NOTAS DE AULA

TÓPICOS EM ESTATÍSTICA PARA ENGENHARIA

CHARLES CASIMIRO CAVALCANTE
VINÍCIUS SILVA OSTERNE RIBEIRO

NOTAS DE AULA

TÓPICOS EM ESTATÍSTICA
PARA ENGENHARIA

CHARLES CASIMIRO CAVALCANTE
charles.casimiro@ieee.org
charlescasimiro.github.io

VINÍCIUS SILVA OSTERNE RIBEIRO
vinicius@osterne.com
www.osterne.com

Sumário

CAPÍTULO 1

História da Estatística

1.1 As primeiras ideias (t.b.d.)

1.2 Etimologia (t.b.d.)

1.3 Origens na probabilidade (t.b.d.)

1.4 A estatística hoje (t.b.d.)

CAPÍTULO 2

Conceitos básicos em Probabilidade

É importante começar nossos estudos com a explanação de que tudo o que se estuda em Estatística tem, como base fundamental, a teoria da Probabilidade que, por sua vez, tem como base fundamental a teoria dos conjuntos.

A teoria da Probabilidade permite que possamos, por exemplo, modelar populações, experimentos, acontecimentos ou realizar previsões com dados e informações que apresentam comportamento não determinístico (estocástico ou aleatório).

Na subseções a seguir apresentamos os conceitos sobre experimentos aleatórios, espaço amostral, eventos, sigma-álgebra, definições de probabilidade, propriedades da probabilidade, probabilidade condicional (regra do produto de probabilidades, teorema da probabilidade total, eorema de Bayes e ndependência de eventos) e lema de Borel-Cantelli.

2.1 Experimentos aleatórios

Definicão 2.1.1. ***(Experimentos aleatórios)*** *Experimentos aleatórios são experimento que, ao serem repetidos nas mesmas condições, não produzem o mesmo resultado. Por outro lado, experimentos que, ao serem repetidos nas mesmas condições, produzem o mesmo resultado são chamado de experimentos determinísticos.*

2.2 Espaço amostral, evento e sigma-álgebra

Definicão 2.2.1. ***(Espaço amostral)*** *Espaço amostral ou espaço amostral universal, Ω, é o conjunto de todos os resultados possíveis de um experimento aleatório.*

Definição 2.2.2. ***(Eventos)*** *Qualquer subconjunto do espaço amostral $\mathcal{S}$ que constitui um campo de Borel $\mathcal{F}$.*

Definição 2.2.3. ***(Eventos mutuamente exclusivos)*** *Quando a ocorrência de um impossibilita a ocorrência do outro.*

Exemplo 2.2.1. *Exemplo: Dado*

$$\left.\begin{array}{l} A = \{par\} \\ B = \{impar\} \end{array}\right\rangle A \cdot B = \emptyset \quad \textbf{(eventos mutuamente exclusivos)}$$

É importante definir a **sigma-álgebra associado ao espaço amostral** (às vezes citados nos livros como σ-álgebra).

Definição 2.2.4. ***(Sigma-álgebra)*** *Uma família de subconjuntos de Ω é chamada de σ-álgebra (ou campo de Borel), denotado por $\mathcal{B}$, se satisfazer as três seguintes propriedades:*

a. $\varnothing \in \mathcal{B}$

b. Se $A \in \mathcal{B}$*, então* $A^c \in \mathcal{B}$

c. Se $A_1, A_2, ... \in \mathcal{B}$*, então* $\cup_{i=1}^{\infty} A_i \in \mathcal{B}$

Uma σ-álgebra é, portanto, o conjunto de todos os subconjuntos do espaço amostral (Ω), incluindo o próprio espaço amostral. Assim, se Ω tem n elementos, então existem 2^n conjuntos em σ-álgebra. Atente-se que isso ocorre quando temos um conjunto contável. Quando Ω não for contável, será difícil descrever a σ-álgebra, entretanto ela é escolhida para conter qualquer conjunto que seja de interesse.

Exemplo 2.2.2. *Se $X = \{a, b, c, d\}$, uma possível σ-álgebra em X é $\Sigma = \{\oslash, \{a, b\}, \{c, d\}, \{a, b, c, d\}\}$, no qual $\oslash$ é o conjunto vazio e Σ é a notação para o conjunto das sigma-álgebra.* •

Mas qual o motivo de estudar a sigma-álgebra, ao invés de todos os subconjuntos?. Existem algumas explicações para isso e elencamos duas a seguir:

i. O espaço amostral pode conter um grau de detalhamento superior ao que estamos interessados no momento;

ii. Queremos associar cada evento A com uma probabilidade numérica, $\mathbb{P}(A)$, porém nosso conhecimento sobre $\mathbb{P}$ pode não se estender para todos os subconjuntos de Σ.

2.3 Definições de probabilidade: subjetiva, frequentista e axiomática

A definição subjetiva de probabilidade refere-se a uma tentativa para lidarmos com eventos históricos únicos, que não podem ser repetidos, carecendo, assim, de interpretação frequencial. Em sentido não rigoroso,

a probabilidade subjetiva pode ser interpretada com a chance que uma pessoa atribuiria a aposta em um evento.

A definição frequentista de probabilidade refere-se ao fato de que se repetirmos um experimento aleatório n vezes e anotarmos o número de vezes a qual um resultado de seu interesse (um evento A, por exemplo) ocorreu, então a frequência relativa de A nas n repetições do experimento é dada por:

$$f_{n,A} = \frac{n(A)}{n}. \tag{2.1}$$

Essa frequência relativa $f_{n,A}$, definida na classe dos subconjuntos do espaço amostral, satisfaz as seguintes condições:

i. $0 \leq f_{n,A} \leq 1$

ii. $f_{n,\Omega} = 1$

iii. Se A e B forem eventos mutualmente excludentes, então:

$$f_{n,A\cup B} = f_{n,A} + f_{n,B}$$

Alguns livros (Introductory Statistics, Wonnacott e Wonnacott, 1980) admitem que uma frequência relativa de um evento tenderá para um valor limite dado por:

$$\mathbb{P}(A) = \lim_{n\to\infty} \frac{n(A)}{n} \tag{2.2}$$

É possível que a frequência $f_{n,A}$ se comporte de maneira não esperada, isto é, podemos jogar um dado um grande número de vezes e o lado do número cinco, por exemplo, persiste em aparecer, tornando a probabilidade para o lado cinco igual tendendo a um.

É necessário qualificar, portanto, afirmando que o limite ocorre com grande probabilidade, mas não com certeza lógica. Então, se utilizarmos o limite anteriormente como definição para probabilidade, estaríamos utilizando o conceito de probabilidade para definir probabilidade, formando um ciclo vicioso. Com o objetivo de romper este ciclo, devemos apelar para o enfoque axiomático.

Definicão 2.3.1. ***(Probabilidade)*** *Uma função $\mathbb{P}$, definida na σ-álgebra $\mathcal{F}$ de subconjuntos de Ω e com valores em $[0,1]$, é uma probabilidade se satisfaz os Axiomas de Kolmogorov:*

1. $\mathbb{P}(\Omega) = 1$;
2. *Pata todo subconjunto* $A \in \mathcal{F}$, $\mathbb{P}(A) \geq 0$;
3. *Para toda sequência* $A_1, A_2, ... \in \mathcal{F}$, *multualmente exclusivos, temos*

$$\mathbb{P}\left(\cup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

2.4 Propriedades da probabilidade

Dado $(\Omega, \mathcal{F}, \mathbb{P})$, considere que os os conjuntos mencionados abaixo são eventos nesse espaço de probabilidade. Dessa forma, temos:

i. $\mathbb{P}(A) = 1 - \mathbb{P}(A^c)$;

ii. Sendo A e B dois eventos quaisquer, vale $\mathbb{P}(B) = \mathbb{P}(B \cap A) + \mathbb{P}(B \cap A^c)$;

iii. Se $A \subset B$, então $\mathbb{P}(A) \leq \mathbb{P}(B)$;

iv. $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$;

v. $\mathbb{P}\left(\cup_{i=1}^{\infty} A_i\right) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

2.5 Probabilidade condicional

No início do estudo da Probabilidade, são apresentados conceitos os quais não existem restrições para o espaço amostral, ou seja, ele é sempre o mesmo e o cálculo das probabilidades, consequentemente, é **incondicional**.

Entretanto, em muitos casos, é necessária que uma atualização desse espaço amostral seja feita, pelo fato de algum elemento ter sido retirado dele, ocosionando, portanto, sua redução. É a partir dessa ideia que surge o que chamamos de **probabilidade condicional**, cuja definição formal é apresentada a seguir.

Definicão 2.5.1. *Se A e B são eventos em Ω (espaço amostral) e $\mathbb{P}(B) > 0$, então a probabilidade condicional de A dado B, denotada por $\mathbb{P}(A|B)$, é dada por*

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Atente-se para o fato de que podemos dizer que $\mathbb{P}(A|B)$ é uma probabilidade, pois:

- $\mathbb{P}(A|B) = \frac{\mathbb{P}(AB)}{\mathbb{P}(B)} \geq 0$;
- $\mathbb{P}(\mathcal{S}|B) = 1$;
- Para $A \cdot C = \emptyset \Rightarrow \mathbb{P}\left[(A + C)|B\right] = \mathbb{P}(A|B) + \mathbb{P}(C|B)$.

Note que agora temos uma **redução do espaço amostral** para B, ou seja, o que antes considerávamos Ω, agora será restrito à B. Portanto, para o cálculo das probabilidade, vamos considerar $\mathbb{P}(B|B) = 1$.

Para **ilustrar essa definição**, considere uma urna com duas bolas azuis e duas bolas brancas. Suponha que desejamos retirar duas bolas, uma após a outra. Podemos nos perguntar quais os possíveis casos (espaço amostral) para esse experimento. Se adotarmos a notação A_i: a i-ésima bola retirada é de cor azul e B_i a

i-ésima bola retirada é de cor branca (para $i = 1, 2$), então o **espaço amostral desse experimento** pode ser descrito da seguinte forma:

$$\Omega = \{A_1A_2, B_1B_2, A_1B_2, B_1A_2\}.$$

Entretanto, podemos limitar esse espaço amostral condicionando o evento. Considere que seja de interesse calcular a probabilidade de a segunda bola retirada ser de cor branca, dado que a primeira também foi de cor branca. Perceba que, agora, o espaço amostral do nosso interesse não inclui mais os eventos em que a primeira bola retirada foi a de cor azul, e sim somente aqueles em que a primeira bola retirada foi de cor branca, ou seja, ouve uma **redução do espaço amostral**.

Como o exemplo é simples, podemos calcular a probabilidade sem muitos cálculos. Se temos somente uma possibilidade de interesse (a primeira ser branca e a segunda também ser branca: B_1B_2) entre duas possíveis (a primeira ser branca e segunda também ser branca ou ser azul: B_1B_2 ou B_1A_2), então a probabilidade de interesse é $1/2$.

Usando a definição anteriormente apresentada para probabilidades condicionais, podemos resolver esse problema calculando $\mathbb{P}(B_2|B_1)$, ou seja:

$$\mathbb{P}(B_2|B_1) = \frac{\mathbb{P}(B_1 \cap B_2)}{\mathbb{P}(B_1)} = \frac{1/4}{1/2} = \frac{1}{2}.$$

2.5.1 Regra do produto de probabilidades

O desenvolvimento apresentado anteriormente para dois eventos pode ser generalizada com objetivo de denotar a probabilidade da interseção de n eventos por meio das probabilidades condicionais sucessivas. Veja a definição a seguir.

Definição 2.5.2. ***(Regra do produto de probabilidades)*** *Para os eventos $A_1, A_2, ..., A_n$ em $(\Omega, \mathcal{F}, \mathbb{P})$, com $\mathbb{P}\left(\cap_{i=1}^{n} A_i\right) > 0$, a regra do produto de probabilidades é dada por:*

$$\mathbb{P}(A_1B_2...A_n) = \mathbb{P}(A_1)\mathbb{P}(A_2|A_1)\mathbb{P}(A_3|A_1A_2)...\mathbb{P}(A_n|A_1A_2A_{n-1}).$$

2.5.2 Teorema da probabilidade total

O teorema das probabilidades totais permite calcular a probabilidade de um evento A quando se conhece as probabilidades de um conjunto de eventos disjuntos cuja reunião é o espaço amostral, bem como as probabilidades condicionais de A dado cada um deles.

Definição 2.5.3. ***(Teorema da probabilidade total)*** *Considere $B_1, B_2, ..., B_n$ uma partição do espaço amostral Ω (são eventos mutualmente excludentes e sua reunião forma Ω). Considere também A um evento e $\mathbb{P}$ uma probabilidade definida nos eventos de Ω, então:*

$$\mathbb{P}(A) = \sum_{k=1}^{n} \mathbb{P}(A|B_k)\mathbb{P}(B_k) \tag{2.3}$$

A probabilidade condicional, definida na classe dos eventos do espaço amostral, satisfaz as propriedades estabelecidas a seguir:

i. Para todo evento B, $\mathbb{P}(B|A) \geq 0$;

ii. Se $B_1, B_2, ..., B_n$ são eventos mutualmente exclusivos, então:

$$\mathbb{P}\left(\bigcup_{k=1}^{n} B_k|A\right) = \sum_{k=1}^{n} \mathbb{P}(B_k|A);$$

iii. Se Ω denota o espaço amostral, então $\mathbb{P}(\Omega|\Omega) = 1$.

2.5.3 Teorema de Bayes

Definição 2.5.4. ***(Teorema de Bayes)*** *Considere uma partição $A_1, A_2, ..., A_n$ do espaço amostra Ω (note que a partição é finita) e B um evento de Ω, então para $i = 1, 2, ..., n$, temos:*

$$\mathbb{P}(A_i|B) = \frac{\mathbb{P}(B|A_i)\mathbb{P}(A_i)}{\sum_{k=1}^{n} \mathbb{P}(B|A_k)\mathbb{P}(A_k)}. \tag{2.4}$$

A probabilidade dada em (2.4) é conhecida na literatura como probabilidade a posteriori. Além disso, a partir da mesma expressão, temos que:

$$\sum_{k=1}^{n} \mathbb{P}(A_i|B) = 1.$$

2.5.4 Independência de eventos

Definição 2.5.5. ***(Eventos independentes)*** *Sejam A e B dois eventos e suponha que $\mathbb{P}(A) \geq 0$. O evento B é dito ser independente do evento A se:*

$$\mathbb{P}(B|A) = \mathbb{P}(B). \tag{2.5}$$

Então, sendo $\mathbb{P}(B|A)$ diferente de $\mathbb{P}(B)$, dizemos que B depende estatisticamente de A, ou é dependente estatisticamente de A. A dependência estatística é o caso usual, pois é muito mais fácil duas probabilidades serem tanto diferentes do que serem extremamente iguais.

Em outras palavras, a definição anterior aplica, para eventos independentes, o tipo mais simples de regra da multiplicação. Além disso, dado que temos eventos independentes, então podemos concluir que:

1. $\mathbb{P}(A|B) = \mathbb{P}(A)$;

2. $\mathbb{P}(\overline{A}B) = \mathbb{P}(\overline{A}) \cdot \mathbb{P}(B)$;

3. $\mathbb{P}(A\overline{B}) = \mathbb{P}(A) \cdot \mathbb{P}(\overline{B})$ e $\mathbb{P}(\overline{AB}) = \mathbb{P}(\overline{A}) \cdot \mathbb{P}(\overline{B})$.

Ou seja, se A e B são independentes, A e $\overline{B}$ são independentes e $\overline{A}$ e $\overline{B}$ também o são.

De uma maneira mais geral, temos:

Definição 2.5.6. *Sejam $A_1, A_2, \ldots, A_n$ eventos. Eles serão independentes se:*

$$\mathbb{P}(A_{i1}A_{i2}\ldots A_{ik}) = \mathbb{P}(A_{i1})\mathbb{P}(A_{i2})\cdots\mathbb{P}(A_{ik}) \tag{2.6}$$

2.6 Exercícios

1. Uma moeda honesta é lançada até que apareça o mesmo resultados duas vezes seguidas.

   a. Descreva o espaço amostral.

   b. Encontre a probabilidade de que o experimento termine antes de 6 lançamentos.

2. Sabe-se que os eventos B_1, B_2 e B_3 são disjuntos par a par e que sua união é igual ao espaço amostral. Estes eventos têm as probabilidades $\mathbb{P}(B_1) = 0,2$ e $\mathbb{P}(B_2) = 0,3$. Existe um outro evento A tal que $\mathbb{P}(A|B_1) = 0,3$, $\mathbb{P}(A|B_2) = 0,4$ e $\mathbb{P}(A|B_3) = 0,1$. Calcule:

   a. $\mathbb{P}(A)$.

   b. $\mathbb{P}(B_2|A)$.

3. (Meyer, 1983) Uma remessa de 1500 arruelas contém 400 peças defeituosas e 1100 perfeitas. Duzentas arruelas são escolhidas ao acaso (sem reposição) e classificadas.

   a. Qual a probabilidade de que sejam encontradas exatamente 90 peças defeituosas?

   b. Qual a probabilidade de que sejam encontradas ao menos 2 peças defeituosas?

4. (Meyer, 1983) Dez fichas numeradas de 1 a 10 são misturadas em uma urna. Duas fichas, numeradas (X,Y), são extraídas da urna, sucessivamente e sem reposição. Qual é a probabilidade de que X+Y=10.

5. (Meyer, 1983) Um mecanismo complexo pode falhar em 15 estágios. De quantas maneiras poderá ocorrer que ele falhe em 3 estágios?

6. Se $\mathbb{P}(A) = 0,7$ e $\mathbb{P}(A \cup B) = 0,8$, então encontre a $\mathbb{P}(B)$ sabendo que A e B são eventos independentes.

7. Qual a probabilidade de sair menos que três caras em cinco lançamentos de uma moeda honesta?

8. Um inspetor de qualidade extrai uma amostra de dez tubos aleatoriamente de carga muito grande de tubos que se sabe que contém 20% de tubos defeituosos. Qual é a probabilidade de que não mais do que dois tubos extraídos sejam defeituosos?

9. (Mirshawka, 1983) A confiabilidade de um foguete é a probabilidade p que em uma tentativa de lançamento o mesmo com sucesso. Foi estabelecido que a confiabilidade de um certo foguete é $0,9$. Qual opção a seguir indica maior evidência na hipótese de que realmente houve modificação no foguete?

Opção 1: De 121 foguetes modificados testados, 116 tiveram desempenho satisfatório.

Opção 2: De 81 foguetes modificados testados, 78 tiveram desempenho satisfatório.

10. Um engenheiro extrai uma amostra de 15 itens aleatoriamente de um processo de fabricação, no qual é sabido que tal processo produz 35% de itens aceitáveis. Qual é a probabilidade de que dez dos itens extraídos sejam aceitáveis? (R. 0.0450)

11. Acredita-se que 20% dos moradores das proximidades de uma grande siderúrgica têm alergia aos poluentes lançados ao ar. Admitindo que este percentual de alérgicos seja correto (real), calcule a probabilidade de que pelo menos quatro moradores tenham alergia entre 13 selecionados ao acaso.

CAPÍTULO 3

Variáveis aleatórias

Entender o que são variáveis aleatórias é um passo fundamental no estudo da Estatística, pois elas representam as características de interesse em uma população. Para exemplificar, considere que você está sentado na calçada da sua rua contando o número de carros que passam por dia. Se, nesse caso, definirmos por X o número de carros que passam por dia nessa rua, podemos dizer que X é uma variável aleatória.

Essa variável aleatória pode ser discreta ou contínua e tem diversas funções importantes associadas a ela, tais como a função de distribuição acumulada, funçôes de probabilidade e densidade e momentos. Ao longo desse capítulo vamos abordar todos esse pontos com detalhes. Porém, antes de avançar, vamos apresentar a definição formal de variável aleatória.

3.1 Conceito de variável aleatória

Dado um fenômeno aleatório qualquer, com certo espaço de probabilidade, desejamos estudar a estrutura probabilística de quantidades associadas a esse fenômeno.

Definicão 3.1.1. ***(Variável aleatória)*** *Seja $(\Omega, \mathcal{F}, \mathbb{P})$ um espaço de probabilidade. Uma variável aleatória X é qualquer função $X : \Omega \to \mathbb{R}$, tal que:*

$$X^{-1}(I) = \{\omega \in \Omega : X(\omega) \in I\} \in \mathcal{F}, \ \ \forall I \subset \mathbb{R}. \tag{3.1}$$

Traduzindo a definição matemática acima, X é uma variável aleatória se sua imagem inversa para intervalos $I \subset \mathbb{R}$ pertencem a σ-álgebra $\mathcal{F}$. Veja essa representação na Figura 3.1 a seguir.

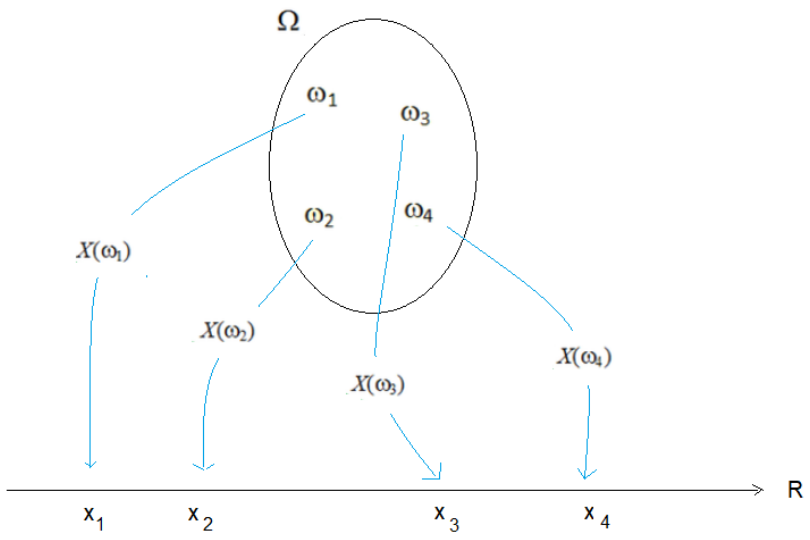

Figura 3.1: Ilustração de uma função de uma variável aleatória.

De maneira menos informal, se considerarmos um experimento e um espaço amostral Ω associado a esse experimento e considerarmos que X é uma função associa cada elemento de Ω a um número real $X(\Omega)$, então X é uma variável aleatória.

3.2 Variáveis aleatórias discretas, contínuas e mistas

As variáveis aleatórias podem ser classificadas em variáveis aleatórias discretas, contínuas e mistas. Essa caracterização prévia da variável é muito importante no processo de modelagem de dados, pois, como veremos mais adiante, existem modelos para cada tipo de variável, sendo ela discreta, contínua ou mista.

Para exemplificar, suponha que em uma lanchonente sejam vendidos 300 pastéis por dia. Assim, se definirmos a variável aleatória X, tal que X é o número de pastéis vendidos em um dia, então X é classificada como uma variável aleatória do tipo discreta.

Definicão 3.2.1. ***(Variável aleatória discreta)*** *Uma variável aleatória é do tipo discreta se assume somente um número enumerável de valores.*

Como sabemos que os valores possíveis para X são $0, 1, 2, ..., 300$ e tais valores não são igualmente prováveis de ocorrer, então X é classificada como uma variável aletória do tipo discreta.

Por outro lado, se nessa mesma lanchonote definirmos como X o tempo de trabalho diário dos funcionários, então X é classificada como uma variável aleatória do tipo contínua.

Definicão 3.2.2. ***(Variável aleatória contínua)*** *Uma variável aleatória é do tipo contínua se ela assume qualquer valor numérico em um determinado intervalo ou série de intervalos. Isto é, uma variável aleatória contínua é uma variável para a qual um conjunto A é um conjunto infinito não enumerável.*

Um exemplo de uma variável aleatória mista pode ser um experimento em que uma moeda é lançada e uma roleta é girada se o resultado do lançamento da moeda for cara. Se o resultado do lançamento da moeda for

cara, X é igual ao valor da roleta. Se o resultado do lançamento da moeda for coroa, X é igual a -1. Há a probabilidade meio de essa variável aleatória ter o valor -1, e meio de ficar no intervalo $[0, 360)$.

Definição 3.2.3. ***(Variável aleatória mista)*** *Uma variável aleatória é do tipo mista se ela assume tanto valores discretos quanto valores em um determinado intervalo. Essas variáveis aleatórias são conhecidas como variáveis aleatórias mistas.*

3.3 Função de probabilidade e função densidade de probabilidade

Conforme alertamos anteriormente, caracterizar uma variável aleatóra em discreta, contínua ou mista é uma passo muito importante para a modelagem estatística. Isso ocorre, pois as respectivas funções de probabilidade recebem nomes diferentes dependendo da sua caracterização.

Se uma varável é do tipo discreta, então ela pode ser modelada pela sua respectiva função de probabilidade. Por outro lado, se uma varável é do tipo contínua, então ela pode ser modelada pela sua respectiva função densidade de probabilidade.

Definição 3.3.1. ***(Função de probabilidade)*** *A função de probabilidade de uma variável aleatória discreta é uma função que atribue probabilidade a cada um dos possíveis valores assumidos pela variável. Assim, considerando X uma variável com valores $x_1, ..., x_n$, temos que*

$$\mathbb{P}(X = x_i) = \mathbb{P}(\{\omega \in \Omega : X(\Omega) = x_i\}), \quad i = 1, ..., n. \tag{3.2}$$

A função de probabilidade de X, no espaço de probabilidade $(\Omega, \mathcal{F}, \mathbb{P})$, deve obedecer às seguintes propriddes:

- $0 \leq \mathbb{P}(X = x_i) \leq 1, \quad \forall i = 1, 2, ...;$
- $\sum_i \mathbb{P}(X = x_i) = 1$, com a soma percorrendo todos os possíveis valores.

Exemplo 3.3.1. *(Magalhães, 2006) Obtenha o valor da constante c, de modo que a função*

$$p(x) = c(x - 2)^2, \;\; x = 3, 4, 5, 6,$$

seja uma função de probabilidade de alguma variável aleatória discreta. •

Com a apresentação da função de probabilidade para a variável aleatória discreta, agora vamos apresentar a a função densidade de probabilidade para a variável aleatória contínua.

Definição 3.3.2. ***(Função densidade de probabilidade)*** *Uma variável aleatória X em $(\Omega, \mathcal{F}, \mathbb{P})$, com função de dsitribuição F, será classificada como contínua, se existir uma função não negativa f tal que:*

$$F(x) = \int_{-\infty}^{x} f(\omega) d\omega, \quad \forall x \in \mathbb{R}, \tag{3.3}$$

com f sendo a função densidade de probabilidade da variável aleatória X.

Assim como a função de probabilidade, a função densidade de probabilidade de X, no espaço de probabilidade $(\Omega, \mathcal{F}, \mathbb{P})$, deve obedecer às seguintes propriedades:

- $f(x) \geq 0, \quad \forall x \in \mathbb{R}$;
- $\int_{-\infty}^{\infty} f(w)dw = 1$.

Exemplo 3.3.2. *Obtenha o valor da constante c, de modo que a função*

$$f(x) = ce^{-cx}\ \mathbb{I}_{[0,\infty)}(x)$$

seja uma função densidade de probabilidade de alguma variável aleatória contínua. •

3.4 Função de distribuição acumulada e função de sobrevivência

Definição 3.4.1. ***(Função de distribuição acumulada)*** *Seja $(\Omega, \mathcal{F}, \mathbb{P})$ um espaço de probabilidade, a função de distribuição acumulada de uma variável aleatória X é definida por:*

$$F_X(x) = P(X \leq x), \qquad \forall x \in \mathbb{R}. \tag{3.4}$$

A função de distribuição acumulada de X, no espaço de probabilidade $(\Omega, \mathcal{F}, \mathbb{P})$, deve obedecer às seguintes propriedades:

- $\lim_{x\to-\infty} F_X(x) = 0$ e $\lim_{x\to\infty} F_X(x) = 1$.
- $F_X(x)$ é uma função não decrescente de x.
- $F_X(x)$ é uma função contínua à direita, isto é, para cada número x_0, $\lim_{x\to-x_0} F_X(x) = F_X(x_0)$.

Exemplo 3.4.1. *Obtenha a função de distribuição acumulada da função densidade de probabilidade dada por*

$$f(x) = 2e^{-2x}\ \mathbb{I}_{[0,\infty)}(x).$$

•

Definição 3.4.2. ***(Função de sobrevivência)*** *Seja $(\Omega, \mathcal{F}, \mathbb{P})$ um espaço de probabilidade. A função de sobrevivência de uma variável aleatória X é definida por:*

$$S_X(x) = P(X > x), \qquad \forall x \in \mathbb{R}. \tag{3.5}$$

Exemplo 3.4.2. *Obtenha a função de sobrevivência da função densidade de probabilidade dada por*

$$f(x) = 2e^{-2x}\ \mathbb{I}_{[0,\infty)}(x).$$

•

3.5 Histograma

O histograma é uma forma simples e rápida de avaliarmos o comportamento da variável em estudo e, assim, realizar associações com distribuições de probabildades conhecidas.

O que, às vezes, não percebemos é que ao construir um histograma, estamos trabalhando com processo de estimação, dado que **esse gráfico consiste em uma estimativa não paramétrica de uma função densidade**.

Nesse sentido, ao utilizar um *software* para gerar esse gráfico, não sabemos como funciona esse processo de construção. Nesta seção, vamos detalhar esse processo.

Podemos resumir a ideia geral da construção desse gráfico em três simples passos, conforme descrevemos abaixo:

Passo 1: Dividir o intervalo dos dados em h classes;

Passo 2: Alocar cada observação em sua respectiva classe;

Passo 3: Calcular a proporção da amostra contida em cada classe e dividir pelo produto entre a largura da classe e o tamanho da amostra.

Essa proporção, calculada no último passo, é representada pelas alturas das barras no histograma, que consiste na estimativa não paramétrica da função densidade de probabilidade.

De um modo mais geral, podemos definir o histograma como uma função $\hat{f}$, representada da seguinte forma:

$$\hat{f}_\lambda(x) = \frac{1}{nh} \sum_{i=1}^{n} \mathbb{I}(x - \gamma_i, h/2), \tag{3.6}$$

sendo n o tamanho amostral, h a largura da classe, γ_i o ponto central da classe da observação x_i e $\mathbb{I}(\cdot)$ a função indicadora do intervalo $[-h/2, h/2]$.

Antes de contuarmos essa explicação, precisamos abrir um parânteses para apresentar uma confusão muito comum que ocorre em algumas análises que usam histograma. Para isso, observe os histogramas da Figura 3.2 construídos para uma amostra de 10 valores da variável aleatória X, tais que $X \sim \mathcal{N}(0, 1)$, gerada no *software* R de acordo com os comandos dados a seguir.

```
> x = rnorm(10,0,1)
> round(x,1)
 [1]  0.5  0.9 -0.8 -0.2 -2.0 -1.7  1.5 -0.4 -0.6  0.7
```

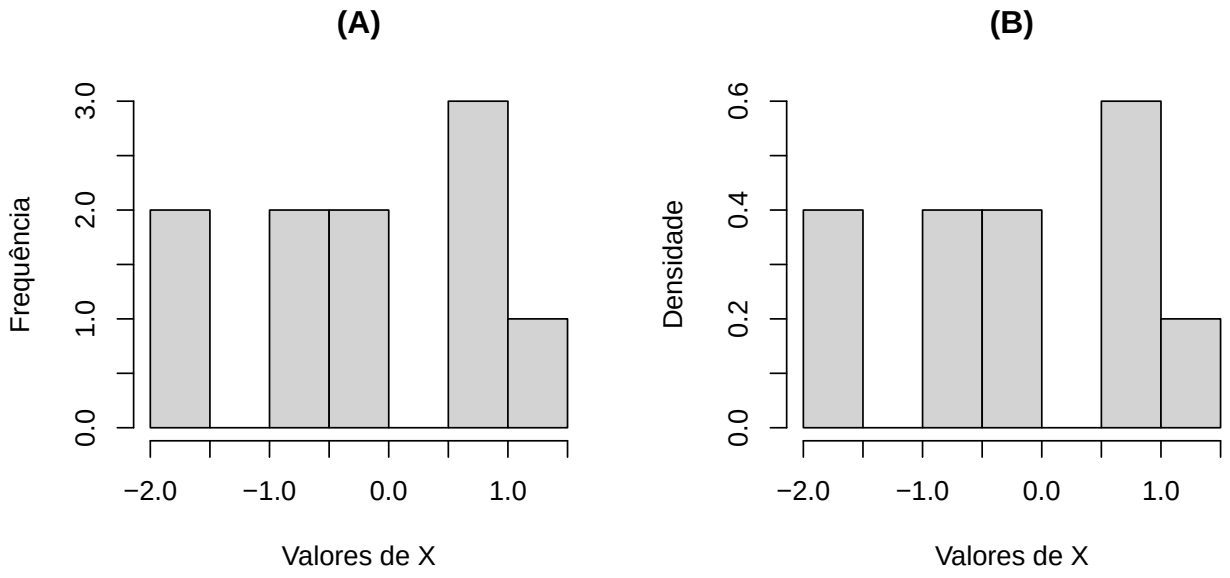

Figura 3.2: Histograma de frequância e de densidade, respectivamente, construído para uma amostra de 10 valores da variável aleatória X, tal que $X \sim \mathcal{N}(0, 1)$.

Na Figura 3.2 (A), temos o histograma de frequências, que considera somente a proporção em relação ao tamanho da classe, representada por h na expressão $\hat{f}_\lambda(x)$. Esse não é o histograma que nos fornece a estimativa da densidade. Já na Figura 3.2 (B), temos os histograma de interesse. Ele considera a proporção em relação ao tamanho da classe e em relação ao tamanho da amostra, representada por h e n, respectivamente, na expressão de $\hat{f}_\lambda(x)$.

Com esse problema de confundimento apresentado e entendido, vamos voltar ao estudo do histograma como função de estimação. Observe que o formato de histograma depende do número de classes que serão utilizadas na sua estimação. Vamos avaliar, portanto, dois histogramas construídos sobre o mesmo conjunto de dados X (com $n = 100$), tais que $X \sim \mathcal{N}(0, 1)$, mas com número de classes diferentes ($h = 5$ e $h = 20$, respectivamente), com a inclusão da curva da densidade conhecida em cada histograma.

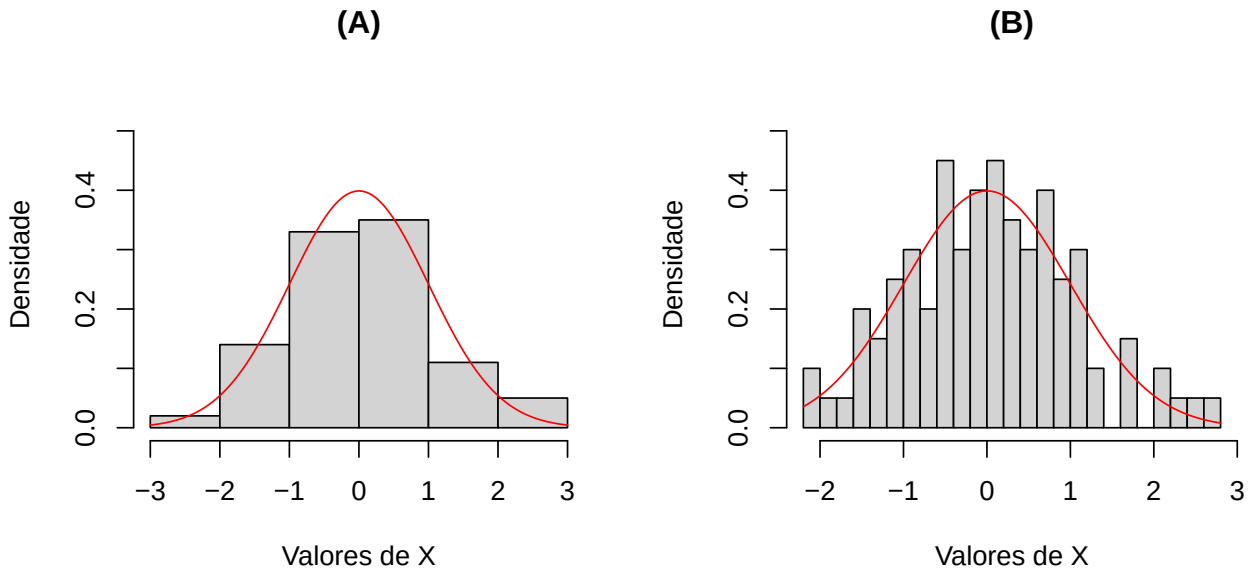

Figura 3.3: Histogramas construídos sobre o mesmo conjunto de dados X (com $n = 100$), tais que $X \sim \mathcal{N}(0, 1)$, mas com número de classes diferentes ($h = 5$ e $h = 20$, respectivamente).

Note que na Figura 3.3 (A), temos um número menor de classes e, portanto, um comportamento sobresuavizado da curva. Já na Figura 3.3 (B), temos um número maior de classes e, portanto, um comportamento subsuavizado da curva.

Dessa forma, podemos observar importância do parâmetro h na estimação da curva de probabilidade, dado que para diferentes valores desse parâmetro, temos diferentes formatos de histogramas. A essa parâmetro damos o nome de **parâmetro de suavização**. Em regressão não paramétrica, esse parâmetro é muito utilizado em diversas abordagens e a compreensão do seu papel no contexto de histogramas é fundamental para o estudo desse tipo de regressão.

Note como ficou mais simples, agora, entendermos o papel da ferramenta mais importante em regressão não paramétrica (o chamado suavizador), conforme alertamos no início desse capítulo.

3.6 Distribuições de probabilidade mais comuns

3.6.1 Distribuição Bernoulli

Definição 3.6.1. ***(Distribuição Bernoulli)*** *Dizemos que uma variável aleatória segue uma distribuição de Bernoulli de parâmetro p, $X \sim Bernoulli(p)$, se ela asssume apenas os valores 0 ou 1. Sua função de probabilidade é dada por*

$$\mathbb{P}(X = 0) = p \quad e \quad \mathbb{P}(X = 1) = 1 - p.$$

Exemplo 3.6.1. *Considere uma caixa com R bolas, sendo a amarelas e $b = R - a$ brancas. Considerando que*

as bolas na caixa são idênticas e apresentam igual probabilidade de serem sorteadas, foram retiradas algumas delas (com reposição) e o objetivo do estudo é avaliar o resultado da primeira extração, sendo o evento de interesse definido pela variável aleatória

$$X : \textit{a primeira bola extraída é amarela.}$$

Dessa forma, temos que:

$$X = \begin{cases} 1, & a\ bola\ é\ amarela \\ 0, & a\ bola\ é\ branca. \end{cases}$$

Facilmente, podemos encontrar a probabilidade da primeira extração ser uma bola de cor amarela, que é dada por:

$$P(X = 1) = \frac{a}{R}.$$

E, também, a probabilidade da primeira extração ser uma bola de cor branca, que é dada por:

$$P(X = 0) = \frac{R - a}{R}.$$

•

3.6.2 Distribuição Binomial

Definicão 3.6.2. ***(Distribuição Binomial)*** *Dizemos que uma variável aleatória segue uma distribuição de Binomial de parâmetros n e p, $X \sim Binomial(n, p)$, quando ela representa o número de sucessos obtidos com a realização de n ensaios Bernoulli independentes. Sua função de probabilidade é dada por*

$$\mathbb{P}(X = x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, ..., n. \tag{3.7}$$

Exemplo 3.6.2. *A probabilidade de um certo componente elétrico estar em condições operacionais satisfatórias é de 0,85. Em uma amostra de cinco componentes, calcula a probabilidade de se encontrar zero itens defeituosos.*

Solução à cargo do leitor.

•

3.6.3 Distribuição Poisson

Definicão 3.6.3. ***(Distribuição Poisson)*** *Dizemos que uma variável aleatória segue uma distribuição de Poisson de parâmetro λ, $X \sim Poisson(\lambda)$, se sua função de probabilidade é dada por*

$$\mathbb{P}(X = x) = \frac{e^{-\lambda}\lambda^k}{k!}, \quad k = 0, 1, \ldots. \tag{3.8}$$

Exemplo 3.6.3. *O número de telefonemas que chegam à uma unidade de atendimento é modelado por um modelo de poisson com taxa de 2 ligações por minuto. Para uma minuto qualquer, calcule a probabilidade de ocorrer pelo menos uma ligação.*

Solução à cargo do leitor.

•

3.6.4 Distribuição Uniforme

Definição 3.6.4. ***(Distribuição Uniforme)*** *Dizemos que uma variável aleatória segue uma distribuição uniforme no intervalo $[a, b]$, $X \sim U[a, b]$, se sua função densidade de probabilidade é dada por:*

$$f(x) = \frac{1}{b-a}, \ \ a \leq x \leq b, \tag{3.9}$$

com a e b parâmetros reais, sendo $a < b$.

Vale ressaltar que a distribuição uniforme também é conhecida como distribuição retangular. Além disso, Se $a = 0$ e $b = 1$ temos a chamada distribuição uniforme padrão.

A distribuição uniforme no intervalo [0,1] é usada para simulação de amostras aleatórias de uma determinada variável aleatória contínua X. Ela é usada como modelo probabilístico em situações nas quais temos certeza que intervalos reais de mesmo comprimento tenham a mesma chance de ocorrer, isto é, $P(x \in [a, b]) = P(x \in [c, d])$, desde que $b - a = d - c$.

3.6.5 Distribuição Exponencial

A distribuição exponencial tem grande atuação na modelagem de problemas que descrevem tempos de vida, seja de indivíduos, produtos ou objetos. Funciona de modo análogo ao uso da distribuição geométrica no caso discreto.

Definição 3.6.5. ***(Distribuição Exponencial)*** *Dizemos que uma variável aleatória segue uma distribuição exponencial de parâmetro λ, $X \sim exp(\lambda)$, se sua função densidade de probabilidade é dada por:*

$$f(x) = \lambda e^{-\lambda x}, \ \ \mathbb{I}_{(0,\infty)}(x). \tag{3.10}$$

Uma das mais importantes leis de falhas é aquela cuja duração até falhar é descrita pela distribuição exponencial. Podemos caracterizá-la de muitas maneiras, mas, provavelmente, a maneira mais simples é supor que a taxa de falhas é constante, isto é:

$$\lambda(t) = \frac{f(t)}{S(t)} = \lambda \tag{3.11}$$

Definição 3.6.6. *Seja T, a duração até falhar, uma variável aleatória contínua, que tome todos os valores não negativos. Então. T terá uma distribuição exponencial se, e somente se, tiver uma taxa de falhas constante.*

A propriedade que afirma que a distribuição exponencial não tem memória funciona no seguinte sentido: suponha que X represente o tempo de vida de algum componente. Suponha também que o componente tenha sobrevivido a 'a' unidades de tempo de operação. Assim a probabilidade que o componente sobreviva a mais 'b' unidades de tempo operação será a mesma que o componente tenha sobrevivido anteriormente a 'b' unidades de tempo de operação. Simplesmente, a informação adicional é esquecida. Abaixo temos a prova dessa propriedade.

$$P(x > a + b|x > a) = P(x > b) \tag{3.12}$$

Prova:

$$\begin{aligned} P(x > a + b|x > a) &= \frac{P(x > a + b \cap x > a)}{x > a} \\ &= \frac{P(x > a + b)}{x > a} \\ &= \frac{S(a + b)}{S(a)} \\ &= \frac{e^{-\lambda(a+b)}}{e^{-\lambda a}} \\ &= \frac{e^{-\lambda a}e^{-\lambda b}}{e^{-\lambda a}} \\ &= e^{-\lambda b} \\ &= P(x > b) \end{aligned}$$

Suponha que a falha em um equipamento tenha ocorrido devido à algum fator aleatório. Seja X_t a variável que representa o número de tais perturbações ocorridas em um determinado intervalo de tempo t, com $X_t > 0$, então podemos admitir que tal situação se trata de um Processo de Poisson. Quer dizer, para qualquer t fixado a variável aleatória X_t tem distribuição de Poisson com parâmetro αt. Sendo T a duração até falhar, então $T > t$ ocorre se, e somente se, não ocorrer perturbação entre [0,t]. Issoacontecerá se, e somente se, $X_t = 0$. Por isso

$$F(t) = 1 - P(X_t = 0) = 1 - e^{-\alpha t}$$

Encontramos, portanto, que a "causa"da taxa de falhas acima envolve uma lei de falhas exponencial.

Comentário: Podemos generalizar o caso acima se desejarmos que a pertubarção ocorra com determinada probabilidade. Agora, $T > t$ se, e somente se, durante [0,t] nenhuma perturbação ocorra, ou uma perturbação ocorra e não resulte em falha, ou duas perturbações ocorram e não resultem em falha, e assim por diante, de modo que possamos contar o número de pertubações e que tenhamos a probalidade disso acontecer, temos

$$\begin{aligned} F(t) &= 1 - \left[e^{-\alpha t} + (\alpha t)e^{-\alpha t}p + (\alpha t)^2\frac{e^{-\alpha t}}{2!}p^2 + ...\right] \\ &= 1 - e^{-\alpha(1-p)^t} \end{aligned}$$

Note que quando o valor do parãmetro b é igual a 1, a distribuição Weibull se reduz a distribuição exponencial de parâmetro a.

3.6.6 Distribuição Normal

Definicão 3.6.7. ***(Distribuição Normal)*** *Dizemos que uma variável aleatória segue uma distribuição Normal (ou gaussiana) de parâmetros μ e σ^2, $X \sim \mathcal{N}(\mu, \sigma^2)$, se sua função densidade de probabilidade é dada por:*

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} exp \frac{(x-\mu)^2}{2\sigma^2} \mathbb{I}_{(-\infty,\infty)}(x). \tag{3.13}$$

Exemplo 3.6.4. *A concentração (em ppm, partícula por milhão) de um poluente em água liberada por uma fábrica tem distribuição $\mathcal{N}(8; 1, 5)$. Qual a probabilidade de que num dado dia a concentração do poluente exceda o limite regulatório de 9 ppm?*

Solução à cargo do leitor.

•

3.6.7 Distribuição Gama

Definicão 3.6.8. ***(Distribuição Gama)*** *Dizemos que uma variável aleatória segue uma distribuição Gama de parâmetros $\beta > 0$ e $\lambda > 0$, $X \sim Gama(\beta, \lambda)$, se sua função densidade de probabilidade é da forma:*

$$f(x) = \frac{\lambda^\beta}{\Gamma(\beta)} x^{\beta-1} e^{-\lambda x}. \tag{3.14}$$

A distribuição exponencial é um caso particular da distribuição Gama. Isso ocorre quando a função densidade de probabilidade apresentada tem o parámetro $b = 1$. Nesse caso, a densidade se reduz à distribuição exponencial de parâmetro λ. Além disso, se o parâmetro β for inteiro positivo, a distribuição gama é chamada de distribuição de Earlang. O parâmetro β passará então a ser denotado po k, sendo k inteiro.

Quando $\lambda = 1/2$ e $\beta = k/2$, sendo k inteiro e positivo, a distribuição gama se reduz a uma distribuição qui-quadrado com k graus de liberdade (veremos a frente o estudo detalhado de tal distribuição).

Um resultado muito utilizado nos cálculos de algumas integrais pode ser obtido a partir da chamada a função gama, que é dada por:

$$\int_0^\infty x^{\beta-1} e^{-\lambda x} dx = \frac{\Gamma(\beta)}{\lambda^\beta}$$

3.6.8 Distribuição qui-quadrado

A distribuição qui-quadrado é uma das distribuições mais utilizadas em estatística inferencial. Ele é utilizada nos chamados testes qui-quadrado para aderência, homogeneidade e indepedência.

Definição 3.6.9. ***(Distribuição qui-quadrado)*** *Dizemos que uma variável aleatória segue uma distribuição qui-quadrado com v graus de liberdade, $X \sim \chi^2(v)$, se sua função densidade de probabilidade é dada por*

$$f(x) = \frac{1}{2^{n/2}\Gamma(n/2)} x^{n/2-1} e^{-x^2/2},$$

com $x > 0$, $n > 0$ e $\Gamma(w) = \int_0^\infty x^{w-1}e^{-x}$, com $w > 0$.

O conceito de **graus de liberdade** pode ser interpretado como o número de valores, de um conjunto de dados qualquer, que podem variar após terem sido impostas certas restrições a esses valores. Para um melhor entendimento, considere que 10 máquinas produzem 100 litros de mel por dia, gerando uma produção diária total de 1000 litros (essa é a restrição). Dessa forma, o conhecimento da produção de 9 máquinas permite saber a produção da décima, pois as 9 produções podem ser escolhidas aleatoriamente, mas a décima deve obedecer à restrição. Portanto, nesse caso, temos $10 - 1 = 9$ graus de liberdade.

Vale resslatar que a soma de n variáveis aleatórias com distribuição normal padrão ao quadrado resulta em uma distribuição qui-quadrado com n graus de liberdade. Ou seja, se $Z \sim \mathcal{N}(0, 1)$, então $X = \sum_{i=1}^n Z_i^2 = \chi^2(n)$.

A **média** e a **variância** de uma variável aleatória X com distribuição qui-quadrado com n graus de liberdade são dadas por:

$$\mathbb{E}(X) = n \quad \text{e} \quad \mathrm{Var}(X) = 2n.$$

3.6.9 Distribuição t de Student

A distribuição t de Student, assim como a qui-quadrado, é uma das distribuições mais utilizadas em estatística inferencial. Ele é utilizada nos chamados testes t para comparação de médias, dentres outros.

Definição 3.6.10. ***(Distribuição t de Student)*** *Dizemos que uma variável aleatória segue uma distribuição t de Student com n graus de liberdade, $X \sim t(n)$, se sua função densidade de probabilidade é dada por*

$$\frac{\Gamma\left(\frac{v+1}{2}\right)}{\sqrt{v\pi}\Gamma(v/2)} \left(1 + \frac{t^2}{v}\right)^{-\left(\frac{v+1}{2}\right)},$$

com $x > 0$, $n > 0$ e $\Gamma(w) = \int_0^\infty x^{w-1}e^{-x}$, com $w > 0$.

O nome dessa distribuição faz referência ao seu autor, que se chamava Student. O fato curioso é que Student é o pseudônimo de William Sealy Gosset, já que William não podia usar seu nome verdadeiro para publicar trabalhos enquanto trabalhasse para a cervejaria Guinness.

3.6.10 Distribuição F de Snedecor

A distribuição F de Snedecor, assim como as distribuições t de Student e qui-quadrado, é uma das distribuições mais utilizadas em estatística inferencial. Ele é utilizada nos chamados testes t para comparação de médias, dentres outros.

Definição 3.6.11. ***(Distribuição F de Snedecor)*** *Dizemos que uma variável aleatória segue uma distribuição F de Snedecor de parâmetros m e n, $X \sim F(m,n)$, se sua função densidade de probabilidade é da forma:*

$$f(x) = \frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}\left(\frac{m}{n}\right)^{\frac{m}{2}-1}\frac{\frac{m}{2}-1}{\left(1+\frac{m}{n}x\right)^{\frac{m+n}{2}}} \quad I_{(0,\infty)}(x). \tag{3.15}$$

A **média** e a **variância** de uma variável aleatória X com distribuição F de Snedecor de parâmetros m e n são dadas por:

$$\mathbb{E}(X) = \frac{n}{n-2} \qquad n > 2 \quad \text{e} \quad \text{Var}(X) = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)} \qquad n > 4.$$

3.7 Momentos

Os momentos de uma variável aleatória são importantes para caracterizar distribuições de probabilidade. Como veremos mais adiante, a distribuição normal (ou gaussiana), por exemplo, é caracterizada apenas pelo seu momento de ordem 1 e pelo seu momento central de ordem 2, que representam a média (medida de tendência central) e a variância (medida de dispersão), respectivamente.

Nas próximas duas subseções, apresentamos a definição do momento de ordem k e do momento centrado de ordem k de uma variável aleatória X.

3.7.1 Momento de ordem k

Definição 3.7.1. ***(Momento de ordem k de uma variável aleatória)*** *O momento de ordem k de uma variável aleatória X é dado por*

$$\mu_k = \mathbb{E}(X^k), \tag{3.16}$$

se $\mathbb{E}(|X^n|) < \infty$.

Dessa forma, se X for uma variável aleatória discreta, então teremos seu momento de ordem k sendo calculado como

$$\mu_k = \mathbb{E}(X^k) = \sum_A x^k \cdot \mathbb{P}(X = x), \tag{3.17}$$

com A representando o suporte de X. E se X for uma variável aleatória contínua, então teremos seu momento

de ordem k sendo calculado como:

$$\mu_k = \mathbb{E}(X^k) = \int_A x^k \cdot f_X(x), \tag{3.18}$$

com A representando o suporte de X.

Exemplo 3.7.1. *Se $X \sim U(a, b)$ então:*

$$E[x^k] = \int_a^b x^k \frac{1}{b-a} dx = \frac{1}{b-a} \int_a^b x^k dx = \frac{b^{k+1} - a^{k+1}}{(b-a)(k+1)},$$

•

Define-se a média de uma variável aleatória X, ou valor esperado de uma variável aleatória, como o momento de ordem 1. Apresentamos isso de uma maneira formal a a seguir.

Definicão 3.7.2. ***(Valor esperado de uma variável aleatória)*** *O valor esperado de uma variável aleatória X é dado por*

$$\mu = \mathbb{E}(X). \tag{3.19}$$

3.7.2 Momento centrado de ordem k

Definicão 3.7.3. ***(Momento centrado de ordem k de uma variável aleatória)*** *O momento de ordem k de uma variável aleatória X é dado por*

$$\mu_k = \mathbb{E}((X - \mu)^k). \tag{3.20}$$

Dessa forma, se X for uma variável aleatória discreta, então teremos seu momento central de ordem k sendo calculado como

$$\mu_k = \mathbb{E}((X - \mu)^k) = \sum_A (x - \mu)^k \cdot \mathbb{P}(X = x), \tag{3.21}$$

com A representando o suporte de X. E se X for uma variável aleatória contínua, então teremos seu momento central de ordem k sendo calculado como:

$$\mu_k = \mathbb{E}((X - \mu)^k) = \int_A (x - \mu)^k \cdot f_X(x), \tag{3.22}$$

com A representando o suporte de X.

Exemplo 3.7.2. *Se $X \sim Bin(n,p)$ então:*

$$\begin{aligned}
\alpha_r &= E[(x-\mu)^r] \\
&= E\left[\sum_{i=0}^{r}\binom{r}{i}(-\mu)^i x^{r-i}\right] \\
&= \sum_{i=0}^{r}\binom{r}{i}(-\mu)^i E\left[x^{r-i}\right] \\
&= \sum_{i=0}^{n}\binom{r}{i}(-\mu)^i \frac{1}{\lambda^i}\frac{(r-i)!}{\lambda^{r-i}} \\
&= \frac{1}{\lambda^r}\sum_{0}^{r}\frac{(-1)^i}{i!}.
\end{aligned}$$

é o seu respectivo momento centrado de ordem k.

•

Define-se a variância de uma variável aleatória X como o momento central de ordem 2. Apresentamos isso de uma maneira formal a a seguir.

Definição 3.7.4. ***(Variância de uma variável aleatória)*** *Sendo $\mu < \infty$, a variância de uma variável aleatória X é dada por*

$$\alpha_2 = \sigma^2 = \mathbb{E}((X-\mu)^2). \tag{3.23}$$

Da mesma forma, define-se o coeficiente de assimetria de uma variável aleatória X como o momento central de ordem 3. Apresentamos isso de uma maneira formal a a seguir.

Definição 3.7.5. ***(Coeficiente de assimetria)*** *Sendo $\mu < \infty$, o coeficiente de assimetria de uma variável aleatória X é dada por*

$$\alpha_3 = \mathbb{E}((X-\mu)^3). \tag{3.24}$$

Para cada valor do coeficnente de assimetria, temos uma interpretação possível, que segue os pontos abaixo:

- Se $\alpha_3 < 0$, então a distribuição é assimétrica negativa;
- Se $\alpha_3 = 0$, então a distribuição é simétrica;
- Se $\alpha_3 > 0$, então a distribuição é assimétrica positiva.

E, ainda, define-se o coeficiente de curtose de uma variável aleatória X como o momento central de ordem 4. Apresentamos isso de uma maneira formal a a seguir.

Definição 3.7.6. ***(Coeficiente de curtose)*** *Sendo $\mu < \infty$, o coeficiente de curtose de uma variável aleatória X é dada por*

$$\alpha_4 = \mathbb{E}((X-\mu)^4). \tag{3.25}$$

Para cada valor do coeficnente de curtose, temos uma interpretação possível, que segue os pontos abaixo:

- Se $\alpha_4 < 3$, então a distribuição é platicúrtica (a distribuição é mais achatada que a distribuição normal);
- Se $\alpha_4 = 3$, então a distribuição é mesocúrtica (a distribuição tem o mesmo achatamento que a distribuição normal);
- Se $\alpha_4 > 3$, então a distribuição é leptocúrtica (a distribuição em questão é mais alta, ou afunilada, e concentrada que a distribuição normal, apresentando caudas pesadas, o que significa dizer que é relativamente fácil obter valores que não se aproximam da média a vários múltiplos do desvio padrão).

3.8 Funções auxiliares

3.8.1 Função geradora de momentos

A função geradora de momentos de uma variável aleatória, como o próprio nome já induz, é uma função cuja derivação permite a obteção dos momentos dessa variável. Podemos caracterizá-la, ainda, como a digital de uma variável aleatória, dado que que ela determina completamente a distribuição de probabilidades.

Definicão 3.8.1. ***(Função geradora de momentos)*** *A função geradora de momentos de uma variável aleatória X é dada por*

$$M_x(t) = \mathbb{E}(e^{tX}), \tag{3.26}$$

desde que a esperança seja finita para t real em algum intervalo $t_0 < t < t_0$, com $t_0 > 0$.

Dessa forma, podemos usar a r-ésima derivada da expressão (3.26) e aplicar no ponto $t = 0$ para obtero r-ésimo momento da respectiva variável.

Definicão 3.8.2. *Seja X uma variável aleatória com função geradora de momentos $M_X(t)$, então o r-ésimo momento da variável aleatória X é obtido a partir de:*

$$\mu'_r = \frac{d^r M_x(t)}{dt_r},$$

aplicado no ponto $t = 0$.

A esperança matemática de uma variável aleatória, por exemplo, pode ser obtida derivando uma vez a função geradora de momentos e aplicando a função no ponto $t = 0$.

Dois pontos importantes precisam ser levantados quando falamos dessa função: (i.) a função geradora de uma soma de variáveis aleatórias independentes é o produto das funções geradoras de cada componenteda soma; e (iv.) a convergência ordinária de uma sequência de funções geradoras corresponde à convergênciadas correspondentes distribuições.

Exemplo 3.8.1. *Vamos encontrar a função geradora de momentos da variável aleatória X com distribuição $Bin(n, p)$, de acordo com a definição apresentada anteriomente.*

Solução:

Já vimos que se $X \sim Bin(n,p)$, então sua função de probabilidade é dada por

$$P(X=x) = \binom{n}{x} p^x (1-p)^{n-x}.$$

Dessa forma, temos:

$$\begin{aligned} M_X(t) &= \sum_{x=0}^{n} e^{tx} P(X=x) \\ &= \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x (1-p)^{n-x} \\ &= \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x (1-p)^{n-x} \\ &= [pe^t + (1-p)]^n \end{aligned}$$

•

3.8.2 Função característica

A função geradora de momentos é uma ferramenta útil dado que que ela determina completamente a distribuição de probabilidades. Entretanto, a integral que define essa função pode nem sempre ser finita.

A partir disso, define-se uma nova transformada que, diferentemente da função geradora de momento, tem a vantagem de sempre existir. Embora apresente essa vantagem, as funções características são um pouco mais complicadas, dado que envolvem números complexos.

Definição 3.8.3. ***(Função característica)*** *A função característica de uma variável aleatória X é dada por*

$$\phi_x(t) = \mathbb{E}(e^{itX}) = \mathbb{E}(cos(tX)) + i\mathbb{E}(sen(tX)) \tag{3.27}$$

para $t \in \mathbb{R}$ e $i = \sqrt{-1}$.

3.9 Exercícios

1. (Magalhães, 2006) Obtenha o valor da constante c, de modo que a função

$$p(x) = c(x-2)^2, \;\; x = 3,4,5,6,$$

seja uma função de probabilidade de alguma variável aleatória discreta.

2. Obtenha o valor da constante c, de modo que a função

$$f(x) = ce^{-cx} \;\; \mathbb{I}_{[0,\infty)}(x)$$

seja uma função densidade de probabilidade de alguma variável aleatória contínua.

3. Obtenha a função de distribuição acumulada da função densidade de probabilidade dada por

$$f(x) = 2e^{-2x}\ \mathbb{I}_{[0,\infty)}(x).$$

4. Obtenha a função de sobrevivência da função densidade de probabilidade dada por

$$f(x) = 2e^{-2x}\ \mathbb{I}_{[0,\infty)}(x).$$

5. A probabilidade de um certo componente elétrico estar em condições operacionais satisfatóriasé de 0,85. Em uma amostra de cinco componentes, calcula a probabilidade de se encontrar zero itens defeituosos.

6. O número de telefonemas que chegam à uma unidade de atendimento é modelado por ummodelo de poisson com taxa de 2 ligações por minuto. Para uma minuto qualquer, calcule a probabilidade deocorrer pelo menos uma ligação.

7. A concentração (em ppm, partícula por milhão) de um poluente em água liberada por uma fábrica tem distribuição $\mathcal{N}(8; 1, 5)$. Qual a probabilidade de que num dado dia a concentração do poluente exceda o limite regulatório de 9 ppm?

8. Mostre, para cada caso a seguir, que a função de probabilidade ou função de probabilidade da variável aleatória X é legítima (somatória ou integral é igual a 1).
   a. $X \sim Bin(n, p)$
   b. $X \sim U(a, b)$
   c. $X \sim exp(\lambda)$
   d. $X \sim N(\mu, \sigma^2)$
   e. $X \sim Weibull(a, b)$
   f. $X \sim Gama(\lambda, \beta)$
   g. $X \sim Beta(a, b)$

9. Encontre a função de distribuição acumulada da variável aleatória X, tal que:
   a. $X \sim Bin(n, p)$
   b. $X \sim U(a, b)$
   c. $X \sim exp(\lambda)$
   d. $X \sim N(\mu, \sigma^2)$
   e. $X \sim Weibull(a, b)$
   f. $X \sim Gama(\lambda, \beta)$
   g. $X \sim Beta(a, b)$

10. Encontre o q-ésimo quantil variável aleatória X, tal que:
    a. $X \sim Bin(n, p)$
    b. $X \sim U(a, b)$
    c. $X \sim exp(\lambda)$
    d. $X \sim N(\mu, \sigma^2)$
    e. $X \sim Weibull(a, b)$
    f. $X \sim Gama(\lambda, \beta)$
    g. $X \sim Beta(a, b)$
11. Encontre a moda variável aleatória X, tal que:
    a. $X \sim Bin(n, p)$
    b. $X \sim U(a, b)$
    c. $X \sim exp(\lambda)$
    d. $X \sim N(\mu, \sigma^2)$
    e. $X \sim Weibull(a, b)$
    f. $X \sim Gama(\lambda, \beta)$
    g. $X \sim Beta(a, b)$
12. Encontre a função de sobrevivência da variável aleatória X, tal que:
    a. $X \sim Bin(n, p)$
    b. $X \sim U(a, b)$
    c. $X \sim exp(\lambda)$
    d. $X \sim N(\mu, \sigma^2)$
    e. $X \sim Weibull(a, b)$
    f. $X \sim Gama(\lambda, \beta)$
    g. $X \sim Beta(a, b)$
13. Encontre o momento de ordem k da variável aleatória X, tal que:
    a. $X \sim Bin(n, p)$
    b. $X \sim U(a, b)$
    c. $X \sim exp(\lambda)$
    d. $X \sim N(\mu, \sigma^2)$
    e. $X \sim Weibull(a, b)$

f. $X \sim Gama(\lambda, \beta)$

Com esse resultado é possível encontrar média e variância dessa distribuição (a cargo do leitor).

g. $X \sim Beta(a, b)$

14. Encontre o momento central de ordem k da variável aleatória X, tal que:

a. $X \sim Bin(n, p)$

b. $X \sim U(a, b)$

c. $X \sim exp(\lambda)$

d. $X \sim N(\mu, \sigma^2)$

e. $X \sim Weibull(a, b)$

f. $X \sim Gama(\lambda, \beta)$

g. $X \sim Beta(a, b)$

15. Encontre o coeficiente de assimetria (interprete seu resultado) da variável aleatória X, tal que:

a. $X \sim Bin(n, p)$

b. $X \sim U(a, b)$

c. $X \sim exp(\lambda)$

d. $X \sim N(\mu, \sigma^2)$

e. $X \sim Weibull(a, b)$

f. $X \sim Gama(\lambda, \beta)$

g. $X \sim Beta(a, b)$

16. Encontre o coeficiente de curtose (interprete seu resultado) da variável aleatória X, tal que:

a. $X \sim Bin(n, p)$

b. $X \sim U(a, b)$

c. $X \sim exp(\lambda)$

d. $X \sim N(\mu, \sigma^2)$

e. $X \sim Weibull(a, b)$

f. $X \sim Gama(\lambda, \beta)$

g. $X \sim Beta(a, b)$

17. Encontre a função geradora de momentos de X, tal que:

a. $X \sim Bin(n, p)$

b. $X \sim U(a, b)$

c. $X \sim exp(\lambda)$

d. $X \sim N(\mu, \sigma^2)$

e. $X \sim Weibull(a, b)$

f. $X \sim Gama(\lambda, \beta)$

g. $X \sim Beta(a, b)$

18. Encontre a função característica de X, tal que:

a. $X \sim Bin(n, p)$

b. $X \sim U(a, b)$

c. $X \sim exp(\lambda)$

d. $X \sim N(\mu, \sigma^2)$.

e. $X \sim Weibull(a, b)$

f. $X \sim Gama(\lambda, \beta)$

g. $X \sim Beta(a, b)$

19. Se $X \sim N(\mu, \sigma^2)$, então qual a distribuição de $Z = \frac{X-\mu}{\sigma}$?

20. Se $Z \sim N(0, 1)$, então qual a distribuição de $X = \mu + \sigma z$?

21. Se $Z \sim N(0, 1)$, então qual a distribuição de $W = Z^2$?

22. Se $X \sim U(a, b)$, então qual a distribuição de $Y = F(x)$?

23. Se $X \sim U(a, b)$, então então qual a distribuição de $Y == a + (b - a)Y$?

24. Se $X \sim Weibull(a, b)$, então qual a distribuição de $Y = ax^b$?

25. Se $X \sim Beta(a, b)$, então qual a distribuição de $Y = 1 - X$?

CAPÍTULO 4

Conceitos básicos em Inferência Estatística

4.1 População e amostra

As duas principais definições básicas em Inferência Estatística são as de população e amostra. Esses são os dois conceitos que são repetidos inúmeras vezes nesse material e que são explanados a seguir.

Definição 4.1.1. ***(População)*** *O conjunto de valores de uma característica (observável) associada a uma coleção de indivíduos ou objetos de interesse é dito ser uma população.*

Definição 4.1.2. ***(Amostra)*** *O conjunto de valores de uma característica (observável) associada a uma coleção de indivíduos ou objetos de interesse é dito ser uma população.*

4.2 Amostra aleatória

Dentro do conceito de amostra, nós podemos ter diferentes métodos de amostragem. O mais utilizado dentro da Inferência é o conceito de Amostra Aleatória Simples, que está explanado a seguir.

Definição 4.2.1. ***(Amostra Aleatória Simples)*** *Seja X uma variável aleatória com distribuição de probabilidade ou função densidade de probabilidade $f(x;\theta)$. Sejam $X_1, X_2, ..., X_n$ variáveis aleatórias independentes e identicamente distribuídas (têm a mesma distribuição), indicando n retiradas de uma população P. Então, $X_1, X_2, ..., X_n$ é uma amostra aleatória simples de P.*

Com a amostra aleatória simples definida, nós podemos assumir que sucessivas coletas desse tipo de amostragem gera o que chamamos de distribuição amostral. Veja a seguir a definição desse conceito.

Definição 4.2.2. ***(Distribuição da amostra)*** *Seja $X_1, X_2, ..., X_n$ uma amostra aleatória de uma variável aleatória com distribuição $f(x_i;\theta)$. A distribuição da respectiva amostra é a distribuição conjunta de $X_1, X_2, ..., X_n$ dada por:*

$$\begin{aligned} f_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) &= f_{X_1}(x_1;\theta) f_{X_2}(x_2;\theta) ... f_{X_n}(x_n;\theta) \\ &= \prod_{i=1}^{n} f_{X_i}(x_i;\theta). \end{aligned}$$

A essa distribuição conjunta dá-se o nome de ***função de verossimilhança****.*

É, portanto, a partir da amostra $X_1, X_2, ..., X_n$ que conseguimos obter informação sobre o parâmetro θ de interesse. Mas como ainda não sabemos o que é um parâmetro, vamos explicar na próxima subseção.

4.3 Parâmetro e espaço paramétrico

Definição 4.3.1. ***(Parâmetro)*** *São características quantitativas da população em estudo, que podem estar presentes em modelos probabilísticos. Em geral são desconhecidos e têm-se o objetivo de estimá-los.*

Vale ressaltar que o parâmetro existe, mesmo que não se tenha um modelo para os dados. A seguir, definimos o intervalo de valores que esse parâmetro pode assumir.

Definição 4.3.2. ***(Espaço Paramétrico)*** *Seja $X_1, X_2, ..., X_n$ uma amostra aleatória de uma variável aleatória com distribuição $f(x_i;\theta)$, em que θ é o parâmetro associado à distribuição. O intervalo Θ de valores possíveis para θ é chamado espaço paramétrico.*

Para que seja possível coletar alguma informação sobre o parâmetro de uma população, precisamos ter informações sobre a amostra a partir de uma estatística e de um estimador. Na subseção a seguir apresentamos a definição de estimador.

4.4 Estatísticas e estimadores

Outras duas definições importantes relacionadas aos conceitos iniciais em Inferência são as de **estatística** e **estimador**. Essas duas definições estão intimamente relacionadas, mas diferem entre si, conforme apresentamos abaixo.

Definição 4.4.1. ***(Estatística)*** *É qualquer função de uma amostra aleatória que não depende do parâmetro. A média, a mediana, o mínimo e o máximo, por exemplo, são exemplos de estatísticas.*

Sendo a estatística uma função da amostra, então é interessante que possamos usar uma estatística para encontrar um estimador para o parâmetro de interesse. Vamos entender, portanto, a definição do estimador.

Definição 4.4.2. ***(Estimador)*** *É qualquer estatística utilizada com objetivo de estimar o parâmetro. A estatística que assuma qualquer valor em Θ é uma estimador para θ.*

A seguir, vamos apresentar algumas propriedades importantes tantos das estatística quanto dos estimadores.

4.5 Estimadores e suas particularidades

4.5.1 Estimador não viciado

Encontrar um estimador razoável para o parâmetro desconhecido θ, ou para uma função $g(\theta)$, é um dos grandes problemas da inferência. Um procedimento utilizado para avaliar um estimador, a qual daremos mais detalhes somente no capitulo sobre estimação pontual, especidicamente na seção sobre avaliação de estimadores pontuais, é o chamado Erro Quadrático Médio (EQM). A seguir apresentamos a definição do EQM (outros detalhes serão discutidos na seção citada anteriormente).

Definição 4.5.1. *Seja $X_1, X_2, ..., X_n$ uma amostra aleatória de tamanho n da variável aleatória com distribuição $f(x_i|\theta)$. Seja $T(X_1, X_2, ..., X_n)$ um estimador para o parâmetro θ. O Erro Quadrático Médio (EQM) de $\widehat{\theta}$ é dado por*

$$EQM(\widehat{\theta}) \quad = \quad \mathbb{E}[(\widehat{\theta} - \theta)^2].$$

Desenvolvendo a expressão dada anteriormente, temos que $\mathrm{EQM}(\widehat{\theta}) = \mathrm{Var}(\widehat{\theta}) - B^2(\widehat{\theta})$, sendo $B(\widehat{\theta}) = \mathbb{E}(\widehat{\theta}) - \theta$ o vício do estimador. Dessa forma, dizemos que um estimador é não viciado, se $B(\widehat{\theta}) = 0$, para todo $\theta \in \Theta$, ou seja, se $\mathbb{E}(\widehat{\theta}) = \theta$.

Além disso, caso $\lim_{n \Rightarrow \infty} B(\widehat{\theta}) = 0$, dizemos que o estimador é **assintoticamente não viciado**, ou seja, à medida que o tamanho da amostra aumenta, o vício do estimador tende a zero e, consequentemente, $E(\widehat{\theta}) = 0$. O EQM de um estimador não viciado, ou assintoticamente não viciado, reduz-se a sua variância, ou seja, $\mathrm{EQM}(\widehat{\theta}) = \mathrm{Var}(\widehat{\theta})$.

Conforme citamos no início dessa subseção, daremos uma explicação mais ampla sobre o Erro Quadrático Médio no capítulo sobre estimação pontual, no qual apresentaremos, também, outros métodos de avaliação de estimadores.

4.5.2 Estimador eficiente

O estimador eficiente é aquele que **atinge o limite inferior da variância dos estimadores não viciados**. Formalmente, temos:

Definição 4.5.2. ***(Eficiência de um estimador)*** *A eficiência de um estimador $\widehat{\theta}$, do parâmetro θ, é dado pelo quociente*

$$e(\widehat{\theta}) = \frac{LI(\theta)}{Var(\widehat{\theta})}, \tag{4.1}$$

no qual $LI(\theta) = \dfrac{1}{n\mathbb{E}\left[\left(\dfrac{\partial \log f(X|\theta)}{\partial \theta}\right)^2\right]}$*, sob certas condições de regularidade [a], é o limite inferior da variância dos estimadores não viciados de θ.*

[a] O suporte $A(x) = \{x, f(x > 0)\}$ deve ser independente de θ e a troca das ordens de derivação e integração, sob a distribuição da variável aleatória X, deve ser possível.

Atente-se que quando o quociente $e(\widehat{\theta})$ é igual a um, **a variância do estimador coincide com o limite inferior da variância dos estimadores não viciados do parâmetro** e, portanto, o estimador é eficiente (essa é a ideia!). Vejamos o exemplo 4.5.1 a seguir.

Exemplo 4.5.1. *Vamos verificar se o estimador $\overline{X}$, obtido a partir de uma amostra de distribuição normal, é eficiente. Para isso, considere que $X_1, X_2, ..., X_n$ é uma amostra aleatória da variável aleatória $X \sim \mathcal{N}(\mu, \sigma^2)$ com σ^2 conhecido. Dessa forma, temos que*

$$\mathbb{E}\left[\left(\frac{\partial \log f(X|\mu)}{\partial \mu}\right)^2\right] = \frac{1}{\sigma^2}.$$

Assim, $LI(\mu) = \dfrac{\sigma^2}{n}$. Se considerarmos o estimador $\overline{X}$ como estimador para μ, temos que $Var(\overline{X}) = \dfrac{\sigma^2}{n}$ e, consequentemente, $e(\overline{X}) = 1$, ou seja, $\overline{X}$ é eficiente para μ. •

Ressalta-se ainda, pelo Exemplo 4.5.1, que $\mathbb{E}\left[\dfrac{\partial \log f(X|\mu)}{\partial \mu}\right] = 0$. Esse resultado (substiuindo o parâmetro μ por um θ genérico) vale, em geral, quando valem as condições de regularidade.

A seguir definimos duas quantidades muito importantes no estudo de inferência: **Função escore** e **Informação de Fisher**.

Definição 4.5.3. ***(Função escore)*** *A quantidade*

$$\frac{\partial \log f(X|\theta)}{\partial \theta}$$

é chamada de funçao escore.

No qual, conforme já citado, sob as condições de regularidade, têm-se que

$$\mathbb{E}\left[\frac{\partial \log f(X|\theta)}{\partial \theta}\right] = 0,$$

ou seja, o valor esperado da função escore é sempre igual a zero.

Definição 4.5.4. ***(Informação de Fisher)*** *A quantidade*

$$I_F(\theta) = \mathbb{E}\left[\left(\frac{\partial \log f(X|\theta)}{\partial \theta}\right)^2\right],$$

é denominada informação de Fisher de θ.

A partir da Definição 4.5.4 e utilizando o fato de que $\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X)$ e $\mathbb{E}(X) = 0$, temos que

$$I_F(\theta) = \mathrm{Var}\left[\frac{\partial \log f(X|\theta)}{\partial \theta}\right].$$

A seguir apresentamos duas propriedades importantes. A primeira estabelece que

$$\mathbb{E}\left[\left(\frac{\partial \log f(X|\theta)}{\partial \theta}\right)^2\right] = -\mathbb{E}\left[\left(\frac{\partial^2 \log f(X|\theta)}{\partial^2 \theta}\right)\right]$$

E a segunda estabelece que a informação total de Fisher de θ correspondente à amostra $X_1, X_2, ..., X_n$ – da variável aleatória X com f.d.p. (ou f.p.) $f(x|\theta)$ e informação de Fisher $I_F(\theta)$ – observada é soma da informação de Fisher das n observações da amostra, que denotamos por $nI_F(\theta)$. Isso ocorre, pois X_i, para $i = 1, 2, ..., n$, têm a mesma informação que X.

Teorema 4.5.1. ***(Desigualdade de Cramer-Rao)*** *Quando as condições de regularidade estão satisfeitas, a variância de qualquer estimador não viciado de* θ *satisfaz a desigualdade*

$$\begin{aligned}
\mathrm{Var}(\widehat{\theta}) &\geq \frac{1}{\mathbb{E}\left[\left(\frac{\partial \log f(X|\theta)}{\partial \theta}\right)^2\right]} \\
&\geq \frac{1}{-\mathbb{E}\left[\frac{\partial^2 \log f(\mathbf{x};\theta)}{\partial^2 \theta}\right]} \\
&\geq \frac{1}{nI_F(\theta)}.
\end{aligned}$$

Em muitos textos sobre o tema, o leitor pode encontrar **Desigualdade da Informação** ao invés de **Desigualdade de Cramer-Rao**. Para ilustrar o Teorema 4.5.1, apresentamos o exemplo seguinte.

Exemplo 4.5.2. *Seja* $X_1, \ldots, X_n$ *uma amostra aleatória de uma distribuição uniforme no intervalo* $[0, \theta]$. *Neste caso, temos que a função densidade de probabilidade* $f(x|\theta)$ *é dada por*

$$f(x|\theta) = \frac{1}{\theta} \qquad \mathbb{I}_{(0,\theta)}(x).$$

Então, temos que

$$\frac{\partial}{\partial \theta} \log(f(x|\theta)) = \frac{\partial}{\partial \theta}[-\log(\theta)] = -\frac{1}{\theta}$$

De forma que

$$\mathbb{E}\left(\left(\frac{\partial}{\partial\theta}\log f(X|\theta)\right)^2\right) = \frac{1}{\theta^2}.$$

Assim, considerando $\widehat{\theta}$ um estimador para θ, segue da Desigualdade de Cramer-Rao, que

$$Var(\widehat{\theta}) \geq \frac{\theta^2}{n}.$$

Para verificarmos a desigualdade, considere o estimador $\widehat{\theta} = X_{(n)}$. Temos que

$$\mathbb{E}(X_{(n)}) = \frac{n}{n+1}\theta.$$

Assim, se fizermos $\widehat{w} = \frac{n+1}{n}X_{(n)}$, concluimos que $\widehat{w}$ é não viciado para θ e

$$Var(\widehat{w}) = \frac{1}{n(n+2)}\theta^2.$$

Agora sim, podemos comparar com a desigualdade, já que temos um estimador não viciado. Porém, note que a Desigualdade de Cramér-Rao não é obedecida neste caso. Isso se justifica pelo fato do Teorema 4.5.1 valer sob as condições de regularidades satisfeitas, situação que não ocorre nesse caso, já que o suporte da distribuição depende do parâmetro de interesse.

Se considerarmos, entretanto, $X_1, X_2, ..., X_n$ uma amostra aleatória de tamanho n da variável aleatória $X \sim Poisson(\theta)$, com função de probabilidade dada por

$$f(x|\theta) = \frac{e^{-\theta}\theta^x}{x!},$$

então $\widehat{\theta} = \overline{X}$, por exemplo, é um estimador não viciado para θ e eficiente, valendo a desigualdade de Cramer-Rao, pois o suporte da distribuição não depende do parâmetro. •

Avaliamos, nesse último tópico, o estudo sobre a eficiência de um estimador. Continuamos o estudo, mas agora para verificar sua otimalidade. Para ser ótimo, segundo o critério do menor EQM, o estimador deve ser função de uma **Estatística Suficiente**. A seguir, apresentamos tal a definição para essa estatística.

4.5.3 Estimador consistente

Os métodos de estimação de máxima verossimilhança e o de método dos momentos produzem, em geral, estimadores consistentes, ou seja, à medida que o tamanho da amostra aumenta, os estimadores ficam tão próximos do parâmetro que está sendo estimado quanto desejado.

Definição 4.5.5. *Seja $X_1, X_2, ..., X_n$ uma amostra aleatória de tamanho n de uma variável aleatória com distribuição $f(x_i;\theta)$. Seja $\widehat{\theta} = \mathbf{T}(X_1, X_2, ..., X_n)$ um estimador do parâmetro θ. Então, $\widehat{\theta}$ será consistente se:*

$$\lim_{n\to\infty} \mathbb{P}(|\widehat{\theta} - \theta|) = 0.$$

Ou seja, a medida que o tamanho da amostra é aumentado, a diferença entre o estimador e o parâmetro é zero. Dessa forma, para n grande, teremos:

$$\begin{aligned} \mathbb{E}(\widehat{\theta}) &= \theta \\ \mathbb{V}(\widehat{\theta}) &= 0. \end{aligned}$$

Observe que consistência está realacionada ao conceito de convergência em probabilidade.

4.6 Exercícios

1. (Bolfarine and Sandoval, 2001) Sejam $X_1, X_2, , X_n$ uma amostra aleatória de tamanho n da distribuição da variável aleatória X com f.d.p. dada por:

$$f(x|\theta) = e^{-(x-\theta)}, \qquad x > \theta, \qquad \theta > 0.$$

   a. Especifique o espaço paramétrico e o suporte associado à distribuição de X.

   b. Verifique se $\widehat{\theta}_1 = \overline{X}$ e $\widehat{\theta}_2 = X_{(1)}$ são estimadores não viciados para θ.

   c. Encontre e compare os EQM dos dois estimadores. Faça um gráfico como função de θ.

2. Seja $X_1, X_2, \cdots, X_n$ uma amostra aleatória de uma população com distribuição normal $N(\mu, \sigma^2)$. Mostre que o estimador $s = \sqrt{\dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2}$ é viesado para σ.

3. (Bolfarine and Sandoval, 2001) Sejam $X_1, X_2, , X_n$ uma amostra aleatória de tamanho n da distribuição da variável aleatória X com f.d.p. dada por:

$$f(x|\theta) = \frac{2x}{\theta^2}, \qquad 0 < x < \theta, \qquad \theta > 0.$$

   a. Especifique o espaço paramétrico e o suporte associado à distribuição de X.

   b. Verifique se $\widehat{\theta}_1 = \overline{X}$ e $\widehat{\theta}_2 = X_{(n)}$ são estimadores não viciados para θ.

   c. Encontre e compare os EQM dos dois estimadores. Faça um gráfico como função de θ.

4. (Bolfarine and Sandoval, 2001) Sejam $X_1, X_2, , X_n$ uma amostra aleatória de tamanho n da distribuição de um variável aleatória $X \sim U(0, \theta)$. Considere os estimadores $\widehat{\theta}_1 = c_1\overline{X}$ e $\widehat{\theta}_2 = c_2 X_{(n)}$.

   a. Encontre c_1 e c_2 que tornam os estimadores não viciados.

   b. Encontre e compare os EQM dos dois estimadores.

5. (Bolfarine and Sandoval, 2001) Sejam $X_1, X_2, , X_n$ uma amostra aleatória de tamanho n da distribuição da variável aleatória $X \sim \mathcal{N}(0, \sigma^2)$. Seja $S^2 = \sum_{i=1}^{n} X_i^2$. Considere o estimador

$$\widehat{\sigma}_c^2 = cS^2.$$

a. Encontre o EQM do estimador acima.

b. Encontre o valor de c que minimiza o EQM em (i).

CAPÍTULO 5

Métodos de estimação

5.1 Método dos momentos

O método de estimação basedo nos momentos da variável aleatória é um método proposto por Pafnuty Chebyshev, entre os anos de 1880 e 1890, e cuja ideia é atribuída a **?**.

A metodologia consiste em igualar os momentos da população (que são funções com parâmetros conhecidos), definidos em termos de valores esperados, aos correspondentes momentos da amostra, cuja solução da equação (ou das equações) é o estimador do parâmetro de interesse. A seguir descrevemos uma apresentação formal para esse método.

Definicão 5.1.1. *Seja $X_1, X_2, ..., X_n$ uma amostra aleatória de uma variável aleatória X com função densidade de probabilidade ou função de probabilidade $f(\mathbf{x}|\theta)$ com p parâmetros. Os estimadores de $\theta = (\theta_1, \theta_2, ..., \theta_p)$, pelo método dos momentos são as soluções das equações, tais que:*

$$\mu_k = m_k, \tag{5.1}$$

com $k = 1, 2, ..., p$, sendo $\mu_k = \mathbb{E}(X^k)$ o momento populacional de ordem k, centrado no zero; e $m_k = \dfrac{\sum_{i=1}^{n} X_i^k}{n}$ o momento amostral de ordem k, também centrado no zero.

Dessa forma, observe que quando uma função de probabilidade tiver mais de um parâmetro, teremos que usar mais de uma equação. Considere, como ilustração, o exemplo a seguir.

Exemplo 5.1.1. *Seja $X_1, X_2, ..., X_n$ uma amostra aleatória de $X \sim Poisson(\lambda)$. Vamos estimar λ pelo método dos momentos.*

Para isso, usamos o primeiro momento da distribuição Poisson, que é dado por:

$$\mu_1 = \mathbb{E}(X) = \lambda$$

Então, podemos estimar o parâmetro da seguinte forma:

$$\begin{aligned} \mu_1 &= m_1 \\ \widehat{\lambda} &= \frac{\sum_{i=1}^{n} X_i}{n} = \overline{X}. \end{aligned}$$

5.2 Método da máxima verossimilhança

Antes de entender como funciona o método, precisamos definir função de verossimilhança.

Definicão 5.2.1. *(**Função de verossimilhança**) Seja $X_1, X_2, ..., X_n$ uma amostra aleatória de tamanho n da variável aleatória X com função densidade de probabilidade ou função de probabilidade $f(\mathbf{x}; \theta)$, com $\theta \in \Theta$, no qual Θ é o espaço paramétrico. A função de verossimilhança de θ correspondente à amostra aleatória observada é dada por*

$$L(\theta|\mathbf{x}) = \prod_{i=1}^{n} f(\mathbf{x}; \theta). \tag{5.2}$$

A partir do conhecimento da funçaõ de verossimilhança, podemos apresentar o chamado estimador de máxima verossimilhança.

Definicão 5.2.2. *(**Estimador de máxima verossimilhança**) O estimador de máxima verossimilhança de θ é o valor $\widehat{\theta} \in \Theta$ que maximiza a função de verossimilhança $L(\theta; \boldsymbol{x})$.*

É muito comum utilizar o logarítmo de $L(\theta; \mathbf{x})$ como função a ser maximizada. Isso ocorre por dois motivos: facilitar os cálculos e ter a certeza de encontrar um ponto de máximo global na maximização, dado que o logarítmo é uma função crescente. O logaritmo natural da função de versossimilhança de θ é dado por:

$$l(\theta; \mathbf{x}) = \log L(\theta; \mathbf{x}). \tag{5.3}$$

Pelo fato do valor de θ que maximiza $L(\theta; \mathbf{x})$ também maximizar $l(\theta; \mathbf{x})$, o estimador de máxima verossimilhança pode ser encontrado como a raiz da equação de verossimilhança

$$l'(\theta; \mathbf{x}) = \frac{\partial l(\theta; \mathbf{x})}{\partial \theta} = 0, \tag{5.4}$$

cuja verificação de ponto máximo é dada pela condição

$$l''(\widehat{\theta}; \mathbf{x}) = \frac{\partial^2 l(\theta; \mathbf{x})}{\partial^2 \theta}|_{\theta=\widehat{\theta}} < 0. \tag{5.5}$$

Considerando situações em que $\theta = (\theta_1, ..., \theta_r)$, isto é, situações as quais a verossimilhança depende de um vetor de parâmetros, os estimadores de máxima verossimilhança de $\theta_1, ..., \theta_r$ podem ser obtidos como soluções das equações

$$\frac{\partial \log L(\theta|\mathbf{x})}{\partial \theta_i} = 0,$$

para $i = 1, ..., r$.

Nos casos em que o suporte da distribuição de X depende de θ ou o máximo ocorre na fronteira de $\mathbf{\Theta}$, o EMV é, em geral, obtido inspecionando o gráfico da função de verossimilhança (de forma semelhante ao caso uniparamétrico). Esse ponto específico será visto mais adiante.

Exemplo 5.2.1. *Considere o vetor* $(12, 15, 9, 10, 17, 12, 11, 18, 15, 13)$ *uma amostra aleatória de uma distribuição normal com média* μ *e variância conhecida e igual a* 4. *Vamos estudar a função de verossimilhança com base nessa amostra.*

Vamos, primeiramente, representar essa amostra obtida de maneira formal. Dizemos, então, que temos $X_1, \ldots, X_10$ *uma amostra aleatória de* $X \sim N(\mu, 4)$. *A densidade para cada observação é dada por*

$$f(x_i) = \frac{1}{2\sqrt{2\pi}} \exp\left\{-\frac{1}{8}(x_i - \mu)^2\right\},$$

e função de verossimilhança e log-verossimilhança dadas, respectivamente, por

$$L(\mu) = \prod_1^{10} f(x_i) \qquad e \qquad l(\mu) = \sum_1^{10} \log(f(x_i)).$$

Assim, aplicando a densidade em questão, temos:

$$l(\mu) = -5\log(8\pi) - \frac{1}{8}\left(\sum_1^{10} x_i^2 - 2\mu\sum_1^{10} x_i + 10\mu^2\right).$$

Qual valor de μ *que maximiza essa função? Podemos derivar a equação e igualá-la a zero (essa seria a forma analítica). Porém, para fins didáticos, vamos considerar vários valores para* μ *e observar o possível estimador por meio de uma representação gráfica. Usando o software R, com o código abaixo, obtemos a Figura 5.1.*

```
> x = c(12, 15, 9, 10, 17, 12, 11, 18, 15, 13)
> mus = seq(10, 16, l=100)
> n = length(x)
> lmu = -5 * log(8*pi) - (sum(x^2) - 2*mus*sum(x) + 10*(mus^2))/8
> plot(mus, lmu, type='l', xlab=expression(mu), ylab=expression(l(mu)))
```

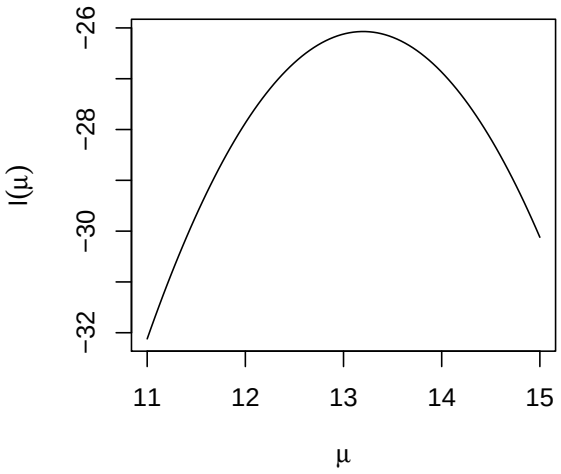

Figura 5.1: Função de verossimilhança para o parâmetro μ da distribuição normal com variância $\sigma^2 = 4$ com os dados do Exemplo 5.2.1.

Pelo gráfico apresentado, podemos observar que o valor de μ que maximiza a função em questão está entre 13 e 14. Mas como precisar com exatidão, dado que não estamos encontrando o estimador pela forma analítica? O software R disponibiliza uma função implementada chamada de `optim`, no qual nos fornece o valor estimado com base na função de log-verossimilhança inserida e no chute assumido. A função nós já temos, e o chute, vamos considerar como sendo $\mu = 13$ (já que sabemos que será bem próximo desse valor). Utilizando o comando citado (veja o código abaixo), obtemos que o valor estimado para μ é, aproximadamente, 13,2.

```
> optim(par = 13, log.vero, x=x)
$par
[1] 13.20059
```

Observações

O estimador obtido por meio do método de Máxima Verososimilhança, aqui denotado por EMV, apresenta uma série de características/observações úteis no estudo da inferência estatística. Vamos apreseetar cada um desses tópicos nas subseções a seguir.

Observação 1

Função de verossimilhança pode não levar a nenhum estimador.

Observação 2

No caso discreto, o EMV de θ pode ser interpretado como o valor de θ que maximiza a probabilidade de se observar a amostra que foi coletada.

Observação 3

O EMV não é único.

Observação 4

Em alguns casos, a expressão (5.4) não pode ser obtida explicitamente. Nessas situações, recorre-se a solução por meio de procedimentos numéricos. Considere $U(\theta)$ a função escore, ou seja,

$$U(\theta) = \frac{\partial l(\widehat{\theta}; \mathbf{x})}{\partial \theta}.$$

Sabendo que para obter o estimador de máxima verossimilhança, $\widehat{\theta}$, temos que fazer

$$U(\widehat{\theta}) = 0,$$

então, utilizando série de Taylor para expandir $U(\widehat{\theta})$ em torno de θ_0, obtém-se

$$U(\widehat{\theta}) = U(\theta_0) + (\widehat{\theta} - \theta_0)U'(\theta_0),$$

então

$$\widehat{\theta} = \theta_0 - \frac{U(\theta_0)}{U'(\theta_0)}.$$

Portanto, a partir disso, temos que o processo iterativo para estimar θ, chamado de Newton-Raphson, é dado por

$$\theta_{j+1} = \theta_j - \frac{U(\theta_j)}{U'(\theta_j)}, \tag{5.6}$$

cujo valor é iniciado om θ_0 e será atualizado conforme nova estimativa. O processo tem seu fim quando $|\theta_{j+1} - \theta_j| < \epsilon$, para ϵ pequeno. Assim, o ponto $\widehat{\theta}$ em que o processo se estabiliza é tomado como o estimador de máxima verossimilhança de θ.

Vale ressaltar que, em alguns casos, a substituição de $U'(\theta_j)$ por $\mathbb{E}(U'(\theta_j))$ representa significativa simplificação no procedimento (visto que o valor de $U'(\theta_j)$ deve ser positivo). Esse método é conhecido como **Escore de Fisher**.

Observação 5

O EMV é **função de uma estatística suficiente**. O teorema a seguir enuncia essa característica.

Teorema 5.2.1. ***O EMV é função de uma estatística suficiente*** *(Bolfarine and Sandoval, 2001) – Sejam $X_1, X_2, ..., X_n$ uma amostra aleatória da variável aleatória X com função densidade de probabilidade (ou função de probabilidade) $f(\mathbf{x}; \theta)$. Seja $T = T(X_1, X_2, ..., X_n)$ uma estatística suficiente para θ. Então, o estimador de máxima verssimilhança $\widehat{\theta}$ (se existir) é função de T.*

Observação 6

Além disso, apresenta o **princípio da invariância** (ou equivariância). O teorema a seguir enuncia essa característica.

Teorema 5.2.2. ***Princípio da invariância*** *(Bolfarine and Sandoval, 2001) – Sejam $X_1, X_2, ..., X_n$ uma amostra aleatória da variável aleatória X com função densidade de probabilidade (ou função de probabilidade) $f(\mathbf{x}; \theta)$. Se $\widehat{\theta}$ é um estimador de máxima verossimilhança de θ, então $g(\widehat{\theta})$ é um estimador de máxima verossimilhança de $g(\theta)$.*

Observação 7

Para grandes amostras (sob condições de regularidade satisfeitas), sua **distribuição assintótica** é conhecida:

$$\sqrt{n}(\widehat{\theta} - \theta) \sim \mathcal{N}\left(0, \frac{1}{I_F(\theta)}\right), \tag{5.7}$$

e

$$\sqrt{n}(g(\widehat{\theta}) - g(\theta)) \sim \mathcal{N}\left(0, \frac{g'(\theta)^2}{I_F(\theta)}\right). \tag{5.8}$$

Observação 8

Podemos encontrar situações em que duas ou mais amostras indepedentes de distribuições dependam de um parâmetro θ em comum.

Considerando um caso mais simples (e que pode ser generalizado) com duas amostras $X_1, ..., X_n$ e $Y_1, ..., Y_n$, então a verossimilhança conjunta é igual ao produto da verossimilhança de cada amostra, ou seja,

$$L(\theta|\mathbf{x}, \mathbf{y}) = L(\theta|\mathbf{x})L(\theta|\mathbf{y}).$$

Isso ocorre, pelo fato das amostras serem independentes.

Se aplicarmos o logaritmo, temos que

$$l(\theta|\mathbf{x}, \mathbf{y}) = l(\theta|\mathbf{x}) + l(\theta|\mathbf{y}).$$

Observação 9

A máxima verossimilhança na família exponencial. Se a distribuição da variável aleatória X pertence à família exponencial unidimensional de distribuições, então o estimador de máxima verossimilhança de θ baseado na amostra $\mathbf{X} = (X_1, X_2, ..., X_n)$ é solução da equação

$$\mathbb{E}(T(\mathbf{X})) = T(\mathbf{X}), \tag{5.9}$$

desde que a solução pertença ao espaço paramétrico correspodente ao parâmetro θ. Esse resultado pode ser estendido para o caso k-paramétrico em que os estimadores de máxima verossimilhança de $\theta_1, ..., \theta_k$ seguem como soluções das equações

$$\mathbb{E}(T_j(\mathbf{X})) = T_j(\mathbf{X}), \qquad j = 1, ..., k. \tag{5.10}$$

5.3 Métodos para avaliação de estimadores pontuais

5.3.1 Erro Quadrático Médio (EQM)

O **Erro Quadrático Médio** (ou EQM) é uma medida que avalia o desempenho de um estimador em relação ao verdadeiro valor do parâmetro. A definição formal para essa medida é descrita a seguir.

Definicão 5.3.1. *Seja $X_1, X_2, ..., X_n$ uma amostra aleatória de tamanho n da variável aleatória com distribuição $f(x_i|\theta)$. Seja $\mathbf{T}(X_1, X_2, ..., X_n)$ um estimador para o parâmetro θ. O Erro Quadrático Médio (EQM) de $\widehat{\theta}$ é dado por*

$$EQM(\widehat{\theta}) \quad = \quad \mathbb{E}[(\widehat{\theta} - \theta)^2].$$

Vale destacar, ainda, que

$$\begin{aligned}
\mathrm{EQM}(\widehat{\theta}) &= \mathbb{E}[(\widehat{\theta} - \theta)^2] \\
&= \mathbb{E}[\widehat{\theta}^2 - 2\widehat{\theta}\theta + \theta^2] \\
&= \mathbb{E}[\widehat{\theta}^2] - 2\theta\mathbb{E}[\widehat{\theta}] + \theta^2 \\
&= (\mathrm{Var}[\widehat{\theta}] + \mathbb{E}^2[\widehat{\theta}]) - 2\theta\mathbb{E}[\widehat{\theta}] + \theta^2 \\
&= \mathrm{Var}(\widehat{\theta}) + [\mathbb{E}(\widehat{\theta}) - \theta]^2 \\
&= \mathrm{Var}(\widehat{\theta}) + B^2(\widehat{\theta}).
\end{aligned}$$

Nesse sentido, o valor $B(\widehat{\theta}) = \mathbb{E}(\widehat{\theta}) - \theta$ é chamado de **Erro Quadrático Médio** vício do estimador θ. Assim, dizemos que um estimador é não viciado, se $B(\widehat{\theta}) = 0$, ou seja, se $E(\widehat{\theta}) = 0$. Além disso, caso

$$\lim_{n \Rightarrow \infty} B(\widehat{\theta}) = 0,$$

dizemos que o estimador é **Erro Quadrático Médio assintoticamente não viciado**, ou seja, à medida que o tamanho da amostra aumenta, o vício do estimador tende a zero e, consequentemente, $E(\widehat{\theta}) = 0$. O EQM de um estimador não viciado, ou assintoticamente não viciado, reduz-se a sua variância, ou seja, $\mathrm{EQM}(\widehat{\theta}) = \mathrm{Var}(\widehat{\theta})$.

Exemplo 5.3.1. *Seja $X_1, X_2, \cdots, X_n$ uma amostra aleatória de uma população com média μ e variância $\sigma^2 < infty$. Então, o estimador não viciado para μ é $\overline{X}$ e para σ^2 é s^2, ou seja*

(i.) $\mathbb{E}(\overline{X}) = \mu$,

(ii.) $\mathbb{E}(s^2) = \sigma^2$.

Demonstração: *Para demonstrar (i), temos que*

$$\mathbb{E}(\overline{X}) = \mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\mathbb{E}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n} n\mathbb{E}(X_1) = \mu.$$

Similarmente, para a variância amostral, temos que

$$\begin{aligned}
\mathbb{E}(s^2) &= \mathbb{E}\left(\frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right]\right) \\
&= \frac{1}{n-1}(n\mathbb{E}(X_1^2) - n\mathbb{E}(\overline{X}^2) \\
&= \frac{1}{n-1}\left(n(\sigma^2+\mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right) = \sigma^2.
\end{aligned}$$

Também vale a pena mencionar que

$$\mathbb{E}(\hat{\sigma}^2) = \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n}(X_i - \overline{X})\right] = \frac{(n-1)}{n}\sigma^2.$$

Assim, podemos concluir que $\overline{X}$ e s^2 são estimadores não enviesados, respectivamente, da média populacional μ e da variância populacional. O estimador σ^2, entretanto, é viciado em σ^2. A seguir, introduziremos o conceito de Erro Quadrado Médio e o avaliaremos em relação aos estimadores de variância populacional.

Vale ressaltar que tanto o EQM quanto a variância de um estimador são inconvenientes para a sua análise direta, pois apresentam unidade de medida igual ao quadrado da usada na medição. Assim, é plausível utilizar três outras medidas (adimensionais e definidas somente quando $\mathbb{E}(\widehat{\theta})$ e θ são positivos), que são bem comuns em estudos sobre estimação:

i. Coeficiente de Variação (CV):

$$\mathrm{CV}(\widehat{\theta}) = \frac{\sqrt{\mathrm{Var}(\widehat{\theta})}}{\mathbb{E}(\widehat{\theta})},$$

ii. Erro Relativo Médio (RLM):

$$\mathrm{RLM}(\widehat{\theta}) = \frac{\sqrt{\mathrm{EQM}(\widehat{\theta})}}{\theta},$$

iii. Vício Relativo (VR):

$$\mathrm{VR}(\widehat{\theta}) = \frac{\sqrt{\mathrm{EQM}(\widehat{\theta}) - \theta}}{\theta}.$$

Outro ponto importante sobre os estimadores, consiste nas sua comparações, ou seja, considere $\widehat{\theta}_1$ e $\widehat{\theta}_2$ dois estimadores não viciados. Se

$$\text{EQM}(\widehat{\theta}_1) < \text{EQM}(\widehat{\theta}_2),$$

então $\widehat{\theta}_1$ é preferível em relação a $\widehat{\theta}_2$.

Nesse sentido, se houver um outro estimador não viciado, denotado por $\widehat{\theta}_3$, com a menor variância dentre todos os estimadores não viciados, então ele é chamado de **estimador não viciado de variância uniformemente mínima (ENVVUM)**.

Exemplo 5.3.2. *Partindo do exemplo 5.3.1, temos que*

$$\begin{aligned} EQM(s^2) &= Var(s^2) = \frac{2\sigma^4}{n-1} \\ EQM(\widehat{\sigma}^2) &= \frac{2\sigma^4}{n-1}\left[1 - \frac{(3n-1)}{2n^2}\right]. \end{aligned}$$

Observe que embora $\widehat{\sigma}^2$ seja viciado, como no exemplo 5.3.1, tem um EQM menor em comparação com s^2.

5.4 Exercícios

1. (Bolfarine and Sandoval, 2001) Sejam $X_1, X_2, , X_n$ uma amostra aleatória de tamanho n da variável aleatória X com função densidade de probabilidade dada por

$$f(x|\theta) = \frac{x}{\theta}e^{-x/\theta} \quad \mathbb{I}_{(0,\infty)}(x) \quad \theta > 0.$$

   a. Encontre o estimador de máxima verossimilhança de θ e verifique se ele é eficiente.

   b. Encontre o estimador de máxima verossimilhança de $V(X)$ e encontre sua distribuição para grandes amostras.

2. (Bolfarine and Sandoval, 2001) Sejam $X_1, X_2, , X_n$ uma amostra aleatória da variável aleatória $X \sim N(\mu, 1)$. Encontre o estimador de máxima verossimilhança de $g(\mu) = \mathbb{P}(X > 0)$.

3. (Bolfarine and Sandoval, 2001) Sejam $X_1, X_2, , X_n$ uma amostra aleatória de tamanho n da variável aleatória X com função densidade de probabilidade dada por

$$f(x;\theta) = \theta x^{\theta-1} \quad \mathbb{I}_{[0,1]}(x) \quad \theta > 0.$$

   Encontre o estimador de máxima verossimilhança de θ.

4. (Bolfarine and Sandoval, 2001) Sejam $X_1, X_2, , X_n$ uma amostra aleatória da variável aleatória X com função densidade de probabilidade

$$f(x;\theta) = \frac{\theta}{x^2} \quad \mathbb{I}_{[\theta,\infty]}(x) \quad \theta > 0.$$

   Encontre o estimador de máxima verossimilhança de θ e de $\mathbb{E}_\theta(1/X)$.

CAPÍTULO 6

Estimação intervalar

6.1 Considerações iniciais

Nesta capítulo vamos abordar o desenvolvimento da estimação de parâmetros por meio de estimativas intervalares. É necessário que o leitor tenha conhecimento prévio sobre probabilidade, principalmente em conceitos como variáveis aleatórias, distribuição de probabilidade e funções auxiliares; inferência, principalmente em conceitos básicos em Inferência, distribuição amostral e estimação pontual; e que tenha conhecimento em consultar tabelas de distribuições, tais como a da normal, da t-Student e da qui-quadrado.

6.2 Motivação para uso de um intervalo de confiança

Os métodos de estimação vistos até o momento são pontuais, ou seja, obtemos o valor da estimativa do parâmetro com base em um ponto. Entretanto, esse ponto pode variar, pois a cada amostra retirada, cada estimativa calculada pode ser diferente e, consequentemente, não conseguimos obter a magnitude do erro da estimativa.

Uma solução interessante, portanto, é realizar uma estimação intervalar, de modo que possamos definir um limite inferior e superior para o parâmetro de interesse. Dessa forma, para um parâmetro θ, desejamos construir um intervalo, tal que

$$\mathbb{P}(\text{LI} < \theta < \text{LS}) = \gamma,$$

sendo LI o limite inferior e LS o limite superior do intervalo, com γ representando o nível de confiança do intervalo.

Entretanto, para construir o intervalo de confiança, precisamos saber quem são as quantidades dadas por LI e LS. Na próxima seção, vamos avaliar essas quantidades.

6.3 Definição de intervalo de confiança

Para entender quem seriam LI e LS, precisamos formalizar a definição para intervalos de confiança e entender melhor esse processo.

Definicão 6.3.1. *Seja $X_1, ..., X_n$ uma amostra aleatória de tamanho n de uma variável aleatória com função densidade (ou de probabilidade) $f(x|\theta)$. Sejam $T_1(\mathbf{X})$ e $T_2(\mathbf{X})$ duas estatísticas tais que $T_1(\mathbf{X}) < T_2(\mathbf{X})$ e $\mathbb{P}(T_1(\mathbf{X}) < \theta < T_2(\mathbf{X})) = \gamma$, $0 < \gamma < 1$, independe de θ. Então, o intervalo $[T_1(\mathbf{X}), T_2(\mathbf{X})]$ é um intervalo para $g(\theta)$, com nível de confiança γ ou $\gamma \times 100\%$.*

A partir dessa definição, precisamos ter atenção em dois pontos: (i.) o estimador intervalar é, portanto, dado por $[T_1(\mathbf{X}), T_2(\mathbf{X})]$; (ii.) quem varia é o intervalo, o parâmetro é fixo.

Portanto, para um parâmetro θ, o intervalo de confiança é dado por

$$\mathbb{P}(T_1(\mathbf{X}) < \theta < T_2(\mathbf{X})) = \gamma,$$

sendo $T_1(\mathbf{X})$ o limite inferior e $T_2(\mathbf{X})$ o limite superior do intervalo, com γ representando o nível de confiança do intervalo.

Assim como na seção anterior, terminamos esta com mais uma pergunta: quem seriam $T_1(\mathbf{X})$ e $T_2(\mathbf{X})$? para que possamos construir o intervalo precisamos utilizar algum método. Na próxima seção, apresentamos alguns desses métodos.

6.4 Métodos para construção de intervalos de confiança

Os métodos mais comuns para construção de intervalo de confinaça são:

- Quantidade pivotal;
- Intervalos bayesianos;
- Intervalos de confinça bootstrap.

Mas também temos métodos mais avançados, tais como:

- Inversão da estatística do teste (é preciso estudar o conteúdo sobre teste de hipóteses);
- Pivotagem da FDA.

Na subseções a seguir, vamos abordar cada um desses métodos, explicando sua origem e como montar o respectivo intervalo de confiança para o parâmetro.

6.4.1 Quantidade pivotal

O método mais utilizado para a construção de intervalos de confiança é o chamado Método da Quantidade Pivotal.

Definicão 6.4.1. ***(Quantidade Pivotal)*** *É uma função de uma estatística (depende da amostra) e do parâmetro, $Q(\mathbf{X}|\theta)$ mas cuja distribuição $f_Q(q)$ não depende do parâmetro desconhecido θ.*

A ideia, portanto, é considerar λ_1 e λ_2, de tal forma que:

$$\mathbb{P}(\lambda_1 < Q(\mathbf{X}|\theta) < \lambda_2) = \gamma;$$

A partir dessa expressão, procura-se isolar o parâmetro para encontrar $T_1(\mathbf{X})$ e $T_2(\mathbf{X})$, tal que:

$$\mathbb{P}(T_1(\mathbf{X}) < \theta < T_2(\mathbf{X})) = \gamma;$$

Para obter a quantidade pivotal, devemos encontrar uma boa estatística associada ao parâmetro de interesse. A partir da distribuição amostral dessa estatística, conseguimos encontrar a quantidade pivotal (já apresentamos as distribuições amostrais de algumas estatísticas em aulas anteriores).

6.4.2 Intervalos bayesianos

O intervalo de confiança no contexto bayesiano é um intervalo de probabilidade a posteriori, usado para fins similares aos dos intervalos de confiança abordados anteriormente, que são construídos na linha da estatística frequentista.

Para ilustrar esse entedimento, podemos considerar que se uma amostra é coletada e o intervalo de credibilidade de 95% para o parâmetro μ é [20;30], isso significa que a probabilidade a posteriori de que μ esteja no intervalo de 20 a 30 é de 0,95.

Vale ressaltar que, em geral, os intervalos de credibilidade bayesianos não coincidem com os intervalos de confiança frequentistas, pois o intervalo de credibilidade incorpora informação contextual específica do problema da distribuição a priori.

Existem vários modos de construir intervalos de credibilidade a partir de uma dada distribuição de probabilidade por parâmetro, tais como: escolher o intervalo mais estreito, o qual, para uma distribuição unimodal envolverá a escolha dos valores de mais alta densidade probabilística, incluindo a moda; escolher o intervalo onde a probabilidade de estar abaixo do intervalo é tão provável quanto estar acima dele; o intervalo incluirá a mediana; escolher o intervalo no qual a média seja o ponto central.

6.4.3 Intervalo de confiança bootstrap

No intervalor de confiança bootstrap, a ideia principal é reamostrar um conjunto de dados, diretamente ou via um modelo ajustado, a fim de criar replicas dos dados, a partir das quais podemos avaliar a variabilidade de quantidades de interesse, sem usar cálculos analíticos.

Esse método de reamostragem pode ser feito de duas formas: por meio do boostrap paramétrico ou por meio do bootstrap não paramétrico.

6.4.4 Pivotagem da FDA (t.b.d)

6.4.5 Inversão da estatística do teste (t.b.d)

6.5 Os intervalos de confiança mais comuns (usando a quantidade pivotal)

Nesta seção, vamos apresentar os intervalos de confiança mais comuns, construídos a partir do **método da quantidade pivotal**. Vamos considerar intervalos tanto para uma amostra, tais como:

- Intervalo de confiança para a média (com variância conhecida);
- Intervalo de confiança para a média (com variância desconhecida);
- Intervalo de confiança para a proporção;
- Intervalo de confiança para a variância com média conhecida;
- Intervalo de confiança para a variância com média desconhecida;

quanto para duas amostras, tais como:

- Intervalo de confiança para diferença de médias (com variâncias conhecidas);
- Intervalo de confiança para diferença de médias (com variâncias desconhecidas);
- Intervalo de confiança para a diferença de proporções;
- Intervalo de confiança para razão de duas variâncias (com médias conhecidas);
- Intervalo de confiança para razão de duas variân-cias (com médias desconhecidas).

Para a construção dos intervalos, vamos utilizar o **método da quantidade pivotal** como base de construção para os respectivos intervalos.

Lembre-se que para construir intervalos usando esse método, precisamos de uma estatística suficiente que seja função da amostra e do parâmetro, mas cuja distribuição não depende do parâmetro desconhecido.

Para essas quantidades, portanto, vamos utilizar as estatísticas apresentadas no capítulo sobre distribuição amostral. Veja que naquele capítulo, separamos cada distribuição amostral para cada caso particular, pois iríamos abordar justamente essa sequência para os intervalos desse capítulo. Veja a seguir a explanação de cada situação.

6.5.1 Intervalo de confiança para a média (com variância conhecida)

Para a construção desse intervalo, vamos considerar $\overline{X} = \sum_{i=1}^{n} X_i/n$ como a estatística suficiente (que também é completa). Como já sabemos que:

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1),$$

então Z é uma quantidade pivotal, pois depende da amostra e do parâmetro e a sua distribuição não depende do parâmetro.

Dessa forma, vamos usar essa quantidade para **construir o intervalo de confiança para a média** utilizando o fato de que $\gamma = \mathbb{P}(\lambda_1 < Q(\mathbf{X}|\theta) < \lambda_2)$, ou seja, vamos incluir a quantidade pivotal dentro dessa inequação e isolar o parâmetro de interesse. Veja esse desenvolvimento a seguir:

$$\begin{aligned}
\gamma &= \mathbb{P}(\lambda_1 < Q(\mathbf{X}|\theta) < \lambda_2) \\
&= \mathbb{P}\left(\lambda_1 < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < \lambda_2\right) \\
&= \mathbb{P}\left(\overline{X} - \lambda_2\frac{\sigma}{\sqrt{n}} < \mu < \overline{X} + \lambda_1\frac{\sigma}{\sqrt{n}}\right),
\end{aligned}$$

sendo $\lambda_1 = \lambda_2 = z_{\alpha/2}$ o quantil de ordem $\alpha/2$ da distribuição normal padrão, com $\gamma = 1 - \alpha$ representando o nível de confiança.

Assim, o intervalo de confiança para μ com variância conhecida e com $\gamma = (1-\alpha)100\%$ de confiança é dado por:

$$IC_\gamma[\mu] = \left[\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}; \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right], \tag{6.1}$$

sendo $z_{\alpha/2}$ o quantil de ordem $\alpha/2$ da distribuição normal padrão.

No caso em que X não tem distribuição normal, ou seja, no caso em que $X \sim f(x|\theta)$, com $\mathbb{E}(X) = \mu$ e $\mathrm{Var}(X) = \sigma^2$, então, a medida que o tamanho da amostra aumenta, temos o mesmo resultado apresentado anteriormente para a estatística Z (com base no Teorema do Limite Central).

A seguir, apresentamos um exemplo para ilustrar os conceitos apresentados nessa seção.

Exemplo 6.5.1. *(Morettin and Bussab, 2017) Uma máquina enche pacotes de café com uma variância igual a $100g^2$ e está programada para encher pacotes com 500g, em média. Suspeita-se que essa máquina esteja desrregulada e tem-se o interesse em saber a nova média μ. Para isso, coletou-se uma amostra de tamanho $n = 25$ e verificou-se que a média resultou em $485g$. Construa um intervalo de confiança com 95% para μ.*

Solução:

$$\begin{aligned}
IC_\gamma[\mu] &= \left[\overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}; \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] \\
IC_{0,95}[\mu] &= \left[485 - 1,96\frac{10}{\sqrt{25}}; 485 + 1,96\frac{10}{\sqrt{25}}\right] \\
&= [481; 489].
\end{aligned}$$

Portanto, com 95% de probabilidade, esse intervalo contém o valor da média populacional μ.

6.5.2 Intervalo de confiança para a média (com variância desconhecida)

Para a construção desse intervalo, vamos considerar $\overline{X} = \sum_{i=1}^{n} X_i/n$ como a estatística suficiente (que também é completa). Como já sabemos que no caso com variância desconhecida temos que:

$$T = \frac{\overline{X} - \mu}{s/\sqrt{n}} \sim t(n-1);$$

então T é uma quantidade pivotal, pois depende da amostra e do parâmetro e a sua distribuição não depende do parâmetro.

Dessa forma, vamos usar essa quantidade para **construir o intervalo de confiança para a média** utilizando o fato de que $\gamma = \mathbb{P}(\lambda_1 < Q(\mathbf{X}|\theta) < \lambda_2)$, ou seja, vamos incluir a quantidade pivotal dentro dessa inequação e isolar o parâmetro de interesse. Veja esse desenvolvimento a seguir:

$$\begin{aligned}
\gamma &= \mathbb{P}(\lambda_1 < Q(\mathbf{X}|\theta) < \lambda_2) \\
&= \mathbb{P}\left(\lambda_1 < \frac{\overline{X} - \mu}{s/\sqrt{n}} < \lambda_2\right) \\
&= \mathbb{P}\left(\overline{X} - \lambda_2\frac{s}{\sqrt{n}} < \mu < \overline{X} + \lambda_1\frac{s}{\sqrt{n}}\right),
\end{aligned}$$

sendo $\lambda_1 = \lambda_2 = t_{\alpha/2}$ o quantil de ordem $\alpha/2$ da distribuição *t-Student*.

Assim, o intervalo de confiança para μ com variância desconhecida e com $\gamma = (1-\alpha)100\%$ de confiança é dado por:

$$IC_\gamma[\mu] = \left[\overline{X} - t_{\alpha/2}\frac{s}{\sqrt{n}}, \overline{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right], \tag{6.2}$$

sendo $\lambda_1 = \lambda_2 = t_{\alpha/2}$ o quantil de ordem $\alpha/2$ da distribuição *t-Student*.

Exemplo 6.5.2. *(Morettin and Bussab, 2017) Em uma amostra de 400 válvulas, verificou-se que a média do tempo de vida dessas válvulas foi de 800h e o desvio-padrão foi de 100 horas. Determine o intervalo de confiança para a média, considerando $\gamma = 99\%$.*

Solução:

$$\begin{aligned}
IC_\gamma[\mu] &= \left[\overline{X} - t_{\alpha/2}\frac{s}{\sqrt{n}}; \overline{X} + t_{\alpha/2}\frac{s}{\sqrt{n}}\right] \\
IC_{0,99}[\mu] = &= \left[800 - 2,58\frac{100}{\sqrt{400}}; 800 + 2,58\frac{100}{\sqrt{400}}\right] \\
&= [787,11; 812,9]
\end{aligned}$$

Conclusão: *Com 99% de probabilidade, esse intervalo contém o valor da média populacional μ;*

6.5.3 Intervalo de confiança para a proporção

Para a construção desse intervalo, vamos considerar $\hat{p} = \sum_{i=1}^{n} X_i/n$ como a estatística suficiente (que também é completa). Como já sabemos que:

$$Z = \frac{\hat{p} - p}{\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n-1}}} \sim N(0,1);$$

então Z é uma quantidade pivotal, pois depende da amostra e do parâmetro e a sua distribuição não depende do parâmetro.

Dessa forma, vamos usar essa quantidade para **construir o intervalo de confiança para a proporção** utilizando o fato de que $\gamma = \mathbb{P}(\lambda_1 < Q(\mathbf{X}|\theta) < \lambda_2)$, ou seja, vamos incluir a quantidade pivotal dentro dessa inequação e isolar o parâmetro de interesse. Veja esse desenvolvimento a seguir:

$$\begin{aligned}
\gamma &= \mathbb{P}(\lambda_1 < Q(\mathbf{X}|\theta) < \lambda_2) \\
&= \mathbb{P}\left(\lambda_1 < \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}} < \lambda_2\right) \\
&= \mathbb{P}\left(\hat{p} - \lambda_2\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}} < p < \hat{p} + \lambda_1\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}\right),
\end{aligned}$$

sendo $\lambda_1 = \lambda_2 = z_{\alpha/2}$ o quantil de ordem $\alpha/2$ da distribuição normal padrão.

Assim, o intervalo de confiança com $\gamma = (1-\alpha)100\%$ de confiança para p é:

$$IC_\gamma[p] = \left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}; \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}\right],$$

sendo $z_{\alpha/2}$ o quantil de ordem $\alpha/2$ da distribuição normal padrão.

Exemplo 6.5.3. *(Morettin and Bussab, 2017) Suponha que em n = 400 provas, obtemos 80 sucessos. Encontre o intervalo de confiança para p, considerando* $\gamma = 0,90$.

Solução:

Pelo enunciado, temos que $\hat{p} = 80/400 = 0,2$*, logo:*

$$\begin{aligned}
IC_\gamma[p] &= \left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}; \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}\right] \\
IC_{0,90}[p] &= \left[0,2 - 1,645\sqrt{\frac{0,2(1-0,2)}{400-1}}; 0,2 + 1,645\sqrt{\frac{0,2(1-0,2)}{400-1}}\right] \\
&= [0,167; 0,233].
\end{aligned}$$

Conclusão: *Com 90% de probabilidade, esse intervalo contém o valor da proporção populacional* p.

6.6 Exercícios

1. Um robô enche sacos de cimento com uma variância igual a $500g^2$ e está programada para encher os sacos com 500kg, em médb. Suspeita-se que essa máquina esteja desrregulada e tem-se o interesse em saber a nova média μ. Para isso, coletou-se uma amostra de tamanho $n = 25$ e verificou-se que a média resultou em $485g$. Construa um intervalo de confiança para μ com 95% para μ.

2. Em uma amostra de 400 válvulas, verificou-se que a média do tempo de vida dessas válvulas foi de 800h e o desvio-padrão foi de 100 horas. Determine o intervalo de confiança para a média, considerando $\gamma = 99\%$.

3. Suponha que em n $=$ 800 provas, obtemos 160 sucessos. Encontre o intervalo de confiança para a proporção p de sucessos, considerando $\gamma = 0,90$.

4. O tempo de vida de equipamentos cirúrgicos produzidos por uma empresa é uma variável aleatória, medida em horas, que se supõe ter distribuição normal com média conhecida e igual a 98 horas. Pretende-se avaliar a variabilidade desse tempo de vida com base na seguinte amostra coletada: 97, 96, 100, 98, 101, 104, 96, 103 e 100. Construa um intervalo de confiança para a variância do peso.

5. O tempo de vida de equipamentos cirúrgicos produzidos por uma empresa é uma variável aleatória, medida em horas, que se supõe ter distribuição normal com média desconhecida. Pretende-se avaliar a variabilidade desse tempo de vida com base na seguinte amostra coletada: 97, 96, 100, 98, 101, 104, 96, 103 e 100. Construa um intervalo de confiança para a variância do peso.

CAPÍTULO 7

Teste de hipóteses

7.1 Motivação para uso de teste de hipóteses

Em muitas situações do cotidiano há interesse em tomar a decisão de aceitar ou rejeitar determinada afirmação baseando-se em um conjunto de evidências. O nosso interesse recai em responder perguntas do tipo:

- Qual decisão devo tomar?
- Eu devo aceitar ou rejeitar uma hipótese?

Para uma melhor entendimento de como podemos utilizar, vamos ilustrar a importância do dos testes de hipóteses a partir de dois exemplos clássicos: decisão do júri e eficiência de vacinas.

No primeiro, o objetivo é decidir se um indivíduo é inocente ou culpado. Nesse exemplo, uma das duas decisões a seguir devem ser tomadas:

- Decisão 1: o indivíduo é inocente;

- Decisão 2: o indivíduo não é inocente.

Na linguagem estatística, essas decisões são chamadas de hipóteses e, para esse caso, podem escritas da seguinte forma:

- H_0 : o indivíduo é inocente;

- H_1 : o indivíduo não é inocente (é culpado).

o outro exemplo, sobre eficiência de vacinas, tem o objetivo é decidir se uma vacina é eficiente ou não. Para isso, duas hipóteses são possíveis:

- Decisão 1: a vacina não é eficiente;

- Decisão 2: a vacina é eficiente;

Conforme já introduzimos, na linguagem estatística, essas decisões são chamadas de hipóteses. Para esse caso, podem escritas da seguinte forma:

- H_0 : a vacina não é eficiente;

- H_1 : a vacina é eficiente.

Tanto no problema da decisão do júriquanto no problema da eficiência de vacinas, temos decisões (que chamamos de hipóteses) que serão testadas com base em evidências, ou seja, com base na amostra coletada. Na seção seguinte, introduzimos os conceitos de teste de hipóteses de maneira mais formal.

7.2 Apresentação dos principais conceitos para testes de hipóteses

Vimos dois exemplos sobre como podemos aplicar o uso dos chamados testes de hipóteses. Mas, de maneira formal, o que é uma hipótese estatística? Vamos para a sua definição.

Definicão 7.2.1. ***(Hipótese estatística)*** *Chamamos de hipótese estatística qualquer afirmação acerca da distribuição de probabilidades de uma ou mais variáveis aleatórias.*

Como o teste depende de uma afirmação feita sobre a distribuição de probabilidade, então considere:

- Uma variável aleatória X com função de densidade (ou de probabilidade) $f(x|\theta)$, com $\theta \in \Theta$;

A ideia por trás do teste é associar os conjuntos Θ_0 e Θ_1, de tal forma que:

- $H_0 : \theta \in \Theta_0$;

- $H_1 : \theta \in \Theta_1$;

sendo H_0 a hipótese nula e H_1 a hipótese alternativa.

Além disso, atribui-se à H_0 a responsabilidade de ser a hipótese de interesse em um teste e se $\Theta_0 = \{\theta_0\}$, dizemos que a hipótese é simples. Caso contrário, dizemos que a hipótese é composta (o mesmo vale para a hipótese alternativa);

Com o conhecimento sobre hipótese estatística, precisamos avançar para o conhecimento sobre a função de decisão, que será útil para a conclusão de um teste de hipóteses.

Definição (Função de decisão). Chamamos de teste de uma hipótese estatística a função de decisão $d : \mathcal{X} \longrightarrow \{d_0, d_1\}$, sendo d_0 a decisão de considerar H_0 como verdadeira e d_1 a decisão de considerar H_1 como verdadeira, com $\mathcal{X}$ denotando o espaço amostral associado à amostra $X_1, X_2, ..., X_n$ coletada.

A função de decisão divide esse espaço em dois conjuntos, a saber

- $\mathbf{A}_0 = \{(x_1, x_2, ..., x_n) \in \mathcal{X} | d(x_1, x_2, ..., x_n) = d_0\}$;

- $\mathbf{A}_1 = \{(x_1, x_2, ..., x_n) \in \mathcal{X} | d(x_1, x_2, ..., x_n) = d_1\}$,

com $\mathbf{A}_0 \cup \mathbf{A}_1 = \mathcal{X}$ e $\mathbf{A}_0 \cap \mathbf{A}_1 = \phi$.

Em $\mathbf{A}_0$ temos os pontos que nos levam a decidir por H_0, então chamaremos esse conjunto de **região de aceitação**; e em $\mathbf{A}_1$ temos os pontos que nos levam a decidir por H_1, então chamaremos esse conjunto de **região de crítica**;

Mas podemos se perguntar: ao decidir por H_0 ou por H_1 podemos estar **comentendo erros**? A resposta é sim, podemos estar comentendo os seguintes erros:

– Erro do tipo I: rejeitar H_0 dado que H_0 é verdadeira;

– Erro do tipo II: não rejeitar H_0 dado que H_0 é falsa;

Associando esses erros a probabilidades, temos:

– $\mathbb{P}(\text{erro tipo I}) = \mathbb{P}(\text{rejeitar } H_0 | H_0 \text{ é verdadeira}) = \alpha$;

– $\mathbb{P}(\text{erro tipo II}) = \mathbb{P}(\text{não rejeitar } H_0 | H_0 \text{ é falsa}) = \beta$.

Vale destacar que o aumento do erro do tipo I leva à diminuição do erro do tipo II. Além disso, procura-se construir as hipóteses nulas e alternativas de tal forma que o erro do tipo I seja o erro mais grave a ser cometido.

Outro conceito importanteem teste de hipóteses é de **função característica da operação**. Essa função é a probabilidade de não rejeitar H_0 em função de μ e é dada por:

$$\beta(\theta) = \mathbb{P}(\text{não rejeitar} H_0 | \theta \in H_1).$$

Outra função importante é a **função poder do teste**, que é a probabilidade de rejeitar H_0 em função de μ e é dada por

$$1 - \beta(\theta) = \mathbb{P}(\text{rejeitar} H_0 | \theta \in H_1).$$

Como a apresentação matemática do conceito é confusa, vamos tentar observar tantos os erros quanto o poder do teste por meio da ilustração abaixo.

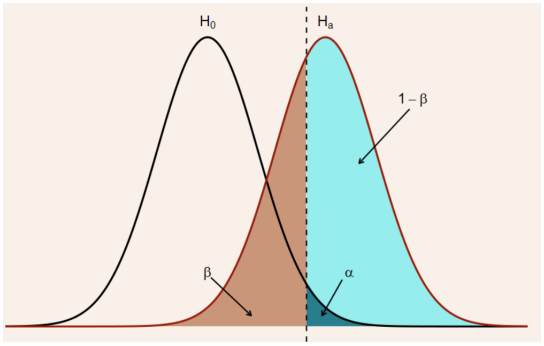

Figura 7.1: Representação do erro do tipo I, erro do tipo II e do poder do teste. Disponível em `https://www.inf.ufsc.br/~andre.zibetti/probabilidade/teste-de-hipoteses.html`.

Atenção

Conceitos que precisam estar em nossa mente a partir de agora:

- Hipótese nula e hipótese alternativa;
- Critério de decisão;
- Região de aceitação e região de rejeição;
- Probabilidades do erro do tipo I e do tipo II;

7.3 Aplicações das definições e dos conceitos

Para ilustrar todos os conceitos sobre teste de hipóteses que foram apresentados na seção anterior, vamos utilizar um exemplo retirado de Morettin and Bussab (2017) e discutir o problema.

Problema para discussão

Uma indústria usa parafusos importados para algumas componentes de suas máquinas e os avalia de acordo com sua resistência à tração (variável de interesse).
Existem dois países que produzem esses parafusos: país A ($\mu_A = 145$, $\sigma_A = 12$) e país B ($\mu_B = 155$, $\sigma_B = 20$).
Haverá um leilão de um lote de parafusos de origem desconhecida, com preço bem abaixo do mercado, e os representantes da indústria querem saber se fazem ou não uma oferta.
Observação: Um pouco antes do leilão, será divulgado a resistência média de uma amostra de 25 parafusos desse lote.

Temos as seguintes informações extraídas do problema:

- Variável de interesse: resistência à tração (X: resistência à tração);
- Parâmetro de interesse: média da resistência à tração (μ);
- Estimador de interesse: média amostral ($\overline{X}$).

Para explorar esse problema, vamos seguir 3 caminhos diferentes, conforme as descrições abaixo:

- Caminho 1:
  - Estabelecer H_0 e H_1 com base em um **critério de decisão** previamente fixado;
  - Encontrar os erros do tipo I e II.
- Caminho 2 (mais comum):
  - Estabelecer H_0 e H_1 com base em um **erro** previamente fixado;

- Encontrar o critério de decisão.

- Caminho 3 (mais comum):
  - Estabelecer H_0 e H_1 **sem o conhecimento** de H_1;
  - Encontrar o critério de decisão com base em H_0.

Seguindo pelo caminho 1:

Vamos definir o critério de decisão de tal forma que

– Se $\overline{X} > 150$, então o lote é do país B;

– Se $\overline{X} \leq 150$, então o lote é do país A;

Consequentemente:

– $H_0 : \overline{X} > 150$ (o lote é do país B);

– $H_1 : \overline{X} \leq 150$ (o lote é do país A);

Com base no critério de decisão, vamos verificar os erros associados a cada decisão, ou seja:

– Optar pelo lote do país A, mas o lote era do país B (erro do tipo I);

– Optar pelo lote do país B, mas o lote era do país A (erro do tipo II);

Não esqueça:

– $\mathbb{P}(\text{erro tipo I}) = \mathbb{P}(\text{rejeitar } H_0 | H_0 \text{ é verdadeira}) = \alpha$;

– $\mathbb{P}(\text{erro tipo II}) = \mathbb{P}(\text{não rejeitar } H_0 | H_0 \text{ é falsa}) = \beta$;

O erro do tipo I considera o fato de rejeitar $H_0 : \overline{X} > 150$, quando H_0 é verdadeira. Sob H_0 verdadeira e pelo Teorema Central do Limite (TCL), $\overline{X} \sim \mathcal{N}(155, 20/\sqrt{25})$. Calculando o erro do tipo I:

$$\begin{aligned} \alpha &= \mathbb{P}(\text{erro tipo I}) \\ &= \mathbb{P}(\overline{X} < 150 | \mu = 155, \sigma = 20) \\ &= \mathbb{P}\left(z < \frac{150 - 155}{20\sqrt{25}}\right) \\ &= 0,1056. \end{aligned}$$

O erro do tipo II considera o fato de não rejeitar $H_0 : \overline{X} > 150$, quando H_0 é falsa. Sob H_1 verdadeira e pelo TCL, $\overline{X} \sim \mathcal{N}(145, 12/\sqrt{25})$. Calculando o erro do tipo II:

$$\begin{aligned} \beta &= \mathbb{P}(\text{erro tipo II}) \\ &= \mathbb{P}(\overline{X} > 150 | \mu = 145, \sigma = 12) \\ &= \mathbb{P}\left(z > \frac{150 - 145}{12\sqrt{25}}\right) \\ &= 0,0187. \end{aligned}$$

Resumo dos resultados obtidos:

– $\alpha = \mathbb{P}(\text{erro tipo I}) = 0,1056$;

– $\beta = \mathbb{P}(\text{erro tipo II}) = 0,0187$;

Considerando o critério de decisão escolhido, estamos cometendo um erro do tipo I com maior probabilidadedo que o erro do tipo II. Esse critério privilegia que o lote de parafusos é do país A, já que o erro do tipo II está sob $H_0 : \overline{X} > 150$ falsa.

Seguindo pelo caminho 2, vamos agora fixar um erro e encontrar o critério de decisão. Neste caso, vamos fixar $\alpha = 0,05$. Consequentemente:

– $H_0 : \overline{X} > \overline{x}_c$ (o lote é do país B);

– $H_1 : \overline{X} \leq \overline{x}_c$ (o lote é do país A);

Fixando $\alpha = 5\%$, sob H_0 e pelo TCL, teremos:

$$\begin{aligned} \alpha = 0,05 &= \mathbb{P}(\text{erro tipo I}) = \mathbb{P}\left(\overline{X} \leq \overline{x}_c | \mu = 155, \sigma = 20\right) \\ &= \mathbb{P}\left(z \leq \frac{\overline{x}_c - 155}{20/\sqrt{25}}\right) = \mathbb{P}\left(z \leq \frac{\overline{x}_c - 155}{4}\right). \end{aligned}$$

Como $0,05 = \mathbb{P}(z \leq -1,645)$, então;

$$\begin{aligned} \frac{\overline{x}_c - 155}{4} &= -1,645 \\ &\Rightarrow \overline{x}_c = 148,2. \end{aligned}$$

O critério de decisão é, portanto, dado por:

– Se $\overline{X} > 148,2$, dizemos que o lote é do país B;

– Se $\overline{X} \leq 148,2$, dizemos que o lote é do país A;

O erro do tipo I para esse critério de decisão é de $\alpha = 0,05$ e o erro do tipo II é dado por:

$$\begin{aligned} \beta &= \mathbb{P}(\text{erro tipo II}) = \mathbb{P}\left(\overline{X} > 148,2 | \mu = 145, \sigma = 12\right) \\ &= \mathbb{P}\left(z \leq \frac{148,2 - 145}{12/\sqrt{25}}\right) = 0,0793; \end{aligned}$$

Observamos que a probabilidade do erro do tipo I é menor, ou seja, esse critério privilegia que o lote de parafusos é do país B.

Seguindo pelo caminho 3, vamos agora estabelecer as hipóteses e definir o critério de decisão com base somente nas informações da hipótese nula (país B).

Nossas hipóteses agora são:

– H_0 : os parafusos são do país B ($\mu = 155$ e $\sigma = 20$);

– H_1 : os parafusos não são do país B;

Como estamos avaliando a resistência média à tração, nosso interesse é sobre o teste:

– $H_0 : \mu = 155$;

– $H_1 : \mu \neq 155$ (podemos adaptar para $H_1 : \mu < 155$ ou $H_1 : \mu > 155$);

Obseve que podemos ter uma infinidades de parâmetros para a hipótese alternativa. Logo, só podemos trabalhar com as informações referentes à H_0, ou seja, considerando o fato de que $\mu = 155$.

A melhor saída é trabalhar com o erro do tipo I (α) que considera H_0 verdadeira. Para isso, fixamos um valor para α e concluímos, com base nas evidências (amostra), se rejeitamos ou não H_0.

E como fica o erro do tipo II? Como temos uma infinidade de valores possíveis para μ, vamos considerar a probabilidade não rejeitar H_0 em função de um valor de μ.

A **função característica** será dada por:

$$\beta(\mu) = \mathbb{P}(\text{não rejeitar } H_0|\mu);$$

No exemplo em questão, temos:

$$\beta(\mu) = \mathbb{P}(\overline{X} = 150|\mu);$$

Se fizermos $\pi(\mu) = 1 - \beta(\mu)$, temos a chamada **função poder do teste**, que é a probabilidade de se rejeitar H_0 em função de μ;

Voltando à ideia inicial do caminho 3, suponha agora que o interesse não está em avaliar se a resistência média do lote de parafusos é maior ou menor que a do país B, mas sim se essa resitência é diferente do país B.

As hipóteses são:

– $H_0 : \mu = 155$;

– $H_1 : \mu \neq 155$;

Assim, a regra de decisão será:

– Se $\overline{x}_{c_1} \leq \overline{X} \leq \overline{x}_{c_2}$, então o lote **é do país B**;

– Se $\overline{X} < \overline{x}_{c_1}$ ou $\overline{X} > \overline{x}_{c_2}$, então o lote **não é do país B**;

Se fixarmos $\alpha = 0,05$, existirão muito valores que satisfazem a primeira condição acima. Vamos focar, portanto, somente naquelas soluções que sejam **simétricas em relação à média**;

Então, sob H_0 e pelo TCL, teremos:

$$\begin{aligned}
\alpha = 0,05 &= \mathbb{P}(\text{erro tipo I}) = \mathbb{P}\left(\overline{X} < \overline{x}_{c_1} \text{ ou } \overline{X} > \overline{x}_{c_2}|\mu = 155, \sigma = 20\right) \\
&= \mathbb{P}\left(z < \frac{\overline{x}_{c_1} - 155}{20/\sqrt{25}} \text{ ou } z > \frac{\overline{x}_{c_2} - 155}{20/\sqrt{25}}\right) \\
&= \mathbb{P}\left(z < \frac{\overline{x}_{c_1} - 155}{4} \text{ ou } z > \frac{\overline{x}_{c_2} - 155}{4}\right).
\end{aligned}$$

Como $0,05 = \mathbb{P}(z \leq -1,96 \text{ ou } z > 1,96)$, então;

$$\begin{aligned} \frac{\overline{x}_{c_1} - 155}{4} &= -1,96 \\ &\Rightarrow \overline{x}_{c_1} = 147,16; \end{aligned}$$

e

$$\begin{aligned} \frac{\overline{x}_{c_2} - 155}{4} &= 1,96 \\ &\Rightarrow \overline{x}_{c_2} = 162,84. \end{aligned}$$

Assim, a regra de decisão será:

– Se $147,16 \leq \overline{X} \leq 162,84$, então o lote é do país B;

– Se $\overline{X} < 147,16$ ou $\overline{X} > 162,84$, então o lote não é do país B;

7.4 Passo a passo para construir um teste de hipóteses

Antes de apresentarmos os testes de hipóteses mais comuns, vamos discutir um passo a passo de como contruir e obter os valores associados a um teste de hipótese.

- Passo 1 – Estabelecer as hipóteses: fixamos a hipótese a ser testada (por exemplo, podemos fazer $H_0 : \mu = \mu_0$). Dependendo da informação que fornece o problema que estamos estudando, a hipótese alternativa pode ter uma das três formas abaixo:

  - Teste bilateral: $H_1 : \mu \neq \mu_0$
  - Teste unilateral à direita: $H_1 : \mu \geq \mu_0$
  - Teste unilateral à esquerda: $H_1 : \mu \leq \mu_0$

- Passo 2 – Calcular, sob a hipótese nula, a estatística do teste (cada teste apresentará sua estatística correspondente, conforme veremos na próxima seção);

- Passo 3 – Fixar o nível de significância α e determinar a região crítica (região de rejeição e região de não rejeição da hipótese nula);

- Passo 4 – Localizar a posição da estatística do teste calculada na região crítica definida, de tal forma que possamos ter uma conclusão. Essa conclusão depende da hipótese alternativa, pois se

  - Se o teste for bilateral, determinamos os pontos críticos $-t_{\alpha/2}$ e $t_{\alpha/2}$ tais que $\mathbb{P}[T \geq t_{\alpha/2}] = \mathbb{P}[T \leq -t_{\alpha/2}] = \alpha/2$ a partir da distribuição t de Student com $n-1$ graus de liberdade.
  - Se o teste for unilateral à direita, determinamos o ponto crítico t_α tal que $\mathbb{P}[T \geq t_\alpha] = \alpha$.
  - Se o teste for unilateral à esquerda, determinamos o ponto $-t_\alpha$ tal que $\mathbb{P}[T \leq -t_\alpha] = \alpha$.

7.5 Os testes de hipóteses mais comuns

7.5.1 Teste de hipóteses para a média (com variância conhecida)

Neste caso estamos interessados em realizar inferência sobre a média populacional μ, com base na amostra $X_1, X_2, \ldots, X_n$ de uma população com distribuição normal e variância conhecida.

As hipóteses a serem testadas serão:

$H_0 : \mu = \mu_0$;

$H_1 : \mu < \mu_0$ ou $H_1 : \mu \neq \mu_0$ ou $H_1 : \mu > \mu_0$.

Conforme vimos no capítulo sobre distribuição amostral, a estatística que podemos utilizar para fazer inferência sobre o parâmetro de interesse é dada por

$$Z_{obs} = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1).$$

A partir do valor obtido de Z_{obs} e do conhecimento sobre α (que nos fornece a região crítica), verificamos se vamos rejeitar ou não a hipótese nula, observando os seguintes casos:

- Se o teste for bilateral: se $Z_{\text{obs}} \geq z_{\alpha/2}$ ou se $Z_{\text{obs}} \leq -z_{\alpha/2}$, rejeitamos H_0. Caso contrário, não rejeitamos H_0.
- Se o teste for unilateral à direita: se $Z_{\text{obs}} \geq z_{\alpha}$, rejeitamos H_0. Caso contrário, não rejeitamos H_0.
- Se o teste for unilateral à esquerda: se $Z_{\text{obs}} \leq -z_{\alpha}$, rejeitamos H_0. Caso contrário, não rejeitamos H_0.

Exemplo 7.5.1. *(Morettin e Bussab, 2017) Uma máquina enche pacotes de café segundo uma distribuição normal com média 500g e variância 400g. Uma amostra de 16 pacotes foi coletada para verificar se a máquina está regulada e obteve-se $\overline{X} = 492g$. Ao nível de 1%, podemos afirmar que a máquina está regulada ou não?*

Solução:

Passo 1 – Definir as hipóteses:

$H_0 : \mu = 500$;

$H_1 : \mu \neq 500$;

Passo 2 – Definir o estimador e a estatística do problema:

$$Z_{obs} = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1);$$

Passo 3 – Sob H_0, fixar o erro do tipo I e calcular a região crítica:

Como $\alpha = 1\%$, então, utilizando a tabela da normal padrão, percebemos que o a região crítica é dada por $Z_{obs} < -2,58$ e $Z_{obs} > 2,58$;

Passo 4 – Calcular o valor da estatística do teste e identificar se esse valor pertence ou não à região crítica. Neste caso, temos que:

$$\begin{aligned} Z_{obs} &= \frac{492-500}{\frac{20}{\sqrt{16}}} \\ &= -1,6; \end{aligned}$$

Ou seja, Z_{obs} não pertence à região crítica, logo não rejeitamos H_0;

Conclusão: *não há evidências para concluir que a máquina está desrregulada.*

7.5.2 Teste de hipóteses para a média (com variância desconhecida)

Neste caso estamos interessados em realizar inferência sobre a média populacional μ, com base na amostra $X_1, X_2, \ldots, X_n$ de uma população com distribuição normal e variância desconhecida.

As hipóteses a serem testadas serão:

$H_0 : \mu = \mu_0$;

$H_1 : \mu < \mu_0$ ou $H_1 : \mu \neq \mu_0$ ou $H_1 : \mu > \mu_0$.

Conforme vimos no capítulo sobre distribuição amostral, a estatística que podemos utilizar para fazer inferência sobre o parâmetro de interesse é dada por

$$Z_{obs} = \frac{\overline{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim \mathcal{N}(0,1).$$

$$T_{obs} = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}} \sim t(n-1);$$

A partir do valor obtido de T_{obs} e do conhecimento sobre α (que nos fornece a região crítica), verificamos se vamos rejeitar ou não a hipótese nula, observando os seguintes casos:

- Se o teste for bilateral: se $T_{\text{obs}} \geq t_{\alpha/2}$ ou se $T_{\text{obs}} \leq -t_{\alpha/2}$, rejeitamos H_0. Caso contrário, não rejeitamos H_0.
- Se o teste for unilateral à direita: se $T_{\text{obs}} \geq t_{\alpha}$, rejeitamos H_0. Caso contrário, não rejeitamos H_0.
- Se o teste for unilateral à esquerda: se $T_{\text{obs}} \leq -t_{\alpha}$, rejeitamos H_0. Caso contrário, não rejeitamos H_0.

Exemplo 7.5.2. *(Morettin e Bussab, 2017) Uma fabricante afirma que seus cigarros contêm não mais que 30mg de nicotina. Uma amostra de 25 cigarros fornece uma média de 31,5 mg e desvio-padrão de 3 mg. No nível de significância de 5%, os dados refutam ou não a afirmação do fabricante?*

Solução:

Passo 1 – Definir as hipóteses:

- $H_0 : \mu = 30$;

- $H_1 : \mu > 30$;

Passo 2 – Definir o estimador e a estatística do problema:

$$T_{obs} = \frac{\overline{X} - \mu_0}{\frac{s}{\sqrt{n}}} \sim t(n-1);$$

Passo 3 – Sob H_0, fixar o erro do tipo I e calcular a região crítica:

Como $\alpha = 5\%$, então, utilizando a tabela da distribuição t-Student, percebemos que a região crítica é dada por $T_{obs} > 1,711$;

Passo 4 – Calcular o valor da estatística do teste e identificar se esse valor pertence ou não à região crítica. Neste caso, temos que:

$$T_{obs} = \frac{31,5 - 30}{\frac{3}{\sqrt{25}}} = 2,5;$$

Ou seja, T_{obs} pertence à região crítica, logo rejeitamos H_0;

***Conclusão:** há evidências para concluir que os cigarros têm mais que 30g de nicotina.*

7.5.3 Teste de hipóteses para a proporção

Neste caso estamos interessados em realizar inferência sobre a proporção p de indivíduos/objetos com certa característica, com base na amostra $X_1, X_2, \ldots, X_n$ de uma população.

As hipóteses a serem testadas serão:

– $H_0 : p = p_0$;

– $H_1 : p < p_0$ ou $H_1 : p \neq p_0$ ou $H_1 : p > p_0$;

Conforme vimos no capítulo sobre distribuição amostral, a estatística que podemos utilizar para fazer inferência sobre o parâmetro de interesse é dada por

$$Z_{\text{obs}} = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \sim \mathcal{N}(0,1).$$

A partir do valor obtido de Z_{obs} e do conhecimento sobre α (que nos fornece a região crítica), verificamos se vamos rejeitar ou não a hipótese nula, observando os seguintes casos:

- Se o teste for bilateral: se $Z_{\text{obs}} \geq z_{\alpha/2}$ ou se $Z_{\text{obs}} \leq -z_{\alpha/2}$, rejeitamos H_0. Caso contrário, não rejeitamos H_0.
- Se o teste for unilateral à direita: se $Z_{\text{obs}} \geq z_{\alpha}$, rejeitamos H_0. Caso contrário, não rejeitamos H_0.
- Se o teste for unilateral à esquerda: se $Z_{\text{obs}} \leq -z_{\alpha}$, rejeitamos H_0. Caso contrário, não rejeitamos H_0.

Exemplo 7.5.3. *(Morettin e Bussab, 2017 - Adaptado) Uma estação de TV afirma que 60% dos televisores estavam ligados no seu programa especial da última segunda-feira. Uma rede competidora deseja contestar essa*

afirmação e, com base numa amostra de 200 famílias, constatou que 104 não assistitam ao programa. Avalie a veracidade da afirmação da estação considerando $\alpha = 0,05$.

Solução passo a passo:

Passo 1 – Definir as hipóteses:

$H_0 : p = 0,6$;

$H_1 : p < 0,6$.

Passo 2 – Definir o estimador e a estatística do problema:

$$Z_{obs} = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \sim \mathcal{N}(0,1)$$

Passo 3 – Sob H_0*, fixar o erro do tipo I e calcular a região crítica:*

Como $\alpha = 5\%$*, então, utilizando a tabela da distribuição normal, percebemos que a região crítica é dada por* $Z_{obs} < -1,645$;

Passo 4 – Calcular o valor da estatística do teste e identificar se esse valor pertence ou não à região crítica. Neste caso, temos que:

$$Z_{obs} = \frac{0,52 - 0,6}{\sqrt{0,24/200}} = -2,31;$$

Ou seja, Z_{obs} *pertence à região crítica, logo rejeitamos* H_0;

Conclusão: *há evidências para concluir que a audiência do programa foi menor que* 60%.

7.6 Outros teste de hipóteses

7.6.1 Testes qui-quadrado: aderência, homogeneidade e indepedência

O teste qui-quadrado pode ser usado nos seguintes casos:

- **Teste de aderência (ou concordância);**
- **Teste de homogeneidade (ou heterogeneidade);**
- **Teste de independência (ou contigência).**

Existem outros testes que também utilizam a estatística com distribuição qui-quadrado, mas, aqui, vamos abordar, em detalhes, somente os três testes citados anteriormente.

Teste de aderência (ou concordância)

Permite verificar a aderência (proximidade) de um conjunto de dados com relação a determinada distribuição de probabilidade. Nesse caso, testamos as hipóteses

$H_0 : p = p_0$

$$H_1 : p \neq p_0,$$

sendo p a população e p_0 uma distribuição especificada. A estatística do teste e a sua respectiva distribuição, nesse caso, são dadas por:

$$\chi^2 = \sum_{i=1}^{s} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2(s-1),$$

sendo s o número de categorias da variável em estudo.

Exemplo 7.6.1. *Podemos usar esse teste para avaliar se um dado é viciado ou não. Para isso, precisamos realizar um experimento de tal forma que se considere sucessivos lançamentos desse dado, com o número de cada face obtida sendo anotada. Abaixo, temos um exemplo para 120 lançamentos.*

Tabela 7.1: Frequências esperadas e observadas do experimento.

Face	1	2	3	4	5	6	Total
Frequência observada	10	25	30	35	10	10	120
Frequência esperada	20	20	20	20	20	20	120

Veja que a tabela, além dos valores observados do experimento, já inclui os valores esperados, ou seja, aqueles valores que esperamos que ocorra, se o dado não fosse viciado. Dessa forma, basta usarmos a estatística apresentada e verificar a validade da hipótese levantada.

Teste de homogeneidade (ou heterogeneidade)

Permite verificar se amostras diferentes em uma série de experimentos semelhantes são homogêneas ou heterogêneas. Nesse caso, testamos as hipóteses

$$H_0 : p_1 = p_2$$

$$H_1 : p_1 \neq p_2,$$

sendo p_1 a população 1 e p_2 a população 2. A estatística do teste e a sua respectiva distribuição, nesse caso, são dadas por:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2(v),$$

sendo $v = (r-1)(s-1)$, com r denotando o número de linhas e s denotando o número de colunas da tabela de observações.

Exemplo 7.6.2. *Podemos usar esse teste para avaliar se o desempenho de uma fábrica X na confecção de parafusos é igual a sua concorrente Y. Para isso, precisamos realizar uma coleta de informações das duas fábricas, por tipo de parafuso. Abaixo, temos um exemplo de coleta.*

Tabela 7.2: Frequências observadas do experimento.

Fábrica	Tipo de parafuso			Total
	A	B	C	
X	35	45	25	105
Y	40	42	10	92
Total	75	87	35	197

Dessa forma, basta usarmos a estatística apresentada e verificar a validade da hipótese levantada.

Teste de independência (ou contigência)

Permite verificar se existe independência entre duas variáveis medidas nas mesmas unidades amostrais (aplicável nos casos em que não se dispõe de uma teoria ou modelo para informar a respeito das probabilidade de ocorrência esperadas nas diferentes classes). Nesse caso, testamos as hipóteses

$H_0 : p_{ij} = p_{i\cdot}p_{\cdot j}$, para todo par (i, j),

$H_1 : p_{ij} \neq p_{i\cdot}p_{\cdot j}$, para algum par (i, j),

sendo p_{ij} a probabilidade de uma observação pertencer às categorias i e j, simultaneamente, e $p_{i\cdot} = \sum_{j=1}^{s} p_{ij}$ e $p_{\cdot j} = \sum_{i=1}^{r} p_{ij}$ as probabilidade marginais, com $i = 1, 2, ..., r$ e $j = 1, 2, ..., s$. A estatística do teste e a sua respectiva distribuição, nesse caso, são dadas por:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \sim \chi^2(v),$$

sendo $v = (r - 1)(s - 1)$, com r denotando o número de linhas e s denotando o número de colunas da tabela de observações.

Exemplo 7.6.3. *Podemos usar esse teste para avaliar se o tipo de acidente (com feridos ou não) independe do motoclista usar ou não o capacete. Abaixo, temos um exemplo de coleta.*

Tabela 7.3: Frequências observadas do experimento.

Frequência	Acidente com ferimentos	Acidente sem ferimentos
Usam capacete	35	45
Não usam capacete	40	42

Dessa forma, basta usarmos a estatística apresentada e verificar a validade da hipótese levantada.

7.7 Exercícios

1. (Morettin and Bussab, 2017) Uma máquina enche pacotes de café segundo uma distribuição normal com média 500g e variância 400g. Uma amostra de 16 pacotes foi coletada para verificar se a máquina está regulada e obteve-se $\overline{X} = 492g$. Ao nível de 1%, podemos afirmar que a máquina está regulada ou não?

2. (Morettin and Bussab, 2017) Uma fabricante afirma que seus cigarros contêm não mais que 30mg de nicotina. Uma amostra de 25 cigarros fornece uma média de 31,5 mg e desvio-padrão de 3 mg. No nível de significância de 5%, os dados refutam ou não a afirmação do fabricante?

3. (Morettin and Bussab, 2017) Uma estação de TV afirma que 60% dos televisores estavam ligados no seu programa especial da última segunda-feira. Uma rede competidora deseja contestar essa afirmação e, com base numa amostra de 200 famílias, constatou que 104 não assistitam ao programa. Avalie a veracidade da afirmação da estação considerando $\alpha = 0,05$.

CAPÍTULO 8

Modelo de regressão linear simples

8.1 Introdução

Em muitos campos científicos, ou até mesmo em situações do cotidiano, temos interesse em investigar se duas ou mais variáveis estão inerentemente relacionadas (mesmo que não exista relação de causa-efeito, podemos investigar quaisquer relacionamento entre variáveis). Um engenheiro, por exemplo, pode estar interessado em saber se a temperatura de um material apresenta alguma relação com sua resistência; por sua vez, um médico pode precisar saber se o tempo de vida de um paciente é diretamente influenciado pelo seu tipo de alimentação; ou um comerciante, que deseja saber se em dias de chuva suas vendas são afetadas ou não.

Em todos esses cenários, podemos aplicar uma das técnicas mais comuns e importantes em análise de dados: a **Análise de regressão**. Ela estabelece um modelo que possa descrever a relação entre as variáveis de estudo. Nesse modelo, temos a variável aleatória Y, chamada de variável resposta, e a variável observável X, chamada de variável explicativa, explanatória ou covariável. Em alguns textos encontramos descrições de variável dependente para Y e variável independente para X, porém essa terminologia é confusa, visto que a utilização de "independência", neste caso, é diferente dos casos em probabilidade e inferência, pelo simples fato dos $\mathrm{X}'s$ não serem necessariamente variáveis aleatórias, logo não podem ser estatisticamente independentes.

Na regressão linear simples há uma forte indicação de que os pontos referentes ao par (X, Y) repousam aleatoriamente dispersos em torno de uma linha reta. Consequentemente, é provável considerar que a média da variável Y esteja relacionada a X pela seguinte relação linear:

$$\mathbb{E}[Y|x] = \beta_0 + \beta_1 x \tag{8.1}$$

É importante lembrar que essa esperança é uma suposição de que a regressão de Y em X é linear, pelo simples fato de não existir nenhuma teoria subjacente para apoiar a relação de linearidade. Porém, trata-se de uma aproximação razoável, uma vez que a relação linear é muito conveniente para se trabalhar.

Assim, para sermos bem formais, devemos escrever:

$$\mathbb{E}[Y|x] \approx \beta_0 + \beta_1 x$$

Entretanto, se começarmos a partir da suposição de que o par (X_i, Y_i) tem uma distribuição normal bivariada, imediatamente segue que a regressão de Y em X é linear.

Voltando a ideia central do estudo, é notável que, pela reta apresentada anteriormente, o valor de y não "cai"exatamente na linha da reta estimada, logo o valor de Y é determinado pela função do valor médio (termo determinístico) mais um termo de erro aleatório (parte aleatória):

$$Y = \beta_0 + \beta_1 x + \epsilon_i \tag{8.2}$$

Devemos fazer a suposição de que os erros seguem uma Normal com média 0 e variância fixa σ^2, ou seja, $\epsilon_i \sim N(0, \sigma^2)$.

Consequentemente, temos:

$$\begin{aligned} Y_i|x_i &= \beta_0 + \beta_1 x + \epsilon_i \\ \mathbb{E}[Y_i|x_i] &= \mathbb{E}[\beta_0 + \beta_1 x + \epsilon_i] \\ &= \beta_0 + \beta_1 x + E[\epsilon_i] \\ &= \beta_0 + \beta_1 x + 0 \\ &= \beta_0 + \beta_1 x \end{aligned}$$

Da mesma forma,

$$\begin{aligned} Y_i|x_i &= \beta_0 + \beta_1 x + \epsilon_i \\ V[Y_i|x_i] &= V[\beta_0 + \beta_1 x + \epsilon_i] \\ &= V[\epsilon_i] \\ &= \sigma^2 \end{aligned}$$

Logo,

$$Y_i|x_i \sim N(\beta_0 + \beta_1 x, \sigma^2)$$

8.2 Estimação dos parâmetros

Para encontrarmos a reta que miniminiza as distâncias entre o valor real, Y_i, e o valor estimado, $\widehat{Y}_i$, devemos minimizar o somatório dos quadrados dos erros. Nessa derivação iremos encontrar os pontos que satisfazem tal objetivo e, consequentemente, serão estes nossas estimativas para os parâmetros. Existem vários métodos para encontrar tais estimadores, porém iremos usar o Método dos Mínimos Quadrados. Para a estimação pelo Método de Máxima Verossimilhança temos resultados idênticos.

O objetivo, portanto, é:

$$\begin{aligned} \epsilon_i &= Y_i - (\beta_0 + \beta_1 x_i) \\ \epsilon_i^2 &= [Y_i - (\beta_0 + \beta_1 x_i)]^2 \\ \sum_{i=1}^{n} \epsilon_i^2 &= \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 x_i)]^2 \end{aligned}$$

Considerando

$$E = \sum_{i=1}^{n} \epsilon_i^2$$

Logo,

$$\begin{aligned} \frac{\partial E}{\partial \beta_0} &= \sum_{i=1}^{n} 2\left[Y_i - (\beta_0 + \beta_1 x_i)\right](-1) \\ \frac{\partial E}{\partial \beta_1} &= \sum_{i=1}^{n} 2\left[Y_i - (\beta_0 + \beta_1 x_i)\right](-x_i) \\ \widehat{\beta_0} &= \overline{Y} - \widehat{\beta_1}\overline{X} \qquad (8.3) \\ \widehat{\beta_1} &= \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{YY}}{\sum_{i=1}^{n} {X_i}^2 - n\overline{X}^2} \qquad (8.4) \end{aligned}$$

Para simplificar a escrita de $\widehat{\beta_1}$, usaremos:

$$\widehat{\beta_1} = \frac{S_{xy}}{S_{xx}}$$

Estimados os valores, encontramos então a reta estimada:

$$\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_i$$

Vamos avaliar agora a validade (saber se o estimador é ou não viciado) e a precisão (variabilidade) dos estimadores dos parâmetros:

$$\begin{aligned} E[\widehat{\beta_1}] &= \beta_1 \\ V[\widehat{\beta_1}] &= \frac{\sigma^2}{S_{xx}} \end{aligned}$$

Prova:

$$
\begin{aligned}
E[\widehat{\beta_1}] &= \frac{S_{xy}}{S_{xx}} \\
&= E\left[\frac{\sum_{i=1}^{n}(X_i-\overline{X})Y_i}{S_{xx}}\right] \\
&= \frac{\sum_{i=1}^{n}(X_i-\overline{X})E[Y_i]}{S_{xx}} \\
&= \frac{\sum_{i=1}^{n}(X_i-\overline{X})(\beta_0+\beta_1 X_i)}{S_{xx}} \\
&= \frac{\beta_0\sum_{i=1}^{n}(X_i-\overline{X})+\beta_1\sum_{i=1}^{n}(X_i-\overline{X})X_i}{S_{xx}} \\
&= \frac{\beta_1\sum_{i=1}^{n}(X_i-\overline{X})X_i}{S_{xx}} \\
&= \beta_1
\end{aligned}
$$

A variância de β_1 segue o mesmo raciocínio. Pode-se, portanto, mostrar também que:

$$
\begin{aligned}
E[\widehat{\beta_0}] &= \beta_0 \\
V[\widehat{\beta_0}] &= \sigma^2\left[\frac{1}{n}+\frac{\overline{x}^2}{S_{xx}}\right]
\end{aligned}
$$

Vista as demonstrações acima, seria interessante avaliar a relação existente entre $\widehat{\beta_0}$ e $\widehat{\beta_1}$, para isso calculemos a covariância entre tais estimadores, porém precisaremos de um Lema e um resultado para obter tal resultado, observe abaixo.

Sejam $Y_1, Y_2, ..., Y_n$ variáveis aleatórias não correlacionadas com $V[Y_i] = \sigma^2$ para todo $i = 1, 2, ..., n$. Suponhamos que $c_1, c_2, ..., c_n$ e $d_1, d_2, ..., d_n$ sejam dois conjuntos de constantes. Então

$$
Cov\left[\sum_{i=1}^{n}c_iY_i, \sum_{i=1}^{n}d_iY_i\right] = Cov\left[\sum_{i=1}^{n}c_id_i\right]V[Y_i]
$$

O que será demonstrado/calculado agora será muito útil para alguns outros resultados.

$$
\begin{aligned}
Cov(\overline{Y}, \widehat{\beta_1}) &= Cov(\sum_{i=1}^{n}\frac{Y_i}{n}, \sum_{i=1}^{n}\frac{(X_i-\overline{X})Y_i}{S_{xx}}) \\
&= Cov(\sum_{i=1}^{n}\frac{1}{n}\frac{(X_i-\overline{X})}{S_{xx}})V[Y_i] \\
&= 0
\end{aligned}
$$

Portanto, a relação existente é:

$$
\begin{aligned}
Cov(\widehat{\beta_0}, \widehat{\beta_1}) &= Cov(\overline{Y}-\widehat{\beta_1}\overline{X}, \widehat{\beta_1}) \\
&= Cov(\overline{Y}, \widehat{\beta_1}) - Cov(\widehat{\beta_1}\overline{X}, \widehat{\beta_1}) \\
&= 0 - \overline{X}V[\widehat{\beta_1}] \\
&= -\frac{\overline{x}\sigma^2}{S_{xx}}
\end{aligned}
$$

A medida que aumentamos o valor da inclinação da reta de regressão, diminuimos o "corte"na reta Y, pois a covariância entre as estimativas é negativa.

8.3 Análise de variância

A principal medida para quantificar o quão bom é um modelo estimado para os dados é a sua variância. Entretanto, usar o somente o termo variância como sendo a principal medida não é o ideal em modelos de regressão linear, visto que vamos trabalhar aqui com o particionamento da variabilidade, ou seja, tal análise desmente seu próprio nome, pois não está preocupada em analisar variâncias, mas sim, a variabilidade das médias ou, simplesmente, a significância do modelo de regressão. Então, caro leitor, parece razoável, para iniciarmos o estudo, comparar os valores de Y_i com a média da variável resposta, pois essas distâncias nos informarão se a reta de regressão é significativa ou não para os dados. Assim, se não houver efeito de regressão o comportamento dos dados pode ser explicado pelo própria reta da média, ou seja, $\overline{Y}$.

Em suma, para análise de variância, iremos comparar Y_i com $\overline{Y}$, ou seja, iremos particionar o seguinte somatório:

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

$$\begin{aligned}
\sum_{i=1}^{n}(Y_i - \overline{Y})^2 &= \sum_{i=1}^{n}(Y_i - (\widehat{Y_i} - \widehat{Y_i}) - \overline{Y})^2 \\
&= \sum_{i=1}^{n}(Y_i - \widehat{Y_i} + \widehat{Y_i} - \overline{Y})^2 \\
&= \sum_{i=1}^{n}(Y_i - \widehat{Y_i})^2 + \sum_{i=1}^{n}(Y_i - \overline{Y})^2 + 2\sum_{i=1}^{n}(Y_i - \widehat{Y_i})(Y_i - \overline{Y})
\end{aligned}$$

Mas a terceira parcela da soma é o mesmo que:

$$\sum_{i=1}^{n}(Y_i - \widehat{Y_i})(Y_i - \overline{Y}) = \sum_{i=1}^{n}(Y_i - \widehat{Y_i})Y_i - \sum_{i=1}^{n}(Y_i - \widehat{Y_i})\overline{Y}$$

Calculando cada parte, temos:

$$\begin{aligned}
\sum_{i=1}^{n}(Y_i - \widehat{Y_i})\overline{Y} &= \overline{Y}\left[\sum_{i=1}^{n}(Y_i - (\widehat{\beta_0} + \widehat{\beta_1}x_i))\right] \\
&= \overline{Y}\left[\sum_{i=1}^{n}Y_i - \sum_{i=1}^{n}\widehat{\beta_0} + \widehat{\beta_1}x_i\right] \\
&= \overline{Y}\left[\sum_{i=1}^{n}Y_i - n\widehat{\beta_0} + \widehat{\beta_1}\sum_{i=1}^{n}x_i\right] \\
&= \overline{Y}\left[\sum_{i=1}^{n}Y_i - n\widehat{\beta_0} + n\widehat{\beta_1}\overline{X}\right] \\
&= \overline{Y}\left[\sum_{i=1}^{n}Y_i - n(\overline{Y} + \widehat{\beta_1}\overline{X}) + n\widehat{\beta_1}\overline{X}\right] \\
&= \overline{Y}\left[\sum_{i=1}^{n}Y_i - n\overline{Y} - n\widehat{\beta_1}\overline{X} + n\widehat{\beta_1}\overline{X}\right] \\
&= 0
\end{aligned}$$

$$\begin{aligned}\sum_{i=1}^{n}(Y_i-\widehat{Y_i})\widehat{Y}_i &= \sum_{i=1}^{n}(Y_i\widehat{Y}_1-\widehat{Y_i}^2)\\ &= \sum_{i=1}^{n}\left[Y_i(\widehat{\beta}_0+\widehat{\beta}_1x_i)-(\widehat{\beta}_0+\widehat{\beta}_1x_i)^2\right]\\ &= 0\end{aligned}$$

Portanto:

$$\sum_{i=1}^{n}(Y_i-\overline{Y})^2=\sum_{i=1}^{n}(Y_i-\widehat{Y_i})^2+\sum_{i=1}^{n}(\widehat{Y_i}-\overline{Y})^2 \tag{8.5}$$

Em palavras, tal resultado é:

- $SQ_{total}=\sum_{i=1}^{n}(Y_i-\overline{Y})^2$
- $SQ_{res}=\sum_{i=1}^{n}(Y_i-\widehat{Y_i})^2$
- $SQ_{reg}=\sum_{i=1}^{n}(\widehat{Y_i}-\overline{Y})^2$

Para a construção do teste de hipóteses que compõe a análise de variância, precisaremos de um teorema que nos informa um resultado muito importante.

Teorema 8.3.1 (Cochram)**.** *Se todas as n observações $Y_1, Y_2, ..., Y_n$, independentes, vêm da mesma distribuição normal com média μ e variância σ^2 e a soma de quadrados total é decomposta em k somas de quadrados SQ_k, cada uma com seus respectivos graus de liberdade, então*

$$\frac{SQ_k}{\sigma^2}$$

são variáveis aleatórias com distribuição quiquadrado, independentes, com gl_k graus de liberdade se

$$\sum_{k=1}^{n}gl_k=gl_{total}$$

Sabe-se que

- $SQ_{total}=\sum_{i=1}^{n}(Y_i-\overline{Y})^2\rightarrow(n-1)g.l.$
- $SQ_{res}=\sum_{i=1}^{n}(Y_i-\widehat{Y_i})^2\rightarrow(n-2)g.l.$
- $SQ_{reg}=\sum_{i=1}^{n}(\widehat{Y_i}-\overline{Y})^2\rightarrow 1g.l.$

Então

$$\begin{aligned}gl_{total} &= gl_{reg}+gl_{res}\\ (n-1) &= 1+(n-2)\end{aligned}$$

Resultado que satisfaz a primeira condição do teorema. Para satisfazer a segunda, devemos supor, sob alguma condição, que os $Y_i's$ são independentes. Para isso considere a hipótese

$$H_0 : \beta_1 = 0$$

Se H_0 for verdadeira, então

$$Y_i = \beta_0 + \epsilon_i$$

Consequentemente

$$\begin{aligned} E[Y_i] &= \beta_0 \\ V[Y_i] &= \sigma^2 \end{aligned}$$

Nos levando a afirmar que

$$Y_i \sim N(\beta_0, \sigma^2) \tag{8.6}$$

Dessa forma, as condições do Teorema de Cochram, sob H_0 verdadeira, são satisfeitas e assim, podemos dizer que:

$$\begin{aligned} \frac{SQ_{reg}}{\sigma^2} &\sim \chi^2(1) \\ \frac{SQ_{res}}{\sigma^2} &\sim \chi^2(n-2) \end{aligned}$$

Mostrado tudo isso, podemos enfim chegar na última parte da análise de variância, o teste F. Pelo Teorema de Cochran, podemos saber a distribuição de:

$$\begin{aligned} \frac{SQ_{reg}}{\sigma^2} &\sim \chi^2(1) \\ \frac{SQ_{res}}{\sigma^2} &\sim \chi^2(n-2) \end{aligned}$$

Consequentemente, sob H_0 verdade, a divisão de de duas quantidade com distrbuição quiquadrado com 1 e n-2 graus de liberdade, respectivamente, tem distribuição F(1,n-2):

$$\frac{\frac{\frac{SQ_{reg}}{\sigma^2}}{1}}{\frac{\frac{SQ_{reg}}{\sigma^2}}{n-2}} = \frac{\frac{SQreg}{1}}{\frac{SQres}{n-2}} = \frac{QMreg}{QMres} \sim F(1, n-2) \tag{8.7}$$

Toda a teoria acima foi desenvolvida para testar a hipótese nula antes definida, isto é, rejeitaremos H_0 se

$$P(F(1, n-2) \geq F_0) < \alpha$$

Em que α é o nível de significância adotado no teste.

Estamos fazendo todas as suposições acima sob H_0 verdade, pois queremos saber se é vantajosa a adoção do modelo linear, ou seja, é observar a redução do resíduo. Se tal redução for muito pequena, os dois modelos

serão praticamente equivalentes, e isso ocorre quando a inclinação é zero ou muito pequena, não compensando usar um modelo mais complexo.

Em resumo, caro leitor, temos a seguinte sucessão de ideias: realmente a soma de quadrados pode ser decomposta (soma de quadrados de resíduos com a soma de quadrados de regressão). Como os $Y_i's$ são independentes, sob a hipótese nula de $\beta_1 = 0$ ser verdade, então, pelo teorema de Cochran, podemos definir que $\frac{SQ_{reg}}{\sigma^2}$ e $\frac{SQ_{res}}{\sigma^2}$, tem distribuição quiquadrado com 1 e $n-2$ graus de liberdade, respectivamente. E, dividindo tais valores, chegamos na estatística F que justamente vai nos informar a veracidade da hipótese nula ser rejeitada ou não. Para sermos mais direto, rejeitar a hipótese nula, isto é, há coeficiente angular, nesse caso, é dizer que a regressão é significativa, até porque sem este parâmetro haveria apenas uma reta constante em β_0. Portanto, a relação linear entre X e Y será significativa se rejeitarmos H_0.
A tabela ANOVA (Análise de Variância) é constituída pelas seguintes quantidades:

Tabela 8.1: ANOVA

Fonte de variação	GL	SQ	QM	F_0
Regressão	1	$\sum_{i=1}^{n}(\widehat{Y_i} - \overline{Y})^2$	$\frac{SQ_{reg}}{1}$	$\frac{QM_{reg}}{QM_{res}}$
Resíduo	n-2	$\sum_{i=1}^{n}(Y_i - \widehat{Y_i})^2$	$\frac{SQ_{res}}{n-2}$	
Total	n-1	$\sum_{i=1}^{n}(Y_i - \overline{Y})^2$		

Tal coeficiente nos informa o quão a variabilidade total é explicada pelo modelo, quanto mais próximo de 1, melhor é o percentual. A medida é denotada por R^2 e é delimitada no intervalo [0,1], sendo definida por:

$$R^2 = \frac{SQ_{reg}}{SQ_{res}} \tag{8.8}$$

O coeficiente de determinação ajustado pelos graus de liberdade é definido por:

$$R^2_{ajust} = 1 - \frac{\frac{SQ_{reg}}{n-2}}{\frac{SQ_{res}}{n-1}} \tag{8.9}$$

Para a SQ_{reg} faremos, primeiramente, um pequeno cálculo para simplificar os passos para essa soma de quadrados.

$$SQ_{reg} = \sum_{i=1}^{n}(\widehat{Y_i} - \overline{Y})^2$$

Mas sabemos que:

$$\begin{aligned} \widehat{Y_i} &= \widehat{\beta_0} + \widehat{\beta_1}X_i \\ \widehat{Y_i} &= \overline{Y} - \widehat{\beta_1}\overline{X} + \widehat{\beta_1}X_i \\ \widehat{Y_i} - \overline{Y} &= = \beta_1(X_i - \overline{X}) \end{aligned}$$

Substituindo,

$$\begin{aligned} SQ_{reg} &= \sum_{i=1}^{n} (\widehat{Y_i} - \overline{Y})^2 \\ &= \sum_{i=1}^{n} \beta_1 (X_i - \overline{X})^2 \\ &= {\beta_1}^2 S_{xx} \end{aligned} \tag{8.10}$$

Feito isso, podemos calcular sua esperança.

$$\begin{aligned} SQ_{reg} &= \widehat{\beta}_1^2 S_{xx} \\ E[SQ_{reg}] &= E[\widehat{\beta}_1^2 S_{xx}] \\ &= S_{xx} E[\widehat{\beta}_1^2] \\ &= S_{xx}[V(\widehat{\beta}_1) + E^2(\widehat{\beta}_1))] \\ &= S_{xx}\left[\frac{\sigma^2}{S_{xx}} + \beta^2\right] \\ &= \sigma^2 + \widehat{\beta}_1^2 S_{xx} \end{aligned} \tag{8.11}$$

Como

$$QM_{reg} = \frac{SQ_{reg}}{1}$$

Então

$$\begin{aligned} SQ_{reg} &= QM_{reg} \\ E[QM_{reg}] &= \sigma^2 + \widehat{\beta}_1^2 S_{xx} \end{aligned} \tag{8.12}$$

Também podemos calcular a esperança para SQ_{res}:

$$SQ_{res} = \sum_{i=1}^{n} [Y_i - \widehat{Y_i}]^2$$

Utilizando o teorema de Cochram, podemos deduzir que:

$$\begin{aligned} \frac{SQ_{res}}{\sigma^2} &\sim \chi^2(n-2) \\ E\left[\frac{SQ_{res}}{\sigma^2}\right] &= n-2 \\ E\left[\frac{SQ_{res}}{n-2}\right] &= \sigma^2 \\ E[QM_{res}] &= \sigma^2 \end{aligned} \tag{8.13}$$

Portanto, QM_{res} é um estimador não viciado para σ^2. Porém, o que calculamos acima não foi uma desmonstração plausível. Sem o auxílio do teorema, devemos fazer:

$$
\begin{aligned}
QM_{res} &= \frac{SQ_{res}}{n-2} \\
E[QM_{res}] &= E\left[\frac{SQ_{res}}{n-2}\right] \\
E[QM_{res}] &= \frac{E[SQ_{res}]}{n-2}
\end{aligned}
$$

Para descobrir $E[SQ_{res}]$, deve-se seguir o raciocínio:

$$
\begin{aligned}
SQ_{total} &= SQ_{res} + SQ_{reg} \\
SQ_{res} &= SQ_{total} - SQ_{reg} \\
E[SQ_{res}] &= E[SQ_{total}] - E[SQ_{reg}]
\end{aligned}
$$

Devemos encontrar $E[SQ_{total}]$ para chegarmos na $E[SQ_{res}]$, pois já calculamos $E[SQ_{reg}]$.

$$
\begin{aligned}
SQ_{total} &= \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2 \\
E[SQ_{total}] &= E\left[\sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2\right] \\
&= E\left[\sum_{i=1}^{n} Y_i^2\right] - E\left[n\overline{Y}^2\right] \\
&= \sum_{i=1}^{n} E[Y_i^2] - nE[\overline{Y}^2] \\
&= \sum_{i=1}^{n} \left[V[Y_i] + E^2[Y_i]\right] - n[V[\overline{Y}] + E^2[\overline{Y}]] \\
&= \sum_{i=1}^{n} \left[\sigma^2 + (\beta_0 + \beta_1 x_i)^2\right] - \left[\frac{\sigma^2}{n} + (\beta_0 + \beta_1 \overline{X})^2\right] \\
&= (n-1)\sigma^2 + {\beta_1}^2\left[\sum_{i=1}^{n} X_i^2 - n\overline{X}\right]
\end{aligned}
$$

E assim:

$$
\begin{aligned}
E[SQ_{res}] &= E[SQ_{total}] - E[SQ_{reg}] \\
&= (n-1)\sigma^2 + {\beta_1}^2\left[\sum_{i=1}^{n} X_i^2 - n\overline{X}\right] - \sigma^2 - {\beta_1}^2 S_{xx} \\
&= (n-2)\sigma^2
\end{aligned} \tag{8.14}
$$

Como já citamos, QM_{res} é um estimador não viciado para a variância:

$$
\begin{aligned}
E[QM_{res}] &= \frac{E[SQ_{res}]}{n-2} \\
&= \sigma^2
\end{aligned}
$$

Podemos agora substituir tal estimativa para encontrar mais alguns resultados importantes como, por exemplo, o erros padrões abaixo:

$$\begin{aligned} V[\widehat{\beta}_1] &= \frac{\sigma^2}{S_{xx}} \\ \widehat{V[\beta_1]} &= \frac{\widehat{\sigma}^2}{S_{xx}} \\ &= \frac{QM_{res}}{S_{xx}} \\ V[\widehat{\beta}_0] &= \frac{\sigma^2 \sum_{i=1}^{n} {X_i}^2}{nS_{xx}} \\ \widehat{V[\widehat{\beta}_0]} &= \frac{\widehat{\sigma}^2 \sum_{i=1}^{n} {X_i}^2}{nS_{xx}} \\ &= \frac{QM_{res} \sum_{i=1}^{n} {X_i}^2}{nS_{xx}} \end{aligned}$$

8.4 Teste de hipóteses

Sabemos que

$$\widehat{\beta}_0 \sim N\left(\beta_0, \frac{\sigma^2 \sum_{i=1}^{n} {X_i}^2}{nS_{xx}}\right)$$

Considerando $H_0 : \beta_0 = \beta_0^*$ verdade:

$$\widehat{\beta}_0 \sim N\left({\beta_0}^*, \frac{\sigma^2 \sum_{i=1}^{n} {X_i}^2}{nS_{xx}}\right)$$

Então:

$$\frac{\widehat{\beta}_0 - {\beta_0}^*}{\sqrt{\frac{\sigma^2 \sum_{i=1}^{n} {X_i}^2}{nS_{xx}}}} \sim N(0,1) \tag{8.15}$$

Como não conhecemos σ^2, vamos recorrer a

$$\begin{aligned} \frac{\frac{\widehat{\beta}_0 - {\beta_0}^*}{\sqrt{\frac{\sigma^2 \sum_{i=1}^{n} {X_i}^2}{nS_{xx}}}}}{\sqrt{\frac{\frac{SQ_{res}}{\sigma^2}}{n-2}}} &\sim t(n-2) \\ \frac{\widehat{\beta_0} - {\beta_0}^*}{\sqrt{\frac{QM_{res} \sum_{i=1}^{n} X_i^2}{nS_{xx}}}} &\sim t(n-2) \end{aligned} \tag{8.16}$$

E, portanto, rejeitaremos H_0 se

$$P[t(n-2) \leq |t_0|] + P[t(n-2) \geq |t_0|] \leq \alpha$$

Para β_1 temos o mesmo raciocínio. Sabemos que

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$$

Considerando $H_0 : \beta_1 = \beta_1^*$ verdade:

$$\widehat{\beta}_1 \sim N\left(\beta_1^*, \frac{\sigma^2}{S_{xx}}\right)$$

Logo

$$\frac{\widehat{\beta}_1 - {\beta_1}^*}{\sqrt{\frac{\sigma^2}{S_{xx}}}} \sim N(0,1)$$

Utilizando novamente o artificio:

$$\begin{aligned} \frac{\frac{\beta_1 - {\beta_1}^*}{\frac{\sigma^2}{S_{xx}}}}{\sqrt{\frac{\frac{SQ_{res}}{\sigma^2}}{n-2}}} &\sim t(n-2) \\ \frac{\widehat{\beta_1} - {\beta_1}^*}{\sqrt{QM_{res} S_{xx}}} &\sim t(n-2) \end{aligned} \tag{8.17}$$

E, portanto, rejeitaremos H_0 se

$$P[t(n-2) \leq |t_0|] + P[t(n-2) \geq |t_0|] \leq \alpha$$

8.5 Intervalos de confiança

Para estabelermos um intervalo de confiança com $(1-\alpha)\%$ de confiança, devemos ter uma quantidade pivotal e uma distribuição, que não depende do parâmetro, para esta quantidade pivotal.

Para β_0

Sabemos que a quantidade pivotal para esse caso é:

$$\frac{\widehat{\beta}_0 - \beta_0}{\sqrt{QM_{res}\frac{\sum_{i=1}^n X_i^2}{nS_{xx}}}} \sim t(n-2)$$

Logo

$$I.C._{1-\alpha}[\beta_0] = \left[\widehat{\beta}_0 \pm t_{(1-\frac{\alpha}{2})}(n-2)\sqrt{QM_{res}\frac{\sum_{i=1}^n X_i^2}{nS_{xx}}}\right] \tag{8.18}$$

Para β_1

Sabemos que a quantidade pivotal para esse caso é:

$$\frac{\widehat{\beta}_0 - \beta_0}{\sqrt{\frac{QM_{res}}{S_{xx}}}} \sim t(n-2)$$

Logo

$$I.C._{1-\alpha}[\beta_1] = \left[\widehat{\beta}_1 \pm t_{(1-\frac{\alpha}{2})}(n-2)\sqrt{\sqrt{\frac{QM_{res}}{S_{xx}}}}\right] \tag{8.19}$$

Digamos que x_0 seja um valor específico da variável preditora. Primeiro, considere estimar a média da população Y associada com x_0. Depois faremos isso para a variância, partindo, em ambos os casos, de:

$$\begin{aligned} \widehat{Y}_i &= \widehat{\beta}_0 + \widehat{\beta}_1 x_0 \\ (\widehat{Y}_i | x = x_0) &= \widehat{\beta}_0 + \widehat{\beta}_1 x_0 \end{aligned}$$

$$\begin{aligned} E[\widehat{Y}_i | x = x_0] &= E[\widehat{\beta}_0 + \widehat{\beta}_1 x_0] \\ &= E[\widehat{\beta}_0] + x_0 E[\widehat{\beta}_1 \\ &= \beta_0 + \beta_1 x_0 \end{aligned}$$

$$\begin{aligned} V[\widehat{Y}_i | x = x_0] &= V[\widehat{\beta}_0 + \widehat{\beta}_1 x_0] \\ &= V[\overline{y} - \widehat{\beta}_1 \overline{x} + \widehat{\beta}_1 x_0] \\ &= V[\overline{y} + \widehat{\beta}_1 (x_0 - \overline{x}) \\ &= V[\overline{y}] + V[\widehat{\beta}_1 (x_0 - \overline{x})] + 2cov(\overline{y}, \widehat{\beta}_1 (x_0 - \overline{x})) \\ &= V[\overline{y}] + (x_0 - \overline{x})^2 V[\widehat{\beta}_1] + 2(x_0 - \overline{x}) cov(\overline{y}, \widehat{\beta}_1) \\ &= \frac{\sigma^2}{n} + (x_0 - \overline{x})^2 \frac{\sigma^2}{S_{xx}} \\ &= \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right] \end{aligned}$$

Então

$$\widehat{Y}_i | x = x_0 \sim N\left(\beta_0 + \beta_1 x_0, \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]\right)$$

$$I.C.[\beta_0 + \beta_1 x_0] = \left[(\widehat{Y}_i | x = x_0) \pm t_{1-\frac{\alpha}{2}}(n-2)\sqrt{QM_{res}\left[\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}}\right]}\right] \tag{8.20}$$

O comprimento do intervalo é mais curto se x_0 estiver mais próximo de $\overline{x}$ e minimizado em $x_0 = \overline{x}$.

Um tipo de inferência a qual não falamos até agora é a previsão de uma variável aleatória, que até o presente não for observada, Y, um tipo de inderência que é de interesse em uma regressão. Assim,

$$\begin{aligned} E[\widehat{Y}_0 - Y_0] &= E[\widehat{Y}_0] - E[Y_0] \\ &= E[\widehat{\beta}_0 + \widehat{\beta}_1 x_0] - E[\beta_0 + \beta_1 x_0 + e_i] \\ &= \beta_0 + \beta_1 x_0 - \beta_0 - \beta_1 x_0 \\ &= 0 \end{aligned}$$

Para a variância devemos saber que $cov(\overline{Y}_0, Y_0)$ é zero, pois como Y_0 não pertence ao conjunto de observações $Y_1, Y_2, ..., Y_n$ utilizadas para estimar os parâmetros, então $\overline{Y}_0$ e Y_0, por suposição, são independentes, ou seja, zero.

$$\begin{aligned} V[\widehat{Y}_0 - Y_0] &= V[\widehat{Y}_0] + V[Y_0] - 2cov(\widehat{Y}_0, Y_0) \\ &= \sigma^2 \left[\frac{1}{n} + \frac{(x_0 - \widehat{x})^2}{S_{xx}} \right] + \sigma^2 \\ &= \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right] \end{aligned}$$

Então

$$\widehat{Y}_0 - Y_0 \sim \left[0, \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right] \right]$$

$$I.C._{(1-\alpha)}[\widehat{Y}_0 - Y_0] = t_{(1-\frac{\alpha}{2})}(n-2) \pm \sqrt{QM_{res} \left[1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{S_{xx}} \right]} \tag{8.21}$$

8.6 Técnicas de diagnóstico

Ajustar um modelo requer várias suposições. A estimação dos parâmetros do modelo requer a suposição de que os erros sejam variáveis aleatórias não correlacionadas com média zero e variância constante. Testes de hipóteses e estimação do intervalo requerem que os erros sejam normalmente distribuídos. Assim, consideramos que a ordem do modelo esteja correta. Porém, o estatístico deve sempre duvidar da validade dessas suposições e conduzir análises para examinar a adequação do modelo que esta testando. A principal análise para isso é o estudo dos resíduos. Os resíduos, como sabemos, são definidos por

$$e_i = y_i - \widehat{y}_i$$

Vamos então aos passos para investigação:

I. Investigação de homocedasticidade - Variância constante

Graficamente $(e_i x X_i)$, se a variância não é constante, teremos comportamentos em que a variância aumenta com o aumento de x; variância diminui com o aumento de x ou variância aumenta e depois diminui com o aumento de x.

II. Investigação de normalidade dos dados

Para investigar a suposição de normalidade devemos comparar os quantis teóricos com os quantis observados. Para isso, devemos ordenar os resíduos de forma crescente e plotar o gráfico que deverá ser uma reta:

$$e(i) \ \text{x} \ \Phi^{-1}\left(\frac{i-1/2}{n}\right)$$

Os testes existentes são o de Sapiro-Wilk e Kolmogorov-Smirnof. Em ambos, a hipótese nula é de normalidade dos dados.

III. Adquacidade dos modelos

Às vezes, observando apenas o gráfico de dispersão, não é possível percebermos que o modelo de regressão linear é adequado. Para melhorar esta forma de comparação, faz-se os gráficos de:

$$X_i \ \text{x} \ e_i$$

$$\widehat{Y}_i \ \text{x} \ e_i$$

Se a dispersão tiver formato de curva ou qualquer outra forma que não se assemelha a uma reta, então o modelo não está adequado. Na seção sobre o uso do R no estudo de regressão, comentaremos mais sobre resíduos e sobre alguns gráficos importante para essa análise.

8.7 Outros modelos lineares simples

Modelos linearizados

Quando aplicamos análise de regressão ao estudo da relação funcional entre duas variáveis, o problema da especificação consiste em determinar a forma matemática da função que será ajustada. Mostraremos agora que existem modelos não-lineares que se transformam em funções lineares por anamorfose, isto é, por substituição dos valores de uma ou mais variáveis por funções destas variáveis. Veja um exemplo:

Exemplo 8.7.1. *Para o modelo abaixo podemos apenas aplicar o logaritmo para termos funções lineares:*

$$\begin{aligned} Y_i &= \beta_0 {x_i}^{\beta_1} e_i \\ ln(Y_i) &= ln(\beta_0 {x_i}^{\beta_1} e_i) \\ ln(Y_i) &= ln(\beta_0) + \beta_1 ln(x_i) + ln(e_i) \end{aligned}$$

Assim:

$$\begin{aligned} Y_i^* &= ln(Y_i) \\ X_i^* &= ln(X_i) \end{aligned}$$

Obs: Se aplicarmos a exponencial no parãmetro, teremos o verdadeiro valor da estimativa, porém esse estimador não tem as mesmas propriedades já ditas até aqui.

•

Transformação de Box-Cox

Realizamos uma transformação na variável com o objetivo de estabilizar a variância e deixar os dados com comportamento normal, ou seja, estaremos adequando o modelo de modo a ficar com homocedasticidade e normalidade quanto aos resíduos. Veja os casos abaixos:

- Quando a variável resposta se refere a contagem (distribuição de Poisson, por exemplo, em que a esperança é proporcional a variância) usaremos a transformação:

$$Y^* = \sqrt{Y}$$

- Quando os dados da variável resposta refere-se a proporção usaremos:

$$Y^* = arcsen\sqrt{Y}$$

- Em outros casos usaremos:

$$Y^* = lnY$$

A proposta para essa transformação é encontrar o valor de uma constate λ para usarmos em:

$$Y^* = Y^\lambda$$

Usando o método de máxima verossimilhança para encontrar tal valor, obtém-se:

$$\begin{cases} \frac{Y^\lambda - 1}{\lambda Y^{\lambda-1}} & , \lambda \neq 0 \\ Y^* lnY & , \lambda = 0 \end{cases}$$

Sendo

$$Y^* = ln^{-1}\left[\frac{1}{n\sum_{i=1}^{n} lnY_i}\right]$$

De maneira geral, utiliza-se

$$\begin{cases} Y^\lambda & , \lambda \neq 0 \\ lnY & , \lambda = 0 \end{cases}$$

Em programas estatísticos, especificamente o R, o comando para fornecer o valor de λ nos diponibiliza o gráfico de sua função de verossimilhança nos informando o intervalo de confiança para tal constante. Desse modo, se o zero pertence ao intervalo, usamos o logaritmo dos dados, mas caso não esteja, usamos os dados elevado ao valor de λ.

Modelo de regressão linear simples passando pela origem

$$Y_i = \beta_1 x_i + e_i \tag{8.22}$$

Usamos esse modelo quando não rejeitamos a hipótese nula $H_0 : \beta_0 = 0$ para o modelo

$$Y_i = \beta_0 + \beta_1 x_i + e_i$$

Atente-se que não usamos esse modelo pelo motivo de termos um par $(0, 0)$ nos dados. Nos só utilizamos quando a hipótese não for rejeitada.

Estimação do parâmetro

$$\begin{aligned}
\sum_{i=1}^{n} {e_i}^2 &= \sum_{i=1}^{n} (Y_i - \beta_1 x_i)^2 \\
\sum_{i=1}^{n} {e_i}^2 &= E \\
\frac{\partial E}{\partial \beta_1} &= \sum_{i=1}^{n} 2((Y_i - \beta_1 x_i))(x_i)(-1)
\end{aligned}$$

Igualando a zero:

$$\begin{aligned}
\sum_{i=1}^{n} (Y_i - \beta_1 x_i)(x_i) &= 0 \\
\sum_{i=1}^{n} x_i y_i - \widehat{\beta_1} \sum_{i=1}^{n} {x_i}^2 &= 0 \\
\widehat{\beta_1} &= \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} {x_i}^2}
\end{aligned} \tag{8.23}$$

Propriedades dos estimadores

$$
\begin{aligned}
\widehat{\beta_1} &= \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} {x_i}^2} \\
E[\widehat{\beta_1}] &= E\left[\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} {x_i}^2}\right] \\
&= \frac{1}{\sum_{i=1}^{n} {x_i}^2} \sum_{i=1}^{n} x_i E[Y_i] \\
&= \frac{1}{\sum_{i=1}^{n} {x_i}^2} \sum_{i=1}^{n} x_i (\beta_1 x_i) \\
&= \frac{1}{\sum_{i=1}^{n} {x_i}^2} \beta_1 \sum_{i=1}^{n} {x_i}^2 \\
&= \beta_1 \\
V[\widehat{\beta_1}] &= V\left[\frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} {x_i}^2}\right] \\
&= \frac{1}{\sum_{i=1}^{n} {x_i}^4} \sum_{i=1}^{n} {x_i}^2 V[Y_i] \\
&= \frac{\sigma^2}{\sum_{i=1}^{n} {x_i}^2}
\end{aligned}
$$

Análise de Variância

Até agora tivemos o mesmo raciocínio em comparação com ao modelo $Y_i = \beta_0 + \beta_1 x_i + e_i$. Porém, a análise de variância para este modelo tem raciocínio diferente. O leitor já deve saber que no modelo com intercepto utilizamos o valor da amostra, Y_i, em comparação com a média amostral, $\overline{Y}$. Isso ocorre, porque se o modelo não for ideal aos dados, todos os pontos estarão na reta $\overline{Y}$, dessa forma, a soma de quadrados total se refere a soma de todos os pontos da amostra com a média com o intuito de análisar a significância da regressão. Como neste modelo não têm-se o intercepto, a soma de quadrados total será a distância dos pontos amostrais em relação ao eixo das abcissas, ou seja, neste modelo teremos:

$$
\begin{aligned}
SQ_{total} &= \sum_{i=1}^{n} (Y_i - 0)^2 \\
&= \sum_{i=1}^{n} {Y_i}^2
\end{aligned}
$$

Desenvolvendo:

$$\begin{aligned} SQ_{total} &= \sum_{i=1}^{n} Y_i^2 \\ &= \sum_{i=1}^{n} [(Y_i - \widehat{Y_i}) + \widehat{Y_i}]^2 \\ &= \sum_{i=1}^{n} (Y_i - \widehat{Y_i})^2 + \sum_{i=1}^{n} \widehat{Y_i}^2 + 2\sum_{i=1}^{n} (\widehat{Y_i} + \widehat{Y_i})\widehat{Y_i} \\ &= \sum_{i=1}^{n} (Y_i - \widehat{Y_i})^2 + \sum_{i=1}^{n} \widehat{Y_i}^2 \\ SQ_{total} &= SQ_{res} + SQ_{reg} \end{aligned}$$

Em que, para este caso

- SQ_{total} tem n graus de liberdade
- SQ_{res} tem $(n-1)$ graus de liberdade
- SQ_{reg} tem 1 grau de liberdade

As condições do teorema de Cochran foram atendidas, então:

$$\begin{aligned} \frac{SQ_{reg}}{\sigma^2} &\sim \chi^2(1) \\ \frac{SQ_{res}}{\sigma^2} &\sim \chi^2(n-1) \end{aligned}$$

Portanto, para o teste de significância do teste, teremos uma distribuição $F(1, n-1)$. Observe a tabela:

Tabela 8.2: ANOVA

Fonte de variação	GL	SQ	QM	F
Regressão	1	$\sum_{i=1}^{n} \widehat{Y_i}^2$	$\frac{SQ_{reg}}{1}$	$\frac{QM_{reg}}{QM_{res}}$
Resíduo	$n-1$	$\sum_{i=1}^{n} (Y_i - \widehat{Y_i})^2$	$\frac{SQ_{res}}{n-1}$	
Total	$n-1$	$\sum_{i=1}^{n} Y_i^2$		

Coeficiente determinação

Neste caso não calculamos R^2 para os dois casos e comparamos. Ao invés disso, comparamos o valor do quadrado médio do resíduo do modelo com intercepto com o modelo sem intercepto. Aquele que tiver menor valor é o modelo que esxplica melhor a variablidade dos dados.

8.8 Exercícios

1. Ajuste um modelo de regressão linear simples, a partir do conjunto de dados disponível em `https://www.ime.usp.br/~giapaula/trees.dat`, com a primeira coluna sendo a variável resposta e a segunda sendo a variável explicativa. Avalia a significância dos parâmetros e interprete os resultados.

2. Ajuste um modelo de regressão linear simples, a partir do conjunto de dados disponível em `https://www.ime.usp.br/~giapaula/imoveis.dat`, com a primeira coluna sendo a variável resposta e a segunda sendo a variável explicativa. Realize análise de diagnóstico com as ferramentas apresentadas nesse capítulo.

3. Ajuste um modelo de regressão linear simples, a partir do conjunto de dados disponível em `https://www.ime.usp.br/~giapaula/vendas.dat`, com a primeira coluna sendo a variável resposta e a segunda sendo a variável explicativa. Interprete os resultados da tabela ANOVA.

CAPÍTULO 9

Modelo de regressão linear múltiplo

9.1 Introdução

Considere, para o prosseguimento desse capítulo, algumas mudanças nas notações. Para o vetor da variável reposta, teremos $\mathbf{Y}$; representando a matriz de covariáveis $\mathbf{X}_{i1}, \mathbf{X}_{i2}, ..., \mathbf{X}_{ik}$, tem-se $\mathbf{X}$. Para o vetor de parâmetros do modelo de regressão linear múltipla, passaremos a usar β, e para o vetor de erros (ε_i), usaremos ε. Para as esperanças e demais cálculos, usaremos sempre vetores.

De acordo com as notações definidas anteriormente, o modelo de regressão passa a ser

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

Na qual o vetor **Y**, de ordem nx1, é dado por

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix}$$

A matriz **X**, de ordem nxp, sendo $p = k + 1$ o número de parâmetros, é dada por

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ 1 & X_{31} & X_{32} & \cdots & X_{3k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{ik} \end{bmatrix}$$

E, finalmente, a matriz de parâmetros, de ordem px1, e a matriz de erros, com ordem nx1, dadas por

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$$

$$\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Em suma, o modelo de regressão linear múltipla é dado por

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + ... + \beta_k X_{1k} + \varepsilon_1 \\ \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + ... + \beta_k X_{2k} + \varepsilon_2 \\ \beta_0 + \beta_1 X_{31} + \beta_2 X_{32} + ... + \beta_k X_{3k} + \varepsilon_3 \\ \vdots \\ \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + ... + \beta_k X_{nk} + \varepsilon_n \end{bmatrix}$$

Passaremos então a escrever as suposições de outra forma, ou seja, sua distribuição será agora uma distribuição n-variada:

$$\varepsilon_i \sim N_n(\mathbf{0}, \sigma^2 I)$$

Sendo $\mathbf{0}$ o vetor de 'zeros' e I a matriz indentidade. A conclusão consequente dessa suposição para os erros é:

$$\begin{aligned} \mathbf{E}[\mathbf{Y}] &= \mathbf{E}[\mathbf{X}\beta + \varepsilon] \\ &= \mathbf{X}\beta + E[\varepsilon] \\ &= \mathbf{X}\beta \\ \mathbf{V}[\mathbf{Y}] &= \mathbf{V}[\mathbf{X}\beta + \varepsilon] \\ &= \mathbf{V}[\varepsilon] \\ &= \sigma^2 \mathbf{I} \end{aligned}$$

Então o vetor $\mathbf{Y}$ tem distribuição

$$\mathbf{Y} \sim N_n(\mathbf{X}\beta, \sigma^2 \mathbf{I})$$

9.2 Estimação dos parâmetros

Da mesma forma como na regressão linear simples, vamos estimar os parâmetros pelo Método dos Mínimos Quadrados. No modelo antes estudado tínhamos que minimizar $\sum_{i=1}^{n} {\varepsilon_i}^2$, agora teremos que minimizar tal valor na forma matricial, isto é

$$\sum_{i=1}^{n} \varepsilon_i = \varepsilon^T \varepsilon$$

Como $\varepsilon_i = (\mathbf{Y} - \mathbf{X}\beta)$, então

$$\begin{aligned}
\varepsilon^T\varepsilon &= (\varepsilon^T\varepsilon)^T(\varepsilon^T\varepsilon) \\
&= \mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}\beta - (\mathbf{X}\beta)^T\mathbf{Y} + (\mathbf{X}\beta)^T(\mathbf{X}\beta) \\
&= \mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}\beta - \beta^T\mathbf{X}^T\mathbf{Y} + \beta^T\mathbf{X}^T\mathbf{X}\beta
\end{aligned}$$

Mas

$$\mathbf{Y}^T\mathbf{X}\beta = \beta^T\mathbf{X}^T\mathbf{Y}$$

Então

$$\varepsilon^T\varepsilon = \mathbf{Y}^T - 2\mathbf{Y}^T\mathbf{X}\beta + \beta^T\mathbf{X}^T\mathbf{X}\beta$$

Essa simplificação nos ajuda na derivação em relação ao vetor de parâmetros, isto é

$$\frac{d(\varepsilon^T\varepsilon)}{d\beta} = \frac{d}{d\beta}[\mathbf{Y}^T - 2\mathbf{Y}^T\mathbf{X}\beta + \beta^T\mathbf{X}^T\mathbf{X}\beta]$$

Nessa derivação de matrizes, precisamos dos seguintes resultados

$$\begin{aligned}
\frac{d\mathbf{a}^T\mathbf{X}}{d\mathbf{X}} &= \mathbf{a} \\
\frac{d\mathbf{X}^T\mathbf{a}\mathbf{X}}{d\mathbf{X}} &= 2\mathbf{a}\mathbf{X}
\end{aligned}$$

Então, respectivamente, temos os valores

$$\begin{aligned}
\frac{d\mathbf{Y}^T\mathbf{X}\beta}{d\beta} &= (\mathbf{Y}^T\mathbf{X}) \\
&= (\mathbf{X}^T\mathbf{Y}) \\
\frac{d\beta^T\mathbf{X}^T\mathbf{X}\beta}{d\beta} &= (\mathbf{Y}^T\mathbf{X}) \\
&= 2(\mathbf{X}^T\mathbf{X})\beta
\end{aligned}$$

Igualando a zero

$$\begin{aligned}
-2\mathbf{X}^T\mathbf{Y} + 2(\mathbf{X}^T\mathbf{X})\widehat{\beta} &= 0 \\
(\mathbf{X}^T\mathbf{X})\widehat{\beta} &= \mathbf{X}^T\mathbf{Y} \\
(\mathbf{X}^T\mathbf{X})^{-1}(\mathbf{X}^T\mathbf{X})\widehat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\
\widehat{\beta} &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}
\end{aligned}$$

Vamos agora estimar o vetor de parâmetros utilizando o Estimador de Máxima Verossimilhança. Como

$$Y \sim N_n(X\beta, \sigma^2 I)$$

A função densidade da Normal Multivariada

$$Y \sim N_n\left(\mu, \sum\right)$$

É dada por

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(2\pi)^{n/2}|\sum|^{1/2}} exp\left[\frac{-1}{2}(\mathbf{Y}-\mu)^T \sum^{-1}(\mathbf{Y}-\mu)\right].$$

Onde $\sum$ é o determinante da matriz de variâncias e covariâncias. A função de verossimilhança é, então, dada por:

$$L(\beta, \sigma^2) = \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}} exp\left[\frac{-1}{2\sigma^2}(\mathbf{Y}-\mathbf{X}\beta)^T(\mathbf{Y}-\mathbf{X}\beta)\right]$$

Aplicando o logaritmo, temos:

$$l(\beta, \sigma^2) = -\frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\sigma^2) - \frac{1}{2}(\mathbf{Y}-\mathbf{X}\beta)^T(\mathbf{Y}-\mathbf{X}\beta)$$

Se derivarmos em relação ao vetor de parâmetros chegaremos a mesma expressão encontrada pelo Método dos Mínimos Quadrados, isto é,

$$\beta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

Porém, vamos utilizar a expressão do logaritmo acima para calcular a estimativa de σ^2:

$$\frac{\partial l(\beta, \sigma^2)}{\partial \sigma} = -\frac{n}{2\sigma^2} + \frac{(\mathbf{Y}-\mathbf{X}\beta)^T(\mathbf{Y}-\mathbf{X}\beta)}{2\sigma^4}$$

Igualando a zero:

$$\begin{aligned} -\frac{n}{2\sigma^2} + \frac{(\mathbf{Y}-\mathbf{X}\beta)^T(\mathbf{Y}-\mathbf{X}\beta)}{2\sigma^4} &= 0 \\ \frac{(\mathbf{Y}-\mathbf{X}\beta)^T(\mathbf{Y}-\mathbf{X}\beta)}{2\sigma^4} &= \frac{n}{2\sigma^2} \\ \sigma^2 &= \frac{(\mathbf{Y}-\mathbf{X}\beta)^T(\mathbf{Y}-\mathbf{X}\beta)}{n}. \end{aligned}$$

Vamos demonstrar agora que o vetor de parâmetros estimados anteriormente é não viciado para os parâmetros. Vamos também calcular o vetor de variâncias.

$$\begin{aligned} E[\widehat{\beta}] &= E[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}] \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{E}[\mathbf{Y}] \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\beta \\ &= \beta \end{aligned}$$

Para a variância, devemos saber um resultado simples

$$\mathbf{V}[\mathbf{AY}] = \mathbf{A}\mathbf{V}[\mathbf{Y}]\mathbf{A}^T$$

Logo, as variâncias temos

$$\begin{aligned} V[\widehat{\beta}] &= \mathbf{V}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}] \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}[\mathbf{Y}][(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]^T \\ &= (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}^T\mathbf{X})^{-1} \end{aligned}$$

A matriz de variâncias e covariâncias fica

$$\mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ 1 & X_{31} & X_{32} & \cdots & X_{3k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{ik} \end{bmatrix}$$

9.3 Análise de Variância

Para o modelo

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + +\beta_k X_{ik} + \varepsilon_i$$

Testaremos as seguintes hipóteses para avaliar a significância do modelo

H_0:$\beta_1 = \beta_2 = ... = \beta_k = 0$

H_1:pelo menos um $\beta_i \neq 0$

Se pelo menos um parâmetro for significativo então o modelo faz sentido. A seguir mostraremos as partes da análise de variância da forma como já estamos acostumados para depois mostrar a forma matricial, dessas partes.

Parte I: Soma de quadrados total

$$\begin{aligned} SQ_{total} &= \sum_{i=1}^{n}(Y_i - \overline{Y})^2 \\ &= \sum_{i=1}^{n} Y_i^2 - n\overline{Y}^2 \end{aligned}$$

Mostrando cada parte matricialmente:

$$\sum_{i=1}^{n} Y_i^2 = Y^T Y$$

Bem como

$$\begin{aligned}\overline{Y} &= \frac{1}{n}\mathbf{1'Y} \\ &= \frac{1}{n}\mathbf{Y}^T\mathbf{1'} \\ \overline{Y}^2 &= \frac{1}{n^2}\mathbf{Y}^T\mathbf{1'1}^T\mathbf{Y}\end{aligned}$$

Então $n\overline{Y}^2$ é dado por

$$\begin{aligned}n\overline{Y}^2 &= \frac{n}{n^2}\mathbf{Y}^T\mathbf{11}^T\mathbf{Y} \\ &= \frac{1}{n}\mathbf{Y}^T\mathbf{11}^T\mathbf{Y}\end{aligned}$$

Assim

$$\begin{aligned}SQ_{total} &= \mathbf{Y}^T\mathbf{Y} - \frac{1}{n}\mathbf{Y}^T\mathbf{11}^T\mathbf{Y} \\ &= \mathbf{Y}^T\left[-\frac{1}{n}\mathbf{11}^T\right]\mathbf{Y}\end{aligned}$$

Parte II: Soma de quadrados dos resíduos

No método linear simples, tínhamos que desenvolver

$$SQ_{res} = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$$

Porém no modelo linear múltiplo, temos que desevolver

$$SQ_{res} = (\mathbf{Y} - \mathbf{X}\widehat{\beta})^T(\mathbf{Y} - \mathbf{X}\widehat{\beta})$$

Assim,

$$\begin{aligned}SQ_{res} &= Y^TY - Y^TX\widehat{\beta} - (X\widehat{\beta})^TY + (X\widehat{\beta})^TX\widehat{\beta} \\ &= Y^TY - 2\widehat{\beta}^TX^TY + \widehat{\beta}^TX^TX\widehat{\beta}.\end{aligned}$$

Sabendo que

$$\widehat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$$

Podemos substituir na expressão da soma de quadrados

$$\begin{aligned}SQ_{res} &= \mathbf{Y}^T\mathbf{Y} - 2[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}]^TX^T\mathbf{Y} + [(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}]^T\mathbf{X}^T\mathbf{X}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}] \\ &= \mathbf{Y}^T\mathbf{Y} - 2\mathbf{Y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} + Y^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= \mathbf{Y}^T\mathbf{Y} - \mathbf{Y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= \mathbf{Y}^T[I - \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]\mathbf{Y}\end{aligned}$$

Observação: a expressão acima encontramos a expressão da matriz H, muito utilizada nos conceitos aprofundados de regressão.

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$$

Parte III: Soma de quadrados de regressão

Como já calculamos duas partes da decomposição da soma de quadrados, faremos agora apenas a subtração:

$$SQ_{reg} = SQ_{total} - SQ_{res}$$

Assim

$$\begin{aligned} SQ_{reg} &= \mathbf{Y}^T\left[-\frac{1}{n}11^T\right]\mathbf{Y} - \mathbf{Y}^T[I-\mathbf{H}]\mathbf{Y} \\ &= \mathbf{Y}^T\mathbf{Y} - \frac{1}{n}\mathbf{Y}^T11^T\mathbf{Y} - \mathbf{Y}^T\mathbf{Y} + \mathbf{Y}^THY \\ &= \mathbf{Y}^T\mathbf{H}\mathbf{Y} - \frac{1}{n}\mathbf{Y}^T11^T\mathbf{Y} \\ &= Y^T\left[\mathbf{H} - \frac{1}{n}11^T\right] \end{aligned}$$

9.4 Teste de hipóteses

Nos testes individuais dos parâmetros, estamos interessados em saber se determinado parâmetro é igual a determinado valor, isto é, estamos interessados em testar as seguintes hipóteses:

$H_0 : \beta_j = \beta_{j_0}$

$H_1 : \beta_j \neq \beta_{j_0}$

Como todo teste, precisamos encontrar a quantidade pivotal. Como os β_{j_0} são funções de variáveis aleatórias com distribuição Normal, então podemos assumir que

$$\beta_{j_0} \sim N(\beta_j, var(\beta_j))$$

Pois, como já demonstramos

$$E[\underline{\widehat{\beta}}] = \underline{\beta}$$

Tendo a distribuição, encontramos a quantidade pivotal

$$\frac{\beta_j - \beta_{j_0}}{\sqrt{var(\widehat{\beta})}} \sim N(0,1)$$

Sabemos que a matriz de variâncias e covariâncias é dada por

$$\mathbf{V}(\underline{\widehat{\beta}}) = \begin{bmatrix} v(\widehat{\beta}_0) & cov(\widehat{\beta}_0,\widehat{\beta}_1) & cov(\widehat{\beta}_0,\widehat{\beta}_2) & \cdots & cov(\widehat{\beta}_0,\widehat{\beta}_k) \\ cov(\widehat{\beta}_0,\widehat{\beta}_1) & var(\widehat{\beta}_1) & cov(\widehat{\beta}_1,\widehat{\beta}_2) & \cdots & cov(\widehat{\beta}_1,\widehat{\beta}_k) \\ cov(\widehat{\beta}_0,\widehat{\beta}_2) & cov(\widehat{\beta}_1,\widehat{\beta}_2) & var(\widehat{\beta}_2) & \cdots & cov(\widehat{\beta}_2,\widehat{\beta}_k) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ cov(\widehat{\beta}_0,\widehat{\beta}_k) & cov(\widehat{\beta}_1,\widehat{\beta}_k) & cov(\widehat{\beta}_2,\widehat{\beta}_k) & \cdots & var(\widehat{\beta}_k) \end{bmatrix}$$

Mas se estamos trabalhando com matrizes, como extrair a variância individual do parâmetro? Simples, usaremos o seguinte artificio

$$V(\underline{\beta}) = \sigma^2 C_{jj}$$

Sendo C_{jj} o elemento de ordem $j+1$ da diagonal da matriz $(\mathbf{X}^T\mathbf{X})^{-1}$ Com isso, podemos chegar em

$$\frac{\beta_j - \beta_{j0}}{\sqrt{\sigma^2 C_{jj}}} \sim N(0,1)$$

Como não temos conhecimento sobre a variância, utilizamos o Teorema de Cochram

$$\frac{SQ_{res}}{\sigma^2} \sim \chi^2$$

Para conseguirmos, finalmente, a quantidade pivotal para o teste

$$\frac{\beta_j - \beta_{j0}}{\sqrt{QM_{res} C_{jj}}} \sim t(n-p)$$

Assim, para um nível de significância α, rejeita-se H_0 se

- Teste bilateral

$$|t_0| > t_{1-\frac{\alpha}{2}}(n-p)$$

- Teste unilateral

$$\begin{aligned} t_0 &< t_\alpha(n-p) \\ t_0 &< t_\alpha(n-p) \end{aligned}$$

Definido da mesma forma que no modelo simples, ou seja,

$$R^2 = \frac{SQ_{reg}}{SQ_{res}}$$

O valor alto do coeficiente de determinação, a medida que aumentamos o número de variáveis, não significa que tais variáveis são signifitativas para o modelo, e sim porque o modelo está 'inchado'. Por esse motivo, usamos o coeficiente de determinação ajustado:

$$R^2 = 1 - \frac{\frac{SQ_{res}}{n-p}}{\frac{SQ_{total}}{n-1}}.$$

9.5 Intervalo de confiança

No modelo de regressão linear simples, o intervalo de confiança para o valor esperado era calculado por meio de um dado valor X_0 e então encontrava-se esperança e variância do valor esperado. Agora, faremos o mesmo, porém em liguagem matricial. Ao invés de estar disponível o valor de X_0, agora teremos o vetor abaixo, já que temos um modelo múltiplo:

$$\mathbf{X}_0 = \begin{bmatrix} 1 & X_{01} & X_{02} & \cdots & X_{0k} \end{bmatrix}^T$$

Então para obter o valor esperado da expressão $E[Y|X_0] = \widehat{\beta}_0 + \widehat{\beta}_1 X_{01} + \widehat{\beta}_2 X_{02} + ... + \widehat{\beta}_k X_{0k}$ na forma de matriz, basta fazermos:

COLOCARR

Ou seja,

$$\widehat{E}[Y|X_0] = {X_0}^T \widehat{\beta}$$

Para a construção do intervalo de confiança precisamos da esperança e da variância dessa estimativa:

$$\begin{aligned} E[{X_0}^T \underline{\widehat{\beta}}] &= {X_0}^T E[\beta] \\ &= {X_0}^T \beta \\ V[{X_0}^T \widehat{\beta}] &= {X_0}^T V(\widehat{\beta}) X_0 \\ &= {X_0}^T (\mathbf{X}^T \mathbf{X})^{-1} \sigma^2 X_0 \end{aligned}$$

Dessa forma a quantidade pivotal fica:

$$\frac{\widehat{E}[Y|X_0] - E[Y|X_0]}{\sqrt{{X_0}^T (X^T X)^{-1} \sigma^2 X_0}} \sim N(0,1)$$

Utilzando novamente o Teorema de Cochran e dividindo pela qui quidrado, chegamos em:

$$\frac{\widehat{E}[Y|X_0] - E[Y|X_0]}{\sqrt{QMres {X_0}^T (X^T X)^{-1} X_0}} \sim t(n-p)$$

Assim, o intervalo fica definido como

$$I.C.[E[Y|X_0]] = \left[{X_0}^T \underline{\widehat{\beta}} \pm t_{1-\frac{\alpha}{2}}(n-p) \sqrt{QMres {X_0}^T (X^T X)^{-1} X_0} \right]$$

Diferentemente do Intervalo de Confiança, tal intervalo representa um valor de Y que não está presente na amostra, por isso denota-se $Y_0|X_0$. Para encontrarmos um intervalo para essa quantidade devemos utilizar o seguinte artifício:

$$D = \widehat{Y}_0|X_0 - Y_0|X_0$$

Então, seguiremos os passos da construção de tal intervalo, ou seja, calcularemos a esperança e a variância de D.

$$\begin{aligned} E[D] &= E[\widehat{Y}_0|X_0 - Y_0|X_0] \\ &= E[X_0^T \underline{\widehat{\beta}} - (X_0^T \underline{\beta} + \varepsilon_i)] \\ &= E[X_0^T \underline{\widehat{\beta}}] - E[X_0^T \underline{\beta}] - E[\varepsilon_i] \\ &= X_0^T \beta - X_0^T \beta \\ &= 0 \\ V[D] &= V[\widehat{Y}_0|X_0 - Y_0|X_0] \\ &= V[\widehat{Y}_0|X_0] + V[Y_0|X_0] - 2Cov[Y_0|X_0, Y_0|X_0] \\ &= V[X_0^T \widehat{\beta}] + V[X_0^T \beta + \varepsilon_i] - 0 \\ &= V[X_0^T \widehat{\beta}] + \sigma^2 \\ &= \sigma^2 [1 + X_0^T (X^T X)^{-1} X_0] \end{aligned}$$

A quantidade pivotal fica, portanto

$$\begin{aligned}
\frac{\widehat{Y}_0|X_0 - Y_0|X_0 - E[\widehat{Y}_0|X_0 - Y_0|X_0]}{\sqrt{Var[\widehat{Y}_0|X_0 - Y_0|X_0 - E[\widehat{Y}_0|X_0 - Y_0|X_0]]}} &\sim N(0,1) \\
\frac{\widehat{Y}_0|X_0 - Y_0|X_0 - 0}{\sqrt{\sigma^2[1 + X_0^T(X^TX)^{-1}X_0]}} &\sim N(0,1) \\
\frac{\widehat{Y}_0|X_0 - Y_0|X_0}{\sqrt{\sigma^2[1 + X_0^T(X^TX)^{-1}X_0]}} &\sim N(0,1)
\end{aligned}$$

Como devemos estimar a variância, utilizaremos o Teorema de Cochran para substituir σ^2, ou seja, ficaremos com

$$\frac{\widehat{Y}_0|X_0 - Y_0|X_0}{\sqrt{QM_{res}[1 + X_0^T(X^TX)^{-1}X_0]}} \sim N(0,1)$$

E o intervalo é dado por:

$$I.C.[Y_0|X_0] = \left[X_0^T\widehat{\beta} \pm t_{1-\frac{\alpha}{2}}(n-p)\sqrt{QM_{res}(1 + X_0^T(X^TX)^{-1}X_0)}\right]$$

9.6 Técnicas de diagnóstico

No modelo clássico, as suposições são adotadas sobre a fonte de variação e a ela associa-se normalidade, homocedasticidade e independência. Após o ajuste do modelo é necessário verificar se essas suposições estão sendo obedecidas ou não (**avaliação do ajuste**), bem como verificar a existência de pontos remotos (*outliers*), pontos influentes e/ou pontos de alavanca (**análise de sensibilidade**).

Nos modelos clássicos, a avaliação do ajuste utiliza a análise de resíduos para validar determinadas suposições, tais como:

i. Homocedasticidade;

ii. Normalidade;

iii. Independência dos erros;

iv. Existência de pontos discrepantes.

Considerando o modelo clássico

$$\mathbf{y} = \mathbf{X}\beta + \epsilon,$$

com $\epsilon \sim \mathcal{N}_n(\mathbf{0}, \sigma^2\mathbf{I})$, cujo estimador de β é dado por $\widehat{\beta} = (\mathbf{X}^\top\mathbf{X}^{-1}\mathbf{X}^\top\mathbf{y})$, então temos que $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\beta} = \mathbf{X}(\mathbf{X}^\top\mathbf{X}^{-1}\mathbf{X}^\top\mathbf{y}) = \mathbf{H}\mathbf{y}$ (a matriz $\mathbf{H}$ é chamada de matriz *hat* ou matriz chapéu). A partir disso, podemos definir três tipos de resíduos: ordinário, estudentizado internamente e estudentizado externamente.

O **resíduo ordinário** é definido por

$$\epsilon = \mathbf{y} - \widehat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y} = (\mathbf{I} - \mathbf{H})\epsilon,$$

logo, $\widehat{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2(\mathbf{I} - \mathbf{H}))$, ou seja, enquanto a fonte de variação é suposta independente e com mesma variância, os resíduos do ajuste, no entanto, apresentam variâncias diferentes, pois sua distribuição depende de σ^2 e da matriz $\mathbf{H}$. Assim, considerar $\epsilon_i = \widehat{\epsilon}_i$ pode não ser adequado devido a essa heterogeneidade.

Uma alternativa a isto, é construir resíduos que não dependam dessa quantidade, pois assim podemos realizar comparações entre os mesmos. Se σ^2 for conhecido, podemos padronizar o resíduo ordinário dividindo-o pelo seu desvio padrão, $\sqrt{\sigma^2(1-h_{ii})}$em que h_{ii} denota o i-ésimo elemento da diagonal principal de $\mathbf{H}$). Dessa forma, a distribuição dos resíduos padronizados não depende mais da variância. Se σ^2 for desconhecido, dividimos por $\sqrt{s^2(1-h_{ii})}$, e chamamos essa quantidade de **resíduo estudentizado internamente**:

$$\widehat{\epsilon}^* = \frac{\widehat{\epsilon}_i}{\sqrt{s^2(1-h_{ii})}} = \frac{\widehat{\epsilon}_i}{\sqrt{\mathsf{QM}_{\mathsf{res}}(1-h_{ii})}}, \qquad i = 1, 2, ..., n.$$

Os resíduos estudentizados internamente resolvem o problema das variâncias distintas, porém um valor discrepante pode alterar profundamente a variância residual. Além disso, tem-se o fato de que o numerador e o denominador do resíduo são variáveis dependentes (Demétrio, 2002).

Para garantir essa independência, define-se o **resíduos estudentizados externamente**:

$$\widehat{\epsilon}^{\mathsf{e}} = \frac{\widehat{\epsilon}_i}{\sqrt{s^2_{(i)}(1-h_{ii})}},$$

sendo $s_{(i)}$ o quadrado médio do resíduo com a ausência da i-ésima observação.

Análise sensibilidade refere-se ao estudo do comportamento do modelo ajustado quando o mesmo sofre algumas perturbações. O objetivo, portanto, é investigar pontos atípicos, sendo estes denominados de pontos remotos (outliers), pontos de alavanca e pontos influentes.

Os **pontos remotos** são observações que não se ajustam bem ao modelo e são detectadas por um afastamento com relação a Y. Esse ponto pode ser de alavanca ou influente.

Os **pontos de alavanca** não afetam o ajuste, são observações extremas de $\mathbf{X}$, a matriz de covariáveis do modelo. A detecção desses pontos é feita observado a diagonal principal da matriz $\mathbf{H}$, assim, se $h_{ii} = 1$, então $\widehat{\mathbf{y}} = \mathbf{y}$, ou seja, a i-ésima observação tem influência total no seu valor predito. O critério de alta alavancagem é dado pelo fato de que $\sum_{i=1}^{n} h_{ii} = p$, p o número de covariáveis do modelo. Assim, a alavancagem média é dada por $\frac{\sum_{i=1}^{n} h_{ii}}{n}$ que é o mesmo que $\frac{p}{n}$. O critério estabelecido para um ponto ter alta alavancagem é se $h_{ii} = \frac{2p}{n}$.

Os **pontos influentes**, ao contrário dos de alavanca, afetam o ajuste do modelo, pois indicam afastamento com relação a $\mathbf{X}$ e a $\mathbf{y}$. Este ponto pode ou não ser um ponto remoto e pode ou não ser um ponto de alavanca. Cook (1977) sugere que a influência de determinada observação é identificada quando o modelo é ajustado com a sua ausência do conjunto de dados. Para a detecção desse ponto utiliza-se a distância de Cook e é uma análise de **influência global**.

Algumas técnicas gráficas para análise de diagnóstico são:

i. Gráfico dos resíduos versus a ordem de coleta dos dados: avaliar a hipótese de independência dos dados.

ii. Gráfico dos resíduos versus valores ajustados: verifica a homoscedasticidade do modelo (espera-se um comportamento aleatório dos resíduos em torno no zero) e linearidade do modelo (espera-se que não apresente tendência);

Além disso, temos:

i. Gráfico dos resíduos estudentizados versus valores ajustados: verifica se existem outliers em Y;

ii. Gráfico dos resíduos padronizados versus valores ajustados: verifica se existem outliers em Y;

iii. Gráfico de alavancagem (Diagonal da Matriz H - *leverage*): verifica se existem *outliers* em X;

iv. Gráfico dos resíduos estudentizados ordenados (observados) versus quantis da normal padrão (teóricos): verifica normalidade (recomenda-se utilizar envelope simulado).

Para a análise formal dos resíduos, podemos realizar os seguintes testes:

i. Testes de Normalidade para os resíduos;

ii. Teste de Durbin-Watson para testar independência dos resíduos;

iii. Teste de Breusch-Pagan e Goldfeld-Quandt para testar se os resíduos são homoscedásticos;

iv. Teste de falta de ajuste para verificar se o modelo ajustado é realmente linear.

9.7 Outros modelos

Especifique um modelo linear heterocedástico e obtenha os estimadores através do Método dos Mínimos Quadrados Generalizados.

Em muitos casos, ao analisarmos os resíduos de um modelo de regressão linear, ao visualizarmos que estes não apresentam a característica de variância constante, temos uma das suposições do modelo não atendidas. Quando isso acontece, dizemos que o modelo apresenta heterocedasticidade nos erros (resíduos), ou ainda que o modelo é heterocedástico. Alguns efeitos causados por essa falha na suposição do modelo são:

Os erros padrões dos estimadores, obtidos pelo Método dos Mínimos Quadrados Ordinários, são incorretos e portanto a inferência estatística não é valida. Não podemos mais dizer que os Estimadores de Mínimos Quadrados Ordinários são os melhores estimadores de variância mínima para β, embora ainda possam ser não viciados.

Por que usar pesos?

Suponhamos que a variância seja não constante, isto é,

$$\text{Var}(Y_i) = \sigma_i^2, \quad \text{para } i = 1, \ldots, n.$$

tomamos, por exemplo, pesos de forma que

$$w_i \propto \frac{1}{\sigma_i^2}, \quad i = 1, \ldots, n.$$

Com isso, as estimativas de Mínimos Quadrados Ponderados (MQP) tem erros padrão menores do que as estimativas de Mínimos Quadrados Ordinários (MQO). Como dito anteriormente, as estimativas de MQO são incorretos, em relação as estimativas de MQP.

A avaliação da hipótese de homoscedasticidade dos resíduos, é feita através das estatísticas de Cochran, Brown-Forsyte (Levene), Breusch-Pagan e Goldfeld-Quandt.

Neste momento, consideramos o modelo de regressão linear simples e vamos denotar por σ_i^2 a variância relacionada ao i-ésimo erro ε_i,, A suposição do modelo é que $\varepsilon_i \sim N(0, \sigma_i)$ independentes. Observe que estamos considerando que a variância σ_i^2 depende da i-ésima observação, podendo ser não constante ao longo das observações. O modelo descrito é da forma:

$$Y_i = \beta_{w0} + \beta_{w1} X_i + \varepsilon_i, \quad \text{para } i = 1, \ldots, n,$$

em que,

- Y_i é a i-ésima observação da variável resposta;
- X_i é a i-ésima observação da covariável constante e conhecida;
- β_{w0} e β_{w1} são os parâmetros desconhecidos da regressão;
- ε_i é o i-ésimo erro, consideramos $\varepsilon_i \sim N(0, \sigma_i^2)$ para $i = 1, 2, \ldots, n$ e n é o número de observações.

Podemos obter os estimadores dos coeficientes da regressão considerando o método de máxima verossimilhança ou pelo método dos mínimos quadrados. A seguir, descrevemos a estimação pelo método de máxima verossimilhança. Para isto, substituímos σ^2 por σ_i^2 devidamente e obtemos a expressão:

$$L(\beta_{w0}; \beta_{w1} | y, x) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left\{ -\frac{\left(Y_i - (\beta_{w0} + \beta_{w1} X_i)\right)^2}{2\sigma_i^2} \right\} \quad .$$

Supomos o peso w_i, inversamente proporcional a variância σ^2, sendo:

$$w_i = \frac{1}{\sigma_i^2} \quad .$$

e então, obtemos a função verossimilhança da seguinte forma:

$$L(\beta_{w1}; \beta_2 | y, x) = \prod_{i=1}^{n} \frac{\sqrt{w_i}}{\sqrt{2\pi}} \prod_{i=1}^{n} \exp\left\{ -\frac{1}{2} w_i \left(Y_i - (\beta_{w0} + \beta_{w1} X_i)\right)^2 \right\}$$

$$= \prod_{i=1}^{n} \left(\frac{w_i}{2\pi} \right)^{1/2} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} w_i \left(Y_i - (\beta_{w0} + \beta_{w1} X_i)\right)^2 \right\} \quad .$$

Obtemos os estimadores dos coeficientes da regressão maximizando em relação a β_{w0} e β_{w1}. Porém, podemos perceber que a função de verossimilhança é inversamente proporcional ao termo exponencial, portanto, maximizar equivale a minimizar o termo:

$$Q_w = \sum_{i=1}^{n} \varepsilon_{w_i}^2 = \sum_{i=1}^{n} w_i \left(Y_i - (\beta_{w0} + \beta_{w1} X_i)\right)^2 \quad .$$

que é soma dos desvios ponderados do método dos mínimos quadrados ponderados.

Os estimadores $\widehat{\beta}_{w0}$ e $\widehat{\beta}_{w1}$ são conhecidos como estimadores de mínimos quadrados ponderados. Notamos que esses estimadores, coincidem com os estimadores de mínimos quadrados ordinários quando consideramos a suposição de homocedasticidade, que implica em pesos (w_i) iguais.

As observações de maior variância têm menos influência sobre os estimadores de β_{w0} e β_{w1}, e as de menor variância têm mais influência. Isso é devido ao fato de que as observações de menor variância apresentam informações mais pertinentes a respeito da $\mathbb{E}[Y|X_i], \quad i = 1, \ldots, n$.

Calculamos os estimadores de mínimos quadrados ponderados derivando Q_w em relação aos parâmetros e igualando a zero para obter o ponto de mínimo, ou seja:

$$\frac{\partial Q_w}{\partial \beta_{w0}} = 2\sum_{i=1}^{n} w_i \left(Y_i - (\beta_{w0} + \beta_{w1} X_i)\right) = 2\sum_{i=1}^{n} w_i Y_i - 2\beta_{w0}\sum_{i=1}^{n} w_i - 2\beta_{w1}\sum_{i=1}^{n} w_i X_i = 0$$

$$\frac{\partial Q_w}{\partial \beta_{w1}} = 2\sum_{i=1}^{n} w_i \left(Y_i - (\beta_{w0} + \beta_{w1} X_i)\right) X_i = 2\sum_{i=1}^{n} w_i Y_i X_i - 2\beta_{w0}\sum_{i=1}^{n} w_i X_i - 2\beta_{w1}\sum_{i=1}^{n} w_i {X_i}^2 = 0$$

Desta forma, obtemos o sistema:

$$\begin{cases} \displaystyle\sum_{i=1}^{n} w_i Y_i = \beta_{w0}\sum_{i=1}^{n} w_i + \beta_{w_{w1}}\sum_{i=1}^{n} w_i X_i \\ \displaystyle\sum_{i=1}^{n} w_i Y_i X_i = \beta_{w0}\sum_{i=1}^{n} w_i X_i + \beta_{w1}\sum_{i=1}^{n} w_i {X_i}^2 \end{cases}$$

Com isso, a solução das equações são dadas por:

$$\beta_{w0} = \frac{\displaystyle\sum_{i=1}^{n} w_i Y_i - \beta_{w1}\sum_{i=1}^{n} w_i X_i}{\displaystyle\sum_{i=1}^{n} w_i} \quad \text{e} \quad \beta_{w1} = \frac{\displaystyle\sum_{i=1}^{n} w_i Y_i X_i - \frac{\displaystyle\sum_{i=1}^{n} w_i Y_i \sum_{i=1}^{n} w_i X_i}{\displaystyle\sum_{i=1}^{n} w_i}}{\displaystyle\sum_{i=1}^{n} w_i {X_i}^2 - \frac{\displaystyle\left(\sum_{i=1}^{n} w_i X_i\right)^2}{\displaystyle\sum_{i=1}^{n} w_i}}$$

Para facilitar a notação, denotamos $\overline{Y}_w = \dfrac{\displaystyle\sum_{i=1}^{n} w_i Y_i}{\displaystyle\sum_{i=1}^{n} w_i}$ e $\overline{X}_w = \dfrac{\displaystyle\sum_{i=1}^{n} w_i X_i}{\displaystyle\sum_{i=1}^{n} wi}$ as médias ponderadas de Y e X, respectivamente. Afim de facilitar os cálculos, vamos reescrever o estimador de mínimos quadrados ponderados de β_{w1} da seguinte maneira:

$$\widehat{\beta}_{w1} = \frac{\displaystyle\sum_{i=1}^{n} w_i (X_i - \overline{X}_w)(Y_i - \overline{Y}_w)}{\displaystyle\sum_{i=1}^{n} w_i (X_i - \overline{X}_w)^2} =$$

$$= \frac{\displaystyle\sum_{i=1}^{n} w_i X_i Y_i - \sum_{i=1}^{n} w_i X_i \overline{Y}_w - \sum_{i=1}^{n} w_i \overline{X}_w Y_i + \sum_{i=1}^{n} w_i \overline{X}_w \overline{Y}_w}{\displaystyle\sum_{i=1}^{n} w_i (X_i - \overline{X}_w)^2} =$$

$$= \frac{\sum_{i=1}^{n} w_i Y_i X_i - \sum_{i=1}^{n} w_i X_i \left(\frac{\sum_{i=1}^{n} w_i Y_i}{\sum_{i=1}^{n} w_i} \right) - \sum_{i=1}^{n} w_i \overline{X}_w Y_i + \sum_{i=1}^{n} w_i \overline{X}_w \left(\frac{\sum_{i=1}^{n} w_i Y_i}{\sum_{i=1}^{n} w_i} \right)}{\sum_{i=1}^{n} w_i (X_i - \overline{X}_w)^2} =$$

$$= \frac{\sum_{i=1}^{n} w_i Y_i X_i - \frac{\sum_{i=1}^{n} w_i Y_i \sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i} - \sum_{i=1}^{n} w_i \overline{X}_w Y_i + \frac{\sum_{i=1}^{n} w_i \sum_{i=1}^{n} \overline{X}_w w_i Y_i}{\sum_{i=1}^{n} w_i}}{\sum_{i=1}^{n} w_i (X_i - \overline{X}_w)^2} =$$

$$= \frac{\sum_{i=1}^{n} w_i Y_i X_i - \sum_{i=1}^{n} w_i Y_i \overline{X}_w - \sum_{i=1}^{n} w_i Y_i \overline{X}_w + \sum_{i=1}^{n} w_i Y_i \overline{X}_w}{\sum_{i=1}^{n} w_i (X_i - \overline{X}_w)^2}$$

$$= \frac{\sum_{i=1}^{n} w_i Y_i X_i - \sum_{i=1}^{n} w_i Y_i \overline{X}_w}{\sum_{i=1}^{n} w_i (X_i - \overline{X}_w)^2} = \frac{\sum_{i=1}^{n} w_i (X_i - \overline{X}_w) Y_i}{\sum_{i=1}^{n} w_i (X_i - \overline{X}_w)^2} \quad .$$

Logo, os estimadores de mínimos quadrados ponderados são dadas por:

$$\widehat{\beta}_{w0} = \overline{Y}_w - \widehat{\beta}_{w1} \overline{X}_w \quad \text{e} \quad \widehat{\beta}_{w1} = \frac{\sum_{i=1}^{n} w_i (X_i - \overline{X}_w) Y_i}{\sum_{i=1}^{n} w_i (X_i - \overline{X}_w)^2}$$

Os valores de $\widehat{\beta}_{w0}$ e$\widehat{\beta}_{w0}$ obtidos são denominados Estimadores de Mínimos Quadrados Ponderados (EMQP).

O modelo de regressão linear simples ponderado ajustado é dado por

$$\widehat{Y}_i = \widehat{\beta}_{w0} + \widehat{\beta}_{w1} X_i \quad i = 1, \ldots, n$$

em que $\widehat{Y}$ é um estimador pontual da média da variável Y para um valor de x, ou seja,

$$\widehat{\mathbb{E}(Y|X_i)} = \widehat{\beta}_{w0} + \widehat{\beta}_{w1} X_i, \quad i = 1, \ldots, n$$

9.8 Exercícios

1. Ajuste um modelo de regressão linear simples, a partir do conjunto de dados disponível em `https://www.ime.usp.br/~giapaula/trees.dat`. Avalie a significância dos parâmetros e interprete os resultados.

2. Ajuste um modelo de regressão linear simples, a partir do conjunto de dados disponível em `https://www.ime.usp.br/~giapaula/imoveis.dat`. Realize análise de diagnóstico com as ferramentas apresentadas nesse capítulo.

3. Ajuste um modelo de regressão linear simples, a partir do conjunto de dados disponível em `https://www.ime.usp.br/~giapaula/vendas.dat`. Interprete os resultados da tabela ANOVA.

Referências Bibliográficas

Bolfarine, H. and Sandoval, M. C. (2001). *Introdução à inferência estatística*, volume 2. SBM.

Magalhães, M. N. (2006). *Probabilidade e variáveis aleatórias*. Edusp.

Meyer, P. L. (1983). Probabilidade: Aplicações à estatística.(2ª edição). *Livros Técnicos e Científicos Editora SA*.

Mirshawka, V. (1983). *Probabilidades e estatística para engenharia*. Nobel.

Morettin, P. A. and Bussab, W. O. (2017). *Estatística básica*. Saraiva Educação SA.