UNIVERSIDADE FEDERAL DO RECÔNCAVO DA BAHIA
CENTRO DE CIÊNCIAS EXATAS E TECNOLÓGICAS

PROFMAT

ALINE DE OLIVEIRA MOREIRA

# UM ESTUDO SOBRE REGRESSÃO LINEAR E REGRESSÃO LOGÍSTICA

CRUZ DAS ALMAS

2021

ALINE DE OLIVEIRA MOREIRA

# UM ESTUDO SOBRE REGRESSÃO LINEAR E REGRESSÃO LOGÍSTICA

Dissertação apresentada ao PROFMAT do Centro de Ciências Exatas e Tecnológicas da Universidade Federal do Recôncavo da Bahia como parte dos requisitos para a obtenção do título de Mestre em Matemática.

Orientadora: Julianna Pinele Santos Porto

CRUZ DAS ALMAS

2021

FICHA CATALOGRÁFICA

M838e

Moreira, Aline de Oliveira.

Um estudo sobre regressão linear e regressão logística / Aline de Oliveira Moreira._ Cruz das Almas, Bahia, 2021.

59f.: il.

Dissertação (Mestrado) – Universidade Federal do Recôncavo da Bahia, Centro de Ciências Exatas e Tecnológicas, Mestrado Profissional em Matemática – PROFMAT.

Orientadora: Prof. Dra. Julianna Pinele Santos Porto.

1.Matemática – Análise de regressão. 2.Matemática – Aprendizado demáquina. 3. Matemática – Regressão logística. I.Universidade Federal do Recôncavo da Bahia, Centro de Ciências Exatas e Tecnológicas. II.Título.

CDD: 519.536

Ficha elaborada pela Biblioteca Universitária de Cruz das Almas - UFRB.
Responsável pela Elaboração - Antonio Marcos Sarmento das Chagas (Bibliotecário - CRB5 / 1615).
(os *dados para catalogação foram enviados pela usuária via formulário eletrônico*).

ALINE DE OLIVEIRA MOREIRA

**UM ESTUDO SOBRE REGRESSÃO LINEAR E REGRESSÃO LOGÍSTICA**

Dissertação apresentada ao Programa de Mestrado em Matemática em Rede Nacional (PROFMAT) do Centro de Ciências Exatas e Tecnológicas (CETEC) da Universidade Federal do Recôncavo da Bahia (UFRB) como requisito parcial para obtenção do grau de Mestre em Matemática.

Trabalho Aprovado em : 16 / 08 / 2021

BANCA EXAMINADORA

Julianna Pinele Santos Porto

**Profa. Dra. Juianna Pinele Santos Porto**
Orientadora - UFRB

João Tiago Assunção Gomes

**Prof. Dr. João Tiago Assunção Gomes**
Examinador - UFRB

Paula Aparecida Kikuchi

**Profa. Dra. Paula Aparecida Kikuchi**
Examinadora - UEMS

Cruz das Almas, Bahia
Agosto de 2021

# Agradecimentos

Agradeço à minha família e amigos, pelo estímulo e pela companhia nos momentos divididos e pela compreensão nos momentos de ausência.

Agradeço à minha orientadora, Professora Doutora Julianna Pinele Santos Porto, por sua orientação, paciência e dedicação.

Agradeço à Sociedade Brasileira de Matemática e à Universidade Federal do Recôncavo da Bahia pela oportunidade de realização do PROFMAT.

Agradeço ainda a todos os Professores e colegas com quem tive contato durante o período em que cursei do PROFMAT.

*"Sua tarefa é descobrir seu trabalho e, então, com todo o coração, dedicar-se a ele."*

***Buda***

# Resumo

Neste trabalho é apresentada a fundamentação teórica sobre os assuntos de regressão linear e regressão logística, conceitos que formam a base para compreensão dos processos usados em aprendizado de máquina, campo da ciência muito explorado na atualidade. Em regressão linear busca-se ajustar em uma reta dados quantitativos disponíveis, e através do modelo obtido com tal ajuste fazer previsões futuras ou estimar valores para novas situações de interesse. Em regressão logística busca-se desenvolver um modelo através do qual seja possível classificar os dados disponíveis, e através do modelo obtido com tal ajuste fazer previsões futuras ou estimar valores para novas situações. São apresentados os procedimentos operacionais para o desenvolvimento de modelos de regressão linear simples e de regressão logística simples, bem como as bases estatísticas para sua construção e validação. Este trabalho serve de fonte bibliográfica para professores do ensino médio que queiram aprender e aplicar em sala de aula os conceitos aqui desenvolvidos. Ainda é apresentada uma sequência didática em que se propõe a aplicação de conceitos de regressão linear.

**Palavras-chave**: Aprendizado de máquina. Regressão linear. Regressão logística.

# Abstract

On this written essay there are going to be presented the theoretical foundation on the topics regarding linear regression and logistic regression, concepts that form the basis for understanding the processes used in machine learning, a field of science that is currently widely explored. In linear regression, one seeks to adjust the available quantitative data in a straight line, and by using the model obtained, make future predictions or estimate values for situations of interest. In logistic regression, the aim is to develop a model through which it is possible to classify the available data, and by employing the model obtained, being able to make future predictions or estimate values for new situations. The operational procedures for the development of both simple linear regression and simple logistic regression models will be presented, as well as the statistical bases for their construction and validation. This work can be used as a bibliographic source for high school teachers who want to learn and apply the concepts developed here in the classroom. A didactic sequence is also presented in which it is proposed the application of linear regression concepts presented here.

**Keywords**: Machine learning. Linear regression. Logistic regression.

# Sumário

# 1 Introdução

Aprendizado de máquina é um subcampo da ciência da computação que tem crescido muito nos últimos anos, e atualmente é uma realidade em nossas vidas (BURKOV, 2019). Pode-se definir aprendizado de máquina, de forma simplificada, como um processo de desenvolvimento de algoritmos baseados em uma série de exemplos, chamados de dados de treinamento, a fim de obter um modelo estatístico que resolva um determinado problema.

Faz-se necessário o uso de aprendizado de máquina em casos em que não é possível escrever diretamente um programa para resolver o problema ou em situações em que os dados sofrem mudanças no tempo, ou seja, situações em que são necessários algoritmos capazes de se adaptar às circunstâncias. Assim, pode-se concluir que aprendizado de máquina não se resume apenas a uma solução para problema de banco de dados, mas também parte do campo da inteligência artificial, pois é desejável que o sistema tenha a habilidade de aprender quando exposto a um ambiente sujeito a mudanças (BURKOV, 2019).

Atualmente, há uma abundância de dados disponíveis, no entanto, os dados armazenados são úteis quando transformados em informações possíveis de serem usadas para fazer previsões como por exemplo da compra de um produto. Um grande volume de dados é processado a fim de se obter um modelo simplificado com alto valor agregado (tal como um modelo preditivo de elevada acurácia) (ALPAYDIN, 2014).

Acredita-se que há padrões nos dados, apesar de nem sempre o processo gerador do dado ser conhecido. Nem sempre é possível identificar o processo em sua totalidade, mas pode-se construir uma aproximação boa e útil a partir da qual podem ser observados padrões e regularidades. Através de tais padrões é possível entender melhor o processo e, inclusive, fazer previsões acuradas, assumindo que o futuro não diferirá muito do momento em que os dados foram coletados (ALPAYDIN, 2014).

Uma vez estabelecido um algoritmo de aprendizagem, obtém-se uma informação mais simples do que o dado, o que requer menos espaço de armazenamento (pode-se dizer que foi feita uma compressão) e menor manipulação para seu processamento (ALPAYDIN, 2014).

Por meio da aplicação dos métodos de aprendizado de máquina, busca-se descrever a relação entre a variável resposta, também chamadas de variáveis dependentes ou de saída, e um conjunto de variáveis explicativas, também chamadas de independentes (JR; LEMESHOW; STURDIVANT, 2013).

Pode-se falar em três abordagens em aprendizado de máquina dependendo do tipo de situação a ser tratada, a saber: aprendizado supervisionado, aprendizado não supervisionado e aprendizado por reforço. No aprendizado supervisionado, os valores corretos são fornecidos por um "professor" e o objetivo é estruturar o algoritmo que relaciona a entrada à saída. Já no aprendizado não supervisionado não há "professor" e apenas tem-se os dados de entrada, assim o objetivo do algoritmo é encontrar padrões nos dados. Os algoritmos capazes de avaliar se um conjunto de ações é ou não favorável para atingir um objetivo, com base em sequências passadas de ações a fim de elaborar suas diretrizes são chamados métodos de aprendizado por reforço (BRAGA; FERREIRA; LUDERMIR, 2007).

Os problemas em que a saída é um valor real, ou seja, uma variável contínua, são chamados de problema de regressão. Casos em que o objetivo da análise é designar cada vetor de entrada a um número finito de categorias discretas são chamados de problema de classificação. Ambos os problemas de regressão e classificação são problemas de aprendizado supervisionado, em que a partir de dados de entrada e saída, busca-se encontrar um algoritmo que relacione tais dados. Em regressão, a saída é dada por um valor, um número (como o preço de uma casa); já em classificação a saída é dada por uma categoria, por exemplo cliente de alto ou baixo risco, no caso de avaliação de clientes para empréstimos.

As áreas de aplicação dos conhecimentos de aprendizado de máquina são muitas: análise financeira para aprovação de crédito, reconhecimento de fraudes bancárias, mercado de ações, diagnóstico médico, otimização de rede de telecomunicações a partir de padrões de ligações e também em todos os campos da ciência (por exemplo, na astronomia e biologia, onde se tem um grande volume de dados inviáveis de serem analisados manualmente).

Praticamente em todos os campos da ciência se faz necessário o ajuste de dados a modelos. A partir de experimentos são feitas observações e coletas de dados das quais se busca extrair informações com uso de modelos que as expliquem. Este processo de extração de regras gerais a partir de dados particulares é chamado de indução. Devido ao grande volume de dados, a análise manual seria custosa e feita por um número reduzido de pessoas capacitadas para esse fim. Portanto, há um crescente interesse em algoritmos computacionais capazes de analisar dados e extrair informações deles, ou seja, aprender com eles.

Em aprendizado de máquina se usam teorias de estatística para a construção de modelos matemáticos a partir de um conjunto de dados de treinamento a fim de se otimizar um critério de performance. O modelo obtido pode ser preditivo (usado para prever comportamentos futuros) ou descritivo (usado para extrair informações dos dados de treinamento ou exemplos).

A seguir, serão descritas a partir de exemplos uma série de situações em que se

pode usar algoritmos de aprendizado de máquina (ALPAYDIN, 2014).

**Exemplo 1.1. Reconhecimento facial**
Uma aplicação de algoritmo de reconhecimento de padrões é o de reconhecimento facial. O desafio vem das diferentes possibilidades de ângulos, iluminação e cortes de cabelo, por exemplo. Através da análise de diferentes amostras de imagens de uma determinada pessoa o programa aprende os padrões de estrutura e simetria, e passa a reconhecer por meio da busca de tais padrões em imagens fornecidas. No problema de reconhecimento facial, a entrada é uma imagem, as classes são compostas pelas pessoas a serem identificadas e o algoritmo deve associar as imagens à identidade da pessoa. Este é um programa relativamente complexo dadas as variações de posições, iluminação e ainda a possibilidade de omissão de algumas características (ocultadas, por exemplo, por óculos e barba).

**Exemplo 1.2. Análise da cesta de compra**
Um exemplo de uso de aprendizado de máquina no varejo seria a análise da cesta de compras, que consistiria em encontrar relação entre produtos comprados. Ou seja, seria encontrar uma regra de associação do tipo: se usualmente um cliente que compra o produto X também compra o Y, então um cliente que compra o produto X e não compra o Y é um potencial cliente de Y. Uma potencial ação seria deixar os dois produtos mais próximos. Essa regra de associação poderia sem entendida como uma probabilidade condicional da forma P(X|Y) (lê-se probabilidade de X dado Y) .

**Exemplo 1.3. Uso no sistema financeiro**
Um exemplo de problema de classificação é o cálculo do score de crédito de cada cliente feito pelos bancos, que é um dado usado na decisão de fornecer, ou não, empréstimo. A partir de informações sobre o cliente (tais como idade, se possui poupança, profissão, histórico financeiro) faz-se sua classificação em alto ou baixo risco. Tal classificação baseia-se em registros de clientes passados em que os dados pessoais, bem como a informação de se o empréstimo foi ou não pago, são usados para inferir tal regra de associação, previsão. Um outro uso de aprendizado de máquina no sistema financeiro seria a identificação de dados que não seguem os padrões (outliers), o que pode ser usado, por exemplo, na detecção de operações fraudulentas.

**Exemplo 1.4. Reconhecimento ótico de caracteres**
Em reconhecimento de padrões também pode-se usar algoritmos de aprendizado de máquina. Pode-se citar o caso de reconhecimento ótico de caracteres (números ou letras, por exemplo), em que se busca identificá-los a partir de imagens. Um uso de tais tecnologias pode ser feito para reconhecimento de letras e palavras a partir de dados escritos à mão. Como a escrita à mão varia de pessoa para pessoa, as palavras podem estar escritas em letras grandes ou pequenas, com caneta ou lápis, ou seja, há muitas imagens que podem corresponder ao mesmo caractere. Como não há uma definição única, por exemplo, da letra

A, tomam-se vários exemplos de escrita e define-se o que é A a partir de características em comum, regularidades advindas de tais exemplos de treinamento.

Uma situação em que se usa este tipo de algoritmo é o reconhecimento de dígitos de CEP escritos à mão nos Estados Unidos da América (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

**Exemplo 1.5. Veículos autônomos**
Outra situação passível de ser resolvida através do uso de métodos aprendizado de máquina seriam os carros autônomos. As entradas seriam os dados obtidos através de sensores e GPS, por exemplo, e a saída seria o ângulo de rotação imprimido ao volante para a navegar sem colidir com objetos e desviar da rota (sendo, portanto, um problema de regressão). Dados para treinamento do algoritmo podem ser obtidos a partir da análise do comportamento de um motorista.

**Exemplo 1.6. Robôs autônomos**
Exemplificando a abordagem de aprendizado por reforço, pode-se citar um robô navegando por uma região com o objetivo de chegar a um ponto específico. Ele escolherá uma direção para se mover a cada instante dentre todas as direções possíveis. Após algumas tentativas, espera-se que ele aprenda a atingir seu objetivo a partir do ponto inicial através de uma sequência de ações no menor tempo e sem colisões com obstáculos.

**Exemplo 1.7. Aplicações em vídeo games**
Jogadores de vídeo game geram grande volume de dados, uma vez que suas ações são salvas. Além de ações e comportamentos, também são importantes os dados de compras de produtos virtuais feitas no jogo. Informações extraídas desses dados são de grande importância para os desenvolvedores a fim de maximizar tais compras. Algoritmos de aprendizado de máquina podem ser usados com o objetivo de se traçar um perfil de cada jogador para poder oferecer produtos mais prováveis de serem adquiridos por cada perfil (BERTENS et al., 2018).

Em jogos multijogadores devido a interação entre os jogadores as ações e erros são imprevisíveis, permitindo oportunidades de ação. Em jogos elaborados para apenas um jogador, essa interação não está presente e o oponente padrão é mais previsível (uma vez que segue uma série de algoritmos manualmente codificados). Uma forma de contornar esse problema e desenvolver um oponente desafiador e balanceado para cada jogador é fazer uso de algoritmos de aprendizado de máquina, usando um conjunto de dados ou exemplos de jogos praticados por um jogador experiente a fim de desenvolver um conjunto de regras e ações (GEISLER, 2004).

Neste trabalho serão apresentados dois algoritmos de aprendizado de máquina: regressão linear simples e regressão logística simples. Apesar de serem algoritmos mais

básicos, não são menos importantes, uma vez que o raciocínio empregado em seu desenvolvimento serve de ponto de partida para algoritmos mais rebuscados e complexos. Ao longo do texto, busca-se trazer uma abordagem mais palpável através da contextualização com exemplos para que professores possam fazer uso dos conteúdos aqui abordados com alunos do ensino médio.

Uma breve apresentação da organização do trabalho se faz necessária. No Capítulo 2, apresenta-se regressão linear simples que é uma ferramenta utilizada em problemas de aprendizagem supervisionada nos quais a variável resposta é do tipo quantitativa. Ainda no Capítulo 2, será apresentado o procedimento operacional para a construção do modelo de regressão linear e ainda para a avaliação de sua acurácia. No Capítulo 3 abordam-se conceitos relacionados a regressão logística simples que é uma ferramenta utilizada em problemas de aprendizagem supervisionada nos quais a variável resposta é do tipo classificatória, bem como procedimentos operacionais para a obtenção do modelo de regressão logística e sua análise. Por fim, no Capítulo 4 será apresentada uma sugestão de sequência didática para a construção de um modelo de regressão linear aplicando conceitos vistos no Capítulo 2.

# 2 Regressão Linear Simples

Regressão linear é uma ferramenta simples e útil na predição de respostas quantitativas, por isso é uma abordagem usada em aprendizado supervisionado e serve de ponto de partida para generalizações e abordagens mais complexas.

Caso o leitor não esteja bem familiarizados com alguns termos estatísticos utilizados ao longo do texto, no Apêndice A estão apresentados alguns conceitos. Ainda há necessidade de conhecimento prévio dos seguintes conceitos: otimização de funções de duas variáveis, pontos críticos, distribuição condicional, variância, distribuição t de Student,$valor - p$ e gráfico de quantis. Para familiarização com tais conceitos, sugere-se consulta de tais conceitos materiais (GUIDORIZZI, 1998), (DEVORE, 2010) e (MORETTIN; BUSSAB, 2010).

Como um exemplo de problema em que se pode usar a abordagem de regressão linear, ao longo desta seção será trabalhada a questão da relação entre a área destinada a um determinado produto na prateleira de um supermercado com a receita de vendas desse produto.

**Exemplo 2.1.** (Vendas em um supermercado) Seja $x_i$ o valor, em metros quadrados, da área destinada a um determinado produto em uma prateleira e seja $y_i$, em centenas de reais, a receita de vendas desse produto. Neste caso, $x_i$ é a variável independente e $y_i$ é a variável dependente. Os dados das observações em ordem crescente de área de prateleira estão na Tabela 1 e foram amostrados com frequência semanal.

Algumas perguntas interessantes podem ser feitas quando se tem um conjunto de dados que se deseja relacionar:

1. Quão forte é a relação entre os dados que se deseja relacionar? Contextualizando: a receita de vendas do produto será muito ou pouco impactada pela variação na área de prateleira destinada a ele? Ou seja, sendo destinada uma certa área para exposição do produto, com qual exatidão poderia ser estimada a receita de vendas?

2. Quão acurada é a estimativa obtida? Para o exemplo, qual o aumento na venda do bem para cada metro quadrado destinado a alocação do item estudado? Com que exatidão esse incremento pode ser previsto?

3. Qual a acurácia das previsões feitas usando o modelo obtido? Para um determinado estado (posição da prateleira, espaço ocupado pelo item analisado, nível de preço) qual a exatidão na previsão da receita de vendas?

Tabela 1 – Receita de vendas (y) de um produto de acordo com a área de prateleira (x).

| Observação($i$) | $x_i$ | $y_i$ |
|---|---|---|
| 1 | 0,40 | 1,52 |
| 2 | 0,42 | 1,71 |
| 3 | 0,48 | 1,38 |
| 4 | 0,51 | 1,48 |
| 5 | 0,57 | 2,02 |
| 6 | 0,60 | 2,33 |
| 7 | 0,70 | 2,00 |
| 8 | 0,75 | 2,30 |
| 9 | 0,75 | 2,24 |
| 10 | 0,78 | 2,13 |
| 11 | 0,84 | 2,50 |
| 12 | 0,95 | 3,30 |
| 13 | 0,99 | 2,98 |
| 14 | 1,03 | 2,97 |
| 15 | 1,12 | 3,55 |
| 16 | 1,15 | 3,68 |
| 17 | 1,20 | 4,26 |
| 18 | 1,25 | 4,18 |
| 19 | 1,25 | 4,32 |
| 20 | 1,28 | 3,71 |
| 21 | 1,30 | 4,77 |
| 22 | 1,34 | 3,62 |
| 23 | 1,37 | 4,49 |
| 24 | 1,40 | 4,25 |
| 25 | 1,43 | 4,60 |
| 26 | 1,46 | 4,68 |
| 27 | 1,49 | 4,27 |
| 28 | 1,55 | 4,84 |
| 29 | 1,58 | 4,71 |
| 30 | 1,60 | 5,42 |

Fonte: Autor.

4. Há relação linear entre o conjunto de dados disponíveis? No exemplo estudado, uma pergunta imediata seria: há relação linear entre a área ocupada pelo produto e a receita de vendas?

5. A relação entre os dados é linear? Analisando o gráfico de dispersão observa-se uma relação entre a área de prateleira destinada ao item e sua venda como sendo aproximadamente linear?

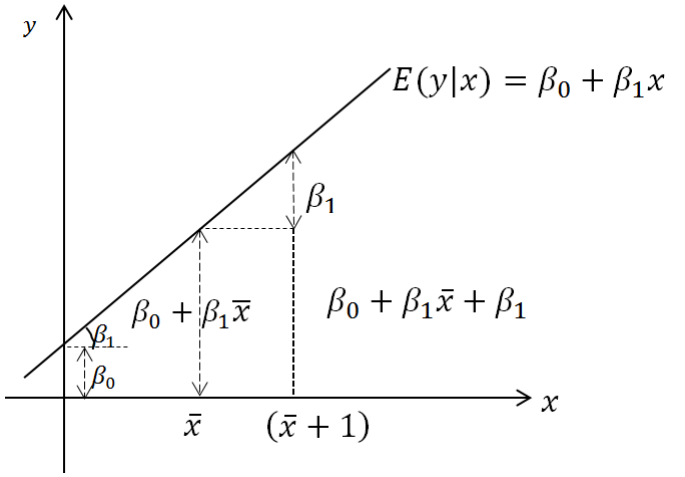
O modelo de regressão linear simples é uma abordagem estatística direta usada na predição de uma resposta $y$ em relação a uma única variável preditiva $x$, em que se assume uma relação linear aproximada entre $x$ e $y$. Matematicamente, escreve-se essa relação da

seguinte forma

$$y \approx \beta_0 + \beta_1 x, \tag{2.1}$$

em que $\beta_0$ e $\beta_1$ são chamados de coeficientes do modelo ou de parâmetros do modelo.

Figura 1 – Representação do modelo de Regressão Linear.




Fonte: Autor.

Uma interpretação dos parâmetros $\beta_0$ e $\beta_1$ pode ser vista na Figura 1. É possível notar que o valor estimado para $y_i$ quando $x_i = 0$ é de $y_i = \beta_0$. Para incrementos de uma unidade na variável de entrada é esperado um incremento de $\beta_1$ na variável de saída.

Uma forma de se inferir que a relação entre as variáveis dependente e independente é linear consiste na construção do diagrama de dispersão. Para construir o diagrama, dispõem-se os pares ordenados $(x_i, y_i)$ dos dados de treinamento em um plano cartesiano de tal forma que no eixo das abcissas corresponda às variáveis independentes e o eixo das ordenadas corresponda às variáveis dependentes (CORREA, 2003). Para o problema proposto no Exemplo 2.1, o diagrama de dispersão é apresentado na Figura 2.

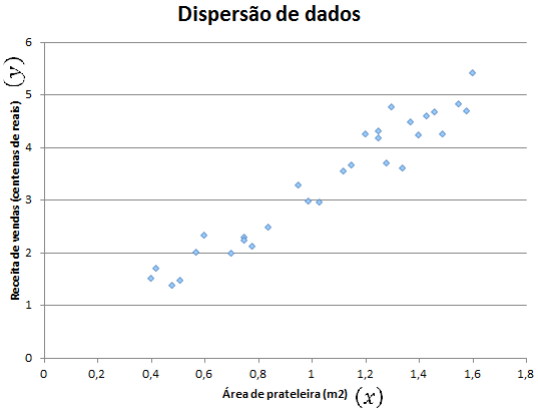
Observando o diagrama de dispersão apresentado na Figura 2 é possível notar uma tendência de relação linear entre os dados de treinamento. Procede-se então à construção do modelo de regressão linear para obter os estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ para os coeficientes $\beta_0$ e $\beta_1$, respectivamente. Ao longo do texto será usada a notação com acento circunflexo " ˆ " para indicar os valores estimados ou preditos.

Os dados de treinamento são utilizados para determinar estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ dos parâmetros $\beta_0$ e $\beta_1$. Com estes estimadores é possível se fazer predições para situações futuras tomando

$$\hat{y} \approx \hat{\beta}_0 + \hat{\beta}_1 x,$$

em que $\hat{y}$ indica a previsão de $Y$ dado $X = x$.

Figura 2 – Representação do Diagrama de Dispersão para o problema apresentado no Exemplo 2.1.




Fonte: Autor.

Para cada par ordenado dos dados amostrais $(x_i, y_i), i = 1, ..., n$ pode-se definir o i-ésimo erro ou desvio $e_i$

$$e_i = y_i - (\beta_0 + \beta_1 x)$$

como a diferença entre a i-ésima resposta observada e o i-ésimo valor de resposta dada pelo modelo linear.

Ainda, define-se a soma dos quadrados dos erros ou desvios $(SQ)$ para uma amostra contendo $n$ pares da forma $(x_i, y_i)$ como

$$SQ(\beta_0, \beta_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x)]^2. \tag{2.2}$$

Note que $SQ$ é uma função dos parâmetros $\beta_0$ e $\beta_1$. Para cada valor atribuído a estes parâmetros será encontrado um resultado para essa soma.

Há muitas formas de estimar os coeficientes do modelo de regressão linear, a fim de se obter coeficientes que apresentem baixos valores de desvios. A abordagem que foi usada nesse texto para a estimação dos coeficientes do modelo linear envolve a minimização dos erros quadrados entre os valores preditos e observados.

## 2.1 Método dos Mínimos Quadrados

Da definição de $SQ$ dada na equação (2.2) tem-se que a soma dos quadrados dos erros pode ser escrita como

$$SQ = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2.$$

Como já mencionado, é possível notar que $SQ$ é uma função de duas variáveis: $\beta_0$ e $\beta_1$. A fim de encontrar os estimadores que torna mínimo $SQ$, primeiramente serão encontrados os pontos críticos $\hat{\beta}_0$ e $\hat{\beta}_1$ da função. Os pontos críticos são aqueles em que as derivadas parciais de primeira ordem são iguais a zero. A saber

$$\frac{\partial(SQ)}{\partial\beta_0}(\hat{\beta}_0,\hat{\beta}_1)=\sum_{i=1}^{n}[-2][y_i-(\hat{\beta}_0+\hat{\beta}_1x_i)]=0, \tag{2.3}$$

$$\frac{\partial(SQ)}{\partial\beta_1}(\hat{\beta}_0,\hat{\beta}_1)=\sum_{i=1}^{n}[y_i-(\hat{\beta}_0+\hat{\beta}_1x_i)][-2x_i]=0. \tag{2.4}$$

Desenvolvendo as equações (2.3) e (2.4), tem-se que

$$\begin{aligned}
&\frac{\partial(SQ)}{\partial\beta_0}(\hat{\beta}_0,\hat{\beta}_1)=0\\
&\sum_{i=1}^{n}[y_i-(\hat{\beta}_0+\hat{\beta}_1x_i)][-2]=0\\
&\sum_{i=1}^{n}y_i-\sum_{i=1}^{n}\hat{\beta}_0-\sum_{i=1}^{n}\hat{\beta}_1x_i=0\\
&\sum_{i=1}^{n}y_i-n\hat{\beta}_0-\hat{\beta}_1\sum_{i=1}^{n}x_i=0\\
&\hat{\beta}_0=\frac{1}{n}\sum_{i=1}^{n}y_i-\hat{\beta}_1\frac{1}{n}\sum_{i=1}^{n}x_i
\end{aligned}$$

$$\hat{\beta}_0=\bar{y}-\hat{\beta}_1\bar{x}. \tag{2.5}$$

Além disso

$$\begin{aligned}
&\frac{\partial(SQ)}{\partial\beta_1}(\hat{\beta}_0,\hat{\beta}_1)=0\\
&\sum_{i=1}^{n}[y_i-(\hat{\beta}_0+\hat{\beta}_1x_i)][-2x_i]=0\\
&\sum_{i=1}^{n}y_ix_i-\sum_{i=1}^{n}\hat{\beta}_0x_i-\sum_{i=1}^{n}\hat{\beta}_1x_i^2=0\\
&\sum_{i=1}^{n}y_ix_i-\hat{\beta}_0\sum_{i=1}^{n}x_i-\hat{\beta}_1\sum_{i=1}^{n}x_i^2=0\\
&\sum_{i=1}^{n}y_ix_i-(\bar{y}-\hat{\beta}_1\bar{x})(n\bar{x})-\hat{\beta}_1\sum_{i=1}^{n}x_i^2=0\\
&\hat{\beta}_1=\frac{\sum_{i=1}^{n}y_ix_i-n\bar{y}\bar{x}}{\sum_{i=1}^{n}x_i^2-n\bar{x}^2},
\end{aligned} \tag{2.6}$$

em que as médias amostrais $\bar{x}$ e $\bar{y}$ são

$$\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i \quad \text{e} \quad \bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i.$$

Rearranjando os termos do numerador e do denominador de $\hat{\beta}_1$ na equação (2.6), ou seja, fazendo

$$\begin{aligned}\sum_{i=1}^{n} y_i x_i - n\bar{y}\bar{x} &= \sum_{i=1}^{n} y_i x_i - \bar{y}n\bar{x} - \bar{x}n\bar{y} + n\bar{y}\bar{x} \\ &= \sum_{i=1}^{n} y_i x_i - \bar{y}\sum_{i=1}^{n} x_i - \bar{x}\sum_{i=1}^{n} y_i + \sum_{i=1}^{n} \bar{y}\bar{x} \\ &= \sum_{i=1}^{n} [y_i x_i - x_i\bar{y} - \bar{x}y_i + \bar{y}\bar{x}] \\ &= \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})\end{aligned}$$

e

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

Segue que os estimadores encontrados através da minimização dos erros quadrados são

$$\hat{\beta}_1 = \frac{\Sigma_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\Sigma_{i=1}^{n}(x_i - \bar{x})^2} \tag{2.7}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}. \tag{2.8}$$

Para mostrar que os pontos críticos encontrados são os pontos de mínimo da função quadrática equação (2.2) observa-se que

$$\begin{aligned}SQ &= \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 \\ &= \sum_{i=1}^{n} [y_i + (\hat{\beta}_0 - \hat{\beta}_0) + (\hat{\beta}_1 x_i - \hat{\beta}_1 x_i) - (\beta_0 + \beta_1 x_i)]^2 \\ &= \sum_{i=1}^{n} \left[(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) + (\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1)x_i\right]^2 \\ &= \sum_{i=1}^{n} \left[e_i + (\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1)x_i\right]^2 \\ &= \sum_{i=1}^{n} \left[e_i^2 + 2((\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1)x_i)e_i + ((\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1)x_i)^2\right] \\ &= \sum_{i=1}^{n} e_i^2 + 2(\hat{\beta}_0 - \beta_0)\sum_{i=1}^{n} e_i + 2(\hat{\beta}_1 - \beta_1)\sum_{i=1}^{n} x_i e_i + \sum_{i=1}^{n} [(\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1)x_i]^2.\end{aligned}$$

Das equações (2.3) e (2.4) segue que

$$2(\hat{\beta}_0 - \beta_0)\sum_{i=1}^{n} e_i = 0$$
$$2(\hat{\beta}_1 - \beta_1)\sum_{i=1}^{n} x_i e_i = 0.$$

Logo, para minimizar $SQ$ basta minimizar a parcela

$$\sum_{i=1}^{n}[(\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1)x_i]^2. \tag{2.9}$$

O mínimo valor da equação (2.9) é zero, já que se trata de uma função quadrática. Logo

$$(\hat{\beta}_0 - \beta_0) + (\hat{\beta}_1 - \beta_1)x_i = 0$$

para todo $i = 1, \ldots, n$. A equação da reta é constante e igual a zero, e isso apenas é possível se seus coeficientes forem nulos. Isso implica que

$$\hat{\beta}_0 - \beta_0 = 0 \Rightarrow \hat{\beta}_0 = \beta_0 \quad \text{e} \quad \hat{\beta}_1 - \beta_1 = 0 \Rightarrow \hat{\beta}_1 = \beta_1$$

Portanto, os valores de $\hat{\beta}_1$ e $\hat{\beta}_0$ dados pelas equações (2.7) e (2.8) são pontos de mínimo.

O modelo de regressão ajustado é representado por

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i, i = 1, ..., n. \tag{2.10}$$

Para o problema apresentado no Exemplo 2.1 os estimadores encontrados foram $\hat{\beta}_1 = 3,08$ e $\hat{\beta}_0 = 0,10$. A equação que representa a relação entre as variáveis dependente e independente é dada por

$$\hat{y}_i = 0,10 + 3,08x_i. \tag{2.11}$$

Como a variável $y$ está mensurada em centenas de reais, uma interpretação dos valores encontrados pela aplicação da equação (2.11) é a seguinte: caso a área de prateleira destinada ao item seja nula (por exemplo, caso em que o produto esteja em display apenas próximo ao caixa), haverá um total de 10 reais em receita com a venda deste produto. Além disso, a cada incremento de uma unidade na quantidade de área destinada para a disposição do produto, haverá um aumento de 308 reais em receita de venda do item.

Pode-se, agora definir cada resíduo $\hat{e}_i$ para o modelo representado pela equação (2.10)

$$\hat{e}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i, i = 1, ..., n.$$

A partir da definição de resíduo, é possível se definir a soma dos quadrados dos resíduos ($SQRes$)

$$SQRes = \sum_{i=1}^{n} \hat{e}_i^{\,2} = \sum_{i=1}^{n}(y_i - \hat{y}_i) \tag{2.12}$$

É importante observar que se o valor do estimador $\hat{\beta}_1$ for nulo, a resposta esperada para previsões futuras de $\hat{y}$ será $\bar{y}$. Dessa forma, o modelo linear se reduz a um modelo mais simples.

## 2.2 Análise do Modelo

Assume-se que a verdadeira relação entre as variáveis aleatórias $X$ e $Y$ seja linear e do tipo

$$y = \beta_0 + \beta_1 x + \epsilon, \tag{2.13}$$

em que $\epsilon$ é um termo de erro aleatório de média zero independente de X. O termo de erro engloba todas as inadequações relacionadas à suposição de relação linear entre as variáveis $X$ e $Y$: possibilidade de haver erro nas medições de $X$ e $Y$, influência de outras variáveis preditoras além de $X$ e, inclusive, a possibilidade de que a suposição de relação linear entre essas grandezas ser falsa.

A equação (2.13) representa a linha de regressão populacional, e os coeficientes são os que melhor relacionam as grandezas $X$ e $Y$. Mas, como na maioria dos eventos estudados, os dados populacionais não são conhecidos, os coeficientes $\beta_0$ e $\beta_1$ não podem ser calculados. Já os coeficientes estimados pelo método de mínimos quadrados, que representam a reta de mínimos quadrados, sempre podem ser calculados mas dependem dos dados amostrais usados na sua estimação (o tamanho da amostra, quais dados amostrais foram usados).

A última afirmação fica mais compreensível quando, por exemplo, se calculam os coeficientes das retas de mínimos quadrados para alguns subconjuntos dos dados amostrais. Para exemplificação, foram tomadas três amostras de tamanho $n = 10$.

Tabela 2 – Receita de vendas ($y_i$) de um produto de acordo com a área de prateleira ($x_i$) para várias amostras.

| Amostra 1 | | Amostra 2 | | Amostra 3 | |
|---|---|---|---|---|---|
| $x_i$ | $y_i$ | $x_i$ | $y_i$ | $x_i$ | $y_i$ |
| 0,40 | 1,52 | 0,84 | 2,50 | 1,30 | 4,77 |
| 0,42 | 1,71 | 0,95 | 3,30 | 1,34 | 3,62 |
| 0,48 | 1,38 | 0,99 | 2,98 | 1,37 | 4,49 |
| 0,51 | 1,48 | 1,03 | 2,97 | 1,40 | 4,25 |
| 0,57 | 2,02 | 1,12 | 3,55 | 1,43 | 4,60 |
| 0,60 | 2,33 | 1,15 | 3,68 | 1,46 | 4,68 |
| 0,70 | 2,00 | 1,20 | 4,26 | 1,49 | 4,27 |
| 0,75 | 2,30 | 1,25 | 4,18 | 1,55 | 4,84 |
| 0,75 | 2,24 | 1,25 | 4,32 | 1,58 | 4,71 |
| 0,78 | 2,13 | 1,28 | 3,71 | 1,60 | 5,42 |

Fonte: Autor.

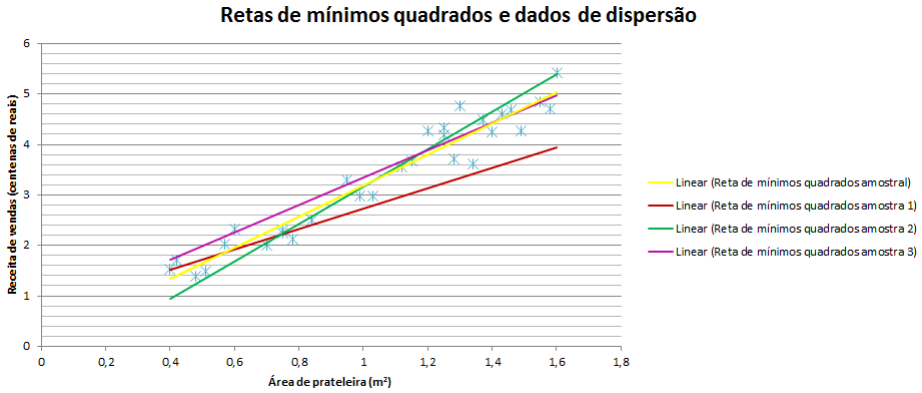
Os estimadores encontrados para cada amostra estão na Tabela 3. A Figura 3 ilustra a representação gráfica das retas de mínimos quadrados.

Tabela 3 – Estimadores calculados para cada amostra.

| Estimador | Amostra 1 | Amostra 2 | Amostra 3 |
|---|---|---|---|
| $\hat{\beta}_1$ | 2,02 | 3,71 | 2,72 |
| $\hat{\beta}_0$ | 0,71 | -0,55 | 0,62 |

Fonte: Autor.

Figura 3 – Retas de mínimos quadrados para os dados amostrais completos, amostra 1, amostra 2 e amostra 3.




Fonte: Autor.

É possível notar que os valores estimados de resposta ($\hat{y}$) são diferentes para cada reta de regressão de mínimos quadrados referentes a cada uma das amostras. Pode-se chegar à conclusão de que o valor estimado de resposta é dependente do conjunto de dados usados na estimação dos parâmetros do modelo de regressão. Em alguns casos, o valor estimado de resposta foi subestimado e em outros, superestimado. Isto pode ser observado na Tabela 4 na qual são apresentados os estimadores de resposta a partir do conjunto de dados completo e de seus subconjuntos (amostras 1, 2 e 3).

Tabela 4 – Estimadores de resposta para dois valores de dados amostrais.

| x | y | $\hat{y}$ | $\hat{y}_1$ | $\hat{y}_2$ | $\hat{y}_3$ |
|---|---|---|---|---|---|
| 0,84 | 2,50 | 2,69 | 2,40 | 2,56 | 2,90 |
| 0,99 | 2,98 | 3,15 | 2,71 | 3,12 | 3,31 |

Fonte: Autor.

Apesar dos estimadores dependerem da amostra, mostra-se a seguir que para um conjunto de dados grande, na média, esse estimadores equivalem ao parâmetro estimado.

### 2.2.1 Hipóteses de um Modelo Linear Simples

O modelo de regressão linear simples é fundamentado sobre certas hipóteses, sob as quais é possível mostrar que os estimadores dados pelos métodos dos mínimos quadrados são os melhores estimadores não enviesados. Caso o leitor não esteja bem familiarizados com alguns termos estatísticos utilizados ao longo do texto, no Apêndice A estão apresentados alguns conceitos. A seguir essas hipóteses são apresentadas

1. O modelo é linear nos parâmetros, isto é, os coeficientes $\beta_0$ e $\beta_1$ da equação (2.1) fazem parte de equação de forma linear.
2. A amostragem é aleatória.
3. Os erros são normalmente distribuídos.
4. A esperança condicional do erro é igual a zero, isto é, em termos estatísticos, $E(e_i|X_i) = 0$.
5. A variabilidade dos erros é constante, isto é, $E(e_i^2) = \sigma^2$. Isso significa que a sua variância é a mesma para qualquer X.

Sob as hipóteses acima, vale o seguinte resultado:

**Teorema 2.2.** *(Teorema de Gauss-Markov) Os estimadores dos métodos dos mínimos quadrados são não enviesados e, dentre todos os estimadores não enviesados, apresentam a menor variância (V). Ou seja, se $\hat{\beta}$ é um dos estimadores do parâmetro $\beta$ dado pelo métodos dos mínimos quadrados, tem-se que*

$$E(\hat{\beta}) = \beta.$$

*Além disso, se $\bar{\beta}$ é um outro estimador para o modelo de regressão linear, tem-se que*

$$V(\hat{\beta}) < V(\bar{\beta}).$$

A demostração do Teorema acima pode ser encontrada em (KUTNER et al., 2005).

### 2.2.2 Ajuste do Modelo

A fim de avaliar o modelo obtido, será feita a análise de quão ajustado aos dados é o modelo linear. Esta análise é feita comparando o modelo linear com um modelo mais simples, o modelo de grau zero. O modelo mais simples é obtido quando previsão de resposta é dada pela média dos valores amostrais da variável dependente, ou seja, quando $\hat{\beta}_1 = 0$.

Um primeiro critério de medição de ajuste do modelo é comparar a soma dos quadrados dos resíduos ($SQRes$), dado na equação (2.12), com a soma dos quadrados dos resíduos do modelo mais simples, a qual dada por

$$SQTot = \sum_{i=1}^{n} [y_i - \bar{y}]^2$$

e é chamada de soma de quadrados total. Quando os resíduos são pequenos, tem-se uma indicação de que o modelo está produzindo bons resultados (MORETTIN; BUSSAB, 2010).

Para o problema apresentado no Exemplo 2.1, os valores encontrados foram

$$SQRes = 2,66 \quad \text{e} \quad SQTot = 42,34.$$

Isso mostra uma redução sensível de cerca de 40 unidades.

Um segundo critério para medir o ajuste do modelo é comparar as suas respectivas variâncias residuais. A importância de se comparar os valores estimados de variância para cada modelo reside no fato de que o número de graus de liberdade de cada modelo é diferente. Isto é, o modelo mais simples tem menos parâmetros que o modelo linear. Vale ainda lembrar que a variância mede a dispersão dos dados em torno da média, valores menores de variância indicam mais acurácia na estimação. Caso haja uma diminuição da variância quando comparado ao modelo linear, pode-se concluir que o modelo linear é mais adequado por estar melhor ajustado aos dados.

Para o modelo linear, o estimador $\hat{\sigma_e}^2$ não enviesado da variância residual $\sigma_e^2$ é dado por

$$\hat{\sigma_e}^2 = \frac{SQRes}{n-2}.$$

Para o modelo em que as previsões futuras são feitas pela média, o estimador não enviesado da variância residual $\hat{\sigma}^2$ é

$$\hat{\sigma}^2 = \frac{SQTot}{n-1}.$$

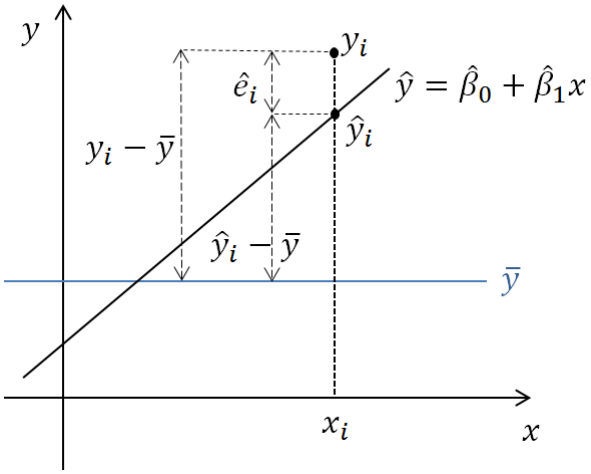
Para os dados do Exemplo 2.1, os valores encontrados para a variância residual foram $\hat{\sigma_e}^2 = 0,09$ e $\hat{\sigma}^2 = 1,46$. Novamente, observa-se uma redução significativa dos valores e pode-se concluir que é mais vantajoso adotar o modelo de linear para a previsão de respostas.

É interessante mencionar o valor que representa a diminuição da soma dos quadrados dos resíduos do modelo linear quando comparada com a soma de quadrados total. Esse valor é chamado soma dos quadrados devido à regressão ($SQReg$) e descrito por

$$SQReg = SQTot - SQRes.$$

Uma melhor visualização do que representa cada um desses resíduos pode ser observado na Figura 4. O resíduo gerado pela estimação através do modelo de regressão linear é $\hat{e}_i$, enquanto que o resíduo gerado pelo modelo mais simples é dado por $y_i - \bar{y}$. Desta forma, $\hat{y}_i - \bar{y}$ pode ser compreendido como uma medida de melhoria no valor predito pelo modelo de regressão linear quando comparado ao modelo mais simples. Note que $\hat{e}_i$, $y_i - \bar{y}$ e $\hat{y}_i - \bar{y}$ representam a i-ésima parcela de $SQRes$, $SQTot$ e $SQReg$, respectivamente. Para um detalhamento maior, ver (MORETTIN; BUSSAB, 2010).

Figura 4 – Representação dos resíduos.




Fonte: Autor.

Ainda através da análise da Figura 4, pode-se entender $SQRes$ como uma medida da dispersão que não pode ser explicada pelo modelo de regressão encontrado e $SQReg$ como sendo uma medição da dispersão que é explicada pelo modelo após executada a regressão linear.

Uma outra forma de analisar o ganho com a introdução do modelo linear é por meio da estatística $R^2$. A estatística pode ser entendida como o ganho percentual de quão mais ajustado é esse modelo quando comparado com a previsão da resposta dada pela média, e é calculada da seguinte forma

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

Desta forma, a estatística $R^2$ propicia uma alternativa de medida de qualidade de ajuste do modelo, uma vez que usa a razão entre a qualidade do ajuste amostral $SQReg$ e o quadrado das somas dos erros total $SQTot$. Assim, os valores encontrados através dessa estatística estarão entre 0 e 1, o que facilita sua interpretação. Uma estatística $R^2$ próxima de 1 indica que grande parte da dispersão pode ser explicada pelo modelo. Enquanto, um valor de $R^2$ próximo de 0, indica que a regressão não foi capaz de explicar a variabilidade na resposta.

Para o problema do Exemplo 2.1, o valor de $SQReg$ e da estatística $R^2$ são respectivamente $39,68$ e $0,94$. Pode-se concluir que o modelo de regressão linear explica 94% da variabilidade total.

Na Tabela 5 são reunidas algumas das estatísticas que permitem avaliar o modelo de regressão linear desenvolvido para o problema apresentado no Exemplo 2.1. Comparando os valores de $SQTot$ e $SQRes$ nota-se que o modelo com a variável independente contribui para a redução dos erros de predição. Na verdade, o impacto na redução dos erros de predição devido à introdução da variável preditiva no modelo de regressão pode ser observado através da análise de $SQReg$ de forma absoluta, e através de $R^2$ de forma percentual.

Tabela 5 – Grandezas relacionadas ao modelo de Regressão Linear.

| Quantidade | Valor |
|---|---|
| $SQTot$ | 42,34 |
| $SQRes$ | 2,66 |
| $SQReg$ | 39,68 |
| $R^2$ | 0,94 |

Fonte: Autor.

### 2.2.3 Intervalos de Confiança para os Parâmetros

Passa-se agora para a análise dos estimadores encontrados pelo método dos mínimos quadrados e também dos parâmetros $\beta_1$ e $\beta_0$. Para tanto, os valores esperados de cada estimador $E(\hat{\beta}_0)$ e $E(\hat{\beta}_1)$ e suas respectivas variâncias $V(\hat{\beta}_0)$ e $V(\hat{\beta}_1)$ são

$$E(\hat{\beta}_0) = \beta_0;$$
$$V(\hat{\beta}_0) = \sigma_e^2 \left[ \frac{\sum_{i=1}^{n}(x_i^2)}{n\sum_{i=1}^{n}(x_i - \bar{x})^2} \right];$$

$$E(\hat{\beta}_1) = \beta_1;$$
$$V(\hat{\beta}_1) = \sigma_e^2 \left[ \frac{1}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \right].$$

Para maiores detalhes dos cálculos dessas propriedades (MORETTIN; BUSSAB, 2010).

Das hipóteses de que o erro e a variável resposta tenham distribuições normais ambas com variância estimada $\sigma_e^2$, tem-se que as estatísticas

$$t(\beta_0) = \frac{\hat{\beta}_0 - \beta_0}{\hat{\sigma_e}^2} \sqrt{\frac{n\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sum_{i=1}^{n}(x_i^2)}};$$
$$t(\beta_1) = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma_e}^2} \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

têm distribuição t de Student com $(n-2)$ graus de liberdade. Com essas estatísticas é possível fazer algumas afirmações sobre os parâmetros $\beta_0$ e $\beta_1$ usando intervalos de confiança. A saber

$$IC(\beta_0;\gamma) = \hat{\beta}_0 \pm t_\gamma(n-2)\hat{\sigma}_e\sqrt{\frac{\sum_{i=1}^{n}(x_i^2)}{n\sum_{i=1}^{n}(x_i-\bar{x})^2}};$$
$$IC(\beta_1;\gamma) = \hat{\beta}_1 \pm t_\gamma(n-2)\hat{\sigma}_e\sqrt{\frac{1}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

em que $\gamma$ é o nível ou coeficiente de confiança e $(n-2)$ é o número de graus de liberdade da distribuição t bicaudal.

Voltando ao problema apresentado no Exemplo 2.1, para $\gamma = 0,95$, $(n-2) = 28$ com valor $t_{0,95}(28) = 2,048$, há 95% de probabilidade de que o valor do parâmetro $\beta_1$ esteja no intervalo $[2,77, 3,34]$. Ainda, sobre $\beta_0$ há 95% de probabilidade de que seu verdadeiro valor esteja no intervalo $[-0,24, 0,44]$. Assim, pode-se concluir que caso a área de prateleira dedicada à exposição do item seja nula será esperada uma receita de venda deste produto no intervalo de $-24$ a 44 reais, mas como o valor de receita negativa não é possível, espera-se que a receita de vendas deste produto esteja entre 0 e 44 reais. Ainda, para cada incremento de $1m^2$ de área de prateleira dedicada à exposição do produto, espera-se um aumento médio na receita de vendas entre 277 e 334 reais.

### 2.2.4 Intervalos de Confiança e de Predição

Outro ponto interessante a ser analisado é a diferença entre intervalo de confiança para a predição de valores médios da variável resposta com base no modelo de regressão encontrado e intervalo de predição para uma observação de resposta dado um valor de variável independente.

Quando fala-se em intervalo de confiança para a previsão da resposta, está se pensando na distribuição amostral da variável resposta e o intervalo de confiança é dado por

$$IC\left(E(Y|x);\gamma\right) = \hat{y}_i \pm t_\gamma(n-2)\hat{\sigma}_e\sqrt{\frac{1}{n}+\frac{(x_i-\bar{x})^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}}.$$

A fim de maior aprofundamento sobre os cálculos, buscar (MORETTIN; BUSSAB, 2010). Do cálculo do intervalo de confiança para a previsão da resposta se obtém o valor médio esperado de resposta, dada uma entrada da variável independente.

Para o problema apresentado no Exemplo 2.1, e supondo que a área destinada ao produto tenha sido de $x = 0,75m^2$, a estimativa pontual de receita de venda calculada pela equação (2.11) é de $\hat{y} = 2,41$, em centenas de reais. Considerando como resposta procurada a receita média de venda do item, em centenas de reais, para um nível de

confiança $\gamma = 0,95$, o intervalo de confiança calculado é $]2,26, 2,56[$. Em outras palavras, para semanas em que a área de prateleira destinada ao produto tenha sido de $x = 0,75m^2$, espera-se em média que a receita de venda com determinado produto pertença ao intervalo 226 a 256 em reais.

O intervalo de predição para uma futura observação, para um dado nível $x_f$, é dado por

$$IP\left(Y_f;\gamma\right) = \hat{y_f} \pm t_\gamma \hat{\sigma_e} \sqrt{1 + \frac{1}{n} + \frac{(x_f - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}},$$

e pode ser entendido como o intervalo em que se espera encontrar o valor de uma resposta dado um valor de entrada.

Para o problema de área de prateleira destinada a um produto em um supermercado (Exemplo 2.1) e ainda supondo que a área destinada à mercadoria tenha sido de $x = 0,75m^2$, o intervalo de predição com 95% de confiança é $]1,76, 3,06[$. Isto significa que caso em uma determinada semana o valor de área de prateleira destinado à mercadoria tenha sido de $x = 0,75m^2$ o valor provável de receita de vendas deste produto nesta semana pertença ao intervalo de 176 a 306 reais.

Vale ressaltar que o intervalo de predição será sempre maior do que o intervalo de confiança para um determinado valor de variável independente, como pôde ser observado pelos cálculos da situação em que $x = 0,75m^2$. Isso ocorre pois com o cálculo do intervalo de confiança encontram-se valores de resposta esperada média a dado valor de variável independente, enquanto que com o cálculo do intervalo de predição encontram-se valores a que provavelmente pertencerá a resposta para uma única observação.

### 2.2.5 Teste de Hipóteses

Outro procedimento que se pode realizar para validação do modelo de regressão linear calculado é o teste de hipóteses. Na realização deste teste faz-se uso do cálculo do desvio padrão estimado para cada um dos estimadores $\hat{\beta_0}$ e $\hat{\beta_1}$ a fim de se avaliar a relevância da relação linear entre $X$ e $Y$. O teste de hipótese mais comum envolve testar a hipótese nula da seguinte forma

$H_0$: não há relação linear entre X e Y

*versus* a hipótese alternativa

$H_a$ : há relação linear entre X e Y.

Matematicamente, essas relações correspondem a

$$H_0 : \beta_1 = 0,$$
$$H_a : \beta_1 \neq 0.$$

Caso a hipótese nula seja verdadeira, a regressão linear toma a forma $Y = \beta_0 + \epsilon$ e a variável independente $X$ não está associada a $Y$, sendo mais interessante prever valores futuros de $y$ através de sua média amostral $\bar{y}$.

O teste de hipótese acima proposto será usado para determinar se o estimador $\hat{\beta}_1$ é suficientemente diferente de zero de tal forma a garantir que o parâmetro $\beta_1$ é diferente de zero. Essa análise é dependente da acurácia de $\hat{\beta}_1$, ou seja, depende do desvio padrão $\hat{\sigma}(\hat{\beta}_1)$ de $\hat{\beta}_1$, dado por $\sqrt{V(\hat{\beta}_1)}$. Se $\hat{\sigma}(\hat{\beta}_1)$ é pequeno, então mesmo valores pequenos de $\hat{\beta}_1$ indicam fortes evidências de que $\beta_1$ é não nulo. Caso $\hat{\sigma}(\hat{\beta}_1)$ seja grande, então $\hat{\beta}_1$ deve ser grande em valor absoluto a fim de que possamos rejeitar a hipótese nula $H_0$.

O procedimento prático para a realização dessa análise é supor que o parâmetro $\beta_1$ seja nulo. Assim, será feira a análise do valor $t$, calculado por

$$t = \frac{\hat{\beta}_1 - 0}{\hat{\sigma}(\hat{\beta}_1)}.$$

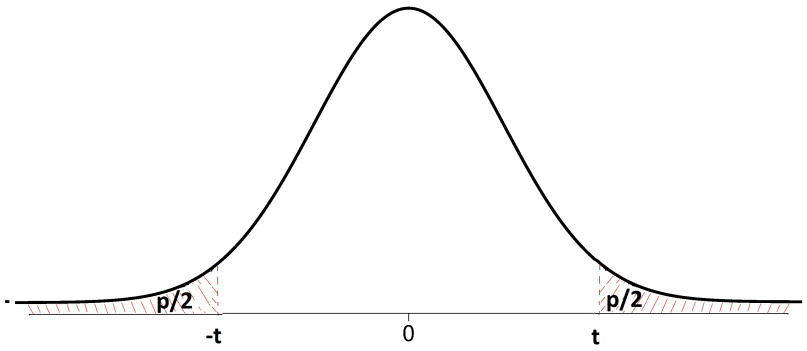
A estatística $t$, que tem distribuição t de Student com $(n-2)$ graus de liberdade, pode ser interpretada como uma medição da distância de $\hat{\beta}_1$ do valor zero (valor do parâmetro $\beta_1$, por suposição) em quantidade de desvios padrão. Quanto maior for o valor encontrado para $t$, maior a evidência de que a hipótese nula é falsa, ou seja, menor é a probabilidade de $\beta_1$ ter valor nulo.

Basta agora consultar a tabela de distribuição t de Student bicaudal com $(n-2)$ graus de liberdade para o cálculo da probabilidade de se observar valores iguais ou maiores do que $|t|$, encontradas por exemplo em (DEVORE, 2010). A Figura 5 traz uma representação de como a estatística $t$ está relacionada com o $valor - p$. É possível notar que o $valor - p$ está associado à área sob a curva da distribuição t de Student. Uma descrição de como estes valores se relacionam é dizer que a probabilidade de se observar valores maiores do que ou iguais a $|t|$ é igual ao $valor - p$.

Baixos valores do $valor - p$ indicam uma baixa probabilidade de que se possa observar uma associação tão grande entre as variáveis analisadas devido ao acaso, ou seja, é baixa a probabilidade de que não haja relação entre as grandezas preditivas e predita. Ou seja, para baixos valores do $valor - p$ pode-se inferir que há uma relação entre as grandezas predita e preditiva e pode-se rejeitar a hipótese nula.

Grande parte das vezes não será possível encontrar o $valor - p$ exato para $t$ pelo uso das tabelas, mas para um nível de significância $\alpha$ de interesse (exemplos de valor de

Figura 5 – Representação da relação entre o valor de t e do $valor - p$ associado.




Fonte: Autor.

significância comumente buscados são 0,05 ou 0,01) é possível encontrar o respectivo valor crítico $t_\alpha$ e comparar com o valor $t$ encontrado. Caso o valor $t$ seja muito maior do que $t_\alpha$ para um dado $\alpha$ da distribuição t com (n-2) graus de liberdade, pode-se concluir que o $valor - p$ será muito menor do que $\alpha$.

Nesse trabalho será usada a escala de significância de Fischer, apresentada na Tabela 6, a fim de se testar a hipótese nula $H_0$ contra a hipótese alternativa $H_a$. Caso o $valor - p$ seja bem pequeno há fortes evidências contra $H_0$, ou seja, há fortes razões para a rejeição de $H_0$.

Tabela 6 – Escala de Fischer.

| valor-p | 0,10 | 0,05 | 0,025 | 0,01 | 0,005 | 0,001 |
|---|---|---|---|---|---|---|
| força da evidência | limítrofe | moderada | substancial | forte | muito forte | fortíssima |

Fonte: (EFRON; GOUS, 1997)

### 2.2.6 Análise dos Resíduos

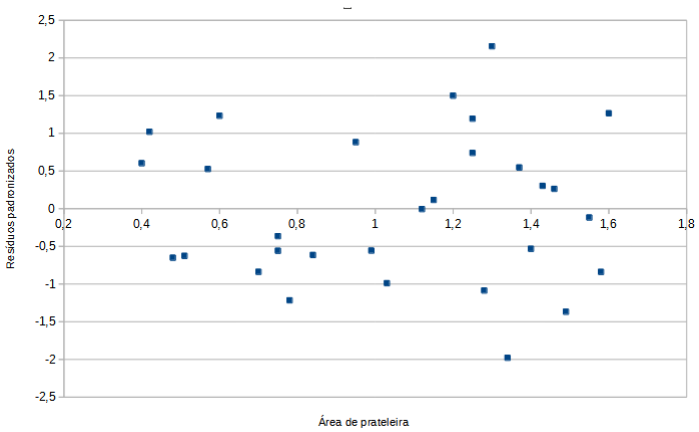
Uma outra análise possível de ser feita sobre os dados para a verificação de distribuição dos erros de forma aleatória em torno do zero, é a elaboração de gráficos de resíduos. Plota-se o gráfico cuja abcissa é dada pela variável independente e a ordenada pelo respectivo valor de resíduo estimado $\hat{e}_i$, ou resíduo padronizado $\hat{z}_i$

$$\hat{z}_i = \frac{\hat{e}_i}{\hat{\sigma}_e}.$$

Para o problema de área de prateleira de um item de supermercado, o gráfico de resíduos padronizado pode ser visto na Figura 6. Pode-se concluir da análise do gráfico que os resíduos estão distribuídos aleatoriamente em torno do zero. Caso fosse observada

região com acúmulo de pontos, regiões de alta densidade, isto seria um indicador de que a hipótese de variabilidade constante dos erros não é satisfeita para o conjunto de dados amostrais. Caso fosse observada tendência na distribuição dos dados (uma distribuição quadrática, por exemplo), isto seria um indicador da possibilidade de inadequação da hipótese de que as variáveis independente e dependente se relacionem de forma linear. Caso houvesse presença de valores *outliers*, isto seria indicador da necessidade de verificação da possibilidade de haver problemas com os métodos de medição das grandezas, ou da necessidade da adoção de métodos mais robustos.

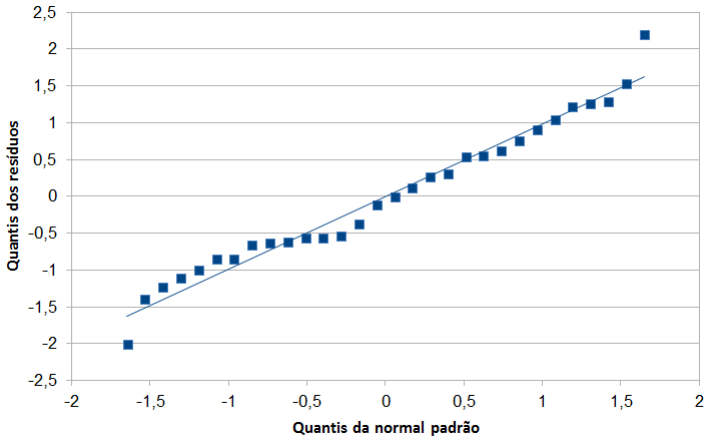
Figura 6 – Representação dos resíduos padronizados.




Fonte: Autor.

A fim de se verificar a hipótese de distribuição normal dos resíduos, faz-se a análise do gráfico de quantis, $q \times q$, na Figura 7, em que se faz comparação dos valores dos resíduos normalizados com valores de uma distribuição normal padrão. Caso os pontos estejam suficientemente próximos da reta e com distribuição simétrica em torno dela, pode-se concluir que a hipótese de distribuição normal dos erros é satisfeita.

Figura 7 – Gráfico $q$x$q$ para os resíduos do modelo.




Fonte: Autor.

É possível notar do gráfico $q \times q$ que os resíduos estão normalmente distribuídos. Para mais informações, consulte (MORETTIN; BUSSAB, 2010).

Serão retomadas e respondidas as perguntas motivadoras do início da seção.

1. Quão forte é a relação entre os dados que se deseja relacionar?
   Para responder esta pergunta, pode-se fazer a análise da estatística $R^2$, como visto na Seção 2.2.5. Para o problema analisado tem-se que $R^2 = 0,94$, de onde é possível concluir que o modelo de regressão calculado explica 94% da variabilidade total. Assim, conclui-se que há sim uma forte relação entre a área de prateleira ocupada pelo produto e sua receita de vendas.

2. Quão acurada é a estimativa obtida?
   A fim de se avaliar a acurácia das estimativas analisam-se os intervalos de confiança para cada um dos parâmetros estimados, como visto na Seção 2.2.3.

3. Qual a acurácia das previsões feitas usando o modelo obtido?
   Pode-se calcular os intervalos de confiança (resposta média) e de predição (intervalo para uma observação) para um dado valor de variável preditiva, como visto na Seção 2.2.4.

4. Há relação entre a área ocupada pelo produto e a receita de vendas?
   Para responder a pergunta procede-se à análise do $valor - p$ encontrado para o valor $t$, como visto na Seção 2.2.5. A hipótese nula $H_0 : \beta_1 = 0$ deverá ser rejeitada. Assim, pode-se concluir que há sim relação entre a área de prateleira ocupada pelo item e sua receita de vendas.

5. A relação entre os dados é linear?
   Com o interesse de responder a esta indagação, faz-se a análise do gráfico de resíduos padronizados, como visto na Seção 2.2.6. Avaliando a Figura 6 é possível perceber que não há padrão de comportamento dos erros padronizados, ou seja, que estão aleatoriamente distribuídos. Assim, pode-se concluir que a relação entre as variáveis independente e dependente é linear. Outro indicador usado para avaliar se há tendência de relação linear entre os dados é o gráfico de dispersão. Ao se observar a Figura 2 é possível notar uma tendência de relação linear entre os dados.

# 3 Regressão Logística Simples

Para problemas em que a variável resposta é do tipo categórica ou qualitativa é adequado se fazer uso de técnicas de predição específicas para modelos de classificação. Um possível método de classificação que pode ser usado em problemas com esta natureza são modelos de regressão logística. Neste tipo de tratamento, pertencente a abordagem de aprendizado supervisionado (pois há um conjunto de dados rotulados de treinamento), a predição se dá através do cálculo de probabilidade de pertencimento da resposta ($y_i$) a cada uma das classes da variável dependente, dado um valor de variável independente ($x_i$) (JAMES et al., 2013).

Para uma melhor compreensão dos conceitos aqui abordados, há necessidade de conhecimento prévio dos seguintes conceitos: função logito, função de verossimilhança e distribuição Qui-Quadrado. Para familiarização com tais conceitos, sugere-se consulta de tais conceitos materiais (DEVORE, 2010) e (MORETTIN; BUSSAB, 2010).

Exemplos de problemas em que se poderia aplicar essa abordagem estatística são a classificação de tumores em maligno ou benigno, classificação de clientes em alto ou baixo risco, e predição de pertencimento a grupo de risco de uma doença dado o estado do paciente.

No desenvolvimento deste trabalho serão considerados apenas problemas em que a variável resposta seja do tipo dicotômica ou binária, ou seja, pode assumir apenas duas classes. Essas classes serão descritas como tendo valor $y = 1$ para a resposta rotulada como resposta de interesse (cliente de alto risco ou tumor maligno, por exemplo) e $y = 0$ caso a resposta não seja a de interesse (cliente de baixo risco ou tumor benigno, por exemplo).

Ao longo desta seção será trabalhado o problema apresentado por (TSUCHIYA, 2002).

**Exemplo 3.1.** (Presença de doença cardíaca em pacientes de acordo com faixa etária) Se busca uma relação entre a idade dos pacientes e a presença ($y = 1$) ou ausência ($y = 0$) de doença cardíaca. Os dados disponíveis são a quantidade de pacientes em que há presença ou ausência da doença, a partir de onde se calculará a frequência de incidência em cada faixa etária. Os dados usados no desenvolvimento do modelo são os apresentados na Tabela 7.

Deseja-se obter um modelo por meio da aplicação do método de regressão logística que relacione as variáveis dependente e independente da seguinte forma

$$P(Y = 1|x = x_i)$$

Tabela 7 – Frequência de incidência de doença cardíaca de acordo com a faixa etária dos pacientes.

| Faixa etária (anos) | Ausência | Presença | Frequência de incidência |
|---|---|---|---|
| 20-29 | 9 | 1 | 0,10 |
| 30-34 | 13 | 2 | 0,13 |
| 35-39 | 9 | 3 | 0,25 |
| 40-44 | 10 | 5 | 0,33 |
| 45-49 | 7 | 6 | 0,46 |
| 50-54 | 3 | 5 | 0,63 |
| 55-59 | 4 | 13 | 0,76 |
| 60-69 | 2 | 8 | 0,80 |
| Total | 57 | 43 | 0,43 |

Fonte: (TSUCHIYA, 2002).

que pode ser lido como a probabilidade de que a variável resposta pertença à categoria $Y = 1$ dado que a variável de entrada (independente) assuma o valor $x = x_i$ e seja representado apenas por $p(x)$. Buscam-se valores entre 0 e 1 pela equação acima, e através de critérios convenientes para a escolha do limiar será feita a classificação de pertencimento ou não à classe $Y = 1$. Por exemplo, na classificação de tumores nas categorias maligno ($Y = 1$) e benigno ($Y = 0$) em função do seu tamanho ($x$), uma possível escolha de limiar é 0,5. Dada esta escolha, pode-se entender que para valores de $P(Y = 1|x = x_i) > 0,5$ o tumor será classificado como sendo maligno e $P(Y = 1|x = x_i) \leq 0,5$ o tumor será classificado como benigno. Caso seja mais conveniente uma abordagem mais conservadora, uma possibilidade seria a escolha de um limiar com valor 0,3.

Uma forma de modelar a probabilidade $P(Y = 1|x)$ é através da função logística

$$p(x) = \frac{e^{\beta_0 + x\beta_1}}{1 + e^{\beta_0 + x\beta_1}},$$

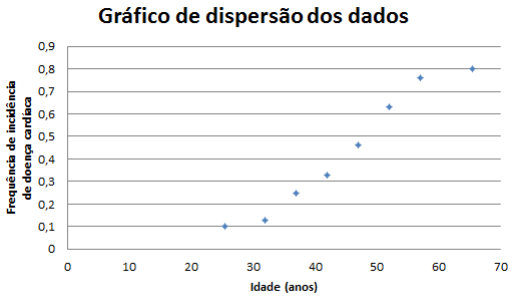
também representada por

$$p(x) = \frac{1}{1 + e^{-(\beta_0 + x\beta_1)}}. \tag{3.1}$$

A escolha pela distribuição logística é dada pelo fato de que a resposta obtida através da transformação está entre 0 e 1 e, por isto, poder ser interpretada como uma probabilidade (JR; LEMESHOW; STURDIVANT, 2013).

A Figura 8 contém o diagrama de dispersão para o problema apresentando no Exemplo 3.1. Ao se analisar a Figura 8 é possível observar uma relação entre as variáveis independente e dependente próxima à relação entre variáveis que se relacionam através da função logística. Nota-se uma aproximação gradativa de 0 e de 1 em seus extremos e ainda é possível notar que incrementos unitários na variável de entrada levarão a incrementos cada

Figura 8 – Representação do Diagrama de Dispersão para o problema apresentado no Exemplo 3.1.




Fonte: Autor.

vez menores na variável de saída quanto mais próximos das extremidades forem tomados estes incrementos. Então, há evidências de que a relação entre as variáveis apresentadas no Exemplo 3.1 pode ser modelada pela função logística.

Uma forma de interpretação da probabilidade obtida pelo modelo de regressão logística é por meio do cálculo da chance $Z$ de ocorrência do evento (FÁVERO et al., 2009). Calcula-se a chance através da razão

$$Z = \frac{p(x)}{1 - p(x)}. \tag{3.2}$$

Substituindo a expressão de $p(x)$ expressa na equação (3.1) na equação (3.2) segue que

$$\begin{aligned} Z &= \frac{\dfrac{1}{1 + e^{-(\beta_0 + x\beta_1)}}}{1 - \dfrac{1}{1 + e^{-(\beta_0 + x\beta_1)}}} \\ &= \frac{\dfrac{1}{1 + e^{-(\beta_0 + x\beta_1)}}}{\dfrac{1 + e^{-(\beta_0 + x\beta_1)} - 1}{1 + e^{-(\beta_0 + x\beta_1)}}} \\ &= \frac{1}{e^{-(\beta_0 + x\beta_1)}}. \end{aligned}$$

Por isso

$$Z = e^{\beta_0 + x\beta_1}. \tag{3.3}$$

A chance calculada através da equação (3.3) pode assumir valores de 0 a infinito, e pode ser entendida como a razão entre a probabilidade de ocorrência do evento de interesse $Y = 1$ *versus* da sua não ocorrência. Normalmente, lê-se a chance na forma fracionária a fim de facilitar sua compreensão. De forma mais concreta, dado $p(x) = 0,8$ calcula-se

o valor da chance $Z = 4$, que pode ser entendido como sendo uma situação em que a chance de ocorrência do evento será de 4 para 1. Um outro exemplo seria o caso em que $p(x) = 0,2$ com valor de chance $Z = \frac{1}{4}$, situação em que a chance de ocorrência do evento será de 1 para 4.

Calculando o logaritmo natural da chance, dada pela equação (3.3), tem-se

$$g(x) = \beta_0 + x\beta_1. \tag{3.4}$$

A equação (3.4) representa a função chamada logito (PAULINO et al., 2011) e é linear em X. Assim, é possível notar que o aumento de uma unidade na variável independente levará ao aumento de $\beta_1$ unidades na função logito e ao produto de $e^{\beta_1}$ na função de chance. O impacto no cálculo da probabilidade $p(x+1)$ não é obtido de forma tão fácil e depende do valor de $p(x)$ a partir do qual foi feito o incremento na variável independente. Apesar disso, para valores positivos de $\beta_1$, um aumento no valor da variável independente levará ao aumento no valor de $p(x)$. Para valores negativos de $\beta_1$, um aumento no valor da variável independente levará a uma diminuição no valor de $p(x)$ (JAMES et al., 2013).

A importância da transformação logito reside no fato de que a função obtida através dela é linear em $x$ e, por isso pode ser continua variando de $-\infty$ a $+\infty$ dependendo da faixa de variação da variável independente (JR; LEMESHOW; STURDIVANT, 2013).

Procede-se então ao cálculo dos estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ dos parâmetros $\beta_0$ e $\beta_1$ da equação (3.1).

## 3.1 Estimadores de Máxima Verossimilhança

A fim de se encontrar os estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ dos parâmetros $\beta_0$ e $\beta_1$ respectivamente, será usado o Método da Máxima Verossimilhança.

A função de verossimilhança pode ser entendida como sendo uma medida de quão provável é a amostra analisada (em relação à população), e uma interpretação dos estimadores encontrados por este método é de que são os que tornam a amostra usada mais provável, mais próxima dos dados da população (MORETTIN; BUSSAB, 2010). Procede-se então a apresentação da função de verossimilhança.

Para cada valor de variável independente $x$ está associada a probabilidade $p(x)$ de que a variável resposta $y$ seja igual a 1. Assim, $1 - p(x)$ pode ser entendido como sendo a probabilidade de que $y$ seja igual a 0. Então, pode-se concluir que a parcela na função de verossimilhança dos pares de treinamento $(x_i, y_i)$ em que $y_i = 1$ é dada por $p(x_i)$. Já para os pares de treinamento $(x_i, y_i)$ em que $y_i = 0$ é dada por $1 - p(x_i)$ (JR; LEMESHOW; STURDIVANT, 2013). Apresentando essa contribuição de uma forma geral, tem-se

$$[p(x_i)]^{y_i}[1 - p(x_i)]^{1-y_i}.$$

Assumindo-se que as observações são independentes, pode-se calcular a função de verossimilhança $l(\beta_0, \beta_1)$ da amostra

$$l(\beta_0, \beta_1) = \prod_{i=1}^{n} [p(x_i)]^{y_i} [1 - p(x_i)]^{1-y_i}. \tag{3.5}$$

Observa-se então que a função de verossimilhança representada pela equação (3.5) é uma função em que as variáveis são $\beta_0$ e $\beta_1$. Os estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ dos parâmetros $\beta_0$ e $\beta_1$ são calculados a fim de maximizar a função de verossimilhança. Para facilitar os cálculos, prefere-se trabalhar com o valor chamado log-verossimilhança $L(\beta_0, \beta_1)$, obtido calculando o logaritmo natural da equação (3.5)

$$L(\beta_0, \beta_1) = \sum_{i=1}^{n} [y_i \log(p(x_i)) + (1 - y_i) \log(1 - p(x_i))]. \tag{3.6}$$

A fim de se encontrar os estimadores que tornam máxima a equação (3.6) ela será diferenciada em relação a $\beta_0$ e $\beta_1$ e igualada a zero (TSUCHIYA, 2002). Desenvolvendo

$$\begin{aligned}
&\frac{\partial L(\beta_0, \beta_1)}{\partial \beta_0}(\hat{\beta}_0, \hat{\beta}_1) = 0 \\
&\sum_{i=1}^{n} \left[ \frac{y_i}{p(x_i)} \frac{\partial [p(x_i)]}{\partial \beta_0} + \frac{(1 - y_i)}{1 - p(x_i)} \frac{\partial [1 - p(x_i)]}{\partial \beta_0} \right] = 0 \\
&\sum_{i=1}^{n} \left[ (y_i)(1 + e^{-(\beta_0 + x_i \beta_1)}) \frac{e^{-(\beta_0 + x_i \beta_1)}}{(1 + e^{-(\beta_0 + x_i \beta_1)})^2} - \frac{(1 - y_i)}{\frac{1 + e^{-(\beta_0 + x_i \beta_1) - 1}}{1 + e^{-(\beta_0 + x_i \beta_1)}}} \frac{e^{-(\beta_0 + x_i \beta_1)}}{(1 + e^{-(\beta_0 + x_i \beta_1)})^2} \right] = 0 \\
&\sum_{i=1}^{n} \left[ (y_i) \left( \frac{e^{-(\beta_0 + x_i \beta_1)}}{1 + e^{-(\beta_0 + x_i \beta_1)}} \right) + (y_i - 1) \frac{1}{1 + e^{-(\beta_0 + x_i \beta_1)}} \right] = 0 \\
&\sum_{i=1}^{n} \left[ (y_i) - \frac{1}{1 + e^{-(\beta_0 + x_i \beta_1)}} \right] = 0.
\end{aligned}$$

Além disso

$$\begin{aligned}
&\frac{\partial L(\beta_0, \beta_1)}{\partial \beta_1}(\hat{\beta}_0, \hat{\beta}_1) = 0 \\
&\sum_{i=1}^{n} \left[ \frac{y_i}{p(x_i)} \frac{\partial [p(x_i)]}{\partial \beta_1} + \frac{(1 - y_i)}{1 - p(x_i)} \frac{\partial [1 - p(x_i)]}{\partial \beta_1} \right] = 0 \\
&\sum_{i=1}^{n} \left[ (y_i)(1 + e^{-(\beta_0 + x_i \beta_1)}) \frac{x_i e^{-(\beta_0 + x_i \beta_1)}}{(1 + e^{-(\beta_0 + x_i \beta_1)})^2} - \frac{(1 - y_i)}{\frac{1 + e^{-(\beta_0 + x_i \beta_1) - 1}}{1 + e^{-(\beta_0 + x_i \beta_1)}}} \frac{x_i e^{-(\beta_0 + x_i \beta_1)}}{(1 + e^{-(\beta_0 + x_i \beta_1)})^2} \right] = 0 \\
&\sum_{i=1}^{n} \left[ (y_i x_i) \left( \frac{e^{-(\beta_0 + x_i \beta_1)}}{1 + e^{-(\beta_0 + x_i \beta_1)}} \right) + (y_i - 1) \frac{x_i}{1 + e^{-(\beta_0 + x_i \beta_1)}} \right] = 0 \\
&\sum_{i=1}^{n} \left[ (y_i x_i) - \frac{x_i}{1 + e^{-(\beta_0 + x_i \beta_1)}} \right] = 0.
\end{aligned}$$

Assim, se obtém as equações:

$$\sum_{i=1}^{n}[y_i - p(x_i)] = 0, \tag{3.7}$$

$$\sum_{i=1}^{n} x_i[y_i - p(x_i)] = 0. \tag{3.8}$$

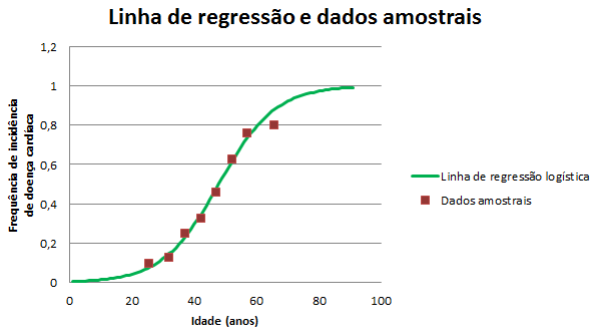
Como as equações (3.7) e (3.8) não são lineares, há necessidade da aplicação de métodos numéricos iterativos para calcular $\hat{\beta}_0$ e $\hat{\beta}_1$ que as maximize. A descrição pormenorizada do procedimento realizado não faz parte do escopo deste trabalho, para um maior detalhamento buscar (BISHOP, 2006). Os valores de $\hat{\beta}_0$ e $\hat{\beta}_1$ podem ser obtidos para um conjunto de dados de treinamento através de softwares estatísticos, tais como o **R**.

Para o problema apresentado no Exemplo 3.1 os valores encontrados dos estimadores são $\hat{\beta}_0 = -5,309$ e $\hat{\beta}_1 = 0,111$, e o modelo de regressão logística, através do qual serão calculadas as probabilidades logísticas estimadas ($\hat{p}(x)$), pode ser escrito como

$$\hat{p}(x) = \frac{1}{1 + e^{(5,309 - 0,111x)}}. \tag{3.9}$$

Para cada valor de variável independente $x_i$ é possível estimar a probabilidade $\hat{p}(x_i)$ de que a variável resposta seja 1 através da equação (3.9). A Figura 9 apresenta o modelo de regressão logística dado pela equação (3.9). É possível notar um bom ajuste aos dados amostrais.

Figura 9 – Modelo de regressão logística obtido com os dados amostrais para o problema apresentado no Exemplo 3.1.




Fonte: Autor.

## 3.2 Análise do Modelo

Os estimadores de variância e erro padrão de $\hat{\beta}_0$ e $\hat{\beta}_1$ poderão ser usados no cálculo de intervalos de confiança dos parâmetros $\beta_0$ e $\beta_1$, e em testes de hipótese para a verificação da acurácia do modelo obtido com o uso dos estimadores.

Uma vez obtidos os estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ a pergunta a ser respondida a partir de então é: o modelo de regressão logística que inclui a variável independente prevê melhor a saída do que um modelo que não a inclui? De forma análoga ao que foi feito para a regressão linear, a pergunta será respondida por meio da comparação de modelos preditivos que incluem e que não incluem a variável independente. Caso o modelo que inclui a variável independente seja mais acurado, de acordo com um critério especificado de medição da acurácia, a variável sob análise será considerada significativa para predições da variável resposta.

### 3.2.1 Teste de Hipótese

De forma análoga ao que foi feito em regressão linear, procede-se ao teste de hipótese para a verificação da relevância da relação entre as variáveis $X$ e $Y$. O teste será estruturado da seguinte forma

$$H_0\text{: não há relação linear entre X e Y,}$$

*versus* a hipótese alternativa

$$H_a\text{: há relação linear entre X e Y.}$$

Matematicamente essas relações correspondem a

$$H_0 : \beta_1 = 0,$$
$$H_a : \beta_1 \neq 0.$$

Caso a hipótese nula seja verdadeira o valor predito será constante e dado por $\hat{p}(x_i) = \bar{y} = n_1/n$, em que $n$ representa o tamanho da amostra e $n_1$ o número de eventos em que a variável resposta é igual a 1. Podem-se usar vários critérios na avaliação do teste de hipóteses, aqui serão citados e apresentados três: estatística G, teste de Wald e teste Score (JR; LEMESHOW; STURDIVANT, 2013).

#### 3.2.1.1 Estatística G

O primeiro critério apresentado será o da estatística $G$. Em regressão linear eram comparados os valores de $SQTot$ com os valores de $SQReg$, para uma medição de quão mais acurado é o modelo que inclui a variável independente através da diminuição dos erros de predição. No caso da função logística a comparação será feita através da diferença do logaritmo natural da função de verossimilhança dos modelos com e sem a variável preditiva sob estudo. Iniciam-se os cálculos com a de razão de verossimilhança ($r$) que é

dada pela expressão

$$r = \frac{l(\text{modelo ajustado})}{l(\text{modelo saturado})}. \tag{3.10}$$

No cálculo da razão de verossimilhança pela equação (3.10) se faz uma comparação entre a verossimilhança do modelo ajustado e a verossimilhança do modelo saturado, que pode ser entendido como contendo tantas variáveis quanto o tamanho da amostra. Assim, a resposta prevista pelo modelo saturado é numericamente igual ao próprio valor observado.

Para calcular o valor utilizado como critério na análise realizada no teste de hipótese, faz-se o cálculo e uso do produto de menos duas vezes o logaritmo natural da razão de verossimilhança, chamado desviância ($D$) (PAULINO et al., 2011). A saber

$$D = -2\log\left[\frac{l(\text{ modelo ajustado})}{l(\text{modelo saturado})}\right]. \tag{3.11}$$

O produto pelo fator (– 2) e a aplicação do logaritmo natural na equação (3.10) para obter a equação (3.11), se deve ao fato de que esta transformação leva a uma distribuição conhecida, a qui-quadrado com um grau de liberdade, ao se comparar os modelos obtidos com e sem a variável independente (JR; LEMESHOW; STURDIVANT, 2013).

Para problemas em que a variável resposta tem natureza dicotômica, a verossimilhança do modelo saturado é 1, pois uma vez que $\hat{p}(x_i) = y_i$

$$y_i = 0 : [p(x_i)]^{y_i}[1 - p(x_i)]^{1-y_i} = [0]^0 * [1 - 0]^{(1-0)} = 1$$

e caso

$$y_i = 1 : [p(x_i)]^{y_i}[1 - p(x_i)]^{1-y_i} = [1^1] * [1 - 1]^{(1-1)} = 1$$

e por isso, o cálculo da verossimilhança em que se usa a Equação (3.5) para o modelo saturado é

$$l(\text{modelo saturado}) = \prod_{i=1}^{n}[p(x_i)]^{y_i}[1 - p(x_i)]^{1-y_i} = 1.$$

Assim, a equação (3.11) fica

$$D = -2\log\left[l(\text{modelo ajustado})\right]. \tag{3.12}$$

O cálculo da desviância em regressão logística é comparável ao cálculo de $SQReg$ em regressão linear e, através deste cálculo, se terá uma medição de quão melhor ajustados aos dados o modelo que contém a variável independente é quando comparado ao modelo que não a contém. A desviância tem um papel importante na medição da qualidade de ajuste do modelo (JR; LEMESHOW; STURDIVANT, 2013).

Subtraindo os valores de desviância dos modelos de regressão logística com e sem a variável independente será obtido um valor de estatística $G$

$$G = D(\text{modelo ajustado sem a variável}) - D(\text{modelo ajustado com a variável}).$$

Logo, da equação (3.12), segue que

$$G = -2\log\left[\frac{l(\text{ modelo ajustado sem a variável})}{l(\text{modelo ajustado com a variável})}\right]. \tag{3.13}$$

Para problemas de classificação em que a variável resposta é dicotômica, para o modelo ajustado sem a variável independente, tem-se a partir da equação (3.7) que o estimador $\hat{\beta}_0$ é dado por

$$\begin{aligned}\sum_{i=1}^{n}[y_i - p(x_i)] &= 0\\ n_1 - n\frac{1}{1+e^{-\hat{\beta}_0}} &= 0\\ \Rightarrow e^{-\hat{\beta}_0} &= \frac{n_0}{n_1}\end{aligned}$$

em que $n$ é o tamanho da amostra, $n_1$ o número de ocorrências em que $y_i = 1$, $n_0$ o número de ocorrências em que $y_i = 0$ e $n$ o número totao de ocorrências e por isso $n = n_0 + n_1$. Finalmente

$$\hat{\beta}_0 = \log\frac{n_1}{n_0}, \tag{3.14}$$

Calculando o log-verossimilhança do modelo ajustado sem a variável, ou seja, para a situação em que $\hat{\beta}_1 = 0$ e $\hat{\beta}_0 = \log\dfrac{n_1}{n_0}$

$$\begin{aligned}L(\hat{\beta}_0, \hat{\beta}_1) =& \sum_{i=1}^{n}[y_i\log(p(x_i)) + (1-y_i)\log(1-p(x_i))]\\ =& n_1\log\left(\frac{1}{1+\frac{n_0}{n_1}}\right) + n_0\log\left(1-\frac{n_1}{n}\right)\\ =& n_1\log\left(\frac{n_1}{n}\right) + n_0\log\left(\frac{n_0}{n}\right)\\ =& n_1\log(n_1) + n_0\log(n_0) - n\log(n)\end{aligned}$$

Com isso, a equação (3.13) fica

$$G = 2\left[\sum_{i=1}^{n}[y_i\log(\hat{p}(x_i)) + (1-y_i)\log(1-\hat{p}(x_i))] - [n_1\log(n_1) + n_0\log(n_0) - n\log(n)]\right]. \tag{3.15}$$

Para o problema apresentado no Exemplo 3.1 o logaritmo natural da verossimilhança para o modelo ajustado com a variável independente é dado como resultado pelo software R e vale $-53,677$. Ainda da Tabela 7, tem-se que $n = 100$, $n_1 = 43$ e $n_0 = 57$. Portanto, o valor da estatística $G$ apresentado na equação (3.15) é

$$\begin{aligned} G &= 2\left[-53,677 - [43\log(43) + 57\log(57) - 100\log(100)]\right] \\ &= 29,31. \end{aligned}$$

O $valor - p$ associado a $G = 29,31$ é muito menor do que $0,005$, uma vez que o valor crítico associado a $\alpha = 0,005$ é de $7,882$ (DEVORE, 2010). Por isso, da Tabela 6 pode-se concluir que a variável independente idade é significativa na predição da presença de doença cardíaca.

#### 3.2.1.2 Teste Wald

O teste Wald é análogo ao que foi feito no modelo de regressão linear, em que se supõe que o parâmetro $\beta_1$ tenha valor nulo, e se faz a análise do valor da estatística de Wald ($W$)

$$W = \frac{\hat{\beta}_1}{\hat{\sigma}(\hat{\beta}_1)} \tag{3.16}$$

que pode ser entendido como uma medição de quanto $\hat{\beta}_1$ está distante do zero em quantidades de desvio padrão estimado para $\hat{\beta}_1$. A distribuição dada por $W$ na equação (3.16) segue uma distribuição padrão normal e a ela está associado um $valor - p$ da distribuição bicaudal que representa a probabilidade de $P(|z| > W)$. Para baixos valores de $valor - p$ entende-se que o valor de $W$ é pouco provável e, por isso, a hipótese nula pode ser descartada. Pode-se usar a Tabela 6 na decisão da rejeição ou não da hipótese nula.

Para o problema apresentado no Exemplo 3.1 e usando os valores apresentados por (TSUCHIYA, 2002), o valor de $\hat{\sigma}(\hat{\beta}_1) = 0,024$ é dado pelo software R. Por isso, o valor da estatística $W = 4,61$ leva a um $valor - p$ muito menor do que $0,001$ donde se conclui que há fortíssimas evidências para a rejeição da hipótese nula. O valor crítico associado ao $valor - p = 0,001$ é de $3,49$ (DEVORE, 2010).

#### 3.2.1.3 Teste de Escores

Outro critério usado para a validação ou rejeição da hipótese nula é dado pelo teste de escores (PAULINO et al., 2011). Como vantagem deste teste pode ser citado seu baixo esforço computacional. O critério será aplicado sobre a distribuição condicional das derivadas parciais do logaritmo da função de verossimilhança. A análise será feita sobre a distribuição de (3.7) dada a (3.8). Para o caso da variável resposta ser dicotômica e

da suposição da hipótese nula de que o parâmetro $\beta_1 = 0$ e ainda da equação (3.14), a estatística do teste de escores, chamada de estatística de escores ($ST$) é calculada

$$ST = \frac{\sum_{i=1}^{n} x_i(y_i - \bar{y})}{\sqrt{\bar{y}(1-\bar{y})\sum_{i=1}^{n}(x_i - \bar{x})^2}}, \tag{3.17}$$

em que $\bar{y} = \dfrac{n_1}{n}$.

A estatística $ST$ segue uma distribuição normal padrão e a análise do $valor - p$ será feita da mesma forma que para o teste Wald.

Para o problema apresentado no Exemplo 3.1 o valor da estatística $ST = 5,14$ que leva a um $valor - p$ muito menor do que $0,001$. Da Tabela 6 pode-se concluir que há evidências fortíssimas para a rejeição da hipótese nula.

De forma análoga ao procedimento que foi realizado em regressão linear, após a determinação da relevância dos estimadores na predição da variável resposta, procede-se à construção de intervalos de confiança para os estimadores e também para a função logito e consequentemente para predições individuais.

### 3.2.2 Intervalos de Confiança para os Parâmetros

Os intervalos de confiança serão construídos da seguinte maneira

$$IC(\beta_1, \alpha) = \hat{\beta}_1 \pm \hat{\sigma}(\hat{\beta}_1) z_{1-\frac{\alpha}{2}}, \tag{3.18}$$

e

$$IC(\beta_0, \alpha) = \hat{\beta}_0 \pm \hat{\sigma}(\hat{\beta}_0) z_{1-\frac{\alpha}{2}} \tag{3.19}$$

em que $z_{1-\frac{\alpha}{2}}$ é o valor critico superior atrelado a $\left(1 - \frac{\alpha}{2}\right)$ que está associado à área sob a curva de distribuição normal de forma análoga à apresentada na Figura 5.

Para o problema apresentado no Exemplo 3.1 serão apresentados os valores de intervalo de confiança dos estimadores. Para $\dfrac{\alpha}{2} = 0,1$, com confiança de 95%, o valor crítico $z$ associado é de $z = 1,96$ e assim

$$\begin{aligned} IC(\beta_1, \alpha) &= 0,111 \pm 0,0241 * 1,96 \\ &= 0,111 \pm 0,047, \end{aligned}$$

$$\begin{aligned} IC(\beta_0, \alpha) &= -5,309 \pm 1,1337 * 1,96 \\ &= -5,309 \pm 2,222. \end{aligned}$$

Assim sendo, há 95% de probabilidade de que o parâmetro $\beta_1$ esteja no intervalo $[0,064, 0,158]$ e de que o parâmetro $\beta_0$ esteja no intervalo $[-7,531, -3,087]$.

### 3.2.3 Intervalo de Confiança para a Função Logito e para as Predições

Dados os estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ é possível calcular o estimador da função logito

$$\hat{g}(x_i) = \hat{\beta}_0 + x_i\hat{\beta}_1$$

com variância

$$V(\hat{g}(x_i)) = \hat{\sigma}^2_{\hat{\beta}_0} + (x_i)^2\hat{\sigma}^2_{\hat{\beta}_1} + 2x_i\hat{\text{Cov}}(\hat{\beta}_0, \hat{\beta}_1). \tag{3.20}$$

Os valores de variância e covariância serão os obtidos pelo uso de softwares como **R** e lidos em uma matriz de covariância. Para um estudo mais aprofundado buscar (JR; LEMESHOW; STURDIVANT, 2013). É possível notar que o estimador de variância da função logito dado pela equação (3.20) depende do valor da variável independente.

Enfim, pode-se representar o intervalo de confiança da função logito como

$$IC(g(x_i), \alpha) = \hat{g}(x_i) \pm \hat{\sigma}_{\hat{g}(x_i)} z_{1-\frac{\alpha}{2}}$$

em que $z_{1-\frac{\alpha}{2}}$ é o valor critico superior atrelado a $\left(1 - \frac{\alpha}{2}\right)$ que está associado à área sob a curva de distribuição normal. Assim sendo, ao se usar os valores limites de $IC(g(x_i), \alpha)$ é possível calcular o intervalo de confiança para $p(x_i)$ dado um valor $(x_i)$ de variável independente.

Para o problema do Exemplo 3.1, os valores dos estimadores de variâncias e covariâncias de $\hat{\beta}_1$ e $\hat{\beta}_0$ estão representados na Tabela 8.

Tabela 8 – Estimadores de variâncias e de covariância de $\hat{\beta}_1$ e $\hat{\beta}_0$ para o problema apresentado no Exemplo 3.1.

| | $\hat{\beta}_1$ | $\hat{\beta}_0$ |
|---|---|---|
| $\hat{\beta}_1$ | 0,000579 | |
| $\hat{\beta}_0$ | -0,026677 | 1,28517 |

Fonte: (JR; LEMESHOW; STURDIVANT, 2013).

Desta maneira, para pacientes com idade de 40 anos

$$\hat{g}(40) = -5,309 + 0,111 * 40 = -0,869$$
$$\hat{p}(40) = e^{(-0,869)}/1 + e^{(-0,869)} = 0,295$$

ou seja, a probabilidade de que haja presença de doença cardíaca em pacientes de 40 anos de idade é em média de $\hat{p}(40) = 0,295$, que representa sua estimativa pontual. Calculando a chance pela equação (3.2) para esta estimativa tem-se que $Z \approx \frac{4}{10}$ o que pode ser interpretado como este evento tendo chance de ocorrência de 4 em 10.

Procede-se ao cálculo de $\hat{\sigma}_{\hat{g}(40)}$ usando a equação (3.20) e os dados apresentados na Tabela 8

$$\begin{aligned}\hat{\sigma}^2_{\hat{g}(40)} &= 1,285 + (40^2) * 0,000579 + 2 * 40 * (-0,0267) \\ &= 0,0754 \\ \hat{\sigma}_{\hat{g}(40)} &= 0,275.\end{aligned}$$

Deste modo, o intervalo de confiança da função logito com 95% de confiança é dado por

$$\begin{aligned}IC(g(40), 0, 1) &= \hat{g}(40) \pm \hat{\sigma}_{\hat{g}(40)} z_{0,95} \\ &= -0,869 \pm 0,275\end{aligned}$$

Portanto, há 95% de probabilidade de que o valor da função logito $g(40)$ esteja no intervalo $[-1,114, -0,594]$. Por fim, efetua-se o cálculo do intervalo de confiança de $p(40)$ usando os limites do intervalo de confiança de $g(40)$. Assim sendo

$$\begin{aligned}e^{(-1,114)}/1 - e^{(-1,114)} \leq p(40) \leq e^{(-0,594)}/1 - e^{(-0,594)} \\ 0,241 \leq p(40) \leq 0,356\end{aligned}$$

Calculando a chance pela equação (3.2) para os limites do intervalo de confiança de $p(40)$ tem-se que

$$p(40) = 0,241 \Rightarrow Z \approx \frac{3}{10}$$

que pode ser interpretado como este evento tendo chance de ocorrência de 3 para 10, e

$$p(40) = 0,356 \Rightarrow Z \approx \frac{5}{10}$$

que pode ser interpretado como este evento tendo chance de ocorrência de 5 para 10.

Logo, procedendo de forma análoga a que foi feita para um paciente de 40 anos é possível realizar a predição de presença de doença cardíada em pacientes de qualquer idade.

# 4 Sequência Didática

Nesta seção será apresentada uma sequência didática através da qual serão aplicados os conceitos abordados no Capítulo 2.

Serão estudados dados de observações de valores de média semanal de publicações de vídeos de um canal do Youtube, bem como a quantidade de visualizações dos vídeos neste mesmo período. A escolha do problema foi feita de modo a estimular o engajamento dos alunos na atividade, visto que se trata de um assunto de grande interesse por parte deles. Seguindo o passo a passo proposto, ao final das atividades o aluno terá calculado o modelo de regressão linear (reta de mínimos quadrados) que relacione as grandezas apresentadas, bem como terá feito a análise do modelo calculado.

É importante que os estudantes tenham contato com esse tipo de conteúdo ainda na escola como forma de consolidar os conceitos já aprendidos e compreender a aplicabilidade de tais conceitos em situações práticas.

## 4.1 Tema

Introdução de conceitos de Regressão Linear como motivação para aprofundamento em Aprendizado de Máquina.

## 4.2 Conteúdo Abordado

Conceitos de Estatística: população, amostra, média, parâmetros, estimadores.

Regressão Linear Simples: diagrama de dispersão, estimação pelo Método dos Mínimos Quadrados, resíduos, $R^2$.

## 4.3 Objetivo

- Identificar relação linear através da análise de gráfico de dispersão.
- Calcular reta de mínimos quadrados.
- Compreender o significado de cada um dos estimadores da reta.
- Compreender os critérios de avaliação de proximidade de cada estimador em relação aos respectivos parâmetros.
- Saber aplicar e interpretar esses critérios em um exemplo.

- Despertar interesse no aprofundamento dos estudos em aprendizado de máquina, por meio da percepção de como este tema está inserido em atividades de seus interesses.

## 4.4 Público Alvo

Alunos do Ensino Médio, por já terem sido apresentados aos conteúdos de funções afins no nono ano.

## 4.5 Recursos Usados

Computador, datashow, quadro.

## 4.6 Descrição das Atividades

- Apresentação dos conceitos de Estatística necessários com ilustração através de exemplos.
- Apresentação do conceito de gráfico de dispersão e inferência de relação linear entre variáveis independente e dependente, aplicando os conceitos no exemplo enunciado.
- Apresentação do método dos mínimos quadrados e cálculo e interpretação de cada parâmetro, aplicando os conceitos no exemplo enunciado.
- Compreensão do impacto da amostra usada nos valores de parâmetros estimados (usando subconjuntos dos dados amostrais do exemplo para perceber impacto).
- Análise de exatidão dos coeficientes e modelo desenvolvidos, aplicando os conceitos no exemplo enunciado.
- Elaboração de relatório com os resultados obtidos das atividades previamente desenvolvidas.

## 4.7 Avaliação

Relatório contendo as atividades desenvolvidas e respostas das perguntas propostas.

## 4.8 Desenvolvimento

A fim de aumentar o número de visualizações em seu canal do Youtube, um Youtuber realizou a coleta dos seguintes pares de dados para avaliar qual estratégia tomar:

média de publicações semanais e número de visualizações dos vídeos do canal, tempo de duração médio dos vídeos e número de visualizações dos vídeos do canal, percentual de tempo de vídeo gasto fazendo propaganda e número de visualizações dos vídeos do canal. Os dados obtidos são apresentados nas tabelas Tabela 9, Tabela 10 e Tabela 11.

Tabela 9 – Dados de média semanal de publicações e centenas de visualizações dos vídeos

| Observação($i$) | Média semanal de publicações | Centenas de visualizações |
|---|---|---|
| 1 | 0,40 | 0,50 |
| 2 | 0,42 | 0,91 |
| 3 | 0,48 | 1,09 |
| 4 | 0,51 | 1,38 |
| 5 | 0,57 | 1,56 |
| 6 | 0,60 | 1,55 |
| 7 | 0,70 | 1,52 |
| 8 | 0,75 | 1,91 |
| 9 | 0,75 | 1,34 |
| 10 | 0,78 | 1,84 |
| 11 | 0,84 | 1,50 |
| 12 | 0,95 | 2,39 |
| 13 | 0,99 | 1,89 |
| 14 | 1,03 | 2,55 |
| 15 | 1,12 | 2,10 |
| 16 | 1,15 | 2,72 |
| 17 | 1,20 | 2,58 |
| 18 | 1,25 | 2,61 |
| 19 | 1,25 | 2,92 |
| 20 | 1,28 | 2,96 |

Fonte: Autor.

Pede-se que o discente proceda às seguintes atividades

- Faça o gráfico de dispersão linear.
  - É possível inferir que há relação linear entre as variáveis?
- Encontre os valores dos estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ pelo método dos mínimos quadrados.
- Calcule os valores dos estimadores para as amostras dadas pelo seguintes subconjuntos de dados: os dez primeiros pares ordenados, os dez últimos pares ordenados, 10 pares ordenados de sua escolha (diferentes dos dois primeiros).
- Para a amostra completa, calcule a soma dos quadrados dos resíduos ($SQRes$), soma dos quadrados total ($SQTot$) e soma dos quadrados da regressão ($SQReg$).
- Estime a variância e o desvio padrão de cada um dos estimadores e explique o significado destes valores.

- Calcule o valor de $R^2$ para o modelo encontrado e explique o significado do valor encontrado.

Espera-se que os gráficos de dispersão elaborados sejam os apresentados nas figuras .... . Da análise das figuras ... espera-se que o aluno perceba que não há relação linear entre as grandezas e por isso não construa o modelo de regressão linear. Da análise da figura... espera-se que o aluno perceba a tendência de relação linear entre as variáveis e por isso proceda à construção do modelo.

Para encontrar os valores dos estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ para a amostra de tamanho 20 e para seus subconjuntos pelo método dos mínimos quadrados serão usadas as equações

$$\hat{\beta}_1 = \frac{\Sigma_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\Sigma_{i=1}^{n}(x_i - \bar{x})^2} \quad \text{e} \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$$

Para calcular a soma dos quadrados dos resíduos ($SQRes$), soma dos quadrados total ($SQTot$) e soma dos quadrados da regressão ($SQReg$)

$$SQRes = \sum_{i=1}^{n}[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 \quad \text{e} \quad SQTot = \sum_{i=1}^{n}[y_i - \bar{y}]^2$$

$$SQReg = SQTot - SQRes$$

A fim de estimar a variância e o desvio padrão de cada um dos estimadores serão usadas as equações

$$\hat{\sigma}_e^2 = \frac{SQRes}{n-2} \quad \text{e} \quad \hat{\sigma}^2 = \frac{SQTot}{n-1}$$

Com o intuito de calcular o valor de $R^2$ faz-se

$$R^2 = \frac{SQReg}{SQTot}$$

Os valores de cada uma das grandezas calculadas estão listados na tabela 12.

Através da análise de $SQRes$, $SQTot$ e de $SQReg$, espera-se que o aluno perceba que o modelo que inclui a variável independente é melhor do que o modelo que não a inclui, dado que $SQRes < SQTot$. Espera-se ainda que ele perceba que $SQReg$ é uma medida de tal melhoria.

Da comparação entre $\hat{\sigma}_e^2$ e $\hat{\sigma}^2$, espera-se que o aluno compreenda que como $\hat{\sigma}_e^2 < \hat{\sigma}^2$, o modelo que inclui a variável independente gerará predições mais acuradas.

Da análise do valor de $R^2$ espera-se que o aluno conclua que o modelo que inclui a variável independente explica 88% da variabilidade na resposta.

Tabela 10 – Dados de tempo em minutos do vídeo publicado e centenas de visualizações dos vídeos

| Observação($i$) | Tempo (min) | Centenas de visualizações |
|---|---|---|
| 1 | 1,00 | 0,62 |
| 2 | 1,00 | 1,00 |
| 3 | 1,20 | 0,89 |
| 4 | 1,80 | 0,82 |
| 5 | 2,10 | 1,07 |
| 6 | 2,70 | 1,19 |
| 7 | 2,70 | 0,83 |
| 8 | 3,00 | 0,98 |
| 9 | 4,00 | 1,40 |
| 10 | 4,00 | 0,55 |
| 11 | 4,00 | 1,53 |
| 12 | 4,20 | 1,51 |
| 13 | 4,50 | 1,53 |
| 14 | 4,50 | 0,90 |
| 15 | 4,80 | 1,25 |
| 16 | 4,80 | 1,34 |
| 17 | 4,80 | 0,94 |
| 18 | 5,10 | 1,97 |
| 19 | 5,40 | 1,30 |
| 20 | 5,70 | 2,15 |
| 21 | 6,00 | 1,28 |
| 22 | 6,50 | 2,06 |
| 23 | 6,90 | 1,53 |
| 24 | 7,00 | 2,05 |
| 25 | 7,50 | 1,98 |
| 26 | 7,50 | 1,96 |
| 27 | 7,80 | 1,75 |
| 28 | 8,20 | 2,19 |
| 29 | 8,40 | 1,80 |
| 30 | 9,50 | 1,16 |
| 31 | 9,50 | 1,16 |
| 32 | 9,50 | 0,98 |
| 33 | 9,90 | 1,76 |
| 34 | 10,30 | 1,71 |
| 35 | 11,20 | 0,64 |
| 36 | 11,50 | 1,33 |
| 37 | 12,00 | 1,06 |
| 38 | 12,50 | 0,19 |
| 39 | 12,50 | 0,31 |
| 40 | 12,80 | 0,03 |

Fonte: Autor.

Tabela 11 – Dados de percentual de tempo de propaganda no vídeo e centenas de visualizações dos vídeos

| Observação($i$) | Percentual de propaganda | Centenas de visualizações |
|---|---|---|
| 1 | 0,11 | 0,69 |
| 2 | 0,13 | 0,91 |
| 3 | 0,18 | 1,83 |
| 4 | 0,20 | 2,38 |
| 5 | 0,25 | 3,44 |
| 6 | 0,27 | 3,80 |
| 7 | 0,31 | 3,96 |
| 8 | 0,33 | 3,83 |
| 9 | 0,35 | 3,57 |
| 10 | 0,37 | 3,05 |
| 11 | 0,38 | 3,01 |
| 12 | 0,39 | 2,76 |
| 13 | 0,40 | 2,46 |
| 14 | 0,40 | 2,31 |
| 15 | 0,41 | 2,19 |
| 16 | 0,45 | 1,39 |
| 17 | 0,46 | 1,18 |
| 18 | 0,47 | 0,96 |
| 19 | 0,48 | 0,86 |
| 20 | 0,50 | 0,56 |
| 21 | 0,54 | 0,21 |
| 22 | 0,57 | 0,09 |
| 23 | 0,60 | 0,04 |
| 24 | 0,68 | 0,00 |
| 25 | 0,70 | 0,00 |
| 26 | 0,74 | 0,00 |
| 27 | 0,78 | 0,00 |
| 28 | 0,80 | 0,00 |

Fonte: Autor.

Tabela 12 – Valores calculados do modelo de regressão linear em que se relaciona o número de publicações semanais com o número de visualizações.

| $\hat{\beta}_0$ | $\hat{\beta}_1$ | $SQRes$ | $SQTot$ |
|---|---|---|---|
| 0,03 | 2,19 | 1,15 | 9,21 |
| $SQReg$ | $\hat{\sigma}_e^2$ | $\hat{\sigma}^2$ | $R^2$ |
| 8,07 | 0,064 | 0,48 | 0,88 |

# 5 Conclusão

Inicialmente mostrou-se a importância e aplicabilidade dos algoritmos de aprendizado de máquina como uma forma de motivação, e também de responder ao questionamento do para quê servem os assuntos abordados em sala de aula na matemática.

Foram apresentados conteúdos de regressão linear e logística como uma forma de mostrar aplicabilidade dos conceitos já aprendidos em sala de aula, tais como probabilidade e funções afins. A apresentação de tais assuntos vem como incentivo para aprofundamento dos pontos já abordados em sala bem como trazer significado para os conceitos aprendidos muitas vezes de forma mais abstrata. Assim, este trabalho serve de base para o aprendizado e compreensão mais ampla do que são algoritmos de aprendizado de máquina bem como trabalha no sentido de uma aprendizagem de conteúdos já ministrados em sala de maneira mais significativa. É possível passar para os alunos uma noção dos cálculos realizados para a construção dos modelos de regressão linear a fim de desmistificar a dificuldade em matemática e também mostrar aplicações como forma de motivação de estudo e aprendizado neste campo da ciência, mostrando de forma concreta aplicações de conceitos muitas vezes abstratos aprendidos em sala de aula.

Sugestões de trabalhos futuros seriam a aplicação da sequência didática proposta e aprofundamento dos estudos em regressão logística, assim como o desenvolvimento de algoritmos em Python e uso de softwares como o R na resolução dos problemas apresentados e ainda como uma ferramenta que permita lidar com bases de dados maiores. Acredita-se que o uso de tais ferramentas possa gerar um maior engajamento e motivação dos alunos. Outra proposta seria a aplicação de algoritmos de regressão logística na predição de aprovação ou reprovação de alunos em disciplinas com altas taxas de reprovação, baseado em notas de disciplinas anteriores ou de testes de nivelamento para geração de planos de ação para mitigação de tais questões, em que os conhecimentos seriam utilizados como ferramenta no próprio processo de ensino.

Enfim, acredita-se que a abordagem destes assuntos em salas de aula do ensino médio possa contribuir para a formação de competências e habilidades mais amplas nos discentes, assim como dar início no processo de alocação profissional dos alunos que se engajarem e se interessarem por este ramo da ciência. Além disso, espera-se que a aplicação de algoritmos de aprendizado de máquina possa servir como ferramenta pelos próprios docentes.

# Referências


ALPAYDIN, E. *Introduction to Machine Learning*. 3. ed. Cambridge: MIT Press, 2014. Citado 2 vezes nas páginas 10 e 12.

BERTENS, P. et al. A machine-learning item recommendation rystem for video games. In: CONFERENCE ON COMPUTATIONAL INTELLIGENCE AND GAMES (CIG), 2018, Maastricht. *2018 IEEE Conference on Computational Intelligence and Games (CIG)*. IEEE, 2018. p. 1–4. Disponível em: <https://ieeexplore.ieee.org/document/8490456>. Acesso em: 16 jul. 2021. Citado na página 13.

BISHOP, C. M. *Pattern Recognition and Machine Learning*. Nova York: Springer, 2006. Citado na página 39.

BRAGA, A. de P.; FERREIRA, A. C. P. de L.; LUDERMIR, T. B. *Redes Neurais Artificiais*: Teorias e aplicações. Rio de Janeiro: LTC, 2007. Citado na página 11.

BURKOV, A. *The Hundred-Page Machine Learning Book*. Quebec: Andriy Burkov, 2019. Citado na página 10.

CORREA, M. S. B. B. *Probabilidade e Estatística*. 2. ed. Belo Horizonte: PUC Minas Virtual, 2003. Citado na página 17.

DEVORE, J. L. *Probabilidade e Estatística para Engenharia e Eiências*. São Paulo: Cengage Learning Edições Ltda, 2010. Citado 5 vezes nas páginas 15, 30, 34, 43 e 57.

EFRON, B.; GOUS, A. Technical Report. *Bayesian and Frequentist Model Selection*. Califórnia, 1997. Citado na página 31.

FÁVERO, L. P. et al. *Análise de Dados: Modelagem Multivariada para Tomada de Decisões*. Rio de Janeiro: Elsevier, 2009. Citado na página 36.

GEISLER, B. Integrated machine learning for behavior modeling in video games. In: THE NINETEENTH NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE, 2004, São José. *Proceedings [...]*. Menlo Park: AAAI Press, 2004. p. 54–62. Citado na página 13.

GUIDORIZZI, H. L. *Um Curso de Cálculo*. Rio de Janeiro: LTC Editora, 1998. v. 2. Citado na página 15.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning*. Nova York: Springer, 2009. (Springer Series in Statistics). Citado na página 13.

JAMES, G. et al. *An Introduction to Statistical Learning*: with applications in r. Nova York: Springer, 2013. (Springer Texts in Statistics). Citado 3 vezes nas páginas 34, 37 e 57.

JR, D. W. H.; LEMESHOW, S.; STURDIVANT, R. X. *Applied logistic regression*. Nova Jersey: John Wiley & Sons, 2013. Citado 6 vezes nas páginas 10, 35, 37, 40, 41 e 45.

KUTNER, M. H. et al. *Applied Linear Statistical Models, 2005*. Nova York: McGraw Hill Irwin, 2005. Citado na página 24.

MORETTIN, P. A.; BUSSAB, W. O. *Estatística básica*. Sao Paulo: Saraiva, 2010. Citado 9 vezes nas páginas 15, 25, 26, 27, 28, 33, 34, 37 e 57.

PAULINO, C. et al. *Glossário Inglês-Português de Estatística*. 2. ed. [S.l.]: Sociedade Portuguesa de Estatística e Associação Brasileira de Estatística, 2011. Citado 3 vezes nas páginas 37, 41 e 43.

REIS, E. A.; REIS, I. A. Relatório Técnico do Departamento de Estatística da UFMG. *Análise descritiva de dados*. Belo Horizonte, 2002. Disponível em: <http://www.est.ufmg.br/>. Citado na página 57.

TSUCHIYA, Í. *Regressão Logística Aplicada na Análise Espacial de Dados Arqueológicos*. Dissertação (Mestrado em Ciências Cartográficas) — Universidade Estadual Paulista (UNESP), Presidente Prudente, 2002. Citado 4 vezes nas páginas 34, 35, 38 e 43.

# Apêndices

# APÊNDICE A – Conceitos Estatísticos Preliminares

Abaixo serão apresentadas definições estatísticas usadas ao longo texto. As definições foram retiradas das referências (DEVORE, 2010), (MORETTIN; BUSSAB, 2010), (REIS; REIS, 2002) e (JAMES et al., 2013).

**Definição A.1** (População)**.** Uma população pode ser definida como o conjunto ou uma coleção bem definida de todos os elementos de interesse em um estudo.

**Definição A.2** (Amostra)**.** Uma amostra é constituída por qualquer subconjunto da população.

**Definição A.3** (Parâmetro)**.** Um parâmetro é uma medida usada para descrever uma característica da população.

**Definição A.4** (Estatística)**.** Uma estatística é uma característica da amostra.

**Definição A.5** (Estimador)**.** Um estimador de um parâmetro $\theta$ é uma função que tem como domínio dados amostrais.

É importante ainda definir a distribuição amostral do estimador $T$ do parâmetro $\theta$.

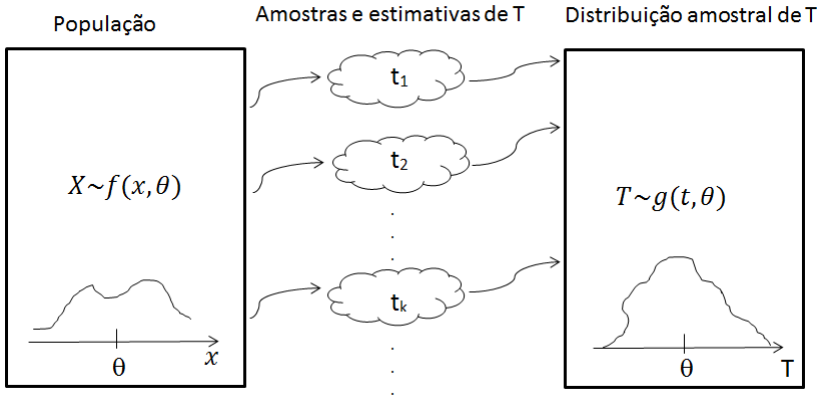
**Definição A.6** (Distribuição Amostral)**.** Ao se tomar todas as amostras possíveis de uma população de acordo com o critério adotado e para cada uma delas se calcular o valor do estimador $T$ (chamado de estimativa), tem-se a distribuição amostral de $T$. Uma representação gráfica pode ser observada na Figura 10.

**Definição A.7** (Distribuição amostral da média)**.** Seja uma população da qual se conhecem os parâmetros de média populacional $\mu$ e variância $\sigma^2$. Seja $\bar{X}$. Tomadas todas as amostras aleatórias simples e com reposição de tamanho $n$ desta população e calculados os valores da estatística média amostral $\bar{X}$, pode-se afirmar

$$E(\bar{X}) = \mu$$
$$\sigma^2(\bar{X}) = \frac{\sigma^2}{n}.$$

**Teorema A.8** (Teorema do Limite Central)**.** *À medida que o tamanho $n$ das amostras aumenta a distribuição amostral das médias tende a uma distribuição normal de média $E(\bar{X}) = \mu$ e de variância $\sigma^2(\bar{X}) = \frac{\sigma^2}{n}$ e tal tendência independe da distribuição da população.*

Figura 10 – Distribuição Amostral de um Estimador




Fonte: Autor.

**Definição A.9** (Viés)**.** Sobre critérios de proximidade de um estimador em relação ao parâmetro estimado, um estimador é chamado de não-viesado se

$$E(T) = \theta.$$

Caso a proposição acima não seja verdadeira, o estimador é chamado viesado e o viés de t $V(T)$ é dado por

$$V(T) = E(T) - \theta.$$

**Definição A.10** (Linha de regressão populacional)**.** A linha de regressão populacional é a melhor aproximação linear da verdadeira relação entre as variáveis independente e dependente.

**Definição A.11** (Diagrama de dispersão)**.** O diagrama de dispersão é um gráfico em que pontos do espaço cartesiano são usados para representar simultaneamente os valores de duas variáveis quantitativas em cada elemento do conjunto de dados amostrais.