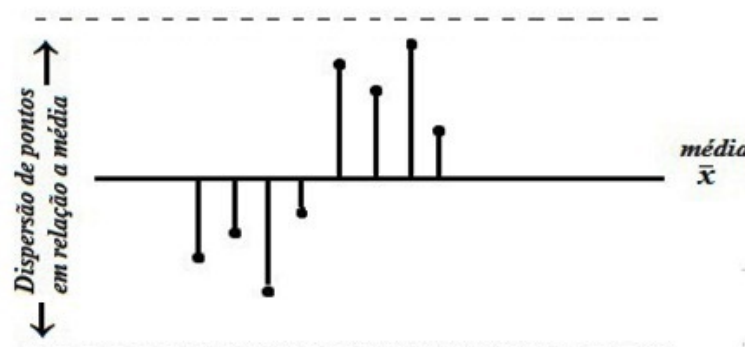
- i. Qual a métrica que leva a escolha da média aritmética?
- ii. Qual a distribuição dos erros que leva a que a média aritmética forneça uma estimativa ótima?

3.2.1 Abordagem não probabilística

A primeira abordagem pretende determinar em que sentido a média aritmética fornece uma aproximação ótima a um conjunto de observações. Neste caso, temos uma questão estatística ou matemática que não envolve *um processo probabilístico*, ou seja, não há preocupação com o caráter aleatório das medidas.

Retomando o exemplo 3.1, vamos considerar a média aritmética (dada por $\bar{x} = 20kg$) como estimativa que representa o conjunto de medidas observadas. Estas medidas estão nas proximidades da média, então queremos determinar um valor que represente o quanto elas estão próximas da estimativa. Veja a figura 3.1 que ilustra a idéia de desvios das medidas observadas (pontos) em relação à média:

Figura 3.1: Dispersão das medidas observadas em relação à média



Fonte: Adaptada de Stevenson [17]

Para determinar um valor que indique a proximidade com a média \bar{x} , poderíamos calcular a soma dos desvios, porém, neste caso, esse valor é zero. Então, é conveniente que façamos a soma dos desvios absolutos, isto é: $\sum_{i=1}^3 |x_i - \bar{x}| = 4$. No entanto, se considerarmos outra estimativa, a mediana \tilde{x} , por exemplo, obteremos: $\sum_{i=1}^3 |x_i - \tilde{x}| = 3$.

Diante disto, parece que a mediana fornece uma estimativa preferível, pois indica uma proximidade maior entre o valor observado e o valor estimado. Mas se observarmos a Tabela 3.2, percebemos que a média aritmética é o valor que minimiza a soma dos quadrados dos desvios, enquanto a mediana é o valor que minimiza a soma dos desvios absolutos:

Tabela 3.2: Média, mediana, desvios absolutos e quadrado dos desvios dos pesos da saca de farinha em estudo.

x_i	\bar{x}	\tilde{x}	$ x_i - \bar{x} $	$ x_i - \tilde{x} $	$(x_i - \bar{x})^2$	$(x_i - \tilde{x})^2$
21	20	21	1	0	1	0
18	20	21	2	3	4	9
21	20	21	1	0	1	0
$\sum_{i=1}^3$	-	-	4	3	6	9

Fonte: Adaptada de Crato [7]

Assim sendo, qual dos dois critérios obtém a melhor aproximação entre valor observado e valor estimado? A esse respeito, Vuolo [18] afirma que a soma dos quadrados dos desvios é uma quantidade mais razoável que a soma dos módulos dos desvios.

De fato, Vuolo [18] diz que se tomarmos três medidas distintas x_1 , x_2 e x_3 , bem como a soma dos desvios absolutos $S = |d_1| + |d_2| + |d_3|$ com $d_i = x_i - \tilde{x}$, de tal modo que a mediana \tilde{x} seja a melhor estimativa que minimiza essa soma, e por outro lado, se admitirmos $x_1 \leq x_2 \leq x_3$, \tilde{x} entre x_1 e x_2 e $|d_1| + |d_3| = (x_3 - \tilde{x}) + (\tilde{x} - x_1) = (x_3 - x_1)$, notamos que $|d_1| + |d_3|$ não depende de \tilde{x} . À vista disto, para minimizarmos S , basta minimizarmos $|d_2|$, o que implica $|d_2| = 0$, ou seja, $\tilde{x} = x_2$.

Segundo Stevenson [17], a mediana divide um conjunto ordenado de valores em dois grupos iguais, dos quais, uma metade terá valores inferiores à mediana e a outra metade valores superiores à mediana. Deste modo, temos que o resultado supra mencionado pode ser generalizado da seguinte forma:

- i. Se tivermos um número ímpar de observações, a mediana deve ser o valor intermediário;
- ii. Se tivermos um número par de observações, a mediana deve ser o valor médio dos dois valores x_i intermediários.

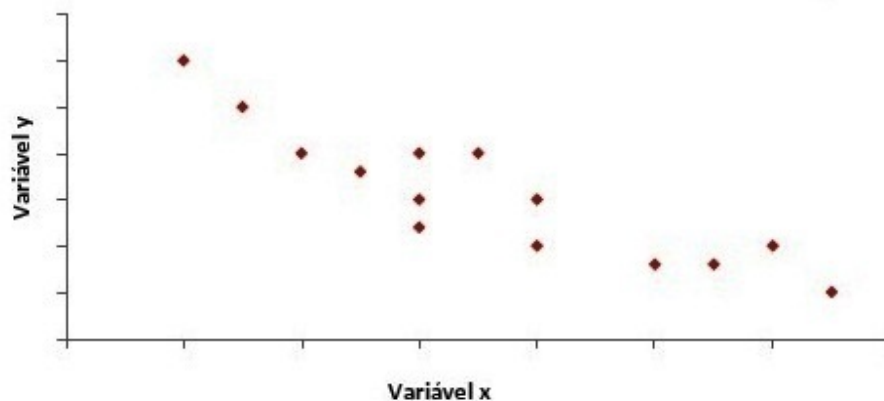
Então, no caso das três medidas, supondo x_2 muito próximo de x_1 , por exemplo, \tilde{x} seria muito próximo de x_1 , o que não é muito razoável quando estamos buscando uma medida central que tipifica um conjunto de medidas. Nesse sentido, a média é uma estimativa preferível e a soma dos quadrados dos desvios parece ser um procedimento melhor que a soma dos desvios absolutos.

Assumindo que o critério dos mínimos quadrados é o melhor procedimento que minimiza os desvios a contar da média, é possível generalizá-lo a outro tipo de estimadores? Ou seja, será possível encontrar processos de estimar parâmetros de modo que as diferenças nas observações (erros) sejam mínimas? De acordo com Crato [7], os estudos de Gauss levaram-no a deduzir um método de combinar observações e, com base nelas, estimar os parâmetros de uma função que minimiza os desvios a contar da função.

Isto significa que, ao determinarmos os parâmetros de uma função, estamos estabelecendo uma estimativa fiável do tipo de relação entre variáveis distintas inerentes ao mesmo objeto observado. Ou seja, estamos ajustando à curva que melhor se aproxima do comportamento dos pontos experimentais.

Supondo que os valores observados sejam oriundos de um processo de medição com duas variáveis x e y , queremos verificar se entre tais variáveis existe uma relação, onde a variável x possa ser caracterizada como independente. De acordo com Vuolo [18], os resultados obtidos das medições são denominados *pontos experimentais*, sendo que cada par de resultados (x, y) pode ser representado como um ponto no gráfico *Y versus X*, que é denominado diagrama de dispersão. Vejamos a figura 3.2:

Figura 3.2: Gráfico de dispersão dos pontos experimentais



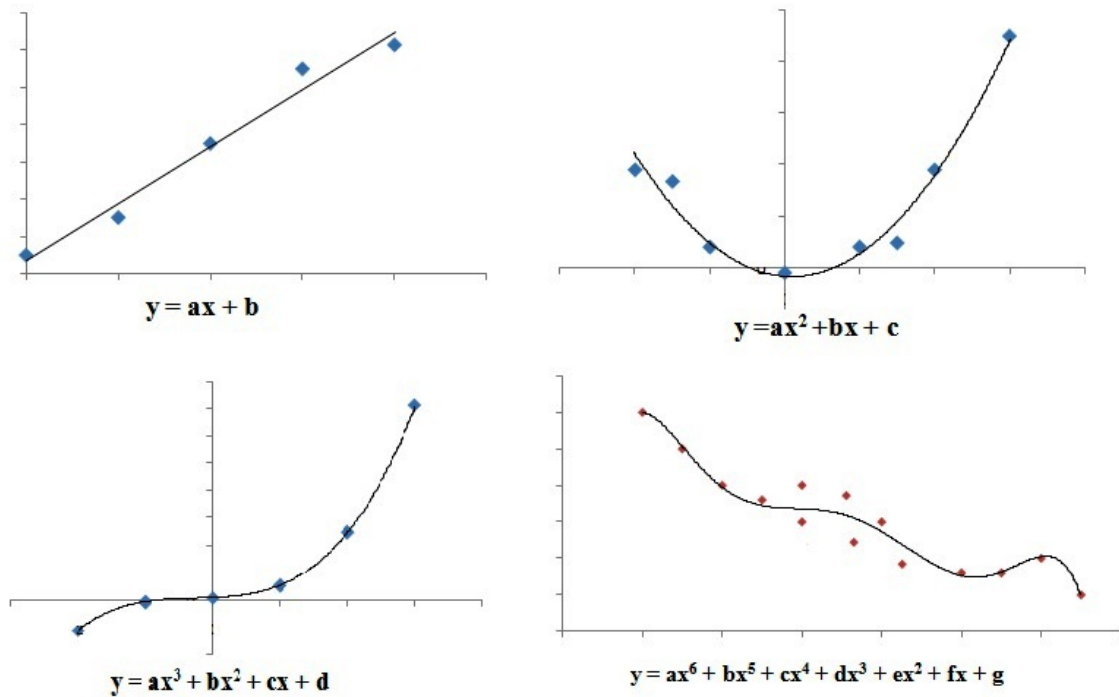
Fonte: Figura gerada pelo autor

Visualizando os pontos experimentais no diagrama de dispersão e supondo que haja uma relação entre as variáveis x e y , esta relação, determina o padrão da disposição dos pontos no gráfico. Logo, ajustar uma curva que se aproxima do comportamento destes pontos, significa ajustar uma função $g(x)$ que melhor se aproxima do comportamento dos dados obtidos experimentalmente.

Entretanto, qual função deve ser ajustada à dispersão dos pontos experimentais? Segundo Ruggiero e Lopes [15], a escolha da função depende de considerações teóricas inerentes ao experimento ou do próprio padrão apresentado pelos pontos. Vejamos a figura 3.3 que ilustra algumas possibilidades de ajuste de funções polinomiais aos

pontos experimentais:

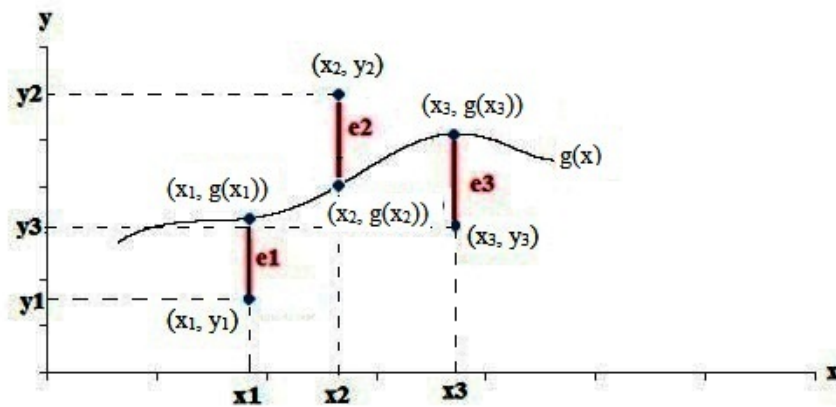
Figura 3.3: Gráficos de funções polinomiais ajustadas



Fonte: Figura gerada pelo autor

Uma vez definida a função $g(x)$ a ser ajustada, precisamos estabelecer um critério para determinação de seus parâmetros que torne mínimo os erros (desvios verticais) de aproximação em cada ponto (x_i, y_i) do gráfico de dispersão. Vejamos a figura 3.4:

Figura 3.4: Representação dos erros (desvios verticais) e_i



Fonte: Figura gerada pelo autor

Da figura 3.4, temos que e_1, e_2 e e_3 representam três dos desvios verticais de cada um dos n pontos até a função $g(x)$. Assim, verificamos que:

$$e_1 = |g(x_1) - y_1|, e_2 = |g(x_2) - y_2|, \dots, e_n = |g(x_n) - y_n|.$$

Como $e_1^2 = |g(x_1) - y_1|^2 = (g(x_1) - y_1)^2$, segue que:

$$\begin{aligned} e_1^2 + e_2^2 + \dots + e_n^2 &= (g(x_1) - y_1)^2 + (g(x_2) - y_2)^2 + \dots + (g(x_n) - y_n)^2 = \\ &= \sum_{i=1}^n (g(x_i) - y_i)^2. \end{aligned}$$

Desta forma, $\sum_{i=1}^n (g(x_i) - y_i)^2$ deve ser mínima para os melhores valores dos parâmetros de $g(x)$. Isto significa que, se os dados se comportam como função polinomial, a melhor função a ser ajustada com forma e números de parâmetros predeterminados é dada pela função:

$$g(x; a_1, a_2, \dots, a_p) = a_1 g_1(x) + a_2 g_2(x) + \dots + a_p g_p(x),$$

onde as funções $g_1(x), g_2(x), \dots, g_p(x)$ são potências de x , ou seja,

$$g_1(x) = x^n, g_2(x) = x^{n-1}, \dots, g_{p-1}(x) = x, g_p(x) = 1$$

e os particulares valores dos parâmetros a_1, a_2, \dots, a_p devem ser tais que minimizam

$$\sum_{i=1}^n (g(x_i; a_1, a_2, \dots, a_p) - y_i)^2.$$

3.2.2 Abordagem probabilística

Crato [7] nos diz que a segunda abordagem de Gauss tem por objetivo determinar as melhores chances de que a estimativa dada pela média seja uma aproximação ótima em um conjunto de valores observados. Desta forma, o problema é estudado conforme um processo probabilístico, onde o caráter aleatório das observações é muito relevante.

De acordo com Stevenson [17], se as medidas observadas mostram aleatoriedade nos resultados, então estas tendem a apresentar diferenças de valores de uma observação para outra. Se as diferenças são influenciadas por fatores (erros) que ocorrem de maneira análoga em observações repetidas um grande número de vezes, então no conjunto de observações existem medidas que podem ser mais prováveis que outras. Nesse sentido, como determinar as medidas mais prováveis?

Retomando o exemplo da medição do peso da saca de farinha, extrapolando as observações para um número maior de medições e supondo que se mantenham as pequenas diferenças de valores, não podemos esperar que as medidas observadas apresentem apenas resultados inteiros, pois existe uma infinidade de valores que essas medidas podem assumir. Esse fato não nos permite pensar na ocorrência da probabilidade de um valor específico, seria pouco razoável, uma vez que as chances seriam muito pequenas. Assim, como podemos determinar a medida mais provável do peso da saca?

Segundo Vuolo [18], podemos estabelecer intervalos de valores com centros e tamanhos dados por medidas específicas. Estas medidas devem ser tais que qualquer valor observado estará incluído em *um e somente um intervalo*. Em vista disso, admitindo que as medidas observadas na medição do peso da saca de farinha ocorram no intervalo (18, 22), podemos estabelecer subintervalos de valores e enquadrar as medidas em agrupamentos de porcentagens de ocorrências, ou seja, numa distribuição de frequências relativas. Esta distribuição “... mostra a proporção de vezes em que as medidas tendem a assumir um dos valores observados” (STEVENSON, 1981)[17]. Vejamos a tabela 3.3:

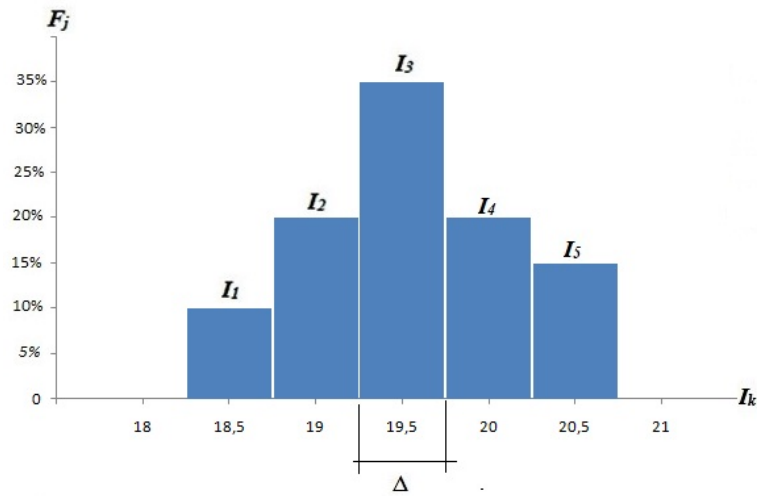
Tabela 3.3: Série de medições, intervalo de valores e frequências relativas.

Série de Medições	Peso (kg)	Série de medições	Peso (kg)	Intervalo de valores	Frequências relativas
1 ^a	18,25	11 ^a	19,25	18,25 a 18,75	10%
2 ^a	19,00	12 ^a	19,25	18,75 a 19,25	20%
3 ^a	18,75	13 ^a	20,00	19,25 a 19,75	35%
4 ^a	19,25	14 ^a	20,25	19,75 a 20,25	20%
5 ^a	19,50	15 ^a	20,50	20,25 a 21,75	15%
6 ^a	19,25	16 ^a	19,50	—	—
7 ^a	19,75	17 ^a	19,50	—	—
8 ^a	19,00	18 ^a	19,75	—	—
9 ^a	18,50	19 ^a	19,75	—	—
10 ^a	19,00	20 ^a	20,75	—	—

Fonte: Tabela gerada pelo autor

Stevenson [17] explica que os intervalos e as respectivas frequências relativas listadas na tabela 3.3 podem ser representados por um histograma. Neste, os intervalos são distribuídos ao longo do eixo horizontal e as frequências ao longo do vertical, onde as fronteiras *das barras* coincidem com os pontos extremos dos intervalos. Vejamos a figura 3.5:

Figura 3.5: Distribuição de frequência dos valores listados na tabela 2.3



Fonte: Figura gerada pelo autor

Da figura 3.5 temos que:

- i. $I_k (k = 1, 2, 3, \dots, n)$ são intervalos de valores;
- ii. $\Delta = \frac{I_n - I_1}{\sqrt{N}}$ é a amplitude ou tamanho do intervalo I_k ;
- iii. N é o número de observações;
- iv. \sqrt{N} com $5 \leq \sqrt{N} \leq 15$ é a quantidade de intervalos da distribuição ¹;
- v. $F_j (j = 1, 2, 3, \dots, m)$ são as frequências das medidas observadas que incidem em cada intervalo.

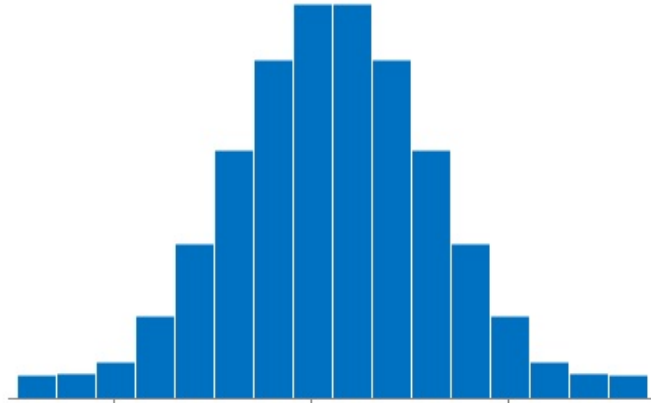
Ao observarmos a figura 3.5, constatamos que o valor mais provável para a medida do peso da saca de farinha parece estar situado próximo ao intervalo que apresenta a maior frequência. A média $\bar{x} = 19,44$ está numa posição intermediária entre os extremos desse intervalo, ou seja: $19,25 < \bar{x} < 19,75$. Então, podemos supor que a *média* é o valor mais provável. Por outro lado, sabemos que o peso do objeto é constante, logo, as diferenças de medidas correspondem a erros nas observações. Como já discutido na seção 3.2.1, os erros são desvios em relação a média que podem ser minimizados conforme o critério dos *mínimos quadrados*.

Isto posto, queremos saber qual será a distribuição dos erros que faz com que a estimativa dada pela média, seja a de maior probabilidade de se aproximar do *verdadeiro peso da saca de farinha*.

¹De acordo com Stevenson [17], a quantidade de intervalos é limitada por $5 \leq \sqrt{N} \leq 15$, pois se for menor que 5 pode ocultar detalhes importantes dos dados, e, se for maior que 15 torna apresentação da distribuição demasiadamente detalhada.

Stevenson [17] salienta que desde o século XVIII astrônomos e outros cientistas observaram que quando se coletava grande número de mensurações, dispondo-as numa distribuição de frequência, elas se apresentavam repetidamente de forma análoga à da figura 3.6.

Figura 3.6: Tendência da distribuição de frequência



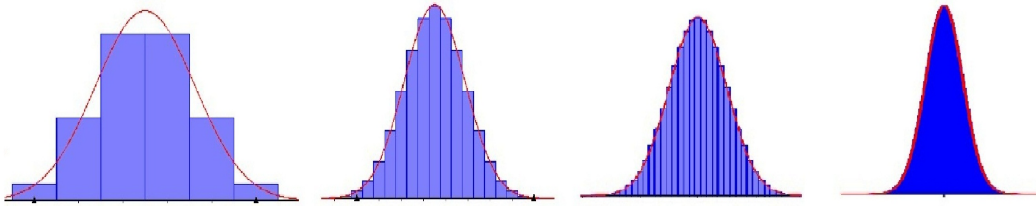
Fonte: Figura gerada pelo autor

A tendência apresentada pela figura 3.6 indica que as medidas observadas distribuem-se em torno da média, onde os valores com maior frequência de ocorrência estão nas suas proximidades, enquanto que os de menor frequência estão mais distantes. A vista disto, podemos supor que a distribuição dos erros minimizados que tem a média como estimativa, segue a tendência apresentada pela figura 3.6.

De fato, Crato [7] relata que Gauss deduziu uma distribuição matemática que justifica o uso da média e do método dos mínimos quadrados, denominada distribuição normal ou Gaussiana. Essa distribuição estabelece uma boa aproximação ao padrão apresentado pelo histograma destacado na figura 3.6.

Corroborando Crato, Vuolo [18] afirma que tal distribuição determina que a grande maioria dos erros aleatórios $e_i (i = 1, 2, \dots, n)$, admitidos como tendo diferentes distribuições, são tais que nenhum e_i particular é muito maior que os demais. Nestas condições, se o erro total é $\sum_{i=1}^n e_i$, então, a distribuição de erros e_i converge para uma distribuição gaussiana, no limite $n \rightarrow \infty$. A figura 3.7 ilustra essa aproximação:

Figura 3.7: Aproximação do histograma à distribuição gaussiana



Fonte: Figura gerada pelo autor

Na figura 3.7, notamos que a convergência descrita por Vuolo [18], implica que quando aumentamos o número de intervalos e diminuimos o seu tamanho, o formato do histograma de distribuição de frequências se aproxima do formato da distribuição gaussiana. Ou seja, intuitivamente temos que a soma das áreas dos retângulos formados pelas barras do histograma é uma aproximação da área da região sob a curva gaussiana e o eixo horizontal.

Em outras palavras, a distribuição gaussiana é definida por uma função $h(y)$ que *informa* a probabilidade de uma variável assumir um dado valor pertencente a um intervalo suficientemente pequeno. Isto é, dado um conjunto $\{y_1, y_2, \dots, y_n\}$ de n valores possíveis que a variável pode assumir, temos que cada valor y_i do conjunto, pode ocorrer com probabilidade $P(y_i) \equiv \Delta P_i$, onde ΔP_i é um intervalo de probabilidade que depende do tamanho do intervalo Δy entre duas medidas, sendo ambos os intervalos infinitesimais.

Assim, a probabilidade ΔP_i é proporcional ao intervalo Δy , e isso significa que a razão entre ΔP_i e Δy é uma quantia definida por $h(y) = \lim_{\Delta y \rightarrow 0} \frac{\Delta P_i}{\Delta y}$. Portanto, se $h(y)$ é conhecida, podemos afirmar que a probabilidade de um valor y_i pertencer ao intervalo Δy é dada por $P(y_i) \cong h(y_i)\Delta y$.

Nesse contexto, ao aproximarmos o histograma da gaussiana, estamos, de acordo com Malta et al.[14]:

- i. tomando um intervalo $[a, b]$ contido no domínio de h ;
- ii. tomando uma partição p para o intervalo, que corresponde a um conjunto finito de valores $\{y_0, y_1, \dots, y_{k-1}, y_k\}$, que satisfaz $y_0 = a < y_1 < \dots < y_{k-1} < b = y_k$ e subdivide o intervalo $[a, b]$ em n subintervalos $[y_{i-1}, y_i]$;
- iii. escolhendo valores c_1, \dots, c_k com $y_{i-1} \leq c_i \leq y_i$, para $1 \leq i \leq k$

de forma que a soma de *Riemann* de h sobre a partição p e à escolha dos valores c_i , que é dada por:

$$S(h, p) = h(c_1) \cdot (y_1 - y_0) + h(c_2) \cdot (y_2 - y_1) + \dots + h(c_i) \cdot (y_i - y_{i-1}) + \dots$$

$$\dots + h(c_k) \cdot (y_k - y_{k-1}) = \sum_{i=1}^k h(c_i) \cdot (y_i - y_{i-1})$$

seja uma boa aproximação da função para intervalos suficientemente pequenos.

Se denotarmos $(y_i - y_{i-1})$ por Δy_i , temos que:

$$S(h, p) = \sum_{i=1}^k h(c_i) \cdot \Delta y_i.$$

Esta soma é a integral de $h(y)$ entre a e b , no limite de $\Delta y_i \rightarrow 0$. Ou seja:

$$\lim_{\Delta y_i \rightarrow 0} S(h, p) = \int_a^b h(y) dy.$$

Em vista disto, podemos afirmar que a probabilidade de uma variável y assumir um valor pertencente ao intervalo $[a, b]$ é dada pela soma das probabilidades para todos os valores de y_i neste intervalo, o que é equivalente a determinar a área sob a curva gaussiana delimitada pelos extremos do intervalo $[a, b]$ e o eixo horizontal. Em outros termos, podemos afirmar que:

$$P(a, b) = \int_a^b h(y) dy.$$

Na igualdade acima, $h(y)$ é uma função denominada densidade de probabilidade. Tal função é definida por

$$h(y) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2} \cdot \left(\frac{y-\bar{y}}{\sigma}\right)^2}, \quad -\infty < y < \infty$$

e apresenta as seguintes características:

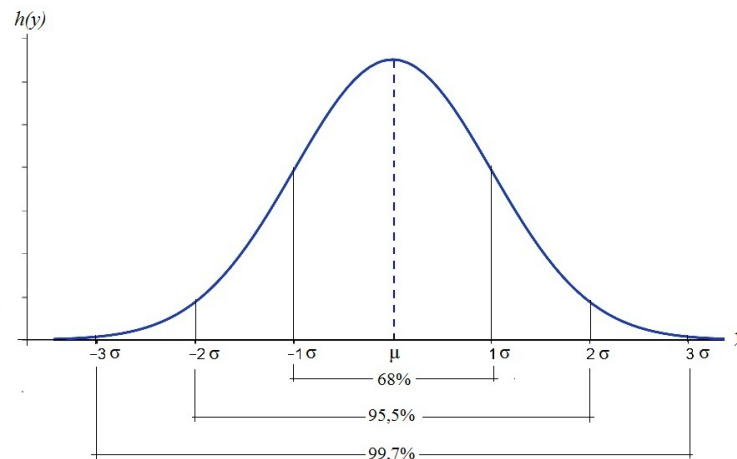
- i. $h(y) \geq 0, \quad \forall y \in \mathbb{R};$
- ii. $\int_{-\infty}^{\infty} h(y) dy = 1;$
- iii. $\lim_{y \rightarrow \pm\infty} h(y) = 0;$
- iv. $h(y)$ fica completamente especificada por dois parâmetros:
 - a. (\bar{y}) média da distribuição normal, que indica a localização do centro da distribuição. É dada por $\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$ e pode assumir qualquer valor na reta real $-\infty < \bar{y} < \infty;$
 - b. (σ) desvio padrão, que indica a variabilidade ou dispersão em relação ao centro. É dado por $\sigma = \sqrt{\frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1}}$, onde n corresponde ao número de observações.

O gráfico de $h(y)$, ilustrado pela figura 3.8, possui algumas características peculiares quanto ao seu formato:

- i. tem a forma semelhante a de um sino;

- ii. é suave e simétrico em relação à média;
- iii. possui ponto de máximo em $h(y = \bar{y}) = \frac{1}{\sigma\sqrt{2\pi}}$;
- iv. $(\bar{y} \pm \sigma)$ são os pontos de inflexão de $h(y)$;
- v. o formato depende da variação dos parâmetros \bar{y} e σ , que implica respectivamente em translação e achatamento da curva;
- vi. apresenta o seguinte padrão de distribuição:
 - a. 68% de sua área está situada entre $\bar{y} \pm \sigma$;
 - b. 95,5% de sua área está situada entre $\bar{y} \pm 2\sigma$;
 - c. 99,7% de sua área está situada entre $\bar{y} \pm 3\sigma$.

Figura 3.8: Distribuição gaussiana



Fonte: Figura gerada pelo autor

Para determinarmos a probabilidade de uma variável y assumir um valor pertencente ao intervalo $[a, b]$ temos que calcular a integral

$$\frac{1}{\sigma\sqrt{2\pi}} \int_a^b e^{-\frac{1}{2}\left(\frac{y-\bar{y}}{\sigma}\right)^2} dy.$$

No entanto, há infinitas possibilidades de combinação para os parâmetros \bar{y} e σ , o que significa infinitas probabilidades se considerarmos todas as distribuições normais possíveis. Contudo, podemos contornar essa dificuldade reduzindo o problema a calcular a integral de $e^{-\frac{z^2}{2}}$.

Fazendo a substituição $z = \frac{y-\bar{y}}{\sigma}$ ($dz = \frac{dy}{\sigma}$), obtemos:

$$\frac{1}{\sigma\sqrt{2\pi}} \int_a^b e^{-\frac{1}{2}\left(\frac{y-\bar{y}}{\sigma}\right)^2} dy = \frac{1}{\sqrt{2\pi}} \int_{\frac{a-\bar{y}}{\sigma}}^{\frac{b-\bar{y}}{\sigma}} e^{-\frac{z^2}{2}} dz,$$

ou seja, uma integral de $e^{-\frac{z^2}{2}}$ no intervalo $[A, B]$, com $A = \frac{a-\bar{y}}{\sigma}$ e $B = \frac{b-\bar{y}}{\sigma}$.

Mas essa integral não tem solução analítica, então devemos procurar uma *solução numérica*. A esse respeito Asano e Colli [2] diz que *em probabilidade, como é muito frequente o uso dessa integral, são adotadas tabelas que podem ser montadas com os métodos de integração numérica*. Uma outra possibilidade, de acordo com Vuolo [18], é consultar *Handbooks* de funções matemáticas que usualmente apresentam tabelas para esta integral.

Da substituição acima, quando utilizamos a variável z , estamos trabalhando com valores relativos ao invés dos valores reais. Nesse sentido, Stevenson [17] argumenta que:

Há uma vantagem em trabalhar com valores relativos. É que, em vez de lidarmos com uma família infinita de distribuições normais, precisamos de apenas uma distribuição normal para todos os problemas. Podemos converter qualquer valor de qualquer distribuição em um valor z . [...] isto nos permite determinar todas as probabilidades da curva normal utilizando uma única tabela padronizada.

O uso da variável z equivale a tomar a média como ponto de referência (origem) e o desvio padrão como medida de afastamento a contar da média (unidade de medida). Vejamos o exemplo a seguir:

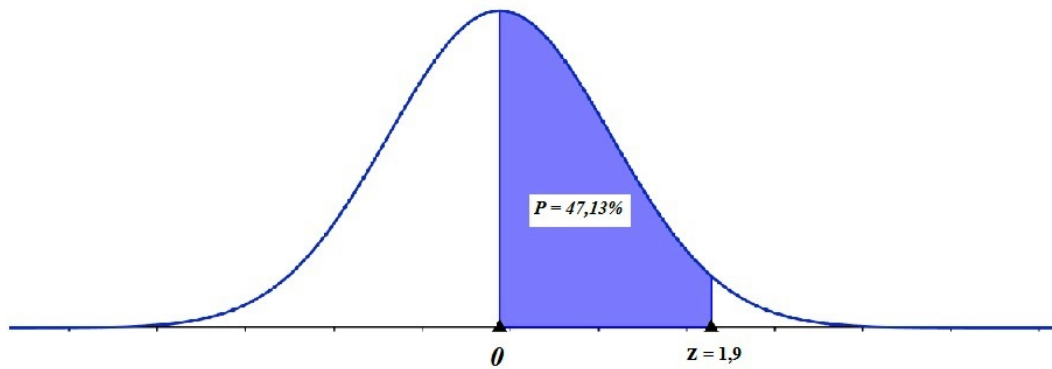
Exemplo 3.2. *Da tabela 3.3 temos que o valor da média é $\bar{y} = 19,44$ e o desvio padrão é $\sigma = 0,63$. Qual deverá ser a probabilidade de y_i variar entre a média e o ponto $y = 20,64$?*

Transformando a variável y na variável z , temos que

$$z = \frac{y_i - \bar{y}}{\sigma} = \frac{20,64 - 19,44}{0,63} \cong 1,9.$$

Consultando a tabela de probabilidades para distribuição normal padronizada [17], constatamos que $P(19,44 \leq y \leq 20,64) = P(0 \leq z \leq 1,9) = 0,4713$, ou seja, há 47,13% de chances de y_i variar entre a média e o ponto $y = 20,64$. A figura 3.9. ilustra essa probabilidade:

Figura 3.9: Área sob a curva normal entre a média e z



Fonte: Figura gerada pelo autor

Do que foi discutido até aqui, em síntese, temos que a distribuição normal é um modelo que aproxima a distribuição dos erros. É determinada por uma função de densidade de probabilidade que toma a média como estimativa para o valor mais provável do conjunto de observações, o desvio padrão como medida de dispersão em relação a média e o critério dos mínimos quadrados como minimizador dos desvios. Tal função nos informa a probabilidade de ocorrência de uma medida pertencente ao conjunto de observações.

Na seção 3.2.1 vimos que o critério dos mínimos quadrados é o melhor procedimento para minimizar os desvios a partir da média. Bem como, observamos que esse critério pode ser generalizado a outros estimadores, o que implica em estimar os parâmetros de uma função que minimiza os desvios a partir da função de modo a obter a melhor aproximação para o conjunto de pontos experimentais.

Nessa perspectiva, queremos obter a melhor aproximação possível para o conjunto de pontos experimentais em termos probabilísticos. Isto é, queremos determinar uma função $g(x)$ para a qual é *máxima a probabilidade* de ocorrer o particular conjunto de pontos, enquanto que os erros ou desvios de aproximação são mínimos.

Segundo Vuolo [18], se considerarmos os pontos experimentais $\{x_i, y_i, \sigma_i\}, i = 1, 2, \dots, n$, com x_i representando a variável independente, y_i variável dependente e σ_i seu desvio padrão, temos que a probabilidade P_i de ocorrer qualquer um dos n pontos é proporcional à função de densidade de probabilidade $h(y)$, ou seja:

$$P_i = \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{1}{2} \left(\frac{\bar{y} - y_i}{\sigma_i} \right)^2},$$

daí segue que a probabilidade P de ocorrer o particular conjunto de pontos é dada por:

$$P = P_1 \cdot P_2 \cdots P_n = \left(\frac{1}{\sqrt{2\pi}} \right)^n \cdot \left(\frac{1}{\sigma_1 \cdot \sigma_2 \cdots \sigma_n} \right) \cdot e^{-\frac{1}{2} \cdot \sum_{i=1}^n \left(\frac{\bar{y} - y_i}{\sigma_i} \right)^2}.$$

Admitindo que a função $g(x)$ estabelece a melhor aproximação aos pontos experimentais de tal forma que a probabilidade P é máxima, então se substituirmos \bar{y} pela função $g(x_i; a_1, a_2, \dots, a_p)$, obtemos:

$$P = \left(\frac{1}{\sqrt{2\pi}} \right)^n \cdot \left(\frac{1}{\sigma_1 \cdot \sigma_2 \cdots \sigma_n} \right) \cdot e^{-\frac{E}{2}},$$

onde

$$E = \sum_{i=1}^n \left[\frac{g(x_i; a_1, a_2, \dots, a_p) - y_i}{\sigma_i} \right]^2.$$

Denotando a probabilidade de ocorrência do conjunto de pontos experimentais por $P(E)$, podemos entender que P é uma função decrescente de E , assim, um máximo de P ocorre quando E é mínimo. Por outro lado, conforme o que já foi discutido na seção 3.2.1, temos que E deve ser mínima para os melhores parâmetros de $g(x)$. Ou seja, os particulares valores dos parâmetros a_1, a_2, \dots, a_p devem ser tais que minimizam

$$\sum_{i=1}^n \left[\frac{g(x_i; a_1, a_2, \dots, a_p) - y_i}{\sigma_i} \right]^2.$$

Portanto, o problema de determinar a melhor aproximação aos pontos experimentais de tal forma que a probabilidade seja máxima e o erro seja mínimo, se reduz a obtenção de parâmetros para uma função, que de acordo com o discutido na seção 3.2.1, possui forma e número de parâmetros predeterminados.

3.3 Problema de mínimos quadrados

De acordo com Vuolo [18], num processo de medição tentamos determinar o valor de uma *grandeza* por meio de medidas experimentais, onde o “valor verdadeiro” da grandeza y_v é desconhecido e as medidas experimentais y_1, y_2, \dots, y_n são aproximações desse valor. Nestas aproximações ocorrem erros definidos por $e_i = y_i - y_v$, $\{i = 1, 2, \dots, n\}$, e estes também são valores desconhecidos, pois dependem do valor verdadeiro.

Assim, se o valor verdadeiro e o erro associado são valores desconhecidos, então podemos afirmar que há uma *incerteza* no melhor valor y que aproxima o valor verdadeiro y_v . Nesse sentido, Vuolo [18] explica que a incerteza é uma medida que indica o quanto esse melhor valor pode diferir do valor verdadeiro em termos probabilísticos. Especificamos esta medida pelo *erro padrão* σ_m , que por sua vez, é definido como o desvio padrão da distribuição dos erros ou desvio padrão do valor médio de n medidas observadas.

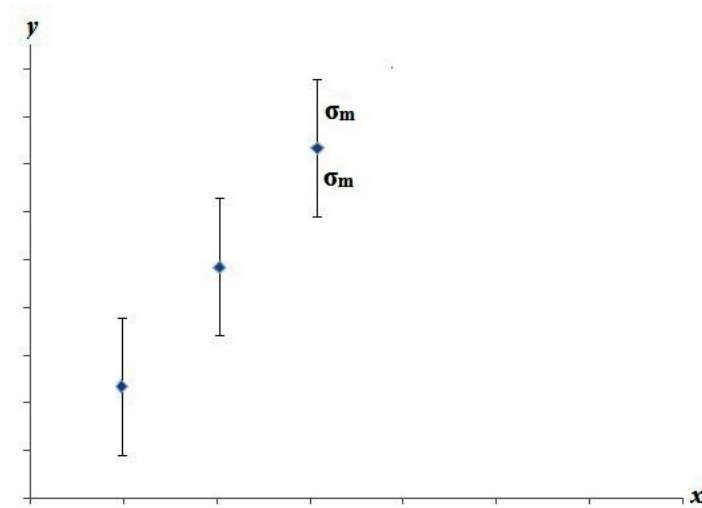
Vuolo [18] diz que o erro padrão pode ser estimado da seguinte forma:

$$\sigma_m = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{\sum_{i=1}^n (\bar{y} - y_i)^2}{n(n-1)}}, \quad (3.1)$$

onde \bar{y} e σ correspondem respectivamente à média e o desvio padrão das n medidas.

Graficamente, podemos representar o erro padrão em cada medida observada por meio de *barras de incerteza* ou *barra de erros*, onde cada barra tem comprimento total $2\sigma_m$ e é a representação do intervalo no qual está contida a medida observada ($y_i \pm \sigma_m$). Vejamos a figura 3.10:

Figura 3.10: Barras de incerteza



Fonte: Figura gerada pelo autor

Ainda de acordo com o mesmo autor, o erro padrão é inerente ao conceito de *intervalo de confiança*, ou seja, o erro num intervalo de confiança refere-se a diferença entre a média amostral e a verdadeira média da população. Desta forma, e de acordo com Stevenson [17], um intervalo de confiança dá um intervalo de valores, centrado na estatística amostral, no qual julgamos com risco conhecido de erro, estar o parâmetro da população.

Em outras palavras, se o nível de confiança de uma afirmativa é a probabilidade P de que esta afirmativa esteja correta, temos que, se a afirmativa for $(y - \delta) < y_v < (y + \delta)$, com $\delta \geq 0$, então $(y - \delta) < y_v < (y + \delta)$ tem a probabilidade P de ser correta. Logo, tal afirmativa define um intervalo de confiança para o valor y_v .

Stevenson [17] diz que o intervalo de confiança pode ser determinado da seguinte forma:

$$\bar{y} \pm z_{\frac{\alpha}{2}} \cdot \sigma_m \quad \text{ou} \quad \bar{y} \pm t_{\frac{(1-\alpha)}{2}} \cdot \sigma_m, \quad (3.2)$$

onde \bar{y} e σ_m correspondem respectivamente à média e o erro padrão, $z_{\frac{\alpha}{2}}$ ou $t_{\frac{(1-\alpha)}{2}}$ são os coeficientes de confiança dados respectivamente pelas tabelas de probabilidade das distribuições normal padronizada e t-student, e α é o nível de confiança desejado. Com relação aos coeficientes de confiança, ressaltamos que geralmente os valores t são preferíveis quando utilizamos uma amostra considerada pequena, ou seja, com número

de observações $n \leq 30$. Vejamos o seguinte exemplo:

Exemplo 3.3. Seja uma amostra com $n = 25$ observações, desvio padrão $\sigma = 1,5$ e média amostral $\bar{y} = 20$. Queremos determinar o intervalo de confiança de 95% para média da população.

Como a amostra é pequena, com $n = 25 < 30$, vamos tomar $\bar{y} \pm t_{\frac{(1-\alpha)}{2}} \cdot \sigma_m$. Para um nível de confiança de 95%, temos que

$$\frac{(1 - \alpha)}{2} = \frac{(1 - 0,95)}{2} = 0,025.$$

Consultando a tabela de probabilidades para distribuição t [17], constatamos que

$$t_{0,025} = 2,064.$$

Calculando o erro padrão, temos que

$$\sigma_m = \frac{\sigma}{\sqrt{n}} = 0,3.$$

Daí segue que

$$t_{0,025} \cdot \sigma_m = 2,064 \cdot 0,3 = 0,62.$$

Então, temos o intervalo $20 \pm 0,62$, ou seja, $19,38 < \bar{y} < 20,62$ é a faixa de valores, com 95% de chances de ocorrência, onde a média da população pode estar.

Isto posto, dado um conjunto de n pontos experimentais $\{x_1, y_1\}, \{x_2, y_2\}, \dots, \{x_n, y_n\}$ e supondo que a variável independente x_i seja isenta de erros, para que possamos discutir o problema de mínimos quadrados, devemos considerar a incerteza σ_i na variável dependente y_i .

Então, tomando os pontos experimentais $\{x_1, y_1, \sigma_1\}, \{x_2, y_2, \sigma_2\}, \dots, \{x_n, y_n, \sigma_n\}$, a função $g(x_i; a_1, a_2, \dots, a_p)$ e considerando o que foi discutido na seção 3.2.2, temos que

$$E = \sum_{i=1}^n \left[\frac{g(x_i; a_1, a_2, \dots, a_p) - y_i}{\sigma_i} \right]^2$$

deve ser mínima para os melhores valores dos parâmetros $\{a_1, a_2, \dots, a_p\}$.

Assumindo E como função erro, queremos determinar os melhores valores para os parâmetros de $E(a_1, a_2, \dots, a_p)$. Isso implica determinar o ponto de mínimo dessa função, o que será feito por meio do cálculo diferencial.

Particularmente, estamos interessados em analisar a função $E : D \subset \mathbb{R}^2 \rightarrow \mathbb{R}$ dada por

$$E(a_1, a_2) = \sum_{i=1}^n \left[\frac{g(x_i; a_1, a_2) - y_i}{\sigma_m} \right]^2,$$

onde $g(x_i; a_1, a_2) = a_1 x_i + a_2$ e $\sigma_m = \sigma_i \equiv \sigma_1 = \sigma_2 = \dots = \sigma_n$, estimada por (3.1), corresponde a incertezas iguais na variável y_i .

Observamos que E é uma função polinomial do segundo grau. Assim, conforme o discutido no capítulo 1, E é contínua e diferenciável em todo ponto de D . Logo, admite derivadas parciais nesses pontos e conseqüentemente deve existir um ponto onde tais derivadas se anulam, ou seja, deve existir um ponto crítico de D em E “candidato” a extremo local.

Isto posto, queremos determinar o ponto crítico de D em E e verificar se é extremo local, ou mais especificamente se é ponto de mínimo local. Caso tenhamos constatado que o ponto investigado é um ponto de mínimo local, precisamos garantir que o mesmo seja um ponto de mínimo global para que de fato E minimize

$$\sum_{i=1}^n \left[\frac{g(x_i; a_1, a_2) - y_i}{\sigma_m} \right]^2.$$

Para determinarmos o ponto crítico de E vamos efetuar os seguintes procedimentos:

i. Derivando $E(a_1, a_2)$ em relação a a_1 , temos

$$\begin{aligned} \frac{\partial E}{\partial a_1}(a_1, a_2) &= \frac{\partial}{\partial a_1} \sum_{i=1}^n \left[\frac{a_1 x_i + a_2 - y_i}{\sigma_m} \right]^2 = \\ &= \sum_{i=1}^n \frac{\partial}{\partial a_1} \left[\frac{a_1 x_i + a_2 - y_i}{\sigma_m} \right]^2 = \sum_{i=1}^n 2 \frac{(a_1 x_i + a_2 - y_i)}{(\sigma_m)^2} x_i = \\ &= 2 \sum_{i=1}^n \left(\frac{x_i}{\sigma_m} \right)^2 a_1 + 2 \sum_{i=1}^n \frac{x_i}{(\sigma_m)^2} a_2 - 2 \sum_{i=1}^n \frac{x_i y_i}{(\sigma_m)^2}. \end{aligned}$$

Desta forma, segue que $\frac{\partial E}{\partial a_1}(a_1, a_2) = 0$ se, e somente se

$$2 \sum_{i=1}^n \left(\frac{x_i}{\sigma_m} \right)^2 a_1 + 2 \sum_{i=1}^n \frac{x_i}{(\sigma_m)^2} a_2 - 2 \sum_{i=1}^n \frac{x_i y_i}{(\sigma_m)^2} = 0,$$

de onde vem

$$2 \sum_{i=1}^n \left(\frac{x_i}{\sigma_m} \right)^2 a_1 + 2 \sum_{i=1}^n \frac{x_i}{(\sigma_m)^2} a_2 = 2 \sum_{i=1}^n \frac{x_i y_i}{(\sigma_m)^2}. \quad (3.3)$$

ii. Analogamente, derivando $E(a_1, a_2)$ em relação a a_2 , temos

$$\begin{aligned} \frac{\partial E}{\partial a_2}(a_1, a_2) &= \frac{\partial}{\partial a_2} \sum_{i=1}^n \left[\frac{a_1 x_i + a_2 - y_i}{\sigma_m} \right]^2 = \\ &= \sum_{i=1}^n \frac{\partial}{\partial a_2} \left[\frac{a_1 x_i + a_2 - y_i}{\sigma_m} \right]^2 = \sum_{i=1}^n 2 \frac{(a_1 x_i + a_2 - y_i)}{(\sigma_m)^2} = \\ &= 2 \sum_{i=1}^n \frac{x_i}{(\sigma_m)^2} a_1 + 2 \sum_{i=1}^n \frac{1}{(\sigma_m)^2} a_2 - 2 \sum_{i=1}^n \frac{y_i}{(\sigma_m)^2} = \end{aligned}$$

$$= 2 \sum_{i=1}^n \frac{x_i}{(\sigma_m)^2} a_1 + \frac{2n}{(\sigma_m)^2} a_2 - 2 \sum_{i=1}^n \frac{y_i}{(\sigma_m)^2}$$

Logo, $\frac{\partial E}{\partial a_2}(a_1, a_2) = 0$ se, e somente se

$$2 \sum_{i=1}^n \frac{x_i}{(\sigma_m)^2} a_1 + \frac{2n}{(\sigma_m)^2} a_2 - 2 \sum_{i=1}^n \frac{y_i}{(\sigma_m)^2} = 0,$$

que resulta em

$$2 \sum_{i=1}^n \frac{x_i}{(\sigma_m)^2} a_1 + \frac{2n}{(\sigma_m)^2} a_2 = 2 \sum_{i=1}^n \frac{y_i}{(\sigma_m)^2}. \quad (3.4)$$

Como y_i e σ_m são valores conhecidos, então as equações (3.3) e (3.4) constituem o seguinte sistema linear nas variáveis a_1 e a_2 :

$$\begin{cases} 2 \sum_{i=1}^n \left(\frac{x_i}{\sigma_m} \right)^2 a_1 + 2 \sum_{i=1}^n \frac{x_i}{(\sigma_m)^2} a_2 = 2 \sum_{i=1}^n \frac{x_i y_i}{(\sigma_m)^2} \\ 2 \sum_{i=1}^n \frac{x_i}{(\sigma_m)^2} a_1 + \frac{2n}{(\sigma_m)^2} a_2 = 2 \sum_{i=1}^n \frac{y_i}{(\sigma_m)^2} \end{cases}$$

Por outro lado, σ_m é uma estimativa para incertezas iguais na variável dependente, então σ_m é uma constante e podemos reescrever o sistema da seguinte maneira:

$$\begin{cases} \frac{2}{(\sigma_m)^2} \sum_{i=1}^n (x_i)^2 a_1 + \frac{2}{(\sigma_m)^2} \sum_{i=1}^n (x_i) a_2 = \frac{2}{(\sigma_m)^2} \sum_{i=1}^n (x_i y_i) \\ \frac{2}{(\sigma_m)^2} \sum_{i=1}^n (x_i) a_1 + \frac{2}{(\sigma_m)^2} n a_2 = \frac{2}{(\sigma_m)^2} \sum_{i=1}^n (y_i) \end{cases},$$

o que corresponde a:

$$\begin{cases} \sum_{i=1}^n (x_i)^2 a_1 + \sum_{i=1}^n (x_i) a_2 = \sum_{i=1}^n (x_i y_i) \\ \sum_{i=1}^n (x_i) a_1 + n a_2 = \sum_{i=1}^n (y_i) \end{cases}. \quad (3.5)$$

O ponto (a_1, a_2) é ponto crítico de E se, e somente se for solução de (3.5). Então, para que possamos determinar tal ponto, vamos reescrever (3.5) como a equação matricial $AX = B$, onde:

$$A = \begin{bmatrix} \sum_{i=1}^n (x_i)^2 & \sum_{i=1}^n (x_i) \\ \sum_{i=1}^n (x_i) & n \end{bmatrix}, \quad B = \begin{bmatrix} \sum_{i=1}^n y_i x_i \\ \sum_{i=1}^n y_i \end{bmatrix} \quad e \quad X = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}. \quad (3.6)$$

Conforme o discutido no capítulo 2, se a matriz A for inversível, então a solução de $AX = B$ é dada por $X = A^{-1}B$, onde

$$A^{-1} = \frac{1}{\det(A)} \bar{A}.$$

Assim, calculando o determinante e a matriz adjunta de A , temos respectivamente

$$\det(A) = n \cdot \sum_{i=1}^n (x_i)^2 - \left(\sum_{i=1}^n x_i \right)^2$$

e

$$\bar{A} = \begin{bmatrix} n & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & \sum_{i=1}^n (x_i)^2 \end{bmatrix}.$$

De acordo com o corolário 2.43, se verificarmos que $\det(A) \neq 0$, constataremos que a matriz A é inversível. Então, tomando a matriz Y , dada por

$$Y = \frac{1}{\left[n \cdot \sum_{i=1}^n (x_i)^2 - \left(\sum_{i=1}^n x_i \right)^2 \right]} \cdot \begin{bmatrix} n & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & \sum_{i=1}^n (x_i)^2 \end{bmatrix},$$

segue que

$$Y = \begin{bmatrix} \frac{n}{n \cdot \sum_{i=1}^n (x_i)^2 - \left(\sum_{i=1}^n x_i \right)^2} & \frac{-\sum_{i=1}^n x_i}{n \cdot \sum_{i=1}^n (x_i)^2 - \left(\sum_{i=1}^n x_i \right)^2} \\ \frac{-\sum_{i=1}^n x_i}{n \cdot \sum_{i=1}^n (x_i)^2 - \left(\sum_{i=1}^n x_i \right)^2} & \frac{\sum_{i=1}^n (x_i)^2}{n \cdot \sum_{i=1}^n (x_i)^2 - \left(\sum_{i=1}^n x_i \right)^2} \end{bmatrix}. \quad (3.7)$$

Calculando o determinante da matriz Y , obtemos

$$\det(Y) = \frac{1}{n \cdot \sum_{i=1}^n (x_i)^2 - \left(\sum_{i=1}^n x_i \right)^2}.$$

Fazendo o produto entre o determinante da matriz A e o determinante da matriz Y , temos

$$\det(A)\det(Y) = \frac{n \cdot \sum_{i=1}^n (x_i)^2 - \left(\sum_{i=1}^n x_i\right)^2}{n \cdot \sum_{i=1}^n (x_i)^2 - \left(\sum_{i=1}^n x_i\right)^2} = 1,$$

o que implica em $\det(A) \neq 0$ e $Y = A^{-1}$. Logo, a matriz A é inversível.

Diante disto, concluímos que

$$\begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \frac{1}{\left[n \cdot \sum_{i=1}^n (x_i)^2 - \left(\sum_{i=1}^n x_i\right)^2\right]} \cdot \begin{bmatrix} n & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & \sum_{i=1}^n (x_i)^2 \end{bmatrix} \cdot \begin{bmatrix} \sum_{i=1}^n y_i x_i \\ \sum_{i=1}^n y_i \end{bmatrix}. \quad (3.8)$$

é solução de $AX = B$.

Desta forma, temos que (3.8) é solução de (3.5) e pelo teorema 2.54 esta solução é única. Portanto, o ponto (a_1, a_2) é ponto crítico de E , e este, é o único candidato a extremo local.

Determinado o ponto crítico (a_1, a_2) , queremos agora verificar se este é ponto de mínimo local. Para tanto, vamos aplicar o teste da derivada segunda em (3.5). Então, calculando as derivadas parciais segundas de E , obtemos:

- i. $\frac{\partial^2 E}{\partial a_1^2}(a_1, a_2) = 2 \sum_{i=1}^n (x_i)^2;$
- ii. $\frac{\partial^2 E}{\partial a_2^2}(a_1, a_2) = 2n;$
- iii. $\frac{\partial^2 E}{\partial a_1 \partial a_2}(a_1, a_2) = 2 \sum_{i=1}^n x_i;$
- iv. $\frac{\partial^2 E}{\partial a_2 \partial a_1}(a_1, a_2) = 2 \sum_{i=1}^n x_i.$

Tomando essas derivadas e considerando o teorema 1.31, temos que a função E tem um valor de mínimo local em $(a_1, a_2) \in D$ se

$$\frac{\partial E}{\partial a_1^2}(a_1, a_2) > 0 \quad e \quad \frac{\partial^2 E}{\partial a_1^2}(a_1, a_2) \frac{\partial^2 E}{\partial a_2^2}(a_1, a_2) - \left[\frac{\partial^2 E}{\partial a_1 \partial a_2}(a_1, a_2) \right]^2 > 0.$$

É imediato que $\frac{\partial E}{\partial a_1^2}(a_1, a_2) = 2 \sum_{i=1}^n (x_i)^2 > 0$ para quaisquer $x_i \neq 0$. Assim, se verificarmos que

$$\frac{\partial^2 E}{\partial a_1^2}(a_1, a_2) \frac{\partial^2 E}{\partial a_2^2}(a_1, a_2) - \left[\frac{\partial^2 E}{\partial a_1 \partial a_2}(a_1, a_2) \right]^2 = 2 \sum_{i=1}^n (x_i)^2 \cdot 2n - \left[2 \sum_{i=1}^n x_i \right]^2 =$$

$$= 4n \sum_{i=1}^n (x_i)^2 - 4 \left[\sum_{i=1}^n x_i \right]^2 > 0,$$

constataremos que (a_1, a_2) é ponto de mínimo local.

Seja então, $x_i \in \mathbb{R}$, com $i = \{1, 2, \dots, n\}$ e $x_1 \neq x_2 \neq \dots \neq x_n$. Da desigualdade das médias temos que

$$\sqrt{\frac{\sum_{i=1}^n (x_i)^2}{n}} > \frac{\sum_{i=1}^n x_i}{n}. \quad (3.9)$$

Elevando ao quadrado ambos os lados da desigualdade (3.9), obtemos

$$\frac{\sum_{i=1}^n (x_i)^2}{n} > \frac{\left(\sum_{i=1}^n x_i \right)^2}{n^2}. \quad (3.10)$$

Multiplicando ambos os lados da desigualdade (3.10) por $4n^2$, temos

$$4n \sum_{i=1}^n (x_i)^2 > 4 \left(\sum_{i=1}^n x_i \right)^2, \quad (3.11)$$

o que corresponde a

$$4n \sum_{i=1}^n (x_i)^2 - 4 \left(\sum_{i=1}^n x_i \right)^2 > 0. \quad (3.12)$$

Assim, conforme teste da derivada segunda, concluímos que o ponto crítico (a_1, a_2) é ponto de mínimo local.

Queremos agora verificar se o ponto de mínimo local (a_1, a_2) é ponto de mínimo global. De acordo com o teorema 1.39, se verificarmos que a função E é convexa, constataremos a globalidade de (a_1, a_2) . Mas, pelo teorema 1.47 a função E é convexa se a matriz hessiana $HE(a_1, a_2)$ é positiva e semidefinida. Por sua vez e conforme o teorema 1.46, a matriz hessiana é positiva e semidefinida se todos os seus menores principais são maiores ou iguais a zero. Então, tomando as derivadas parciais segundas de E , temos que estas são elementos da matriz hessiana

$$HE(a_1, a_2) = \begin{bmatrix} \frac{\partial^2 E}{\partial a_1^2} & \frac{\partial^2 E}{\partial a_1 \partial a_2} \\ \frac{\partial^2 E}{\partial a_2 \partial a_1} & \frac{\partial^2 E}{\partial a_2^2} \end{bmatrix} = \begin{bmatrix} 2 \sum_{i=1}^n (x_i)^2 & 2 \sum_{i=1}^n x_i \\ 2 \sum_{i=1}^n x_i & 2n \end{bmatrix}.$$

Os menores principais de $HE(a_1, a_2)$ são os seguintes determinantes:

- i. $\left| 2 \sum_{i=1}^n (x_i)^2 \right| = 2 \sum_{i=1}^n (x_i)^2,$
- ii. $\left| 2 \sum_{i=1}^n x_i \right| = 2 \sum_{i=1}^n x_i,$

iii. $| 2n | = 2n$

iv.
$$\begin{vmatrix} 2 \sum_{i=1}^n (x_i)^2 & 2 \sum_{i=1}^n x_i \\ 2 \sum_{i=1}^n x_i & 2n \end{vmatrix} = 4n \sum_{i=1}^n (x_i)^2 - 4 \left(\sum_{i=1}^n x_i \right)^2 .$$

É imediato verificar que os determinantes (i), (ii) e (iii) são maiores ou iguais a zero. Para verificarmos se o determinante (iv) é maior ou igual a zero, vamos considerar todo $x_i \in \mathbb{R}$ com $i = \{1, 2, \dots, n\}$, e tomar

$$\sqrt{\frac{\sum_{i=1}^n (x_i)^2}{n}} \geq \frac{\sum_{i=1}^n x_i}{n},$$

dai segue por processo análogo ao realizado para obtenção de (3.12), que

$$4n \sum_{i=1}^n (x_i)^2 - 4 \left(\sum_{i=1}^n x_i \right)^2 \geq 0.$$

Desta forma concluímos que todos os menores principais da matriz hessiana $HE(a_1, a_2)$ são maiores ou iguais a zero. Logo, a matriz é positiva semidefinida, a função E é convexa e por conseguinte o ponto (a_1, a_2) é ponto de mínimo global. Portanto, $E(a_1, a_2)$ minimiza

$$\sum_{i=1}^n \left[\frac{g(x_i; a_1, a_2) - y_i}{\sigma_m} \right]^2$$

e conseqüentemente a_1 e a_2 são os melhores parâmetros de $g(x_i; a_1, a_2)$.

A incerteza na variável dependente y_i implica em incerteza nos parâmetros a_j que determinam a função de ajuste dos dados experimentais. Assim sendo, queremos estimar as incertezas nos melhores parâmetros de $g(x_i; a_1, a_2)$. A equação (3.1) nos fornece uma estimativa de incerteza referenciada nos desvios em relação à média amostral, mas precisamos estimar as incertezas nos parâmetros considerando os desvios em relação a função. Adaptando os cálculos sugeridos por Stevenson [17] podemos estimar as incertezas nos parâmetros da seguinte forma:

i. Incerteza no parâmetro a_1 :

$$\sigma_{a_1} = \left(\frac{\sum_{i=1}^n [g(x_i; a_1, a_2) - y_i]^2}{(n - p)} \right)^{\frac{1}{2}} \cdot \left(\frac{n}{n \sum_{i=1}^n (x_i)^2 - \left[\sum_{i=1}^n x_i \right]^2} \right)^{\frac{1}{2}}, \quad (3.13)$$

ii. Incerteza no parâmetro a_2 :

$$\sigma_{a_2} = \left(\frac{\sum_{i=1}^n [g(x_i; a_1, a_2) - y_i]^2}{(n-p)} \right)^{\frac{1}{2}} \cdot \left(\frac{\sum_{i=1}^n (x_i)^2}{n \sum_{i=1}^n (x_i)^2 - \left[\sum_{i=1}^n x_i \right]^2} \right)^{\frac{1}{2}}. \quad (3.14)$$

Onde n é o número de pontos experimentais e p é o número de parâmetros da função.

Tomando (3.2) e substituindo σ_m por (3.13) e (3.14), podemos estabelecer um intervalo de confiança para os “verdadeiros” valores dos parâmetros a_1 e a_2 . Assim, para um certo nível de confiança α , temos que:

- i. se $n > 30$, então $a_1 \pm z_{\frac{\alpha}{2}} \cdot \sigma_{a_1}$ e $a_2 \pm z_{\frac{\alpha}{2}} \cdot \sigma_{a_2}$
- ii. se $n \leq 30$, então $a_1 \pm t_{\frac{(1-\alpha)}{2}} \cdot \sigma_{a_1}$ e $a_2 \pm t_{\frac{(1-\alpha)}{2}} \cdot \sigma_{a_2}$.

3.4 Avaliação do ajuste

O método dos mínimos quadrados nos fornece os melhores parâmetros de uma função predeterminada a ser ajustada, mas não nos permite inferir sobre a qualidade do ajuste. Assim, de acordo com Vuolo [18] algum critério de avaliação da qualidade do ajuste deve sempre ser utilizado juntamente com o método dos mínimos quadrados, mesmo quando a função a ser ajustada é bem conhecida.

Um critério que podemos usar é a determinação de uma medida que estabelece o quanto as variáveis estão relacionadas. Segundo Stevenson [17] essa medida é denominada *coeficiente de correlação* (r) e pode ser calculada da seguinte maneira:

$$r = \sqrt{1 - \frac{\sum_{i=1}^n [y_i - g(x_i)]^2}{\sum_{i=1}^n (y_i - \bar{y})^2}}, \quad (3.15)$$

onde $[y_i - g(x_i)]^2$ e $[y_i - \bar{y}]^2$ representam respectivamente os desvios dos pontos experimentais em relação a função ajustada e à média.

O valor de r varia de -1 a 1 , essa variação caracteriza a natureza do relacionamento entre as variáveis de tal modo que:

- i. valores próximos de -1 ou 1 indicam relacionamento forte, uma vez que a dispersão dos pontos experimentais em relação a função ajustada é pequena;
- ii. valores próximos de 0 indicam relacionamento fraco, pois sugerem uma maior dispersão dos pontos experimentais em relação a função ajustada;

- iii. valores de $r > 0$ indicam que valores altos (ou baixos) de uma das variáveis, correspondem a valores altos (ou baixos) da outra;
- iv. valores de $r < 0$ indicam que valores altos (ou baixos) de uma das variáveis, correspondem a valores baixos (ou altos) da outra.

Podemos complementar a estatística dada pelo coeficiente de correlação, obtendo uma descrição do grau de relacionamento entre as variáveis por meio do *coeficiente de determinação* (r^2). Este, nos fornece a porcentagem de variação de uma variável que é *explicada* estatisticamente pela variação em outra variável.

O intervalo de variação do coeficiente de determinação é $0 \leq r^2 \leq 1$, em termos percentuais, o intervalo indica que quanto mais próximo r^2 estiver de 1 a variação de uma variável será melhor explicada pela variação da outra, e isso significa, de acordo com Stevenson [17], que a função ajustada é melhor preditor que a média. Inversamente, a variação de uma variável não poderá ser explicada pela variação da outra variável quanto mais próximo r^2 estiver de zero, o que implica segundo o mesmo autor, que a média é melhor preditor que a função ajustada. Vejamos o seguinte exemplo:

Exemplo 3.4. Seja um coeficiente de correlação $r = 0,9$. O sinal de r nos diz que existe um relacionamento positivo entre dois conjuntos de valores pertinentes a duas variáveis. O valor de r sugere que as variáveis tem um relacionamento forte. Por outro lado, $r^2 = 0,81$, o que significa que 81% da variação dos valores das variáveis pode ser explicada pelo relacionamento entre ambas.

4 O experimento

Neste capítulo, inicialmente relataremos a modelagem de um experimento realizado por alunos do ensino fundamental nas aulas da disciplina matemática. Em seguida, aplicaremos o método dos mínimos quadrados no ajuste de curva dos dados coletados de acordo com o que foi discutido nos capítulos anteriores.

4.1 Motivação para realização do experimento

Na qualidade de professor de matemática de ensino fundamental de escola pública municipal, somos regularmente levados a refletir sobre nossos métodos de ensino por meio de ações realizadas por nossos coordenadores pedagógicos. Essa prática ocorre devido às características peculiares do sistema público de ensino, bem como, devido as dificuldades de aprendizagem apresentadas por nossos alunos.

Um pressuposto recorrente destacado pelos coordenadores em suas indagações sobre aprendizagem matemática é o seguinte:

“se os conceitos matemáticos fazem sentido para o aluno, então este ficará motivado a buscar a sua compreensão.”

Assim, *o sentido* para eles, é uma condição necessária para que a aprendizagem matemática ocorra, e esta é estabelecida quando a prática pedagógica permite contextualizações que se aproximem do cotidiano dos alunos. Nessa mesma perspectiva, Lima (2007), citado por Costa [6], entende que *“boas contextualizações são as que, por meio da problematização, envolvam aplicações ou manipulações [...] de informações que sejam reais ou simulem a realidade”*. Ainda de acordo com o mesmo autor, a instrumentação matemática adequada para traduzir as situações contextualizadas é a chamada *modelagem matemática*.

Corroborando as afirmações de Lima, nas citações de Costa, Bassanezi [3] diz que

“a modelagem matemática consiste na arte de transformar problemas da realidade em problemas matemáticos e resolvê-los interpretando suas soluções na linguagem do mundo real. Desta forma, [...] a modelagem é um processo que alia teoria e prática, motivando seu usuário na procura do

entendimento da realidade que o cerca e na busca de meios para agir sobre ela e transformá-la”.

Considerando então, que a modelagem matemática pudesse favorecer a construção de sentido para o aluno no processo de ensino e aprendizagem, direcionamos esforços para implementar uma prática de ensino que fizesse uso de tal processo. E a oportunidade surgiu quando, numa das escolas onde lecionamos matemática, foi realizado junto aos alunos de 9º ano um projeto denominado MOBFOG – Mostra Brasileira de Foguetes. Tal projeto era constituído de dois objetivos:

- i. construção e melhoramento estrutural de foguetes (construídos com garrafas PET e propulsionados a ar comprimido) de modo a atingir a maior distância possível ao longo da trajetória horizontal;
- ii. coleta e tabulação dos resultados de lançamento para envio a OBA - Olimpíada Brasileira de Astronomia.

O projeto foi inicialmente desenvolvido pela professora da disciplina *ciências*, mas na fase de coleta e tabulação dos resultados de lançamento fomos convidados a participar. A princípio iríamos apenas ajudar os alunos (que eram comuns às disciplinas *ciências* e *matemática*) no ordenamento e representação dos resultados.

No entanto, percebemos a possibilidade de usar a *modelagem matemática* como estratégia de ensino do conteúdo que estávamos ministrando na época (função afim). Assim, transformamos essa etapa do projeto MOBFOG num experimento definido da seguinte maneira:

Lançamento de foguete construído com garrafas PET, cuja propulsão se dá por meio de bombeadas numa bomba de ar para bicicletas e o desempenho é determinado pela medida (em metros) da distância percorrida ao longo da trajetória horizontal.

O objetivo do experimento era determinar, a partir dos dados coletados, uma relação funcional entre a distância percorrida pelo foguete ao longo da trajetória horizontal e o número de bombeadas aplicadas para propulsioná-lo, de modo que pudéssemos elaborar um *modelo* para fazer previsões de desempenho dado um número qualquer de bombeadas.

4.2 Relato do experimento

Nas aulas de ciências cada aluno construiu seu foguete com adaptações e melhoramentos estruturais particulares. Então, estabelecemos uma amostra de 10 lançamentos por foguete para que cada aluno determinasse o melhor desempenho de seu foguete. Dos resultados gerados, foi construída uma tabela que destacava os melhores desempenhos de cada foguete, e assim, os alunos puderam eleger o melhor dos foguetes.

Estabelecido o melhor dos foguetes, foi feita uma segunda amostra, onde cada aluno realizou 10 lançamentos com este foguete. Analogamente ao que foi feito na primeira amostra, os resultados foram organizados numa tabela geral dos melhores desempenhos e assim, os alunos conseguiram determinar o melhor desempenho individual com o melhor foguete. A figura 4.1 ilustra um dos lançamentos do melhor foguete.

Figura 4.1: Lançamento do melhor foguete



Fonte: Foto gerada pelo autor

Ordenamos o experimento dessa forma pois observamos que cada aluno apresentava variação no número de bombeadas realizadas de um lançamento para outro, devido a falhas na bomba, bem como havia variação no número de bombeadas de aluno para aluno devido as características físicas de cada um.

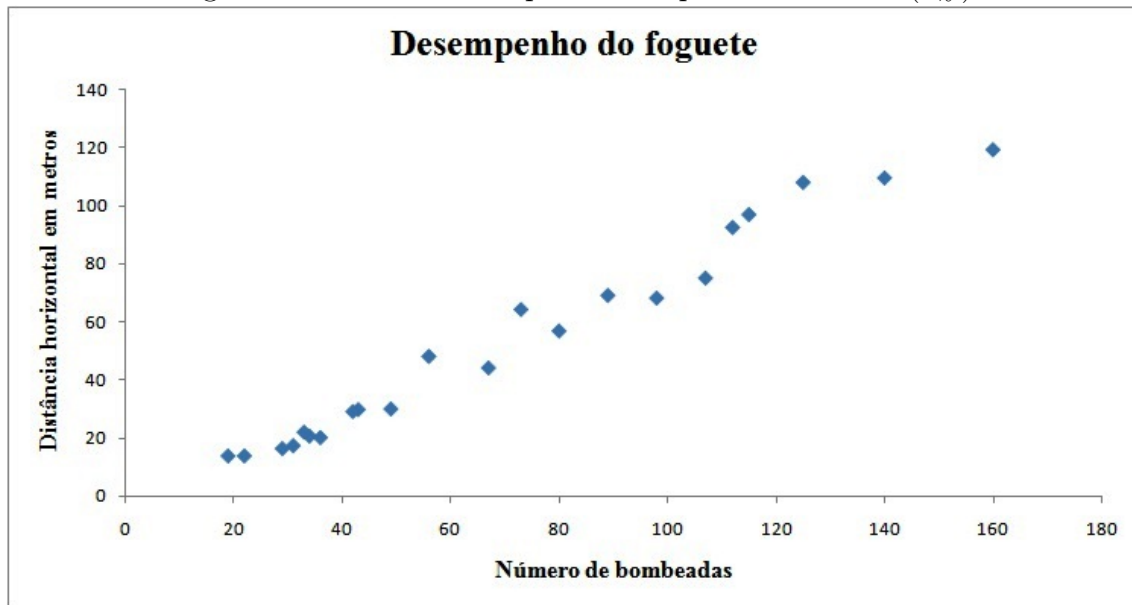
Procedendo dessa maneira, obtivemos um conjunto de dados oriundos dos melhores desempenhos do foguete gerados a partir dos melhores desempenhos individuais dos alunos. Esses dados foram organizados na tabela 4.1:

Tabela 4.1: Desempenho do foguete em função do número de bombeadas

Série de lançamentos	Número de bombeadas	Distância percorrida (em metros)	Série de lançamentos	Número de bombeadas	Distância percorrida (em metros)
1 ^a	98	68,32	12 ^a	89	69,21
2 ^a	43	29,91	13 ^a	140	109,68
3 ^a	160	119,42	14 ^a	112	92,64
4 ^a	22	13,95	15 ^a	19	13,9
5 ^a	34	20,76	16 ^a	36	20,26
6 ^a	73	64,38	17 ^a	42	29,21
7 ^a	31	17,46	18 ^a	56	48,23
8 ^a	67	44,25	19 ^a	115	97,1
9 ^a	80	56,98	20 ^a	49	30,12
10 ^a	29	16,45	21 ^a	125	108,16
11 ^a	107	75,2	22 ^a	33	22,13

Fonte: Tabela gerada pelo autor

A partir dessa tabela, ministramos numa primeira etapa, aula expositiva na sala de informática com o objetivo de mostrar aos alunos as várias formas de representação de dados organizados. Dentre elas, estabelecemos de acordo com nossos objetivos, que o gráfico de dispersão se apresentava como a melhor forma de representação. Denotamos o número de bombeadas pela variável x e o desempenho do foguete pela variável y , e desta forma, obtivemos pares ordenados (x, y) como pontos do gráfico de dispersão Y versus X . Tal gráfico é ilustrado pela figura a seguir:

Figura 4.2: Gráfico de dispersão dos pares ordenados (x,y) 

Fonte: Gráfico gerado pelo autor

Numa segunda etapa da aula, tomamos a figura 4.2 e induzimos uma discussão a respeito do significado da disposição dos pontos no gráfico. Considerando o conteúdo que os alunos estavam estudando naquele momento, queríamos que percebessem alguma relação entre as variáveis. Dentre as várias suposições que fizeram, uma delas chamou atenção, um aluno afirmou que “*a variação dos valores no eixo das abscissas x , parecia provocar uma variação nos valores do eixo das ordenadas y , mas não conseguia entender como isso acontecia.*”

A esse respeito, Rezende (2003) citado por Costa [6], afirma que “*compreender que a variação de uma grandeza depende da variação de outra é um aspecto importante no estudo do conceito de função [...], mas se torna incompleto se não estudarmos como ocorre esta variação...*” Em vista disto, percebemos que estávamos no caminho certo no uso da modelagem como estratégia de ensino, uma vez que esta abordagem favoreceria a discussão e compreensão da *variação* enquanto *relação* entre as variáveis, e isso poderia ajudar os alunos a assimilar melhor os conceitos que estavam estudando.

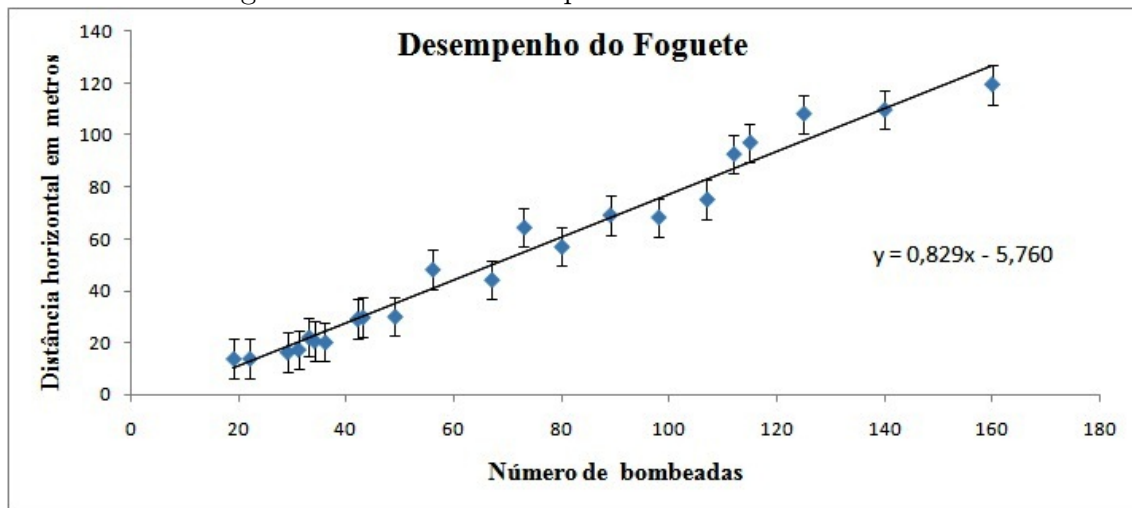
Assim, decidimos mostrar aos estudantes que se existisse uma relação entre as variáveis x e y , esta determinaria o padrão da disposição dos pontos no gráfico de dispersão. Perguntamos a eles se reconheciam algum padrão na disposição dos pontos da figura 4.2, e a resposta foi *negativa*.

Diante disto, explicamos que embora fosse impossível uma reta passar por cada um dos pontos grafados na figura 4.2, o comportamento destes indicava com certa razoabilidade uma *relação linear entre as variáveis*. Então, deveríamos encontrar uma reta que estivesse o mais próximo possível de cada um dos pontos. Esclarecemos ainda,

que esse procedimento é denominado *ajuste de reta aos pontos experimentais* e que os *desvios* (distâncias entre cada ponto e a reta) são erros de aproximação no ajuste.

Findada as explicações e esclarecimentos, resolvemos ensiná-los, usando o software Excel[®], a efetuar o ajuste de reta aos dados experimentais por meio da construção da *linha de tendência*. A figura 4.3 ilustra o resultado que obtiveram:

Figura 4.3: Gráfico de dispersão e linha de tendência



Fonte: Gráfico gerado pelo autor

Da figura 4.3, destacamos aos alunos que devido ao fato de haver diferenças relevantes no desempenho individual quanto ao número de bombeadas aplicadas ao foguete, não tínhamos *certeza* a respeito do desempenho do mesmo. Assim, cada barra vertical na qual estava contido um ponto do gráfico, representava um intervalo de valores possíveis de desempenho do foguete para um valor específico de bombeadas.

A expressão $y = 0,829x - 5,760$ dada pelo Excel, representava a equação da reta de aproximação aos pontos experimentais, que por sua vez expressava uma relação funcional $f(x) = y$ entre as variáveis x e y . Desta forma, podíamos entender que a reta ajustada era a representação gráfica de uma função do tipo $f(x) = ax + b$, com $a = 0,829$ (coeficiente angular da reta) indicando a variação de y por unidade de variação de x e $b = -5,760$ (coeficiente linear) indicando a cota da reta em $x = 0$. Tal função era o modelo matemático que estávamos querendo determinar para fazer previsões de desempenho do foguete dado um número qualquer de bombeadas.

No entanto, salientamos que embora tenhamos afirmado aos alunos que a equação $y = 0,829x - 5,760$ expressava uma relação funcional que nos permitiu formular um modelo matemático para determinação de desempenho do foguete, sabemos que:

“uma relação funcional, obtida através de ajuste de dados, propicia condições para formulação de modelos [...] que devem comportar em seus parâmetros qualidades e significados inerentes ao fenômeno analisado e para isto se faz necessário um estudo mais detalhado do próprio fenômeno.” (BASSANEZI, 2002) [3]

Como o objetivo didático de nosso experimento era envolver os alunos em algum processo de ensino e aprendizagem que colaborasse para que *os conceitos matemáticos estudados por eles naquele momento fizessem sentido*, não efetuamos uma análise mais aprofundada das condições de lançamento do foguete, e portanto não estabelecemos significados empíricos para os parâmetros da função encontrada no ajuste.

Todavia, essa especificidade da modelagem não implicou em prejuízo para o entendimento da concepção matemática da expressão que apresentamos como modelo, nem tão pouco na relação que tentamos estabelecer entre as variáveis observadas no experimento.

Os alunos manifestaram interesse em saber como o Excel formulou a expressão $y = 0,829x - 5,760$. Explicamos a eles que tal expressão era resultado de cálculos realizados de acordo com *método dos mínimos quadrados* elaborado pelo matemático Carl Friedrich Gauss. Esse método consistia em estimar os parâmetros de uma função ajustada aos pontos experimentais de modo que os erros de aproximação do ajuste fossem os menores possíveis.

Justificamos aos alunos que era necessário conhecimentos avançados para entender a parte inicial do processo de cálculo do método dos mínimos quadrados, mas como já haviam estudado, no 8º ano, resolução de sistemas de duas equações do 1º grau com duas incógnitas, poderíamos apresentar a eles o sistema resultante dos cálculos e a partir dele mostrar a determinação dos parâmetros da função de ajuste.

Inicialmente, explicamos o conceito de somatório e sua notação, em seguida apresentamos o sistema na sua forma geral:

$$\left[\sum_{i=1}^{22} (x_i)^2 \right] \cdot a + \left[\sum_{i=1}^{22} (x_i) \right] \cdot b = \left[\sum_{i=1}^{22} (x_i) \cdot (y_i) \right]$$

$$\left[\sum_{i=1}^{22} (x_i) \right] \cdot a + 22 \cdot b = \left[\sum_{i=1}^{22} (y_i) \right]$$

Apresentamos o sistema dessa forma, pois não consideramos as incertezas nas explicações que fizemos para os alunos.

Solicitamos então, que construíssem a seguinte tabela:

Tabela 4.2: Tabela de somatório dos dados do experimento

(i)	(x_i)	(y_i)	(x_i) ²	(x_i) · (y_i)
1	98	68,32	9604	6695,36
2	43	29,91	1849	1286,13
3	160	119,42	25600	19107,2
4	22	13,95	484	306,9
5	34	20,76	1156	705,84
6	73	64,38	5329	4699,74
7	31	17,46	961	541,26
8	67	44,25	4489	2964,75
9	80	56,98	6400	4558,4
10	29	16,45	841	477,05
11	107	75,2	11449	8046,4
12	89	69,21	7921	6159,69
13	140	109,68	19600	15355,2
14	112	92,64	12544	10375,68
15	19	13,9	361	264,1
16	36	20,26	1296	729,36
17	42	29,21	1764	1226,82
18	56	48,23	3136	2700,88
19	115	97,1	13225	11166,5
20	49	30,12	2401	1475,88
21	125	108,16	15625	13520
22	33	22,13	1089	730,29
$\sum_{i=1}^{22}$	1560,00	1167,72	147124,00	113093,43

Fonte: Tabela gerada pelo autor

Em seguida, pedimos que substituíssem os resultados de somatório da tabela 4.2 no sistema geral apresentado para que pudessem reescrevê-lo com os dados do experimento. E assim, obtiveram o seguinte sistema:

$$147124 \cdot a + 1560 \cdot b = 113093,43$$

$$1560 \cdot a + 22 \cdot b = 1167,72$$

Relembramos a eles, numa breve exposição, dois métodos de resolução de sistemas de duas equações do 1º grau com duas incógnitas, denominados *método da substituição* e *método da adição*. Solicitamos então, que escolhessem um dos métodos para que

trabalhassem de maneira homogênea, elegeram o método da substituição e efetuaram seus cálculos. Organizamos os resultados apresentados da seguinte maneira:

Tomando a equação

$$1560 \cdot a + 22 \cdot b = 1167,72. \quad (4.1)$$

Isolando a em (4.1), obtemos

$$a = \frac{1167,72 - 22 \cdot b}{1560}. \quad (4.2)$$

Substituindo (4.2) na equação

$$147124 \cdot a + 1560 \cdot b = 113093,43 \quad (4.3)$$

obtemos

$$147124 \cdot \frac{1167,72 - 22 \cdot b}{1560} + 1560 \cdot b = 113093,43.$$

Que resulta em

$$b = \frac{-2965,457}{514,826} = -5,760. \quad (4.4)$$

Substituindo (4.4) em (4.2) obtemos

$$a = \frac{1167,72 - 22 \cdot (-5,760)}{1560} = 0,829. \quad (4.5)$$

Diante dos resultados (4.4) e (4.5), esclarecemos aos alunos que como tínhamos estabelecido que a função de ajuste dos pontos experimentais era do tipo

$$f(x) = ax + b,$$

onde $f(x) = y$ representava a relação funcional entre as variáveis x e y de nosso experimento, poderíamos concluir que a função procurada era

$$f(x) = 0,829x - 5,760,$$

e esta confirmava o modelo sugerido pelo Excel.

Isto posto, encerramos o experimento solicitando aos alunos que realizassem duas atividades:

- i. Construir, usando a planilha Excel, uma tabela contendo número de bombeadas, valores coletados do desempenho do foguete, valores dados pela simulação de desempenho do modelo e erro absoluto associado. Fazer uma breve explicação dos resultados apresentados na tabela;
- ii. Dada a função $f : \mathbb{R} \rightarrow \mathbb{R}$ tal que $f(x) = 0,829x - 5,760$, responder e fazer o que se pede:

- 1-) Qual é a raiz dessa função? Qual é o seu significado geométrico?
- 2-) Construir o gráfico de f ;
- 3-) Fazer o estudo de sinal da função f .

A atividade (i) tinha por objetivo propiciar aos alunos a possibilidade de inferência da idéia de aproximação e variação, e, a atividade (ii) o de favorecer a análise da função $f(x) = 0,829 \cdot x - 5,76$ de acordo com os conteúdos estudados em sala de aula.

4.3 Ajuste de curva

Do que foi relatado na seção 4.2, vimos que o comportamento dos pontos experimentais indicava uma relação linear entre as variáveis designadas por x e y . Em vista disto e dadas as restrições do nível de escolaridade dos alunos, fizemos o ajuste de reta apresentando a eles apenas uma síntese dos cálculos exigidos pelo método dos mínimos quadrados. Assim sendo, queremos nessa seção realizar o ajuste efetivo dos dados obtidos no experimento.

De acordo com a discussão feita no capítulo 3, encontrar a melhor reta que descreve o comportamento dos pontos experimentais implica em ajustar a função

$$g(x_i; a, b) = ax + b.$$

Para tanto, devemos determinar os valores dos parâmetros a e b de tal forma que minimizem os erros de aproximação dados por:

$$E(a, b) = \sum_{i=1}^n \left[\frac{g(x_i; a, b) - y_i}{\sigma_i} \right]^2 = \sum_{i=1}^n \left[\frac{(ax_i + b) - y_i}{\sigma_i} \right]^2.$$

Conforme o que foi visto em (3.5) tais valores devem satisfazer, necessariamente, às condições:

$$\frac{\partial E}{\partial a} = 0 \iff \sum_{i=1}^n (x_i)^2 a + \sum_{i=1}^n x_i b = \sum_{i=1}^n y_i x_i \quad (4.6)$$

$$\frac{\partial E}{\partial b} = 0 \iff \sum_{i=1}^n x_i a + nb = \sum_{i=1}^n y_i \quad (4.7)$$

Reescrevendo as equações (4.6) e (4.7) em notação matricial, ou seja na forma $AX = B$, onde

$$A = \begin{bmatrix} \sum_{i=1}^n (x_i)^2 & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & n \end{bmatrix}, \quad B = \begin{bmatrix} \sum_{i=1}^n y_i x_i \\ \sum_{i=1}^n y_i \end{bmatrix} \quad e \quad X = \begin{bmatrix} a \\ b \end{bmatrix},$$

temos que a solução geral para ajuste da função é dada por $X = A^{-1}B$, ou seja:

$$\begin{bmatrix} a \\ b \end{bmatrix} = \frac{1}{\left[n \cdot \sum_{i=1}^n (x_i)^2 - \left(\sum_{i=1}^n x_i \right)^2 \right]} \cdot \begin{bmatrix} n & -\sum_{i=1}^n x_i \\ -\sum_{i=1}^n x_i & \sum_{i=1}^n (x_i)^2 \end{bmatrix} \cdot \begin{bmatrix} \sum_{i=1}^n y_i x_i \\ \sum_{i=1}^n y_i \end{bmatrix}. \quad (4.8)$$

Assim, tomando os valores da tabela 4.1 e calculando os somatórios descritos na solução geral (4.8), obtemos os resultados listados na tabela 4.3:

Tabela 4.3: Tabela resumida de somatório dos dados do experimento

$\sum_{i=1}^n (x_i)^2$	$\sum_{i=1}^n x_i$	$\sum_{i=1}^n y_i$	$\sum_{i=1}^n y_i x_i$	n
147124,00	1560,00	1167,72	113093,43	22

Fonte: Tabela gerada pelo autor

Substituindo os valores da tabela 4.3 na equação (4.8), obtemos

$$\begin{bmatrix} a \\ b \end{bmatrix} = \frac{1}{[22 \cdot 147124 - (1560)^2]} \cdot \begin{bmatrix} 22 & -1560 \\ -1560 & 147124 \end{bmatrix} \cdot \begin{bmatrix} 113093,43 \\ 1167,72 \end{bmatrix}.$$

O que resulta em

$$\begin{aligned} a &= 0,829771 \\ b &= -5,760120 \end{aligned}$$

Stevenson [17] afirma que ao tomarmos um número n de observações para determinarmos a função $g(x)$, este número é relativamente pequeno diante de uma população infinita de pares de valores possíveis. Assim, $g(x)$ é uma estimativa da real relação, porém desconhecida, que existe entre as variáveis observadas. Consequentemente, os parâmetros a e b de $g(x)$ também são estimativas dos reais parâmetros populacionais correspondentes.

Em vista disto, precisamos estimar as incertezas e estabelecer os intervalos de confiança para os verdadeiros valores dos parâmetros a e b de $g(x)$. Então, considerando a função $g(x)$, a média \bar{y} dos n pontos experimentais, os valores descritos na tabela 4.1 e calculando alguns somatórios, obtemos a seguinte tabela:

Tabela 4.4: Tabela de alguns somatórios dos dados do experimento, dos desvios em relação a função e à média, do número de observações e do número de parâmetros da função ajustada

$\sum_{i=1}^n (x_i)^2$	$\left(\sum_{i=1}^n x_i\right)^2$	$\sum_{i=1}^n [g(x_i) - y_i]^2$	$\sum_{i=1}^n [\bar{y} - y_i]^2$	n	p
147124,00	2433600,00	622,453593	25757,431527	22	2

Fonte: Tabela gerada pelo autor

Tomando os valores da tabela 4.4 e substituindo em (3.13) e (3.14), obtemos as incertezas em a e b :

$$\sigma_a = \left(\frac{622,453593}{22-2}\right)^{\frac{1}{2}} \cdot \left(\frac{22}{22 \cdot 147124 - 2433600}\right)^{\frac{1}{2}} = 0,029198$$

e

$$\sigma_b = \left(\frac{622,453593}{22-2}\right)^{\frac{1}{2}} \cdot \left(\frac{147124}{22 \cdot 147124 - 2433600}\right)^{\frac{1}{2}} = 2,387744.$$

Queremos estabelecer um intervalo de confiança de 95% para os verdadeiros valores dos parâmetros a e b . Como o número de observações é pequeno, $n = 22 < 30$, de (3.2) vamos tomar

$$t_{\frac{(1-\alpha)}{2}} \sigma_m. \quad (4.9)$$

Então, consultando a tabela de probabilidades para distribuição t [17] e calculando $t_{\frac{(1-\alpha)}{2}}$, com $\alpha = 0,95$ e $(n-p) = 20$ graus de liberdade, temos que

$$t_{\frac{(1-0,95)}{2}} = t_{0,025} = 2,086.$$

Substituindo σ_m por σ_a e σ_b e aplicando em (4.9), obtemos respectivamente:

$$t_{0,025} \cdot \sigma_a = 2,086 \cdot 0,029198 = 0,060907028$$

e

$$t_{0,025} \cdot \sigma_b = 2,086 \cdot 2,387744 = 4,980833984.$$

Assim, temos que o intervalo de confiança de 95% para os verdadeiros valores dos parâmetros a e b são:

$$a \pm 0,060907028 \quad \text{ou} \quad 0,768863972 < a < 0,890678028$$

e

$$b \pm 4,980833984 \quad \text{ou} \quad -10,74095398 < b < -0,779286016.$$

Tomando os valores da tabela 4.4 e aplicando em (3.15), obtemos o coeficiente de correlação

$$r = \sqrt{1 - \frac{622,453593}{25757,431527}} = 0,987843.$$

O sinal e o valor de r sugerem que as variáveis x e y tem um relacionamento positivo e forte, ou seja, a medida que cresce os valores da variável x também cresce os valores da variável y , bem como, a dispersão entre os pontos experimentais e a função ajustada é pequena. Por outro lado, temos que o coeficiente de determinação $r^2 = 0,975834$ nos indica que aproximadamente 97,5834% da variação dos valores de y podem ser explicados pela variação dos valores de x , e isso significa que a função ajustada é melhor preditor que a média.

Portanto, de acordo com o método dos mínimos quadrados, a função

$$g(x) = 0,829771x - 5,760120,$$

com

$$a = (0,8298 \pm 0,061) \quad e \quad b = (-5,7601 \pm 4,981)$$

é a que melhor se ajusta aos dados experimentais.

5 Considerações finais

De acordo com as orientações descritas nos Parâmetros Curriculares Nacionais – PCNs, que enfatiza a resolução de problemas, a compreensão e aplicação de conceitos matemáticos como elementos norteadores da prática pedagógica, realizamos junto aos alunos do ensino fundamental de uma escola pública, a modelagem matemática de um experimento, onde tivemos a oportunidade de implementar uma abordagem de ensino que nos permitiu contextualizar os conteúdos curriculares que estávamos ministrando.

A modelagem do experimento enquanto prática pedagógica, objetivava a determinação de um modelo (oriundo do ajuste de função aos dados observados) para que os alunos pudessem perceber, a partir de uma situação vivenciada na prática, as aplicações dos conceitos matemáticos estudados em sala de aula.

Nesse sentido, uma discussão mais aprofundada com os alunos, sobre o método usado e as condições gerais de ajuste, não faziam parte do contexto pedagógico no qual construímos essa prática de ensino, sobretudo por conta do nível educacional dos alunos (9º ano do ensino fundamental).

A partir desse experimento, no entanto, pudemos efetuar um estudo do método dos mínimos quadrados fundamentado em resultados matemáticos que nos permitiu ampliar nosso entendimento acerca das concepções e aplicabilidade do método.

Assim, da vivência que tivemos na aplicação do experimento e do estudo que realizamos, concluímos que é viável extrapolar essa prática para alunos do ensino médio, de modo a explorar a aplicabilidade de conceitos tais como: funções, operações com matrizes, resolução de sistemas lineares, entre outros que possam ser enquadrados em situações-problema que atendam as restrições do experimento e do processo de cálculo do método.

Referências

- [1] AMORIM, Ronan Gomes de. *Introdução à Análise Convexa - Conjuntos e Funções Convexas*. Dissertação de Mestrado. Goiânia/GO: PROFMAT - UFG, 2013.
- [2] ASANO, Claudio Hirofume. COLLI, Eduardo. *Cálculo numérico - Fundamentos e Aplicações*. São Paulo: Edusp, 2009.
- [3] BASSANEZI, Rodney Carlos. *Ensino-aprendizagem com modelagem matemática: uma nova estratégia*. São Paulo: Contexto, 2002.
- [4] BORTOLOSSI, Humberto José. *Cálculo Diferencial a Várias Variáveis- Uma introdução à Teoria de Otimização*, Rio de Janeiro: Ed. PUC-Rio; São Paulo: Loyola, 2002.
- [5] CAROLI, Alésio de. [et al.]. *Matrizes, Vetores, Geometria analítica: teoria e exercícios*. 17^a ed. São Paulo: Nobel, 1989.
- [6] COSTA, Matheus Moreira. *O ensino das funções exponenciais: uma proposta por meio de contextualização, modelagem matemática e recursos tecnológicos*. Dissertação de Mestrado. Rio Claro/SP: PROFMAT - UNESP, 2016.
- [7] CRATO, N. *O papel dos mínimos quadrados na descoberta dos planetas*. Disponível em: <https://pascal.iseg.utl.pt/ncrato/papers/MinQdSPM.pdf>. 27.01.2014.
- [8] GUELLI, Cid. Augusto [et.al.]. *Álgebra II: análise combinatória, probabilidade, matrizes, determinantes, sistemas lineares*. Matemática moderna. São Paulo: moderna, vol. 6, 197-?.
- [9] GUIDORIZZI, Hamilton Luiz. *Um curso de cálculo*, 5^a ed. Rio de Janeiro: LTC, vol. 2, 2002.
- [10] HLENKA, Vanessa. *Principais propriedades de funções convexas*. Monografia. Curitiba: UFPR, 2006.
- [11] IEZZI, Gelson. [et.al.]. *Matemática 2^o grau, 2^a série: versão azul*. São Paulo: Atual, 1993.

-
- [12] LEITHOLD, Louis. *O cálculo com Geometria Analítica*. Tradução: PATARRA, Cyro de Carvalho. 3^a ed. São Paulo: Editora Harbra, vol. 2, 1994.
- [13] LIMA, Paulo Cupertino de. *Cálculo de Várias Variáveis*. Belo Horizonte: Editora UFMG, 2009.
- [14] MALTA, Iaci. [et al.]. *Cálculo a uma variável*. São Paulo: Ioyola, 2002.
- [15] RUGGIERO, Marcia A. Gomes. LOPES, Vera Lúcia da Rocha. *Cálculo Numérico - Aspectos Computacionais*, 2^a ed. São Paulo: Person Makron Books, 1996.
- [16] SÁ, Fernanda Lúcia. *Estudo dos determinantes*. Caderno DÁ Licença. Rio de Janeiro, v.5, Ano 6, p. 70-84, Dez. 2004.
- [17] STEVENSON, Willian J. *Estatística aplicada à administração*. Tradução: FARIA, Alfredo Alves. São Paulo:Harper Row do Brasil, 1981.
- [18] VUOLO, José Henrique. *Fundamentos da Teoria de Erros*, São Paulo: Editora Edgard Blücher Ltda, 2002.