Universidade Federal de Campina Grande
Centro de Educação e Saúde
Unidade Acadêmica de Física e Matemática
Licenciatura em Matemática

TRABALHO DE CONCLUSÃO DE CURSO

Aplicabilidade de Funções Linear e Quadrática na Estatística

Sérgio Luiz Macêdo do Amaral

Cuité-PB
Julho, 2018

Universidade Federal de Campina Grande
Centro de Educação e Saúde
Unidade Acadêmica de Física e Matemática
Licenciatura em Matemática

Aplicabilidade de Funções Linear e Quadrática na Estatística

Sérgio Luiz Macêdo do Amaral

Trabalho de Conclusão de Curso apresentado a Unidade Acadêmica de Física e Matemática do Curso de Licenciatura em Matemática da Universidade Federal de Campina Grande como requisito parcial para a obtenção do grau de Licenciado em Matemática.

Orientador: Prof. Dr. Jorge Alves de Sousa.

Cuité-PB
Julho, 2018

FICHA CATALOGRÁFICA ELABORADA NA FONTE
Responsabilidade Rosana Amâncio Pereira – CRB 15 – 791

A447a Amaral, Sérgio Luiz Macêdo do.

Aplicabilidade de funções linear e quadrática na estatística. / Sérgio Luiz Macêdo do Amaral. – Cuité: CES, 2018.

70 fl.

Monografia (Curso de Licenciatura em Matemática) – Centro de Educação e Saúde / UFCG, 2018.

Orientador: Dr. Jorge Alves de Sousa.

1. Funções. 2. Regressão Linear Simples. 3. Software R. I. Título.

Biblioteca do CES - UFCG CDU 51

Aplicabilidade de Funções Linear e Quadrática na Estatística

Sérgio Luiz Macêdo do Amaral

Aprovado em _____

Banca Examinadora

Prof. Dr. Jorge Alves de Sousa
Orientador

Prof. Dr. Aluizio Freire da Silva Junior
Examinador Interno

Prof. Ms. Maria de Jesus R. da Silva
Examinador Interno

Por isso mesmo, empenhem-se para acrescentar à sua fé a virtude; à virtude o conhecimento; ao conhecimento o domínio próprio; ao domínio próprio a perseverança; à perseverança a piedade; à piedade a fraternidade; e à fraternidade o amor.

(Biblia Sagrada, 2 Pedro 1:5-7.)

Dedicatória

A Deus pela possibilidade de poder respirar cada dia ao amanhecer. A sua proteção ao trilhar cada jornada, e sua força que me impulsiona na direção das minhas conquistas.

.

Agradecimentos

Depois de tão longa caminhada, lembrando todo o trajeto desse estudo, é com imensa gratidão, amor, carinho e saudade a todos que fizeram parte desse percurso da minha vida que concluo mais essa etapa.

Sou grato a Deus por toda luz que me guiou por esse caminho, iluminando meus passos, que mesmo nos momentos mais difíceis nunca me deixou desanimar, e que diante da fé sempre fosse feita sua vontade. De forma especial agradeço a minha família: A minha esposa Joanna que esta comigo desde o início da minha vida acadêmica apoiando minhas escolhas e a minha filha Beatriz, que me foi inspiração para conclusão desse trabalho. Aos meus pais João e Cleonice, que não mediram esforços na tentativa de zelar pela minha educação na busca de um futuro melhor. As minhas irmãs Joanice e Jussara, pelo incentivo nas horas de desânimo e pela presença nas alegrias de minhas conquistas.

É com imensa gratidão e carinho que venho citar minha primeira professora Gorete, que além do conhecimento, transmitiu a forma carinhosa e respeitosa com a qual desde cedo tratava todos aqueles que tiveram a satisfação e o privilégio dos seus ensinamentos como professora, lições essas que foram muito além da sala de aula.

Aos meus avós maternos Lula e Maria, e paternos, Josias e Maria, estes em memória. As minhas tias Lucimar e Daluz, onde elas representam todos os meus tios, obrigado por todo apoio e carinho.

Não poderia deixar de agradecer a todos os funcionários das escolas Municipal e Estadual de Frei Martinho, onde cursei ensino fundamental e médio respectivamente, aos professores Francinha, Fernandinho, Melânia e a diretora Guia Matos, a qual mesmo após a conclusão do ensino médio sempre se dispos a me ajudar.

Ao meu orientador, Jorge Alves, pela paciência na orientação e incentivo que tornaram possível a conclusão desta monografia. Agradeço por transmitir seus conhecimentos e por fazer da minha monografia uma experiência positiva.

Aos professores Aluizio Freire e Maria de Jesus, por terem aceitado participar como membros na banca dessa monografia. Tenho certeza que suas considerações são de grande valor para esse trabalho.

Aos mestres desta instituição que tão bem me instruíram com tão rico conhecimento

e muito me acrescentaram com ímpar dedicação e amizade.

À Universidade Federal de Campina Grande, pela oportunidade de concretizar a Licenciatura em Matemática. A essa instituição, devo minha vida acadêmica e meu crescimento intelectual, cultural e político.

A todos que participaram direta ou indiretamente para a minha formação, a todos vocês meus agradecimentos com todo amor e carinho.

“A vida não é sobre metas, conquistas e linhas de chegada. É sobre quem você se torna durante a caminhada”.

Conteúdo

Lista de Figuras

Lista de Tabelas

Resumo

Neste trabalho, vislumbramos a conscientização dos futuros docentes dos Cursos de Licenciatura em Matemática para aplicabilidade das funções polinomiais de primeiro e segundo grau, construindo-se um elo com a ferramenta estatística análise de regressão, tendo como exemplos problemas do cotidiano, podendo contribuir significativamente para a abrangência e profundidade de uma prática de ensino interdisciplinar. Inicialmente se explanou as funções de primeiro e segundo grau, que são elementos chave para descrever o mundo real em termos matemáticos. Tais conceitos estão diretamente associados a regressão, por interessar muitas vezes saber o comportamento de duas variáveis, outras vezes estudar como uma variável varia em função de outra, neste cenário a regressão Linear constitui uma tentativa de estabelecer uma equação matemática que descreva o relacionamento entre essas duas variáveis. Para validação implementou-se uma análise no software estatístico R, onde, os resultados foram considerados satisfatórios no contexto avaliado.

Palavras-chave: Função, Regressaõ Linear, Software R.

Abstract

In this work, we envisage the awareness of the future teachers of the Mathematics Degree Courses for the applicability of polynomial functions of first and second degree, building a link with the statistical tool regression analysis, having as examples daily problems and can contribute significantly to the breadth and depth of an interdisciplinary teaching practice.First and second degree functions were explained, which are key elements to describe the real world in mathematical terms.These concepts are directly associated to regression, since it is often interesting to know the behavior of two variables, other times to study how one variable varies with another, in this scenario linear regression is an attempt to establish a mathematical equation that describes the relationship between these two variables.For validation an analysis was implemented in the R statistical software, where the results were considered satisfactory in the context evaluated.

Keywords: Function, Linear Regression, Software R.

Capítulo 1

Introdução

Muitas leis físicas e muitos princípios de engenharia descrevem uma quantidade dependente de outra. Em 1673, essa idéia foi formalizada por Leibniz, que cunhou o termo *função* para indicar a dependência de uma quantidade em relação a outra [7].

Os estudos dos diferentes fenômenos da natureza e a resolução dos diversos problemas técnicos e, por conseguinte, matemáticos, levam-nos a considerar a variação de uma grandeza em correlação com a variação de outra grandeza. Assim quando estudamos um movimento, consideramos o caminho percorrido como uma variável que depende do tempo [10].

Na metade do século XVIII, o matemático suiço Leohnard Euler (pronuncia-se "oiler") concebeu a idéia de denotar funções pelas letras do alfabeto, tornando possível, desse modo, trabalhar com funções sem apresentar fórmulas específicas, gráficos ou tabelas. Para entender a idéia de Euler, pense numa função como sendo um programa de computador que toma uma *entrada* x, opera com ela de alguma forma e produz exatamente uma *saída* y. O programa de computador é um objeto por si só, assim podemos dar-lhe um nome, digamos f. Dessa forma, a função f (o programa de computador) associa uma única saída y a cada entrada x [7].

Para pares de observações coletadas sobre uma série de indivíduos ou objetos, uma correlação positiva entre duas variáveis medidas implica que altos valores em uma variável tendem a estar associados a altos valores em outra variável, e que baixos valores em uma variável tendem estar associados a baixos valores em outra variável. De modo semelhante, se as duas variáveis estiverem correlacionadas negativamente, existe uma relação inversa entre essas duas variáveis [2].

Para entendermos essa relação faremos uso da Estátistica, que é a ciência que fornece os príncipios e os métodos para coleta, organização, resumo, análise e interpretação de dados [18].

Em alguma fase do seu trabalho, o pesquisador depara com o problema de analisar e

entender um conjunto de dados relevante ao seu projeto particular ou objeto de estudos. Ele necessitará trabalhar os dados para transforma-los em informações, para compará-los com outros resultados, ou ainda para julgar a adequação a alguma teoria [3].

O uso da relação entre duas variáveis para prever o valor de uma a partir de outra pode ser formalizado de modo a fornecer as melhores previsões possíveis. Além disso, e talvez não menos importante do que isso, essa metodologia pode ser usada para explicar a variação em uma variável como consequência de sua relação com uma ou com outras variáveis. Regressão Linear Simples é um modelo usado para gerar essas previsões e explicações [2].

A estatística é um dos ramos da matemática que nos proporciona sair um pouco da busca pela exatidão, possibilitando mensurar probabilidades dos acontecimentos, fato este que nos atrai no intuito de melhorar nossa capacidade de compreensão de mundo.

Na busca por uma ferramenta que possa aliar conhecimentos matemáticos a interpretação de fatos, encontramos na regressão linear simples junto as aplicações no software R, uma combinação que viabiliza a síntetise de informações colhidas no cotidiano.

Diante do exposto, o objetivo geral deste trabalho é mostrar a aplicabilidade de funções linear e quadrática na estatística através de regressão linear simples com o auxílio do *software* estatístico R.

Este trabalho está organizado da seguinte maneira:

Capítulo 2: Realizamos uma breve explanação sobre funções, que são elementos-chave para descrever o mundo real em termos matemáticos, fixando conceitos e detalhando funções linear e quadrática.

Capítulo 3: Estudamos Regressão Linear Simples e seus principais pontos.

Capítulo 4: Faremos a análise de dados com implementação dos resultados e discursões com o auxílio do software R.

Capítulo 5: Expomos uma síntese geral do nosso trabalho com explanações contextualizadas e embasadas nas ideias gerais contidas neste material.

Capítulo 2

FUNÇÕES

Um dos conceitos mais fundamentais da matemática é o de função, podendo ser representada por uma equação, uma tabela numérica, um gráfico, ou ainda, ser descrita verbalmente.

Muitas vezes ocorre na prática que o valor de uma quantidade depende do valor de outra. Exemplificando, o salário de uma pessoa pode depender do número de horas trabalhadas; a produção total de uma fábrica pode depender do número de máquinas usadas; a distância percorrida por um objeto pode depender do tempo decorrido desde que ele deixou um dado ponto; o volume do espaço ocupado por um gás sob uma determinada pressão constante depende da temperatura do gás; a resistência de um fio elétrico com comprimento fixo depende de seu diâmetro, e assim por diante. A relação entre tais quantidades é dada frequêntemente por uma *função* [8].

Na Matemática, como em outras ciências, muitas vezes queremos estabelecer uma relação ou correspondência entre dois conjuntos. Suponhamos, por exemplo, que temos dois conjuntos: um conjunto de números, $A = \{1,2,3,4\}$, e um conjunto de quatro pessoas, $B = \{\text{Ari, Rui, Lina, Ester}\}$. Uma relação de A em B pode ser aquela que ao número 1 associa o nome de Ari, ao 2 associa Ester, ao 3 associa Lina e ao 4, Rui. Esquematicamente, usamos a seguinte representação denominada diagrama de flechas Figura 2.1.

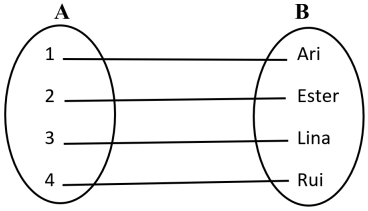

Figura 2.1: Relação entre A e B.

Ou seja, os números em ordem crescente associamos os nomes em ordem alfabética. Outra maneira de representar seria utilizando a notação de pares ordenados:

$$(1,\text{Ari}), \quad (2,\text{Ester}), \quad (3,\text{Lima}), \quad (4,\text{Rui}).$$

Notemos que a correspondência estabelecida determina um conjunto de pares ordenados, que chamaremos:

$$M = \{(1,\text{Ari}),(2,\text{Ester}),(3,\text{Lima}),(4,\text{Rui})\}.$$

É claro que esta não é a ùnica relação que se pode ser estabelecida entre os conjuntos A e B.

Vejamos outros exemplos. Façamos corresponder ao número 1 os indivíduos do sexo masculino e, ao número 2, os indivíduos do sexo feminino. Temos o diagrama Figura 2.2, constituindo o conjunto:

$$N = \{(1,\text{Rui}),(1,\text{Ari}),(2,\text{Ester}),(4,\text{Rui})\}$$

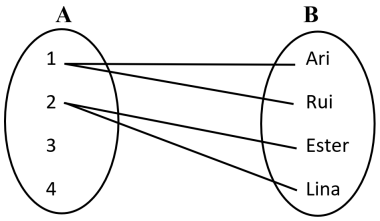

Figura 2.2: Relação entre A e B.

Uma terceira relação que podemos considerar é aquela que associa aos números ímpares o nome Ari e aos pares o nome Lina. Teremos o diagrama da Figura 2.3, constituindo o conjunto:

$$P = \{(1,\text{Ari}),(2,\text{Lina}),(3,\text{Ari}),(4,\text{Lina}).$$

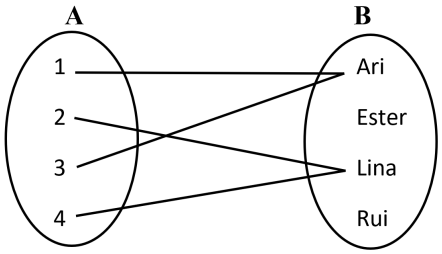

Figura 2.3: Relação entre A e B.

Notemos que os conjuntos M,N e P são formados por pares ordenados cujos primeiros elementos pertencem a A e cujos segundos elementos pertencem a B. Ou seja, todos são subconjuntos do produto cartesiano, isto é, fazem parte da totalidade dos pares ordenados de A por B, logo:

$$M \subset A \text{ x } B, N \subset A \text{ x } B, \text{e } P \subset A \text{ x } B.$$

É possível determinar outras relações de A em B, mas todas serão subconjuntos de A x B.Como A x B tem 16 elementos , e o número de subconjuntos de A x B é 2^{16}, podemos estabelecer, ao todo, 2^{16} relações de A em B. Assim, temos a seguinte definição formal:

S é uma relação de A em B se S for um subconjunto de A x B.

Quando os conjuntos A e B são numéricos, as relações são formadas por pares ordenados de números. Um par ordenado de números reais pode ser representado geometricamente por meio de dois eixos perpendiculares, sendo o horizontal chamado de eixo das abscissas, ou eixo x; e o vertical, de eixo das ordenadas ou eixo y.

Um par ordenado (a,b) pode ser representado colocando-se a no eixo x, e b no eixo y, e traçando-se uma vertical por a e uma horizontal por b. O ponto P de intersecção dessas duas retas é representação do par (a,b), conforme a Figura 2.4.

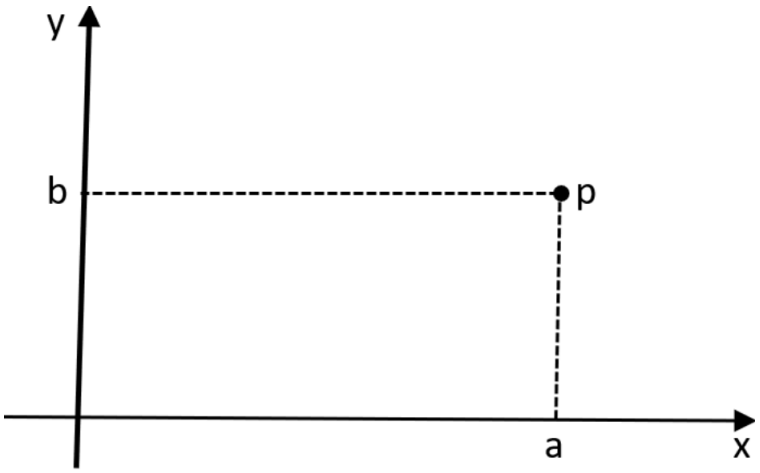

Figura 2.4: Representação geométrica do par ordenado (a, b).

Definida uma relação S de A em B, podemos considerar dois novos conjuntos: o domínio da relação $D(S)$ e o conjunto imagem da relação $Im(S)$.

O domínio de S é o conjunto dos elementos $x \in A$ para os quais existe um $y \in B$ tal que $(x, y) \in S$. O conjunto imagem de S é o conjunto dos $y \in B$ para os quais existe um $x \in A$ tal que $(x, y) \in S$. Em outras palavras, o domínio é o conjunto dos elementos de A que possuem um correspondente de B dado pela relação.

É claro que $D(S)$ é um subconjunto de A e $Im(S)$ é um subconjunto de B. Quando não houver possibilidade de confusão, o domínio e o conjunto imagem são indicados simplesmente por D e IM, respectivamente.

Definição 1. Seja D um dado conjunto de números reais. Uma *função* f definida em D é uma regra, ou lei de correnpondência, que atribui um único número real y a cada valor x de D. O conjunto D dos valores permitidos para x chama-se *domínio* (ou *domínio de definição*) da função, e o conjunto dos valores correspondentes de y chama-se *imagem*. O número y, que é especificado para x pela função f, escreve-se usualmente $f(x)$, de modo que $y = f(x)$, e chama-se *valor* de f em x [12].

Costumamos chamar x de *variável independente* (ou *variável aleatória*), por que ela é livre para assumir qualquer valor do domínio, e chamar y de *variável dependente*, por que seu valor numérico depende da escolha de x.

2.1 O Conceito de Função

Consideremos os seguintes diagramas de flecha que representam relações de A em B Figura 2.5:

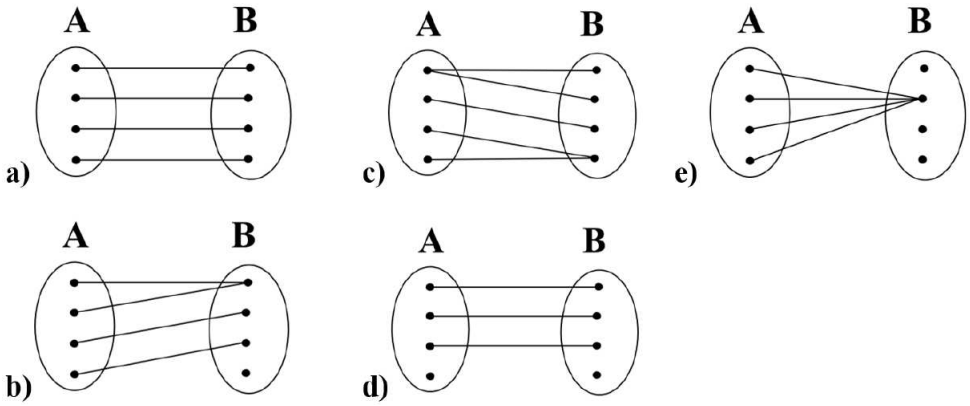

Figura 2.5: Relações entre A e B.

Observe que, entre essas relações, têm em particular importância aquelas que obedecem à definição seguinte:

Uma relação f de A em B é uma função se e somente se:

a) Todo elemento x pertencente a A tem um correspondente y pertencente a B definido pela relação, chamado imagem de x.

b) A cada x pertencente a A não podem corresponder dois ou mais elementos de B por meio de f.

c) Verificamos que as relações (a), (b) e (e) da figura 2.5 são funções de A em B, ao passo de (c) não é função, pois a um dado x pertencente a A correspondem dois elementos em B; d) também não é função, pois existe um elemento de A que não tem corrrespondente em B.

2.2 Funções Reais de uma Variável Real

Se f é uma função com domínio em A e imagem em B, dizemos que f é uma função definida em A com valores em B. Se tanto A como B forem subconjuntos dos reais, dizemos que f é uma função real de variável real [9].

2.3 Normas Elementares para o Estudo de uma Função

2.3.1 Domínio

Quando temos uma função real de uma variável real, de A em B, sabemos que A é um subconjunto dos números reais. Nos exemplos dados anteriormente, a função era definida por uma sentença $y = f(x)$, e os conjuntos A e B eram especificados.

Nas situações em que não é mencionado o domínio, convenciona-se que ele seja formado por todos os valores reais de x para os quais existam imagem de y.

Observemos que em funções envolvendo situações práticas, o domínio é constituído de todos os valores reais de x para os quais tenha significado o cálculo da imagem. Assim, por exemplo, caso tenhamos uma função custo $C(x) = a + bx$, os valores de x não podem ser negativos (não podem ter quantidades negativas). Além disso, caso o produto seja indivisível (por exemplo, quando x é a quantidade de carros), o domínio é constituido apenas de números inteiros não negativos.

2.3.2 Gráfico de Uma Função

Definição 2. Seja f uma função. O gráfico de f é o conjunto de todos os pontos $(x, f(x))$ de um plano coordenado, onde x pertence ao domínio de f [4].

2.3.3 Interceptos

Definição 3. São os pontos de intersecção do gráfico de uma função com os eixos. Os pontos de intersecção com o eixo x têm coordenadas do tipo $(x, 0)$ e são chamados x-interceptos. Os pontos de intersecção com o eixo y têm coordenadas do tipo $(0, y)$ e são chamados de y-interceptos [9].

2.3.4 Funções Crescente e Decrescente

Definição 4. Dizemos que uma função é *crescente* num intervalo [a,b] se à medida que aumenta o valor de x, dentro do intervalo, as imagens correspondentes também aumentam. Em outras palavras, f é crescente num intervalo [a,b] se para quaisquer valores x_1 e x_2 do intervalo, com $x_1 < x_2$, tivermos $f(x_1) < f(x_2)$ [9].

Definição 5. Analogamente, dizemos que uma função f é *decrescente* num intervalo [a,b] se à medida que aumenta o valor de x, dentro do intervalo, as imagens correspondentes vão diminuido. Em outras palavras, f é decrescente num intervalo [a,b] se para quaisquer valores x_1 e x_2 do intervalo, com $x_1 < x_2$, tivermos $f(x_1) > f(x_2)$ [9].

A Figura 2.6 a seguir ilustra essas duas situações.

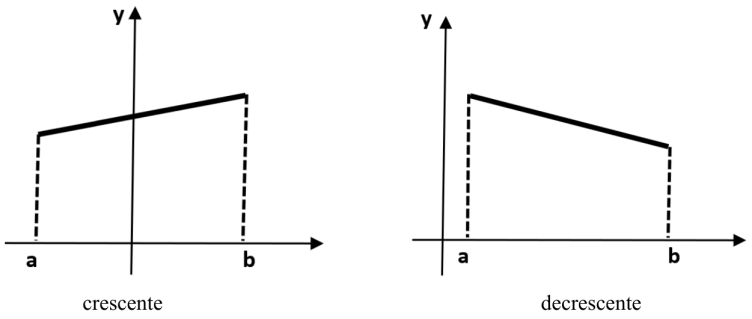

Figura 2.6: Funções crescente e decrescente.

Caso a função tenha a mesma imagem em todos os pontos de um intervalo [a,b], dizemos que a função é constante naquele intervalo.

Uma função que seja crescente ou constante num intervalo é chamada não decrescente naquele intervalo; se uma função for constante ou descrescente num intervalo ela é chamada não crescente naquele intervalo. A Figura 2.7 ilustra funções não decrescente e não crescente.

Figura 2.7: Funções não decrescente e não crescente.

2.3.5 Pontos de Máximo e Mínimo

Seja f uma função definida num domínio D. Dizemos que x_0 é um *ponto de máximo relativo* (ou simplesmente *ponto de máximo*) se existe um intervalo aberto A, com centro x_0 tal que:

$$f(x) \leq f(x_0) \qquad \forall\, x \in A \cap D.$$

Em outras palavras, x_0 é um ponto de máximo relativo se as imagens de todos os valores de x pertencentes ao domínio, situados num intervalo centrado em x_0, forem menores ou iguais à imagem de x_0. A imagem $f(x_0)$ é chamada de valor máximo de f.

Analogamente dizemos que x_0 é um *ponto de mínimo relativo* (ou simplesmente *ponto de mínimo*) se existir um intervalo aberto A, com centro em x_0, tal que:

$$f(x) \geq f(x_0) \qquad \forall\, x \in A \cap D.$$

Em outras palavras, x_0 é um ponto de mínimo relativo se as imagens de todos os valores de x pertencentes ao domínio, situados num intervalo centrado em x_0, forem menores ou iguais à imagem de x_0. A imagem $f(x_0)$ é chamada de valor mínimo de f.

Assim, por exemplo, a função definida no intervalo [a,b] e representada no gráfico a seguir, teremos:

Pontos de máximo:a, x_2, x_4.

Pontos de mínimo:x_1, x_3, b.

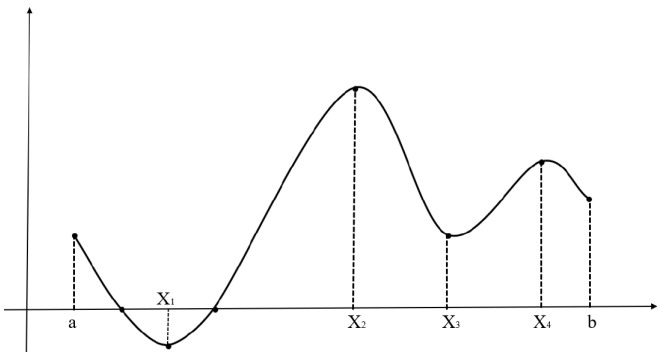

Figura 2.8: Ilustração de Pontos de Máximo e Mínimo.

Por outro lado, dizemos que x_0 é um ponto de máximo absoluto se

$$f(x) \leq f(x_0) \qquad \forall\, x \in D,$$

e x_0 é um ponto de mínimo absoluto se

$$f(x) \geq f(x_0) \qquad \forall\, x \in D.$$

Portanto, a diferença entre um ponto de máximo relativo e máximo absoluto é que o primeiro é um conceito vinculado às vizinhanças do ponto considerado, ao passo que o segundo é ligado a todo o domínio da função. A mesma diferença ocorre entre ponto de mínimo relativo e mínimo absoluto.

Na função representada na Figura 2.8, x_2 é o ponto de máximo absoluto, e x_1 é o ponto de mínimo absoluto.

2.3.6 Estudo do Sinal de uma Função

Estudar o sinal de uma função significa obter valores de x para os quais $y > 0$ ou $y < 0$ ou $y = 0$.

Desse modo, por exemplo, na função definida no intervalo [2,10] e representada na Figura 2.9 a seguir, teremos:

- $y > 0$ para $2 \leq x < 3$ ou para $7 < x \leq 10$;
- $y < 0$ para $3 < x < 7$;
- $y = 0$ para $x = 3$ ou $x = 7$.

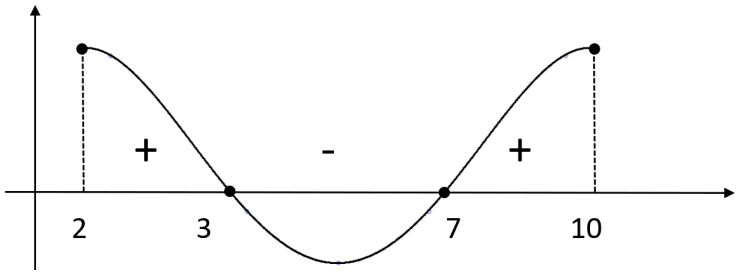

Figura 2.9: Ilustração e representação simbólica do sinal de uma função.

2.4 Função Constante

É toda função do tipo $y = k$, em que k é uma constante real. Verifica-se que o gráfico dessa função é uma reta horizontal, passando pelo ponto de ordenada k.

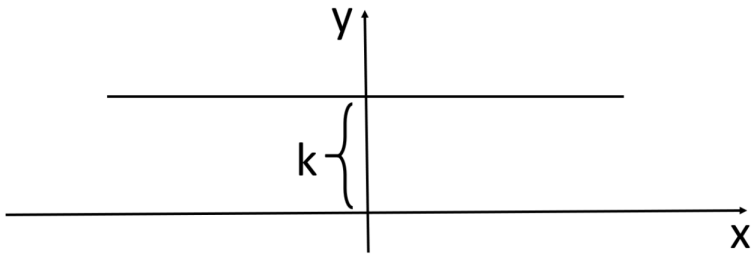

Figura 2.10: Gráfico da função constante $y = k$.

2.5 Função Linear (ou Função do 1º Grau)

Esse tipo de função apresenta um grande número de aplicações.

Definição 6. Uma função é chamada de função linear, (função do 1º grau ou função afim) se sua sentença for dada por $y = m \cdot x + n$, sendo m e n constantes reais com $m \neq 0$.

Verifica-se que o gráfico de uma função do 1º grau é uma reta. Assim, o gráfico pode ser obtido por meio de dois pontos distintos (pois dois pontos distintos determinam uma reta).

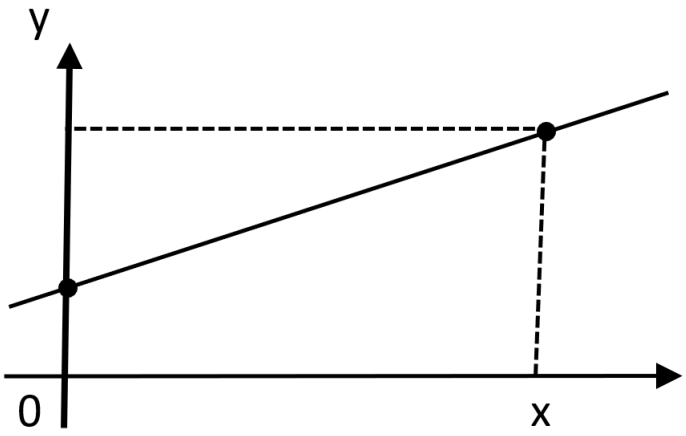

Figura 2.11: Gráfico da função $y = mx + n$.

2.5.1 Observações

1) A constante n é chamada de *coeficiente linear* e representa, no gráfico, a ordenada do ponto de intersecção da reta com o eixo y, como mostra a Figura 2.12. A justificativa para essa afirmação é feita lembrando que, no ponto de intersecção do gráfico da função com o eixo y, a abscissa x vale zero; assim, o ponto de iterseção é da forma $(0, y)$, e, como ele pertence também ao gráfico da função, podemos substituir x por 0 na função $y = m \cdot x + n$. Teremos então:

$$y = m \cdot 0 + n \Rightarrow y = n$$

Portanto o ponto de intersecção do gráfico com o eixo y tem ordenada n.

2) A constante m é chamada de coeficiente angular e representa a variação de y correspondente a um aumento do valor de x igual a 1, aumento esse considerado a partir de qualquer ponto de reta, quando $m < 0$, o gráfico correspondente a uma função decrescente, como podemos observar na Figura 2.12.

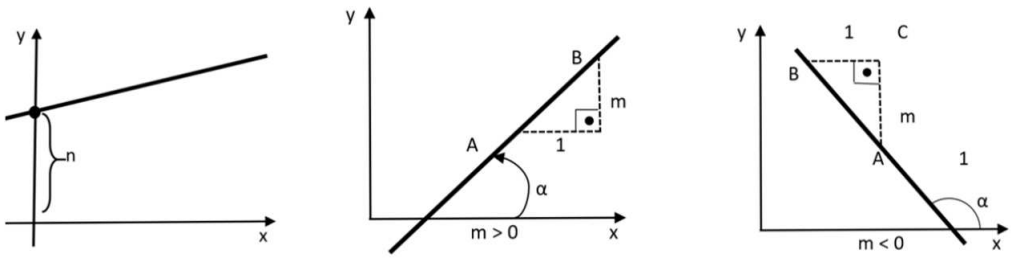

Figura 2.12: Coeficiente linear e angular de uma reta.

A demosntração dessa propriedade é a seguinte.

Seja x_1 a abscissa de um ponto qualquer da reta e seja $x_2 = x_1 + 1$. Sejam y_1 e y_2 as ordenadas dos pontos da reta correspondentes àquelas abscissas. Teremos

$$y_1 = m \cdot x_1 + n \tag{2.1}$$

e

$$y_2 = m \cdot x_2 + n. \tag{2.2}$$

Subtraindo membro a membro as relações (2.2) e (2.1), e tendo em conta que $x_2 = x_1 + 1$, obteremos

$$y_2 - y_1 = m(x_2 - x_1) \Rightarrow y_2 - y_1 = m.$$

Assim, m pode corresponder à variação de y correspondente a uma variação de x igual a 1. Notemos ainda que, se $m > 0$, teremos $y_2 > y_1$ e consequentemente a função será crescente. Por outro lado, se $m < 0$ então $y_2 < y_1$ e consequentemente a função será decrescente. É fácil verificar no triângulo ABC da Figura 2.12 que $m = tg\alpha$, em que α é o ângulo de inclinação da reta em relação ao eixo x.

3) Conhecendo-se dois pontos de uma reta $A(x_1, y_1)$ e $B(x_2, y_2)$, o coeficiente angular m é dado por

$$m = \frac{y_2 - y_1}{x_2 - x_1} = \frac{\Delta y}{\Delta x} \tag{2.3}$$

A demosntração (2.3) é feita considerando-se o triângulo ABC da Figura 2.13.

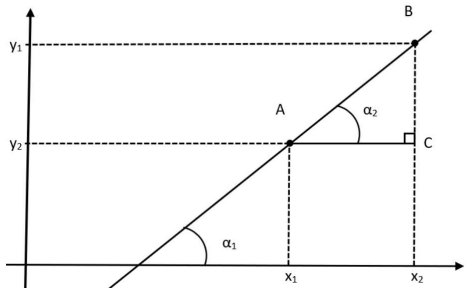

Figura 2.13: Interpretação do coeficiente angular.

Temos:

$$tg\alpha_2 = \frac{\overline{BC}}{\overline{AC}} = \frac{y_2 - y_1}{x_2 - y_1}.$$

Como $\alpha_2 = \alpha_1$, então $tg\alpha_2 = tg\alpha_1$, segue que $m = \frac{y_2 - y_1}{x_2 - x_1}$. A demonstração é análoga se na Figura 2.13 consideremos uma reta de uma função decrescente.

4) Conhecendo um ponto $P(x_0, y_0)$ de uma reta e seu coeficiente angular m, a função correspondente é dada por

$$y - y_0 = m(x - x_0). \tag{2.4}$$

De fato, seja $Q(x, y)$ um ponto genérico da reta, distinto de P como na (Figura 2.14).

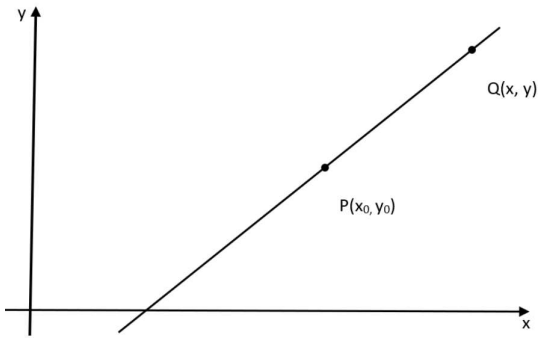

Figura 2.14: Determinação da reta por um ponto e pelo coeficiente angular.

Teremos

$$m = \frac{y-y_0}{x-x_0} \Rightarrow y - y_0 = m(x - x_0)$$

e obtemos (2.4).

2.6 Função Quadrática (ou Função do 2º Grau)

Definição 7. Função quadrática é toda função do tipo $y = ax^2 + bx + c$, em que a, b e c são constantes reais com $a \neq 0$.

O gráfico desse tipo de função é uma curva chamada parábola. A concavidade é voltada para cima se $a > 0$, e voltando para baixo se $a < 0$ Figura 2.15.

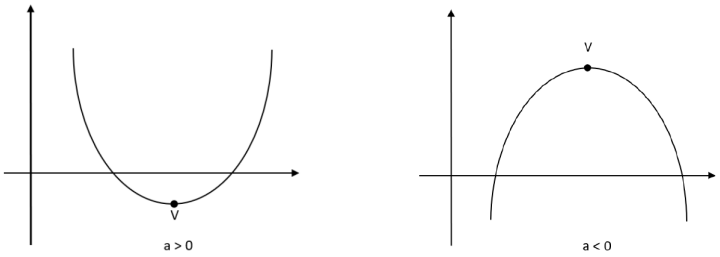

Figura 2.15: Gráfico da função quadrática.

O ponto V da parábola, na Figura 2.15, é chamado *vértice*. Se $a > 0$, a abscissa do vértice é um ponto de mínimo; se $a < 0$, a abscissa de vértice é um ponto de máximo.

Os eventuais pontos de intersecção da parábola com eixo x são obtidos fazendo $y = 0$.Teremos a equação $ax^2 + bx + c = 0$.

Se a equação tiver duas raízes reais distintas ($\Delta > 0$), a parábola interceptará o eixo x em dois pontos distintos; se a equação tiver uma únca raiz real ($\Delta = 0$), a parábola interceptará o eixo x em um único ponto; finalmente, se a equação não tiver raizes reais ($\Delta < 0$), a parábola não interceptará o eixo x. Observe a Figura 2.16.

A intersecção com o eixo y é obtida fazendo-se $x = 0$. Portanto:

$$x = 0 \Rightarrow y = a \cdot 0^2 + b \cdot 0 + c \Rightarrow y = c,$$

ou seja, o ponto de intesecção da parábola com o eixo y é (0, c).

Com relação ao vértice da parábola, indicando por x_v, e y_v a abscissa e a ordenada do vértice, respectivamente, teremos:

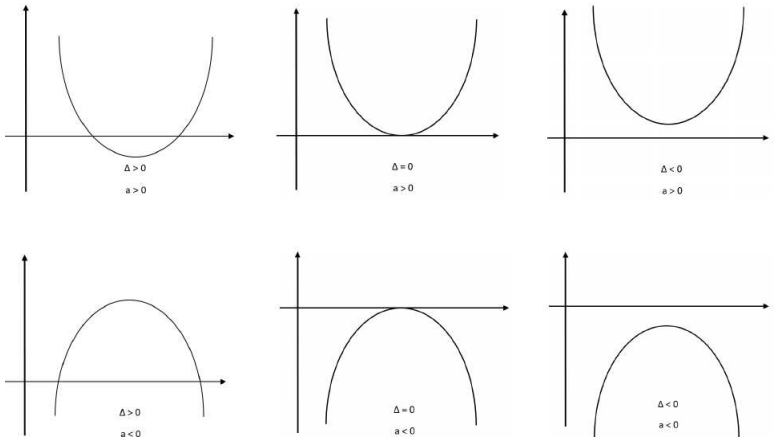

Figura 2.16: Funções quadráticas.

$$x_v = \frac{-b}{2a}, \qquad \text{e} \qquad y_v = f(x_v) = \frac{-\Delta}{4a}.$$

Para demonstrarmos essas relações, vamos proceder da seguinte forma; seja

$$y = ax^2 + bx + c,$$

logo

$$y = a\left(x^2 + \frac{b}{a}x\right) + c.$$

Dentro do parênteses, vamos adicionar $\frac{b^2}{4a^2}$ e, para compensar, vamos subtrair de c o valor $\frac{b^2}{4a^2}$ (Note que o termo adicionado dentro dos parêntesses está multiplicado por a). Dessa forma teremos:

$$y = a\left(x^2 + \frac{b}{a}x + \frac{b^2}{4a^2}\right) + c - \frac{b^2}{4a}.$$

Ou seja, $y = a\left(x + \frac{b}{2a}\right)^2 + c - \frac{b^2}{4a}$, pois o termo entre parênteses é um trinômio quadrado perfeito.

Como o termo entre parênteses é um quadrado, será sempre maior ou igual a zero. Assim:

- Se $a > 0$, a concavidade será para cima, e o ponto de mínimo será sempre aquele para o qual a expressão entre parênteses dá zero, ou seja $x = \frac{-b}{2a}$, e esta é a abscissa do vértice.

- Se $a < 0$, a concavidade será para baixo, e o ponto de máximo será aquele para o qual a expressão entre parênteses dá zero, ou seja $x = \frac{-b}{2a}$, e esta é a abscissa do vértice.

Assim, em qualquer caso, a abscissa do vértice será $x_v = \frac{-b}{2a}$. A justificativa de qualquer $y_v = f(x_v)$ é imediata, pois a ordenada do vértice é a imagem da abscissa do vértice.

Capítulo 3

REGRESSÃO LINEAR SIMPLES

Sempre é interessante conhecer os efeitos que algumas variáveis exercem, ou que parecem exercer, sobre outras. Mesmo que não exista relação causal entre as variáveis podemos relaciona-las por meio de uma expressão matemática, que pode ser útil para se estimar o valor de uma das variáveis quando conhecemos os valores das outras (estas de mais fácil obtenção ou antecessoras da primeira no tempo), sob determinadas condições [6].

Para tais estudos dispomos da análise de regressão, que é uma técnica de modelagem utilizada para analisar a relação entre uma varável dependente Y, e uma ou mais variáveis independentes $X_1, X_2, X_3, ..., X_n$. O objetivo dessa técnica é identificar (estimar) uma função que descreve, o mais próximo possível, a relação entre essas variáveis e assim podermos predizer o valor que a variável dependente Y irá assumir para um determinado valor da variável aleatória X. Este modelo, designado por modelo de **regressão linear simples**, define uma relação linear entre as variáveis em questão.

3.1 Conceito

Vamos conceituar Regressão Linear Simples [3]. Para posteriores aplicações e análise de dados.

Definição 8. Sejam X e Y duas variáveis aleatórias quantitativas, onde a **esperança condicional** de Y, dado que $X = x$, denotada por $E(Y|x)$, é uma função de x, ou seja,

$$E(Y|x) = \mu(x). \tag{3.1}$$

Uma definição similar vale para $E(X|y)$, que será uma função de y. Vamos considerar o caso em que X e Y são definidas sobre uma mesma população P.

Entende-se por população o conjunto de elementos que têm, em comum, determinada característica [17].

Por exemplo, X pode ser idade e Y o tempo de reação ao estímulo, a qual a relação entre as duas variáveis modelamos por,

$$y_{ij} = \mu_i + e_{ij}, \quad i = 1, ..., 5, \quad j = 1, ..., 4, \tag{3.2}$$

onde μ_i é a média do grupo de idade i.

Podemos pensar que o fator idade determina cinco subpopulações (ou estratos) em P e de lá escolhemos cinco amostras aleatórias de tamanhos $n_i = 4, i = 1, ..., 5$.

Todo subconjunto não vazio e com menor número de elementos do que a população constitui uma amostra [17].

Em (3.1), $\mu(x)$ pode ser qualquer função de x. Tanto X (idade) como Y (tempo de resposta ao estímulo) são variáveis aleatórias contínuas (variáveis para as quais possíveis valores pertencem a um intervalo de números reais), e podemos pensar em introduzir um modelo alternativo para y_{ij}, dada a relação X e Y. Observando as médias de Y, segundo os grupo de idade, ou seja, $E(Y|x)$, percebemos que estas aumentam conforme as pessoas envelhecem. A Figura 3.1 mostra os dados observados, onde notamos uma tendência crescente bem como os valores repetidos de Y para cada nível de idade x.

Um modelo razoável para $E(Y|x)$ pode ser

$$E(Y|x) = \mu(x) = \alpha + \beta x, \tag{3.3}$$

ou seja, o tempo médio de reação é uma função linear da idade.

O comportamento conjunto de duas variáveis quantitativas pode ser observado através de um gráfico, denominado gráfico de dispersão [17].

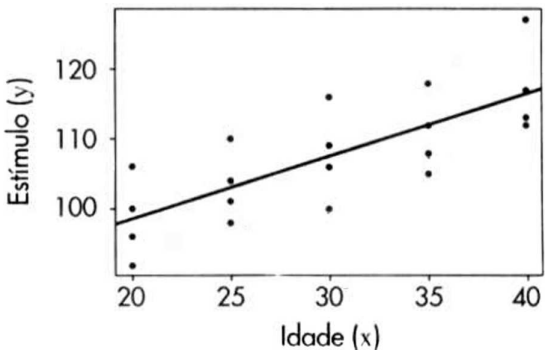

Figura 3.1: Gráfico de dispersão de idade e reação ao estímulo, com reta ajustada.

A forma da função $\mu(x)$ deve ser definida pelo pesquisador, em função do grau de conhecimento teórico que ele tem do fenômeno sob estudo. Um modelo alternativo a (3.2)

seria, então,

$$y_{ij} = \mu(x_i) + e_{ij}, \tag{3.4}$$

como $E(Y|x_i) = \mu(x_i) = \alpha + \beta x_i, i = 1, 2, ..., 5$. Entretanto, a forma usual de escrever o modelo é

$$y_i = \mu(x_i) + e_i, \tag{3.5}$$

onde y_i indica o tempo de reação do i-ésimo indivíduo com x_i anos de idade, e_i é a dispersão dos dados em torno da reta (erro de aproximação), $i = 1, 2, ..., n$, e n é o número total de observações. Teremos então com essa notação, valores repetidos para X, por exemplo, $x_1 = ... = x_4 = 20$. Convém reforçar a idéia que estamos propondo um modelo de comportamento para as médias das subpopulações, logo teremos de estimar os parâmetro envolvidos na função $\mu(x)$, baseados numa amostra de $n = 20$ observações.

No caso de (3.3) o modelo pode ser escrito como

$$y_i = E(Y|x_i) + e_i = \alpha + \beta x_i + e_i, \qquad i = 1, 2, ..., n. \tag{3.6}$$

devendo-se encontrar os valores mais prováveis para α e β, segundo algum critério, a partir de n observações de pares de valores de (X, Y).

Antes de proseguirmos, seria conveniente interpretar os parâmetros envolvidos do modelo (3.3). Sabendo de α, o intercepto, representa o ponto onde a reta corta o eixo das ordenadas, e β, o coeficiente angular, representa o quanto varia a média de Y para um aumento de uma unidade de variável X. Esses parâmetros estão representados de Figura 3.2.

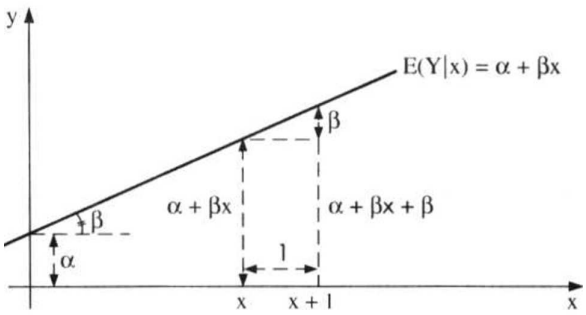

Figura 3.2: Representação do modelo $E(Y|x) = \alpha + \beta x$.

Voltando ao descrito anteriorente, onde X é a idade e Y o tempo de reação, β representa o acréscimo no tempo médio de reação para cada ano de envelhecimento das

pessoas. Aqui α representa o tempo de reação para idade zero (recém-nascido), o que é uma inadequação do modelo.

Observação. Chamamos (3.3) de modelo *linear*, pois este representa uma reta. Todavia, em casos gerais, o termo linear refere-se ao modelo como os *parâmetros entram no modelo*, ou seja, de forma linear. Por exemplo, o modelo

$$E(Y|x) = \alpha + \beta x + \gamma x^2$$

embora graficamente represente uma parábola, é um *modelo linear em* α, β, *e* γ. Por outro lado,

$$E(Y|x) = \alpha e^{\beta x} \tag{3.7}$$

não é um modelo linear em α *e* β.

Vários procedimentos estatísticos são baseados na suposição de que os dados provêm de uma distribuição normal ou então mais ou menos simétrica. Mas, muitas situações de interesse prático, a distribuição dos dados da amostra é assimétrica e pode conter valores atípicos [3].

Determinados modelos não-lineares podem ser transformados em lineares, por meio de transformações das variáveis.

Se quisermos utilizar tais procedimentos, o que se propõe é efetuar uma transformação das observações, de modo a se obter uma distribuição mais simétrica e próxima da normal.

Assim, tomando-se o logaritmo (de base e) em (3.6) obtemos

$$\ln E(Y|x) = \ln(\alpha) + \beta x = \alpha' + \beta x$$

que é linear em α' e β.

Ao lado de um tratamento formal para estudar o modelo (3.6), devemos usar técnicas de análise de dados. Em particular podemos fazer diversos tipos de gráficos *antes* que o modelo seja ajustado, *durante* o processo de ajuste e, finalmente, *depois* que o modelo foi ajustado.

Na Figula 3.1 podemos observar um gráfico que deve ser feito antes de selecionar o modelo. Ou seja, temos um gráfico de dispersão entre as variáveis X (idade) e Y (tempo de reação ao estímulo). Esse tipo de diagrama permite ver qual o tipo de relação existente entre as variáveis, se há valores atípicos, se há valores repitidos, se a variabilidade de Y está aumentando ou não com X etc.

3.2 Estimação dos Parâmetros

Iremos encontrar os estimadores de mínimos quadrados para os parâmetros de modelo linear (3.6), mas o mesmo desenvolvimento pode ser aplicado em modelos mais complexos.

Introduziremos algumas suposições para as variáveis aleatórias envolvidas. A primeira delas é que a variável X é por hipótese controlada e não está sujeita a variações aleatórias. Dizemos que X é uma variável fixa (ou sem erro ou determinística). Segundo, para dado valor x de X, os erros distribuem-se ao redor da média $\alpha + \beta x$ com média zero, isto é

$$E(e_i|x) = 0 \tag{3.8}$$

Em terceiro lugar, devemos supor que os erros tenham a mesma **variabilidade** em torno dos níveis de X, ou seja

$$Var(e_i|x) = \sigma_e^2 \tag{3.9}$$

E em quarto lugar, introduziremos a restrição de que os erros sejam não-correlacionados. Colhida uma amostra de n indivíduos, teremos n pares de valores (x_i, y_i), $i = 1, ..., n$, que devem satisfazer ao modelo (3.6), isto é,

$$y_i = \alpha + \beta x_i + e_i, \qquad i = 1, ..., n. \tag{3.10}$$

Temos então, n equações e $n + 2$ incógnitas $(\alpha, \beta, e_1, e_2, ..., e_n)$. Precisamos introduzir um critério que permita encontrar α e β. Vamos adotar o critério que consiste em encontrar os valores de α e β que minimizam a soma dos erros, dados por

$$e_i = y_i - (\alpha + \beta x_i), \qquad i = 1, ..., n. \tag{3.11}$$

Obtemos, então, a quantidade de informação perdida pelo modelo ou **soma dos quadrados dos erros** (ou desvios)

$$SQ(\alpha, \beta) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \{y_i - (\alpha + \beta x_i)\}^2 \tag{3.12}$$

Para cada valor de α e β teremos um resultado para essa soma de quadrados, e a solução de mínimos quadrados (MQ) é aquela que torna essa soma mínima. Temos, então, o problema de encontrar o mínimo de uma função de duas variáveis, α e β no caso. Derivando em relação a α e β e igualando a zero, observamos que as soluções $\hat{\alpha}$ e $\hat{\beta}$ devem satisfazer

$$n\hat{\alpha} + \hat{\beta} \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i,$$

$$\hat{\alpha} \sum_{i=1}^{n} x_i + \hat{\beta} \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i, \tag{3.13}$$

as quais produzem as soluções

$$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x},$$

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2} \tag{3.14}$$

Substituindo em (3.3), teremos o estimador para a média $\mu(x)$, dado por

$$\widehat{\mu}(x_i) = \widehat{\alpha} + \widehat{\beta}x_i, \qquad i = 1, ..., n. \tag{3.15}$$

que iremos indicar por

$$\widehat{y}_i = \widehat{\alpha} + \widehat{\beta}x_i, \tag{3.16}$$

ou ainda por

$$\widehat{y}_i = \bar{y} - \widehat{\beta}\bar{x} + \widehat{\beta}x_i = \bar{y} + \widehat{\beta}(x_i - \bar{x}), \qquad i = 1, ..., n. \tag{3.17}$$

3.3 Avaliação do Modelo

Nesta seção e nas seguintes estudaremos várias formas de avaliar se o modelo linear postulado é adequado ou não, dadas as suposições que fizemos sobre ele.

3.3.1 Estimador de σ_e^2

Para julgar a vantagem de adoção de um modelo mais complexo (linear ou outro qualquer), vamos usar a estratégia de compará-lo com o modelo mais simples, ou seja,

$$y_i = \mu + e_i, \tag{3.18}$$

A vantagem será sempre medida por meio da diminuição dos erros de previsão, ou ainda, da **variância residual** S_e^2. Para o modelo ajustado (3.16), cada *resíduo* é dado por

$$\widehat{e}_i = y_i - \widehat{y}_i = \widehat{\alpha} - \widehat{\beta}x_i, \tag{3.19}$$

Vários gráficos envolvendo esses resíduos podem ser feitos para avaliar se eles são bons representantes dos verdadeiros e_i desconhecidos, no sentido de que as suposições feitas sobre estes estão satisfeitas.

Quando estes resíduos forem pequenos, temos uma indicação de que o modelo está produzindo bons resultados. Para julgarmos se o resíduo é pequeno ou não, devemos compará-lo com os resíduos do modelo alternativo dados por $y_i - \widehat{y}$. Da dificuldade de

compara-los individualmente, preferimos trabalhar com as respectivas **somas de resíduos quadrático**, dadas por

$$SQTot = \sum_{i=1}^{n} (y_i - \widehat{y})^2 \tag{3.20}$$

e

$$SQRes = \sum_{i=1}^{n} \widehat{e}_i^2 = \sum_{i=1}^{n} (y_i - \widehat{y})^2. \tag{3.21}$$

No entanto, a comparação direta dessas somas de quadrado não nos parece justa, pois o modelo (3.18) tem mais parâmetros do que o modelo (3.19). Vejamos, então, como comparar as **varâncias residuais**. Para o modelo simples (3.19) o **estimador não-viesado** de σ_e^2 é

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \widehat{y})^2 = \frac{SQTot}{n-1} \tag{3.22}$$

Para o modelo (3.2), com I níveis ou subpopulações, o **estimado da variância residual** era

$$S_e^2 = \frac{SQDen}{n-1} = \frac{SQRes}{n-1} \tag{3.23}$$

e I também denota o número de parâmetros desconhecidos do modelo (as médias μ_i). Portanto, de modo geral, pede-se um grau de liberdade para cada parâmetro envolvido no modelo e é natural definir o **estimador** de σ_e^2 num modelo de regressão como sendo

$$S_e^2 = \frac{SQRes}{n-p}, \tag{3.24}$$

onde p é o número de parâmetros do modelo. No caso particular da regressão linear simples, $p = 2$ e

$$S_e^2 = \frac{SQRes}{n-2}, \tag{3.25}$$

será um **estimador não-viesado** de σ_e^2, isto é, $E(S_e^2) = \sigma_e^2$.

3.3.2 Decomposição da Soma de Quadrados

Apesar de passarmos do modelo simples para o modelo de regressão linear, a redução da soma de quadrados é dada por $SQTot - SQRes$. À adoção do segundo modelo e será

indicado por $SQReg$, significando a *soma dos quadrados devida à regressão.* segue-se que

$$SQReg = SQTot - SQRes, \tag{3.26}$$

ou seja,

$$SQTot = SQReg + SQRes, \tag{3.27}$$

Observando a Figura 3.3, notamos que vale a seguinte relação:

$$y_i - \bar{y} = (y_i - \widehat{y}_i) + (\widehat{y}_i - \bar{y}) = \widehat{e}_i + (\widehat{y}_i - \bar{y}), \tag{3.28}$$

Em palavras, o desvio de uma observação em relação à medida pode ser decomposto como o desvio da observação em relação ao valor ajustado pela regressão, mais o desvio do valor ajustado em relação a média.

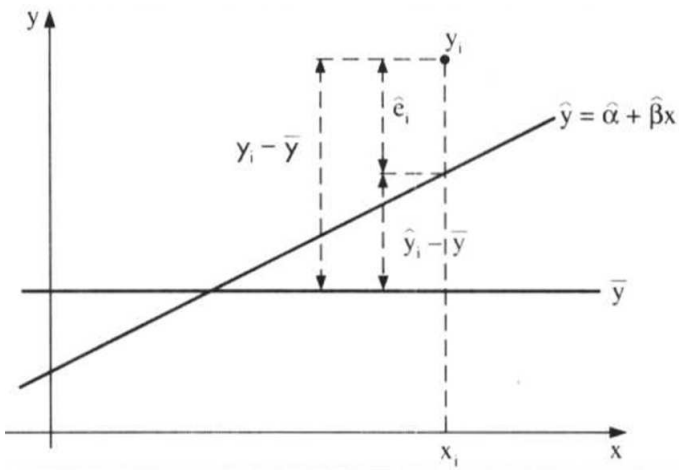

Figura 3.3: Representação gráfica dos diversos desvios.

Elevando-se ao quadrado ambos os membros da igualdade (3.29), tomando-se a soma e observando-se que a soma do duplo produto se anula, obtemos

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}\widehat{e}_i^2. \tag{3.29}$$

ou

$$SQTot = \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2 + SQRes, \tag{3.30}$$

do que deduzimos que

$$SQReg = \sum_{i=1}^{n}(\widehat{y}_i - \bar{y})^2, \tag{3.31}$$

De (3.17) obtemos que

$$\widehat{y}_i - \bar{y} = \widehat{\beta}(x_i - \bar{x})$$

portanto podemos escrever

$$SQReg = \widehat{\beta}^2 \sum_{i=1}^{n}(x_i - \bar{x})^2. \tag{3.32}$$

Daqui se pode observar que quanto maior o valor de $\widehat{\beta}$, maior será a redução da soma dos quadrados dos resíduos.

3.3.3 Tabela de Análise de Variância

Podemos resumir as informações anteriores numa única tabela ANOVA, ilustrada na Tabela 3.1.

Tabela 3.1: Tabela ANOVA para modelo de regressão.

F.V.	g.l.	SQ	QM	F
Regressão	1	SQReg	SQReg=QMReg	QMReg/S_e^2
Resíduo	$n-2$	SQRes	SQRes/$(n-2) = S_e^2$	
Total	$n-1$	SQRTot	SQTot/$(n-1) = S^2$	

Também podemos medir o ajuste do modelo pelo coeficiente de correlação, esta medida varia entre -1 e $+1$, incluive, isto é, $-1 \leq r \leq +1$. Se r assume o valor 1, diz-se que duas variáveis têm *correlação perfeita positiva* e se r assume o valor -1,diz-se que as duas variáveis tem *correlação perfeita negativa*. Se r assume o valor zero, não existe correlação entre as duas variáveis (a correlação é nula) [17].

$$R^2 = \frac{SQReg}{SQTot}. \tag{3.33}$$

$$R = \sqrt{r^2}$$

A estratégia adotada para verificar se compensa ou não utilizar o modelo $y = \alpha + \beta x + e$ é observar a redução no resíduo quando comparado com o modelo $y = \mu + e$. Se a redução for muito pequena, os dois modelos serão praticamente equivalentes, e isso ocorre quando a

inclinação β for zero ou muito pequena, não compensando usar um modelo mais complexo. Estaremos, pois, interessados em testar a hipótese

$$H_0 : \beta = 0. \tag{3.34}$$

o que irá exigir que se coloque uma estrutura de probabilidades sobre os erros. A Figura 3.4 ilustra as duas situações que podem ocorrer. Na Figura 3.4 (a) temos o caso em que claramente a variável auxiliar ajuda a prever a variável resposta. Na situação da Figura 3.4 (b) teremos dúvidas se vale a pena ou não introduzir o modelo mais complexo, ganhamos muito pouco em termos de explicações.

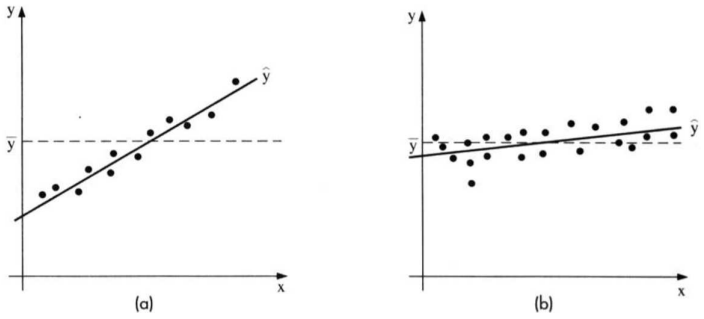

Figura 3.4: Retas ajustadas a dois conjuntos de dados.(a)x explica y;(b)x não explica y.

Para a avaliação final do modelo devemos investigar com mais cuidado o comportamento dos resíduos.

3.4 Propriedades dos Estimadores

Iremos agora estudar as propriedades amostrais dos estimadores $\widehat{\alpha}$ e $\widehat{\beta}$, e para isso é conveniente voltar ao modelo e as suposições adotadas para a variável aleatória Y sob investigação. Lembramos que a variável X é suposta controlada, fixa, e para cada valor x de X teremos associada uma distribuição de probabilidades para Y, como ilustra a Figura 3.5 (a), onde supomos que a dispersão é a mesma para cada nível da variável X. A Figura 3.5 (b) ilustra o caso que será considerado aqui, em que estas distribuições adicionais são normais, como a mesma variância. Note que $E(Y|x)$ é linear, como estamos considerando.

Formalmente, o modelo

$$Y = E(Y|x_i) + e_i = \alpha + \beta x_i + e_i, \quad i = 1, ..., n$$

deve satisfazer as seguintes suposições.

(i) Para cada valor de x_i, o erro e_i tem média zero e variância constante σ_e^2;

(ii) Se $i \neq j$, $Cov(e_i, e_j) = 0$, isto é, para duas observações distintas, os erros são não-correlacionados.

Segue que

$$E(Y_i|x_i) = \alpha + \beta x_i \qquad \text{e} \qquad Var(Y_i|x_i) = \sigma_e^2,$$

e ainda que Y_i e Y_j são não correlacionados, para $i \neq j$.

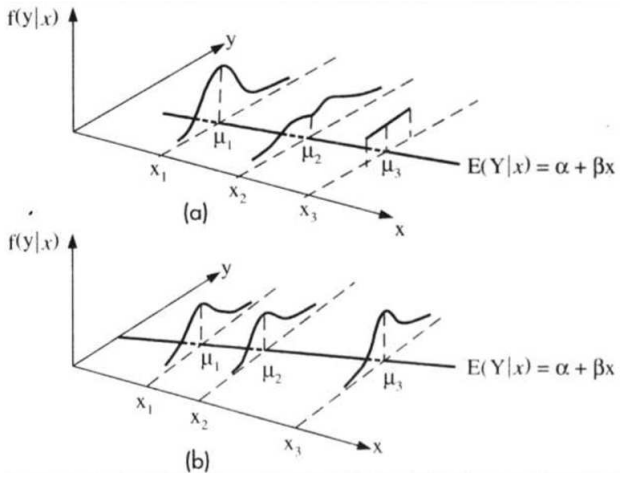

Figura 3.5: (a)médias alinhadas, distribuições com a mesma variâcia; (b)médias alinhadas, distribuições normais com a mesma variância.

3.4.1 Média e Variância dos Estimadores

Vamos obter a média e a variância dos estimadores $\widehat{\alpha}$ e $\widehat{\beta}$, dados em (3.14).

Proposição 1. Para o estimador $\widehat{\beta}$ temos

$$E(\widehat{\beta}) = \beta, \tag{3.35}$$

$$Var(\widehat{\beta}) = \frac{\sigma_e^2}{\sum_{i=1}^n (x_i - \bar{x})^2}, \tag{3.36}$$

Prova. Inicialmente, vamos escrever β de um modo mais conveniente:

$$\widehat{\beta} = \frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i - \bar{Y}\sum_{i=1}^n (x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$
$$= \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = \sum_{i=1}^n \frac{(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} Y_i = \sum_{i=1}^n w_i Y_i$$

onde estamos usando a notação Y (maiúscula) e x (minúscula) para diferenciar o fato de que a primeira está sendo considerada aleatória e a segunda , fixa; e

$$w_i = \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2}, \sum_{i=1}^n w_i = 0.$$

Observe que estamos usando o fato de $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$ e que

$$\begin{aligned}\sum_{i=1}^{n} w_i x_i &= \sum_{i=1}^{n} w_i x_i - \bar{x}\sum_{i=1}^{n} w_i = \sum_{i=1}^{n} w_i(x_i - \bar{x}) \\ &= \sum_{i=1}^{n} \frac{(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}(x_i - \bar{x}) = 1.\end{aligned}$$

Usando a propriedade da esperança e variância de somas de variáveis aleatórias, podemos escrever

$$\begin{aligned}E(\widehat{\beta}) = E(\sum_{i=1}^{n} w_i Y_i) = \sum_{i=1}^{n} w_i E(Y_i) \\ = \sum_{i=1}^{n} w_i(\alpha + \beta x_i) = \alpha\sum_{i=1}^{n} w_i + \beta\sum_{i=1}^{n} w_i x_i = \beta,\end{aligned}$$

o que mostra que o estimador é não-viesado. Para a varância,

$$Var(\widehat{\beta}) = Var(\sum_{i=1}^{n} w_i Y_i) = \sum_{i=1}^{n} w_i^2 Var(Y_i),$$

pois as observações são não-correlacionadas, e, portanto,

$$Var(\widehat{\beta}) = \sum_{i=1}^{n} w_i^2\sigma_e^2 = \sigma_e^2\sum_{i=1}^{n}(\frac{x_i-\bar{x}}{\sum_{i=1}^{n}(x_i-\bar{x})^2})^2 = \sigma_e^2\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{[\sum_{i=1}^{n}(x_i-\bar{x})^2]^2},$$

e o resultado segue.

Proposição 2. Para o estimado $\widehat{\alpha}$ temos:

$$E(\widehat{\alpha}) = \alpha, \tag{3.37}$$

$$Var(\widehat{\alpha}) = \sigma_e^2\frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}, \tag{3.38}$$

Prova. Consideremos os seguintes resultados:

$$Cov(\bar{y}, \widehat{\beta}) = 0, \tag{3.39}$$

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2, \tag{3.40}$$

Como

$$\begin{aligned}\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n}(\alpha + \beta x_i + e_i) \\ = \alpha + \beta\bar{x} + \frac{1}{n}\sum_{i=1}^{n} e_I,\end{aligned}$$

temos que

$$E(\bar{y}) = \alpha + \beta\bar{x} + \frac{1}{n}\sum_{i=1}^{n} E(e_i) = \alpha + \beta\bar{x},$$

dado que x é supostamente fixa e não uma variável aleatória. Também,

$$Var(\bar{y}) = \frac{1}{n^2}\sum_{i=1}^{n} Var(e_i) = \frac{\sigma_e^2}{n}.$$

Temos, então, que

$$E(\widehat{\alpha}) = E(\bar{y} - \widehat{\beta}\bar{x}) = \alpha + \beta\bar{x} - \beta\bar{x} = \alpha,$$

e

$$\begin{aligned} Var(\widehat{\alpha}) = Var(\bar{y} - \widehat{\beta}\bar{x}) &= Var(\bar{y}) + Var(\widehat{\beta}\bar{x}) - 2Cov(\bar{y}, \widehat{\beta}\bar{x}) \\ &= Var(\bar{y}) + x^2 Var(\widehat{\beta}) - 2\bar{x}Cov(\bar{y}, \widehat{\beta}) \end{aligned}$$

3.4.2 Distribuições Amostrais dos Estimadores dos Parâmetros

Para completar o estudo das propriedades dos estimadores, vamos introduzir uma terceira suposição:
(iii) Os erros e_i são variáveis aleatórias com distribuição normal, isto é,

$$e_i \sim N(0; \sigma_e^2), \tag{3.41}$$

o que implica

$$y_i \sim N(\alpha + \beta x_i; \sigma_e^2). \tag{3.42}$$

Como $\widehat{\beta}$ e $\widehat{\alpha}$ são combinações lineares de variáveis aleatórias normais e independentes, temos o seguinte resultado:

Proposição 3. Os estimadores $\widehat{\alpha}$ e $\widehat{\beta}$ têm ambos distribuição normal, com as médias e variâncias dadas pelas Proposições 3.1 e 3.2, isto é,

$$\widehat{\alpha} \sim N\left(\alpha; \frac{\sigma_e^2 \sum x_i^2}{n\sum(x_i - \bar{x})^2}\right), \tag{3.43}$$

$$\widehat{\beta} \sim N\left(\beta; \frac{\sigma_e^2}{\sum(x_i - \bar{x})^2}\right), \tag{3.44}$$

Os resultados acima permitem concluir que

$$\frac{\widehat{\beta} - \beta}{\sigma_e}\sqrt{\sum(x_i - \bar{x})^2} \sim N(0, 1), \tag{3.45}$$

$$\frac{\widehat{\alpha} - \alpha}{\sigma_e}\sqrt{\frac{n\sum(x_i - \bar{x})^2}{\sum x_i^2}} \sim N(0, 1), \tag{3.46}$$

3.4.3 Intervalos de Confiança para α e β

Substituido σ_e por seu estimador S_e em (3.45) e (3.46), sabemos que as estatísticas resultantes terão distribuição t de Student, com $(n-2)$ graus de liberdade, o que permitirá construir **intervalos de confiança** para os parâmetros.

Proposição 4. As estatísticas

$$t(\widehat{\beta}) = \frac{\widehat{\beta} - \beta}{S_e}\sqrt{\sum (x_i - \bar{x})^2} \tag{3.47}$$

e

$$t(\widehat{\alpha}) = \frac{\widehat{\alpha} - \alpha}{S_e}\sqrt{\frac{n\sum (x_i - \bar{x})^2}{\sum x_i^2}} \tag{3.48}$$

têm distribuição t de Student com $(n-2)$ graus de liberdade.

Esses resultados, combinados com os procedimentos de construção de intervalos de confiança, nos leva aos seguintes intervalos para α e β, com γ denotando o coeficiente de confiança e $t_\gamma(n-2)$ denotando o valor obtido, com $(n-2)$ graus de liberdade:

$$IC(\alpha;\gamma) = \widehat{\alpha} \pm t_\gamma(n-2)S_e\sqrt{\frac{\sum x_i^2}{n\sum (x_i - \bar{x})^2}} \tag{3.49}$$

$$IC(\beta;\gamma) = \widehat{\beta} \pm t_\gamma(n-2)S_e\sqrt{\frac{1}{\sum (x_i - \bar{x})^2}} \tag{3.50}$$

Estes intervalos de confiança podem ser usados para testar **hipóteses** do tipo

$$H_0 : \alpha = \alpha_0,$$
$$H_0 : \beta = \beta_0.$$

Em particular, temos o seguinte resultado:

Proposição 5. A estatística para testar $H_0 : \alpha = 0$ é

$$t(\widehat{\alpha}) = \frac{\widehat{\alpha}}{S_e}\sqrt{\frac{n\sum (x_i - \bar{x})^2}{\sum x_i^2}} \tag{3.51}$$

e a estatística para testar $H_0 : \beta = 0$ é

$$t(\widehat{\beta}) = \frac{\widehat{\beta}}{S_e}\sqrt{\sum (x_i - \bar{x})^2} \tag{3.52}$$

cada um tendo a distribuição t Student com $(n-2)$ graus de liberdade.

Observe que

$$[t(\widehat{\beta})]^2 = \frac{\widehat{\beta}\sum(x_i-\bar{x})^2}{S_e^2},$$

e usando o resultado (3.32) podemos escrever

$$[t(\widehat{\beta})]^2 = \frac{SQReg}{S_e^2}, \tag{3.53}$$

que é a estatística F que aparece na tabela ANOVA. Assim, para testar hipótese H_0 : $\beta = 0$, pode-se usar a estística (3.53), que segue uma distribuição $F(1, n-2)$.

Se quiséssemos saber dentro de que intervalo 95% das futuras observações iriam estar, construiríamos o Intervalo de Predição:

3.5 Análise de Resíduos

Para verificar se um modelo é adequado, temos que investigar se as suposições feitas para o desenvolvimento do modelo estão satisfeitas. Para tanto, estudamos o comportamento do modelo usando o conjunto de dados observados, notadamente as discrepâncias entre valores observados e os valores ajustados pelo modelo, ou seja, fazemos uma *análise dos resíduos*.

O i-ésimo resíduo é dado por

$$\widehat{e} = y_i - \widehat{y}_i, \quad i = 1, 2, ..., n. \tag{3.54}$$

Lembramos que já utilizamos este resíduo para obter medidas de qualidade e dos estimadores dos parâmetros do modelo. Agora iremos estudar o comportamento individual e os conjuntos destes resíduos, comparando com as suposições feitas sobre os verdadeiros erros e_i. Existem várias técnicas formais para conduzir esta análise, mas aqui iremos ressaltar basicamente métodos gráficos.

Uma representação gráfica bastante útil é obtida plotando-se pares $(x_i, \widehat{e})$, $i = 1, 2, ..., n$. Outras vezes, é de maior utilidade fazer representação gráfica dos chamados resíduos padronizados.

$$\widehat{z}_i = \frac{y_i - \widehat{y}_i}{S_e} = \frac{\widehat{e}_i}{S_e}, \tag{3.55}$$

plotando-se os pares $(x_i, \widehat{z}_i)$. Observe que a forma dos dois gráficos será semelhante, havendo apenas uma mudança de escala das ordenadas nos dois casos. Por isso, iremos usar a primeira representação, indicando no gráfico a posição do valor S_e.

Outro resíduo usado é o chamado *resíduo estudentizado*, definido por

$$\widehat{r}_i = \frac{\widehat{e}_i}{S_e\sqrt{1-v_{ii}}}, \tag{3.56}$$

onde $v_{ii} = 1/n + (x_i - \widehat{x})^2 / \sum (x_i - \widehat{x})^2$. O denominador de (3.56) é o desvio padrão de $\widehat{e}_i$. Não iremos explorar aqui a análise feita com esse tipo de resíduo.

Obtido o gráfico dos resíduos, precisamos saber como identificar possíveis inadequações. Apresentamos na Figura 3.6 alguns tipos usuais de gráfico de resíduos. A Figura 3.6 (a) é a situação ideal para os resíduos, distribuidos aleatóriamente em torno do zero, sem nenhuma observação muito discrepante.

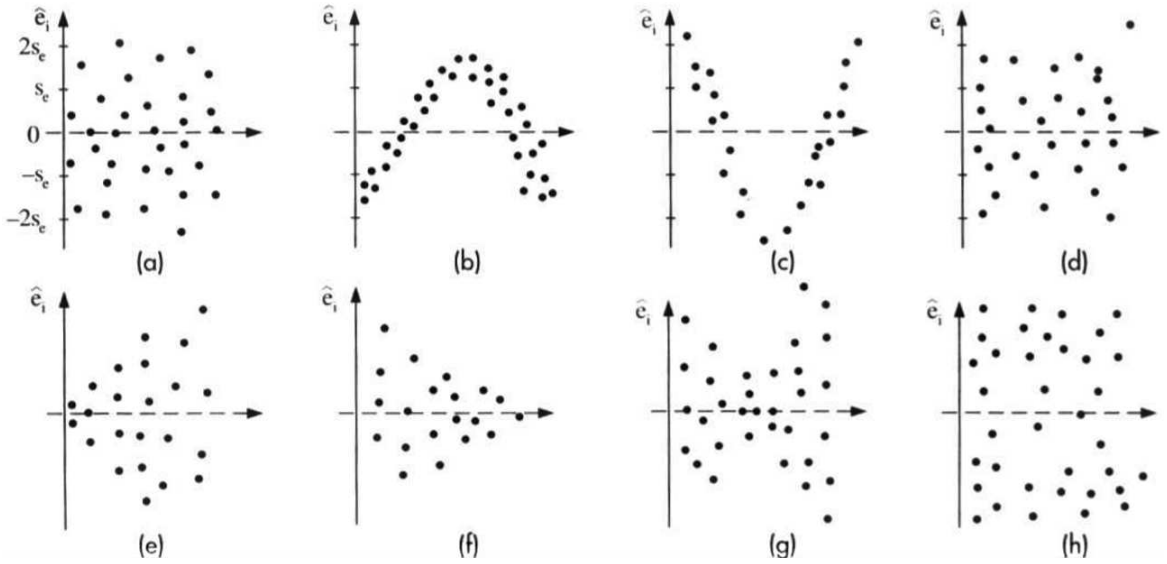

Figura 3.6: Gráfico de resíduos.(a) situação ideal;(b),(c) modelo não-linear;(d) elemento atípico; (e), (f), (g) heterocedasticidade; (h) não-normalidade.

Nas situações (b) e (c) temos possíveis inadequações do modelo adotado, e as curvas sugerem que devemos procurar outras funções matemáticas que expliquem melhor o fenômeno.

A Figura 3.1 (d) mostra a existência de um elemento discrepante, e deve ser investigada a razão desse desvio tão marcante. Pode ser um erro de medida, ou a discrepância pode ser real. Em situações como essa, em que há observações muito diferentes das demais, métodos chamados robustos têm de ser utilizados.

Os casos (e), (f) e (g) indicam claramente que a suposição de homocedasticidade (mesma variância) não esta satisfeita. Em (h), parece haver maior incidência de observações nos extremos, mostrando que a suposição de normalidade não está satisfeita.

Analisando os resíduos e diagnosticada uma possível transgressão das suposições, devemos propor alterações que torne o modelo mais adequado aos dados e as suposições feitas.

A verificação da hipótese de normalidade pode ser realizada fazendo-se um histograma dos resíduos ou um gráfico de quantis.

3.5.1 Gráfico de Quantis

Podemos construir uma representação gráfica dos quantis, chamada *gráfico de quantis*, o que nos ajuda a interpretar um conjunto de dados. Ao se colocar os valores no eixo das abscissas e das ordenadas unem-se os pontos por segmentos de retas.

O Gráfico de quantis pode ser útil para verificar se a distribuição dos dados é simétrica ou (aproximadamente simétrica).

Se os dados forem aproximadamente simétricos, os pontos no topo superior direito do gráfico de quantis comportam-se como os pontos do canto inferior esquerdo. Se os dados forem assimétricos à direita, os pontos do topo superior direito são mais inclinados do que os pontos do canto inferior esquerdo.

3.6 A Normalização de Distribuições Não-Normais Através da Transformação de Box-Cox

Uma das suposições mais frequentes na área de controle estatístico de processo é que variáveis mensuráveis seguem distribuição normal. Manuais tradicionais vão até o ponto de admitir que existe a necessidade de testar hipótese nula de normalidade, mas não explicam o que fazer na eventualidade de rejeição da hipótese. Na literatura, a suposição de normalidade tem sido investigada em vários estudos no tocante a índices de capacidade de processo. Dois trabalhos mais recentes sugerem o uso de distribuições de probabilidade mais gerais as quais incluem a distribuição normal como um caso especial. Castagliola (1996) sugere o uso da distribuição de Burr e Clements (1989) a família de distribuições Pearson. O primeiro trabalho oferece uma curta mais abrangente bibliografia da literatura. Nossa proposta é atacar o problema de não-normalidade pela transformação dos dados originais através da transformação Box-Cox para chegar a valores transformados que demonstrem normalidade ou no mínimo normalidade aproximada.

A transformação de BOX-COX foi introduzida na literatura em 1964 para resolver o problema de estimação de regressões não lineares. Veja Kennedy, 1994, pp, 103-104, onde existe um resumo desta área da Econometria. Box e Jenkins usam esta transformação no software Autobox (1990) para calcular variâncias simples.

3.6.1 Transformação de Box-Cox

A transformação é relativamente simples:

$$y^{(l)} = \frac{(y^l - 1)}{l}$$

Onde l varia entre $(-1, 1)$. O valor transformado dos dados passa através de vários

tipos de equações para cada valor de l. Ver tabela a seguir:

Tabela 3.2: Alguns valores transformados, para determinados valores de λ.

se λ for igual a:	então $\frac{(y^l-1)}{l}$ será igual a:
-1,0	$\frac{1}{y}+1$
-0,5	$\frac{\frac{1}{\sqrt{y}}-1}{-0,5}$
0,0	$\ln y$ (da regra de L'Hospital)
0,5	$\frac{\sqrt{y}-1}{0,5}$
1,0	$y-l$

O resultado do que $y^{(0)} = \ln y$ pode ser demonstrado utilizando a regra de L'Hospital. Portanto, medidas de assimetria e curtose são minimizadas e a distribuição normal aproximada quando valores apropriados de λ são encontrados. O coeficiente de assimetria é definido como:

$$\text{coef. de assim} = \frac{m3}{s^3}$$

Onde $m3$ é o terceiro momento de distribuição e s^3 é o desvio padrão elevado ao cubo. Existe uma relação entre λ e o coeficiente de assimetria representado no gráfico a seguir.

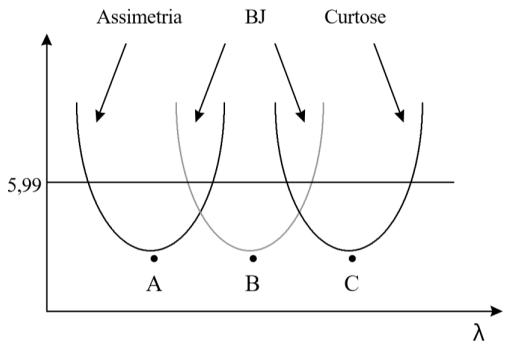

Figura 3.7: A relação entre λ e algumas medidas de não normalidade.

No ponto A, assimetria está no seu valor mínimo dado o valor apropriado de λ.

O coeficiente de curtose é medido pelo quarto momento normalizado pelo desvio padrão elevado a quatro.

$$\text{coef. de curtose} = \frac{m4}{s^4}$$

Desde que o coeficiente de curtose em excesso apresentado é igual a três no caso da distribuição normal, nesse trabalho utilizaremos o coeficente de curtose em excesso:

$$\text{coef. de curtose em excesso} = \left(\frac{m4}{s^4} - 3\right)$$

Existe uma relação entre λ e o coeficiente de curtose em excesso representada no gráfico da Figura 3.7. No ponto C, curtose é minimizado.

A estatística de Bera-Jarque segue a distribuição normal de c^2 sob hipótese nula de normalidade com dois graus de liberdade e é uma medida ponderada dos dois coeficientes apresentados acima;

$$BJ = N^* \sqrt{\left(\frac{\left(\frac{m3}{s^3}\right)^2}{6} + \frac{\left(\frac{m4}{s^4} - 3\right)^2}{24}\right)} \approx c^2_{(2)}$$

Onde N é o tamanho da amostra e num nível de significancia de 5% o valor de $c^2_{(2)} = 5,99$. Em outras palavras, quando BJ é menor que 5,99, a hipótese nula de normalidade é aceita. O gráfico da Figura 3.7 demonstra também a relação entre λ e a estatística de Bera-Jarque. No ponto B a estatítica tem o valor mínimo e talvez menor que 5,99.

Capítulo 4

ANÁLISE DE REGRESSÃO LINEAR NO SOFTWARE R

Existe uma grande utilização da análise de regressão em diversas áreas com o auxilio de recursos computacionais, aliada à disseminação do software estatístico gratuito R, faremos uso dessa tão importante ferramenta que facilita a análise e interpretação de dados.

Para melhor compreensão de sua aplicabilidade iniciaremos com um exemplo. Um investigador deseja estudar a possível relação entre o salário (em anos completos) e o tempo de experiência (em mil reais) no cargo de gerente de agências bancárias de uma grande empresa. Os dados coletados são mostrados na Tabela 4.1.

Tabela 4.1: Salário e tempo de experiência dos gerentes de uma agência bancária.

Salário	Experiência	Salário	Experiência
1.903	0	4.223	23
3.176	17	4.092	20
2.276	8	3.600	18
3.130	15	4.707	27
2.776	9	3.146	11
3.092	15	2.992	10
2.653	8	4.746	29
2.223	5	4.115	23
2.853	13	2.361	4
3.230	20	4.092	22
2.823	11	4.507	25
1.907	1	2.907	9
2.538	6	4.484	25
2.569	7		

Note que são considerados 27 pares de observações correspondentes à variável resposta Salário e à variável explicativa Experiência, para cada um dos gerentes da empresa.

4.1 Leitura de Dados

Inicialmente os dados devem ser organizados como *objetos de dados R*, nesse caso como um *data frame* (planilha). Para isso, é necessario que a tabela se encontre numa estrutura tabular, na qual as colunas representam as variáveis e as linhas representam os indivíduos. Sendo possível digitar direto na linha de comando ou fazer a importação de dados, se eles estiverem em documento excel, como é o nosso caso, devemos salvar o arquivo em formato conveniente para entrada no R, fazendo uso do `read.table`,que é uma ferramenta muito útil. O R pode ler arquivos de texto (ASCII) e também em outros formatos (Excel, SAS, SPSS, etc), e até mesmo acessar bancos de dados SQL. Nesses termos, seja o arquivo de texto *gerentes.csv*, utiliza-se a função read.table para que o nome do arquivo seja lido pelo R.

Algoritmo

```
getwd()
dados<-read.table("C:/Users/sergi/Desktop/TCC_sergio/TCC_R/
gerentes.csv",header=TRUE,sep=";",dec=".")
attach(dados)
dados
```

R

Saída no Terminal

```
> getwd()
[1] "C:/Users/sergi/Documents"
> dados<-read.table("C:/Users/sergi/Desktop/TCC_sergio/TCC_R/
gerentes.csv",header=TRUE,sep=";",dec=".")
> attach(dados)
> dados
   Salario Experiencia
1    1.930           0
2    3.176          17
3    2.276           8
4    3.130          15
.
.
.
```

R

Observe que a função `attach` anexa o objeto *dados* no caminho do software.

4.2 Análise Exploratória

4.2.1 Estatística Descritiva

Uma maneira fácil de obter algumas estatísticas descritivas das variáveis em estudo é através do comando `summary()`, que retorna as estatísticas *mínimo, quartis, média* e *máximo*. Para medir a variabilidade, utilize as funções `var()` e `sd()` para obter a *variância* e o *desvio padrão*.

Saída no Terminal

```
> summary(Salario)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.907   2.611   3.092   3.228   4.092   4.746
> var(Salario)
[1] 0.7370166
> sd(Salario)
[1] 0.8584967
```

R

4.2.2 Diagrama de Dispersão

Para verficar a existência de alguma relação entre *Salário* e *Experiência*, deve-se construir um *Diagrama de Dispersão* para as duas variáveis:

Saída no Terminal

```
> plot(Experiencia,Salario)
```

R

4.2.3 Correlação Linear

Para calcular o Coeficiente de Correlação Linear de Pearson entre as variáveis, utilize a função cor:

Saída no Terminal

```
> cor(Experiencia,Salario)
[1] 0.969241
```

R

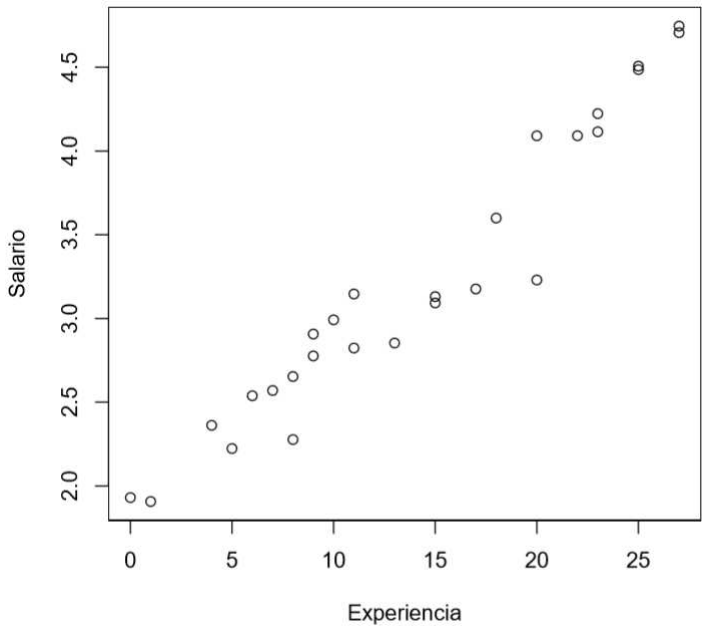

Figura 4.1: Diagrama de Dispersão de Salário versus Experiência.

Observe que o R retornou o valor 0.969241 o que evidencia uma forte relação linear entre as variáveis em estudo. Para avaliar se esse resultado é signicativo, pode-se realizar um *Teste de Hipóteses* para a o Coeciente de Correlação (supondo que as suposições do teste sejam satisfeitas):

Saída no Terminal

```
> cor.test(Experiencia,Salario)

        Pearson's product-moment correlation

data:  Experiencia and Salario
t = 19.691, df = 25, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9328011 0.9860631
sample estimates:
     cor
0.969241
[1] 0.969241
```

R

Como o Valor P do teste (p-value $<$ 2.2e-16) é bem pequeno, conclui-se que o valor do Coeciente de Correlação Linear de Pearson tem significância Estatística.

4.3 Regressão Linear Simples

4.3.1 Ajuste do Modelo de Regressão

Sejam X e Y, respectivamente, as variáveis *Experiência* (explicativa) e *Salário* (resposta). Propõe-se um modelo de regressão linear de primeira ordem, dado pela equação: $Y = \beta_0 + \beta_1 X + \epsilon$, onde β_0 e β_1 são parâmetros desconhecidos e ϵ é o erro aleatório.

Para ajustar um modelo de regressão linear no R utiliza-se a função `lm`:

Saída no Terminal

```
> ajuste=lm(Salario ~ Experiencia)
>
> ajuste

Call:
lm(formula = Salario ~ Experiencia)

Coefficients:
(Intercept)  Experiencia
     1.7921       0.1023
```

R

Note que função `lm()` é chamada com o formato lm(y ~ x), ou seja, a variável resposta é y e a preditora é x, sempre nessa ordem.

O R retorna o valor dos coeficientes de $\widehat{\beta}_0$ e $\widehat{\beta}_1$ estimados via Método de Mínimos Quadrados. Logo, a equação da reta ajustada é dada por $\widehat{Y} = 1,79 + 0,1023X_i$.

Com a função `summary`, diversas medidas descritivas úteis para a análise do ajuste podem ser obtidas:

Saída no Terminal

```
> summary(ajuste)

Call:
lm(formula = Salario ~ Experiencia)

Residuals:
     Min       1Q   Median       3Q      Max
-0.60764 -0.08734  0.06094  0.15571  0.25436

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.792134   0.083877   21.37   <2e-16 ***
Experiencia 0.102275   0.005194   19.69   <2e-16 ***
---
Residual standard error: 0.2155 on 25 degrees of freedom
Multiple R-squared:  0.9394,    Adjusted R-squared:  0.937
F-statistic: 387.7 on 1 and 25 DF,  p-value: < 2.2e-16
```

R

Da execução desse comando, pode-se obter, por exemplo, os erros-padrão (`Std. Error`) das estimativas dos coeficientes de regressão: $\text{EP}(\widehat{\beta}_0) = 0,0839$ e $\text{EP}(\widehat{\beta}_0) = 0,0052$. Além disso, obtém-se o valor do Coficiente de Determinação (`Multiple R-Squared`), $R^2 = 0,9394$.

Com a função `anova`, pode-se construir a Tabela da Análise de Variância:

Saída no Terminal

```
> anova(ajuste)
Analysis of Variance Table

Response: Salario
            Df  Sum Sq Mean Sq F value    Pr(>F)
Experiencia  1 18.0017 18.0017  387.73 < 2.2e-16 ***
Residuals   25  1.1607  0.0464
---
```

R

Da tabela ANOVA, obtém-se o Quadrado Médio (`Mean Sq`) Residual, que é uma estimativa para a variância dos erros (σ^2), ou seja, $s^2 = 0,0464$.

Para esboçar a reta ajustada no diagrama de dispersão, utilize a função `abline`:

Saída no Terminal

```
> windows()
> plot(Experiencia,Salario)
> abline(lm(Salario ~ Experiencia))
```

R

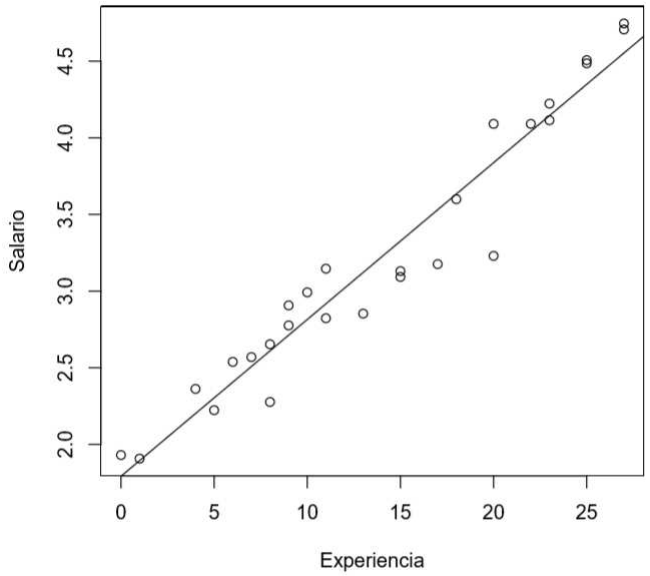

Figura 4.2: Diagrama de Dispersão de Salário versus Experiência com reta ajustada.

4.3.2 Intervalos de Confiança para β_0 e β_1

Para construir os Intervalos de Cofiança (95%) para os coeficientes da regressão, utiliza-se o seguinte comando:

Saída no Terminal

```
> confint(ajuste)
                2.5 %    97.5 %
(Intercept) 1.6193868 1.9648808
Experiencia 0.0915781 0.1129728
```

R

4.3.3 Teste de Hipótese

Para proceder o *Teste F da Significâcnia da Regressão* e os *Testes t individuais*, verifique o *Valor P* para cada caso através da saída da função `summary`:

Saída no Terminal

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.792134   0.083877   21.37   <2e-16 ***
Experiencia 0.102275   0.005194   19.69   <2e-16 ***
---
(...)
F-statistic: 387.7 on 1 and 25 DF,  p-value: < 2.2e-16
```

R

4.3.4 Análise dos Resíduos

Para avaliar as suposições de que os erros possuem variância constante e são não correlacionados entre si, construiremos os gráficos de "Resíduos versus Valores Ajustados da Variável Resposta" e "Resíduos versus Valores da Variável Explicativa":

Saída no Terminal

```
> windows()
> plot(fitted(ajuste),residuals(ajuste),xlab="Valores Ajustados",
ylab="Resíduos")
> abline(h=0)
> windows()
> plot(Experiencia,residuals(ajuste),xlab="Experiencia",
ylab="Resíduos")
> abline(h=0)
```

R

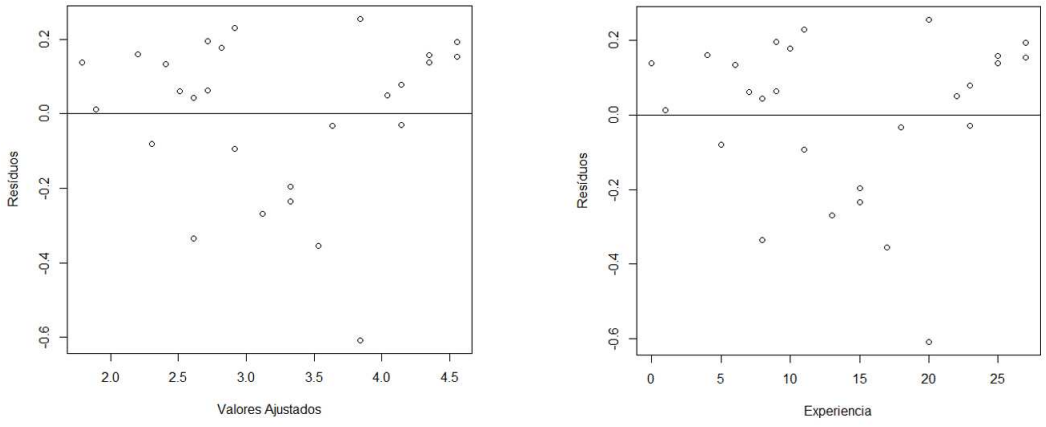

Figura 4.3: Gráfico para Análise de Resíduos.

Para exibir os Valores Ajustados e os Resíduos do ajuste, digite os comandos:

Algoritmo

```
ajuste$residuals
ajuste$fitted.values
```

R

Na Figura 4.3, observa-se a violação da suposição de homocedasticidade dos erros. Para corroborar esse resultado, pode-se dividir o conjunto de dados em duas partes, utilizando a mediana por exemplo, e realizar um teste para comparar as variâncias de cada subconjunto:

Saída no Terminal

```
> median(Experiencia)
[1] 13
>
> var.test(residuals(ajuste)[dados$Experiencia>13],residuals
(ajuste)[dados$Experiencia<13])

        F test to compare two variances

data:  residuals(ajuste)[dados$Experiencia > 13] and residuals
(ajuste)[dados$Experiencia < 13]
F = 2.6672, num df = 12, denom df = 12, p-value = 0.1024
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
 0.8138611 8.7413117
sample estimates:
ratio of variances
          2.667248
```

R

Observe que o *Valor P* do teste `(p-value = 0.1024)` é maior que os níveis de significância mais usuais (0,01; 0,05; 0,10). Portanto, conclui-se que a variância dos dois subconjuntos não é igual, o que implica a heterocedasticidade dos erros.

Outra maneira de avaliar a heterocedasticidade dos erros é realizar algum teste de homocedasticidade. Na biblioteca *lmtest* do *R*, a função bptest realiza o teste de Breusch-Pagan. Ressalta-se, entretanto, que tal teste não é muito poderoso e pode levar à para avaliar a suposição de normalidade dos erros, deve-se construir o gráfico da "Probabilidade Normal dos Resíduos":

Saída no Terminal

```
> windows()
> qqnorm(residuals(ajuste),ylab="Resíduos",xlab="Quantis
teóricos",main="")
> qqline(residuals(ajuste))
```

R

Pela Figura 4.4, observa-se a violação da suposição de que os erros aleatórios têm distribuição Normal. Considere, também o *Teste de Normalidade de Shapiro Wilk*:

Algoritmo

```
shapiro.test(residuals(ajuste))
```

R

Saída no Terminal

```
        Shapiro-Wilk normality test

data:  residuals(ajuste)
W = 0.88528, p-value = 0.006259
```

R

Portanto, como o *Valor P* do teste é pequeno, rejeita-se a hipótese de normalidade dos resíduos e, por consequência, conclui-se que os erros não são normalmente distribuídos.

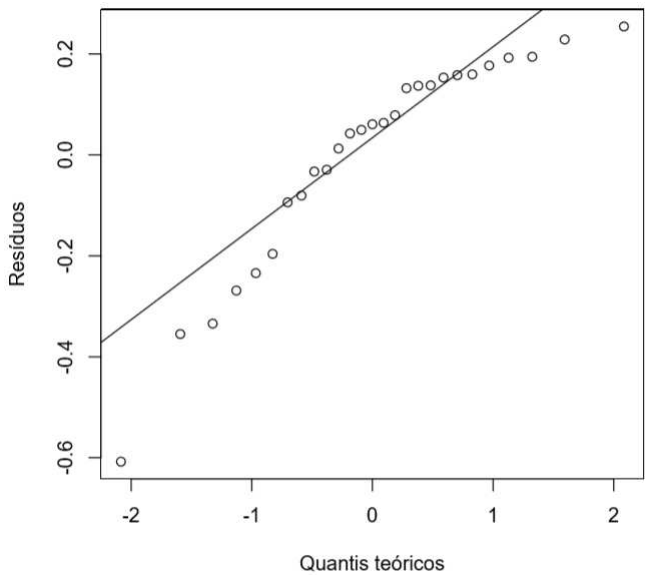

Figura 4.4: Gráfico de Probabilidade Normal dos Resíduos.

4.4 Transformações

4.4.1 Tranformação na Variável Resposta

Consideremos os dados da Tabela 4.1 para análise e avaliação do modelo para uma transformação da variavél resposta.

Algoritmo

```
getwd()
dados<-read.table("C:/Users/sergi/Desktop/TCC_sergio/TCC_R/
gerentes.csv",header=TRUE,sep=";",dec=".")
attach(dados)
dados

plot(Salario ~ Experiencia)
```

R

Saída no Terminal

```
> getwd()
[1] "C:/Users/sergi/Documents"
> dados<-read.table("C:/Users/sergi/Desktop/TCC_sergio/TCC_R/
gerentes.csv",header=TRUE,sep=";",dec=".")
> attach(dados)
> dados
   Salario Experiencia
1    1.930           0
2    3.176          17
3    2.276           8
4    3.130          15

.
.
.
>
> plot(Salario ~ Experiencia)
```

R

Após a leitura dos dados, veja a relação entre as duas variáveis:

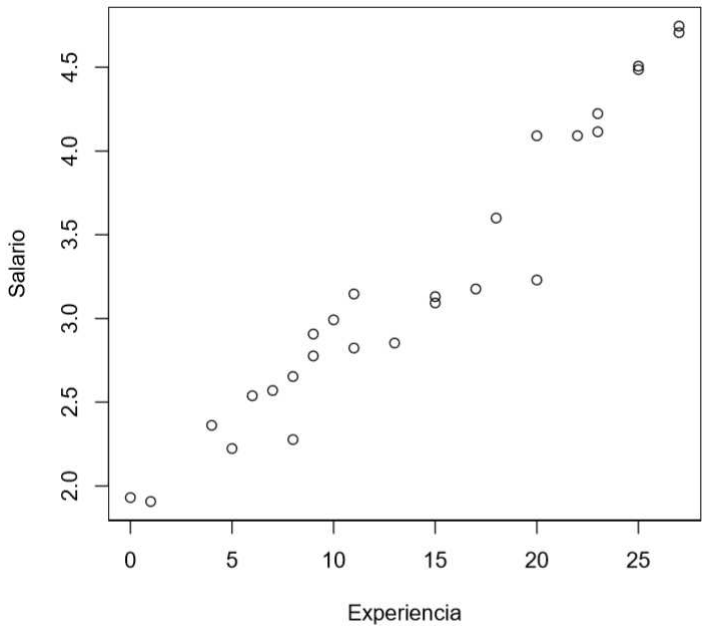

Figura 4.5: Gráficos de dispersão de Salário vs Experiência.

Observe que o diagrama de dispersão da Figura 4.5 mostra uma forte relação crescente não linear entre as medidas de Salário e Experiência. As medidas de Salário apresentam aumento de variabilidade para valores crescentes de Experiência.

Diante disso, podemos concluir que o ajuste do modelo de regressão linear simples com as variáveis na sua forma original é inadequado neste caso. No entanto, tal ajuste é realizado a seguir, a fim de evidenciar sua inadequação na Análise de Resíduos.O modelo linear ajustado que veremos a seguir, de Salário em Experiência é $\widehat{Y} = 1,79 + 0,1023X_i$, com R^2 ajustado = 0,937.

Algoritmo

```
ajuste = lm(Salario ~ Experiencia,dados)
summary(ajuste)

shapiro.test(residuals(ajuste))
```

R

Saída no Terminal

```
Call:
lm(formula = Salario ~ Experiencia, data = dados)

Residuals:
     Min       1Q   Median       3Q      Max
-0.60764 -0.08734  0.06094  0.15571  0.25436

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.792134   0.083877   21.37   <2e-16 ***
Experiencia 0.102275   0.005194   19.69   <2e-16 ***
---
Residual standard error: 0.2155 on 25 degrees of freedom
Multiple R-squared:  0.9394,    Adjusted R-squared:  0.937
F-statistic: 387.7 on 1 and 25 DF,  p-value: < 2.2e-16

> shapiro.test(residuals(ajuste))
        Shapiro-Wilk normality test
data:  residuals(ajuste)
W = 0.88528, p-value = 0.006259
```

R

Os gráficos para análise dos resíduos serão gerados a partir dos comandos a seguir:

Algoritmo

```
windows()
par(mfrow = c(2, 2))
plot(fitted(ajuste), residuals(ajuste),xlab="Valores Ajustados",
ylab="Resíduos")
abline(h=0)
plot(Experiencia, residuals(ajuste),xlab="Experiencia",
ylab="Resíduos")
abline(h=0)
hist(residuals(ajuste),main="",xlab="Resíduos",ylab="Freqüência")
qqnorm(residuals(ajuste),main="",xlab="Quantis teóricos",
ylab="Resíduos")
qqline(residuals(ajuste))
```

R

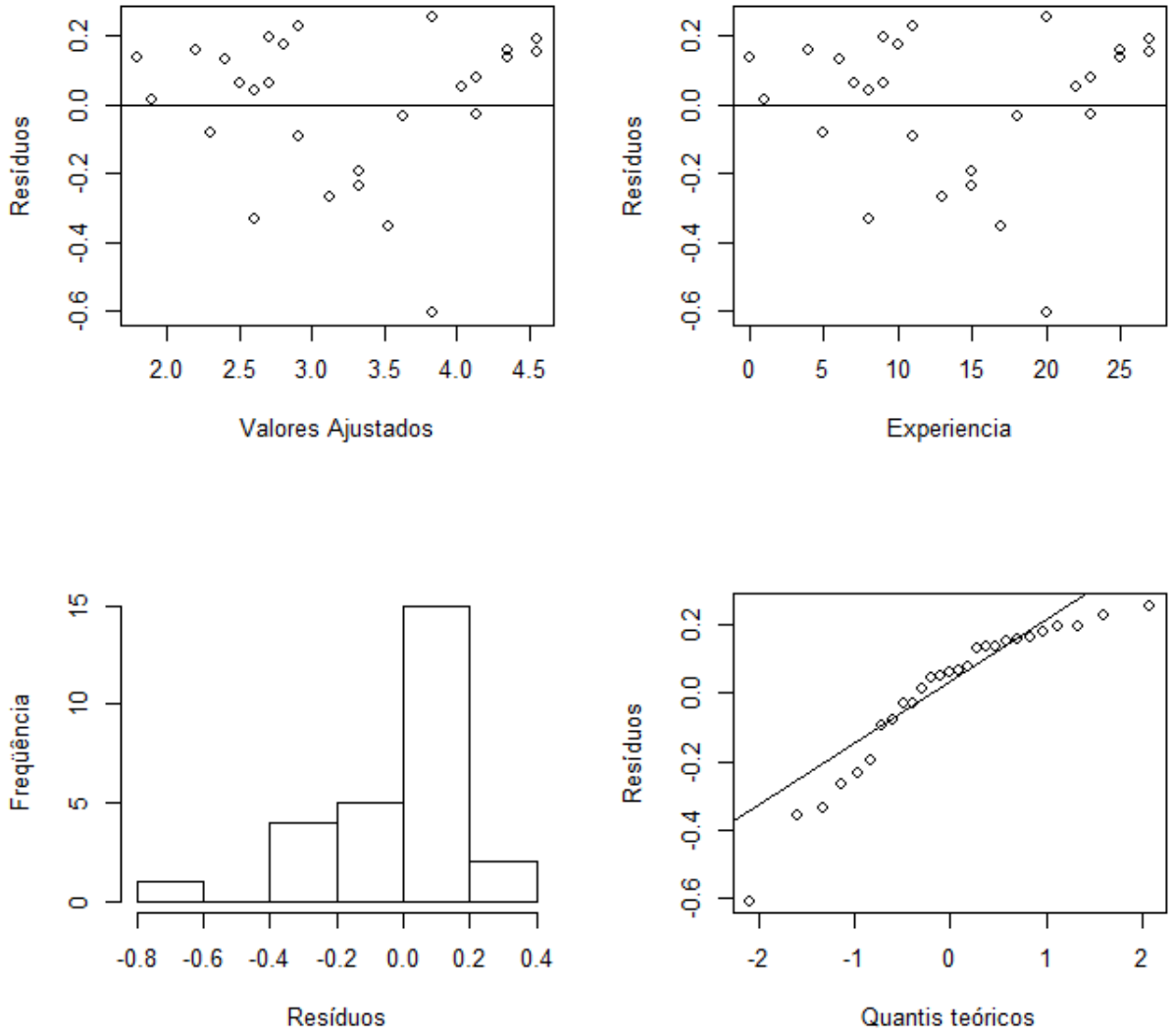

Figura 4.6: Gráfico para Análise dos Resíduos.

Observe da Figura 4.6, apesar que a variância dos erros não é constante e que a normalidade dos erros é violada, suposição que também é rejeitada pelo Teste de normalidade de *Shapiro-Wilk*, cujo *P-valor* é 0,006259.

A fim de solucionar os problemas de variância não-constante e não-normalidade dos erros, deve-se tentar realizar uma transformação na variável resposta. Apesar de ser possível, em muitos casos, selecionar empiricamente a transformação adequada, apresentaremos aqui apenas a técnica mais formal e objetiva. Uma transformação adequada para a variável resposta via *Procedimento de Box Cox* é obtida da seguinte forma:

Algoritmo

```
require(MASS)
windows()
par(mfrow = c(1, 2))
boxcox(ajuste)
boxcox(ajuste, lambda = seq(0.1, 0.5, by = 0.01))
```

R

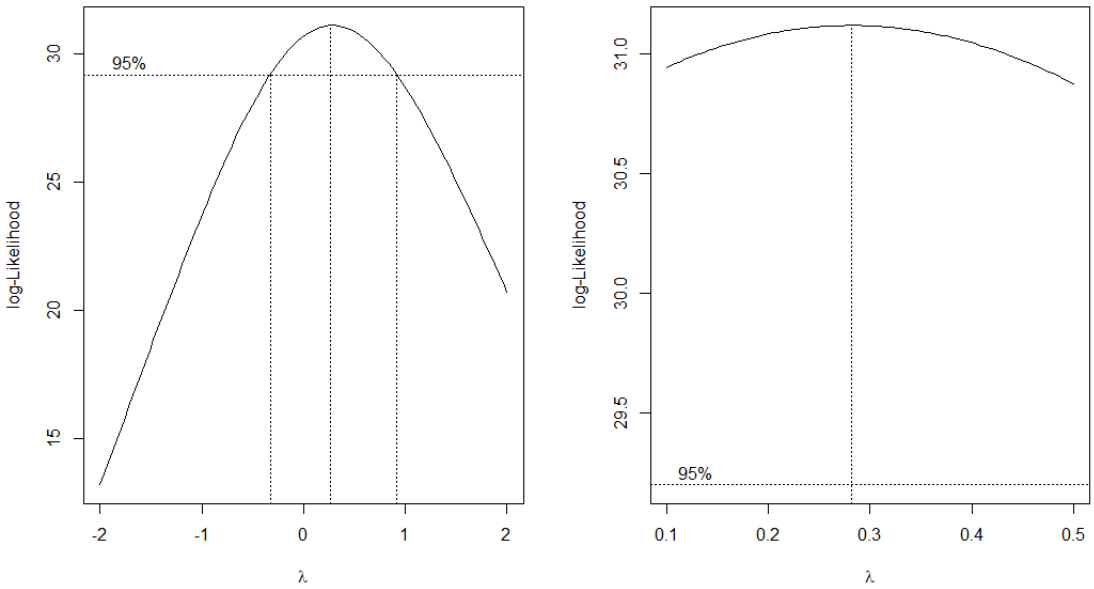

Figura 4.7: Log-verossimilhança e Intervalo de Confiança de 95% para valores de λ da tranformação de Box Cox.

Note que a Figura 4.7 mostra os valores da log-verossimilhança para um intervalo de valores do parâmetro de transformação λ. O máximo da verossimilhança foi atingido com aproximadamente $\lambda = 0,28$, com intervalo de confiança de 95% igual a [-0,35;0,92]. Como esse intevalo não inclui o valor 1, há forte evidência da necessidade de transformação na variável resposta Salário, dado por: $Salario = (Salario^{0,28} - 1)/0,28$. Sendo assim a nova variável transformada Salariotrans deve ser inserida no banco de dados, para que o novo modelo de regressão linear simples seja ajustado.

Saída no Terminal

```
> Salariotrans = (Salario^0.28-1)/0.28
> dadostrans = data.frame(cbind(dados,Salariotrans))
> dadostrans
   Salario Experiencia Salariotrans
1    1.930           0    0.7219384
2    3.176          17    1.3644959
3    2.276           8    0.9248188
4    3.130          15    1.3443734
.
.
.
> plot(Experiencia, Salariotrans,ylab="Salario transformado")
```

R

A Figura 4.8 mostra uma forte relação linear crescente entre as medidas de Salário transformado, via método de Box Cox, versus Experiência, com variabilidade aproximadamente constante.

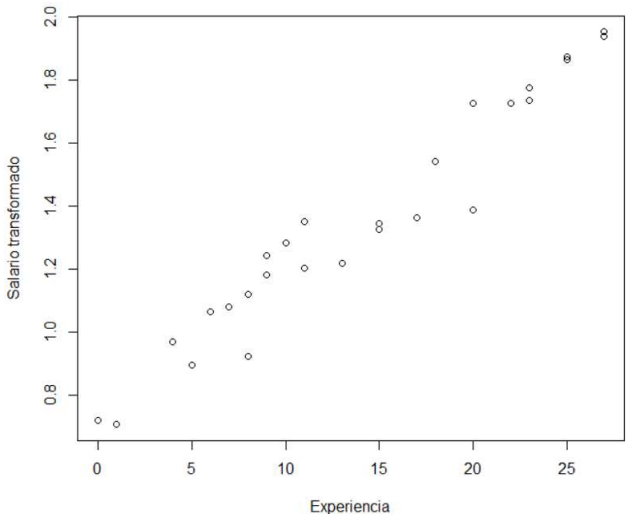

Figura 4.8: Diagrama de dispersão da transformação de *Salario* versus *Experiencia*.

Saída no Terminal

```
> ajuste2 = lm(Salariotrans ~ Experiencia, dadostrans)
> summary(ajuste2)

Call:
lm(formula = Salariotrans ~ Experiencia, data = dadostrans)

Residuals:
     Min       1Q   Median       3Q      Max
-0.22777 -0.05527  0.02371  0.04393  0.13226

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.734534   0.033809   21.73   <2e-16 ***
Experiencia 0.044055   0.002094   21.04   <2e-16 ***
---

Residual standard error: 0.08685 on 25 degrees of freedom
Multiple R-squared:  0.9466,    Adjusted R-squared:  0.9444
F-statistic: 442.8 on 1 and 25 DF,  p-value: < 2.2e-16
```

R

Após ajuste do modelo para obter a nova reta de regresão e gráficos para análise, teremos:

Saída no Terminal

```
> windows()
> par(mfrow = c(2, 2))
> plot(fitted(ajuste2), residuals(ajuste2),xlab="Valores
 Ajustados",ylab="Resíduos")
> abline(h=0)
> plot(Experiencia, residuals(ajuste2),xlab="Experiencia",
ylab="Resíduos")
> abline(h=0)
> hist(residuals(ajuste2),main="",xlab="Resíduos",
ylab="Freqüência")
> qqnorm(residuals(ajuste2),main="",xlab="Quantis teóricos",
ylab="Resíduos")
> qqline(residuals(ajuste2))
```

R

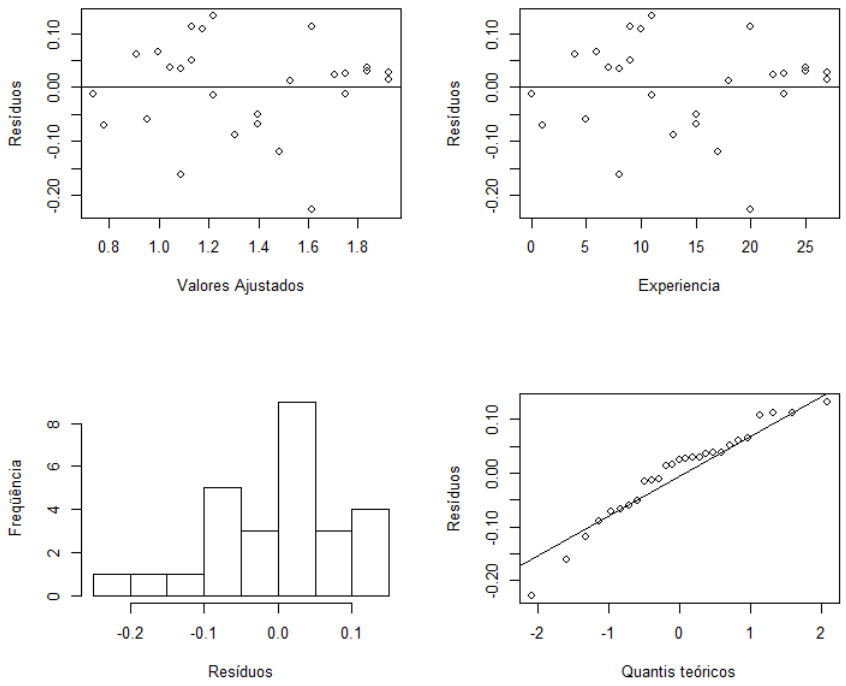

Figura 4.9: Gráficos para Análise dos Resíduos.

Logo, a equação da reta ajustada é dada por $\widehat{Y} = 0,735 + 0,044X_i$, com R^2 *ajustado* = 0,9444 e pela Figura 4.9 observa-se que a suposição de normalidade é aceitável, embora a transformação não tenha solucionado o problema da heterocedasticidade (dispersão dos erros em torno da reta).

Capítulo 5

Conclusões

Neste trabalho falamos sobre a aplicabilidade de funções linear e quadrática na estatística fazendo aplicações com regressão linear simples, mostrando que esse é um modelo usado para gerar previsões e explicações baseado nas relações entre variáveis que estejam correlacionadas.

Visamos a interdisciplinaridade com o intuito do professor, nesse processo de ensino aprendizagem despertar, ou melhor, incentivar o discente a fazer uma leitura de mundo com olhos matemáticos, tornando assim mais fascinante todo esse processo.

Após a apresentação da base teórica onde observamos que a regressão linear parte dos conceitos de função, fixamos alguns pontos sobre regresão linear simples, para assim realizar uma leitura dos dados aplicados no software R.

É inegável como o recurso computacional torna atrativa a forma de trabalhar conteúdos, após a explanação dos conceitos, o software R contribuiu significativamente de forma clara e suscinta para análise e discursão dos dados aplicados.

As aplicações são inúmeras que vão desde situações mais simples, como por exemplo, analisar em nossas casas a relação gasto e receita, observar a média de temperatura mensal de sua cidade para ver a correlação com o consumo de energia de sua casa. Situações comerciais como investimento e lucro, horas trabalhadas e produção, banco de dados com idade e peso, idade e tempo de resposta. Situações mais complexas para fundamentar uma teoria, como por exemplo, a relação investimento em educação versus aumento da criminalidade, desemprego versus índices de consumo. Em fim, existindo a correlação entre valores que possam ser mensurados e tabelados, a regressão linear simples será uma importante ferramenta para a análise e compreensão desses dados.

Bibliografia

[1] AMARAL, Gabriela D.; SILVA, Vanessa L.; REIS, Afonso R.;**Análise de Regressão Linear no Pacote R**. Relatório Técnico. Minas Gerais: UFMG Departammento de Estatística, 2009.

[2] BLAIR, R. Clifford; TAYLOR, Richard A.; **Bioestatística para ciências da saúde**. 4. ed. São Paulo: Pearson Education do Brasil, 2013.

[3] BUSSAB, Wilton de O.; MORETTIN, Pedro A. **ESTATÍSTICA BÁSICA**. 5. ed. São Paulo: Saraiva, 2002.

[4] FLEMMING, Diva M.; GONÇALVES, Mirian B. **Cálculo A**. 6. ed. São Paulo: Pearson Prentice Hall, 2006.

[5] GUIDORIZZI, Hamilton L.; **UM CURSO DE CÁLCULO**. 5. ed. Rio de Janeiro: LTC, 2001.

[6] HOFFMANN, Rodolfo; **Análise de regressão: uma introdução à econometria [recurso eletrônico]**. Piracicaba: ESALQ/USP, 2015.

[7] HOWARD, Anton; BIVENS, Irl; DAVIS, Stephen; **Cálculo**. 8. ed. Porto Alegre: Bookman, 2017.

[8] LEITHOLD, L.; **O CÁLCULO COM GEOMETRIA ANALÍTICA**. 3. ed. São Paulo: editora HARBA ltda, 1994.

[9] MORETTIN, Pedro A.; HAZZAN, Samuel; BUSSAB, Wilton de O. **Cálculo funções de uma e várias variavés**. São Paulo: Saraiva, 2003.

[10] PISKOUNOV, N.; **CÁLCULO DIFERENCIAL E INTEGRAL**. 18. ed. Porto: Livraria Lopes da Silva, 2000.

[11] SAMOHYL, R. Wayne; et al.; **NORMALIZAÇÃO DE DISTRIBUIÇÕES NÃO-NORMAIS ATRAVÉS DA TRANSFORMAÇÃO DE BOX-COX....**

Santa Catarina: Departamento de Engenharia de Produção e Sistemas - Centro Tecnológico, [200-?].

[12] SIMMONS, F. George; **CÁLCULO com Geometria Analítica**. 5. ed. São Paulo: Pearson Makrons Books, 1987.

[13] SWOKOWSKI, Earl W.; **CÁLCULO Com Geometria Analítica**. 2. ed. São Paulo: Makron Books, 1994.

[14] SODRÉ, Ulysses; **LATEX Para Matemáticca com o TexnicCenter**. Versão compiladano dia 21 de Agosto de 2006. Londrina: Depertamento de Matemática-UEL, 2006.

[15] STEWART, James; **CÁLCULO**. 7. ed. São Paulo: Cengage Learning, 2013.

[16] VIEIRA, Sonia; **Bioestatística: tópicos avançados**. 3. ed. Rio de Janeiro: Elsevier, 2010.

[17] VIEIRA, Sonia; **Introdução à bioestatística**. 3. ed. revista e 3. ed. ampliada. Rio de Janeiro: Elsevier, 1980.

[18] VIEIRA, Sonia; **Introdução à bioestatística [recurso eletrônico]**. 4. ed. Rio de Janeiro: Elsevier, 2011.

[19] WEIR, Maurice D.; **CÁLCULO (GEORGE B. THOMAS)**. 11. ed. São Paulo: Addison Wesley, 2009.