1. ESTIMAÇÃO

Um **estimador**, $\hat{\theta}$, do parâmetro θ é uma função qualquer dos elementos da amostra. **Estimativa** é o valor numérico assumido pelo estimador quando os valores observados são considerados. Os **parâmetros** são funções de valores populacionais, enquanto **estatísticas** são funções dos dados amostrais.

Propriedades dos estimadores

1) Não tendenciosidade

Um estimador $\hat{\theta}$ do parâmetro θ é dito um estimador não tendencioso se $E\left(\hat{\theta}\right)=\theta$.

Exemplo: $\bar{X}=\sum_{i}^{n} X_i \Big/ n$ é um estimador não tendencioso da média populacional μ.

Prova

$$E\left(\bar{X}\right)=E\left(\sum_{i}^{n} X_i \Big/ n\right)=\frac{1}{n}E\left(\sum_{i}^{n} X_i\right)=\frac{1}{n}E\left(X_1+X_2+\cdots+X_n\right)=$$

$$=\frac{1}{n}\left(E(X_1)+E(X_2)+\cdots+E(X_n)\right)=\frac{1}{n}\left(\mu+\mu+\cdots+\mu\right)=\frac{1}{n}\left(n\mu\right)=\mu$$

2) Consistência

Um estimador $\hat{\theta}$ é dito um estimador consistente do parâmetro θ se:

i) $\lim_{n\to\infty} E\left(\hat{\theta}\right)=\theta$;

ii) $\lim_{n\to\infty} V\left(\hat{\theta}\right)=0$.

Exemplo: $\bar{X}=\sum_{i}^{n} X_i \Big/ n$

i) $\lim_{n\to\infty} E\left(\bar{X}\right)=\lim_{n\to\infty}\mu=\mu$;

ii) $\lim_{n\to\infty} V\left(\bar{X}\right)=\lim_{n\to\infty}\frac{\sigma^2}{n}=0$.

3) Eficiência

Se $\hat{\theta}_1$ e $\hat{\theta}_2$ são dois estimadores não tendenciosos de θ, então, $\hat{\theta}_1$ é mais eficiente que $\hat{\theta}_2$ se:

$$V\left(\hat{\theta}_1\right) < \mathrm{V}\left(\hat{\theta}_2\right).$$

4) Suficiência ou Precisão

Um estimador é suficiente se contém o máximo de informação com relação ao parâmetro por ele estimado.

$$\text{Quantidade de informação ou precisão}=\frac{1}{V\left(\hat{\theta}\right)}$$

MÉTODOS DE ESTIMAÇÃO

1. <u>Estimadores de momentos</u> (Método dos Momentos)

Considere uma variável aleatória (v.a.) X com densidade $f(x;\theta_1,\theta_2,\ldots,\theta_r)$, a qual depende de r parâmetros. O primeiro momento populacional é definido como $\mu_1 = E(X)$. O k-ésimo momento populacional é $\mu_k = E(X^k)$.

No caso de uma distribuição normal, $X \sim N(\mu,\sigma^2)$, temos k = 2, $\theta_1 = \mu$ e $\theta_2 = \sigma^2$. O primeiro momento é $\mu_1 = E(X)$. O segundo momento é $\mu_2 = E(X^2)$.

Sabemos que $Var(X) = E(X^2) - (E(X))^2$. Portanto o segundo momento populacional é $E(X^2) = Var(X) + (E(X))^2 = \sigma^2 + \mu^2$

Suponha que seja retirada uma amostra $x_1, x_2, \cdots, x_n$ dessa população. Definem-se então os momentos amostrais como sendo, o k-ésimo momento $m_k = \frac{1}{n}\sum_{i=1}^{n} x_i^k$. Tem-se, portanto, que o primeiro momento amostral corresponde à média amostral, $m_1 = \overline{x}$.

Definição: Os estimadores obtidos pelo método dos momentos são soluções da equação: $m_k = \mu_k$. O procedimento de obter os estimadores consiste em substituir os momentos teóricos pelos respectivos momentos amostrais.

Exemplo. Considere uma v.a. X com distribuição normal e parâmetros μ e σ^2. Encontre os estimadores de momentos para os dois parâmetros.

2. <u>Estimadores de mínimos quadrados</u>

O método dos mínimos quadrados consiste em encontrar o estimador para o parâmetro, que minimize a soma dos quadrados dos resíduos do modelo em questão. Isso é feito derivando a equação do resíduo em relação a cada um dos parâmetros, igualando essas derivadas a zero e resolvendo as equações resultantes.

Como exemplo, considere o modelo estatístico de uma regressão linear simples:

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

Os resíduos e a soma de quadrados de resíduos são obtidos do modelo de regressão:

$\varepsilon_i = y_i - \beta_0 - \beta_1 x_i$; $\qquad Q = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$

O método dos mínimos quadrados consiste em minimizar a soma de quadrados do erro ou resíduo do modelo ao longo de todos os *n* pares (x_i, y_i). Para isto, derivamos a expressão (Q) em relação aos parâmetros β_0 e β_1 obtendo-se assim o Sistema de Equações Normais (SEN):

$$(SEN)\begin{cases}\dfrac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) \\ \dfrac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)x_i\end{cases}$$

Igualando essas derivadas a zero e substituindo β_0 e β_1 pelos respectivos estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ tem-se:

$$\begin{cases}-2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0 \\ -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)x_i = 0\end{cases} \Leftrightarrow \begin{cases}\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right) = 0 \quad (\mathrm{A}) \\ \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i\right)x_i = 0 \quad (\mathrm{B})\end{cases} \Leftrightarrow \begin{cases}\sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1\sum_{i=1}^{n} x_i = 0 \\ \sum_{i=1}^{n} x_i y_i - \hat{\beta}_0\sum_{i=1}^{n} x_i - \hat{\beta}_1\sum_{i=1}^{n} x_i^2 = 0\end{cases}$$

Resolvendo esse sistema para os parâmetros, tem-se os estimadores de mínimos quadrados para β_0 e β_1, respectivamente:

$$\hat{\beta}_0 = b_0 = \overline{y} - \hat{\beta}_1\overline{x} \qquad \text{e} \qquad \hat{\beta}_1 = b_1 = \frac{SPxy}{Sxx}$$

3. Estimadores de Máxima Verossimilhança

Esse princípio baseia-se em escolher o estimador que maximiza a probabilidade de se obter a amostra particular observada.

Considere uma v.a. X com densidade $f(x;\theta)$, sendo θ o parâmetro desconhecido, ou o vetor de parâmetros. Retirando-se dessa população uma amostra de tamanho n $(x = x_1, x_2, \cdots, x_n)$, define-se a função de verossimilhança dos dados por,

$$L(\theta; x_1, x_2, \ldots, x_n) = L(\theta; x) = L(\theta \mid x) = f(x_1;\theta)\cdot f(x_2;\theta)\cdot \ldots \cdot f(x_n;\theta)$$

O procedimento para obter os estimadores de máxima verossimilhança consiste em obter a função de verossimilhança e maximizar essa função $L(\theta;x)$, ou o seu logaritmo $\log(L(\theta;x)) = l(\theta;x)$, que geralmente é mais conveniente.

Exemplo. Considere uma v.a. X com distribuição exponencial, com parâmetro β. Encontre o estimador de máxima verossimilhança para o parâmetro dessa distribuição.

2. MODELO DE REGRESSÃO LINEAR MÚLTIPLA

A análise de regressão múltipla é o estudo de como a variável dependente *y* se relaciona com duas ou mais variáveis independentes. Portanto, temos uma regressão linear múltipla quando admitimos que o valor da variável dependente é função linear de duas ou mais variáveis independentes. O modelo estatístico de uma regressão múltipla com k variáveis independentes é:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i$$

$$\text{ou } y_i = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \varepsilon_i$$

$$\text{ou } y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon$$

em que, *y* é a variável dependente; x_1, x_2, ..., x_k são as variáveis independentes; β_0, β_1, β_2, ... , β_k são os parâmetros da regressão; ε é o termo que representa o resíduo ou erro da regressão, ou seja, é um termo de erro aleatório.

O termo β_0 é denominado intercepto, ou coeficiente linear, e representa o valor da intersecção com o eixo *y*. Os termos β_1, β_2, ... , β_k são chamados de coeficientes angulares. O termo de erro é responsável pela variabilidade em *y* que não pode ser explicada pelo efeito linear das variáveis independentes (*k* variáveis).

No ajuste do modelo de regressão múltipla, é conveniente expressar as operações matemáticas por meio de notação matricial. Utilizando a notação matricial o modelo fica

$$Y = X\beta + \varepsilon$$

em que,

$$\underset{n\,x\,1}{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \underset{n\,x\,p}{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{k1} \\ 1 & x_{12} & \cdots & x_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \end{bmatrix}, \quad \underset{p\,x\,1}{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} \text{ e } \underset{n\,x\,1}{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

PRESSUPOSIÇÕES SOBRE O MODELO DE REGRESSÃO LINEAR MÚLTIPLA

As pressuposições sobre o modelo de regressão múltipla são:

i) O erro ε é uma variável aleatória com média, ou valor esperado, igual a zero, ou seja, $E(\varepsilon) = 0$.

Implicação: Para dados valores de x_1, x_2, ..., x_k, o valor esperado, ou média, de *y* é dado por: $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$.

ii) A variância do erro é constante e igual a σ^2 para todos os valores das variáveis independentes x_1, x_2, ..., x_k (condição de homogeneidade de variância).

iii) Os erros são independentes, isto é, o erro de um conjunto de valores em particular das variáveis independentes não está relacionado com o erro de qualquer outro conjunto de valores ($Cov(\varepsilon_i, \varepsilon_j) = 0$ para todo $i \neq j$).

iv) Os resíduos ou erros têm distribuição normal com média zero e variância constante ($I\sigma^2$).

EQUAÇÃO DE REGRESSÃO MÚLTIPLA ESTIMADA

Se os valores de β_0, β_1, β_2, ... , β_k fossem conhecidos, a equação de regressão poderia ser usada para calcular o valor médio de *y* em relação a valores de x_1, x_2, ..., x_k. Mas, geralmente

esses parâmetros são desconhecidos, e devem ser estimados a partir de dados amostrais. A estimação dos parâmetros pode ser obtida por meio do Método de Mínimos Quadrados. Esse método consiste em minimizar a soma de quadrados dos resíduos. O estimador de Mínimos quadrados de β é:

$$\hat{\beta} = (X'X)^{-1} X'Y$$

em que

$$\underset{n\,x\,1}{Y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \underset{n\,x\,p}{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{k1} \\ 1 & x_{12} & \cdots & x_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \cdots & x_{kn} \end{bmatrix}, \quad \underset{p\,x\,1}{X'Y} = \begin{bmatrix} \sum y_i \\ \sum x_{1i} y_i \\ \vdots \\ \sum x_{ki} y_i \end{bmatrix},$$

$$\underset{n\,x\,n}{X'X} = \begin{bmatrix} n & \sum x_{1i} & \cdots & \sum x_{ki} \\ \sum x_{1i} & \sum x_{1i}^2 & \cdots & \sum x_{1i} x_{ki} \\ \sum x_{2i} & \sum x_{1i} x_{2i} & \cdots & \sum x_{2i} x_{ki} \\ \vdots & \vdots & \ddots & \vdots \\ \sum x_{ki} & \sum x_{1i} x_{ki} & \cdots & \sum x_{ki}^2 \end{bmatrix} \text{ e } \underset{p\,x\,1}{\hat{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}.$$

Nota-se que $\hat{\beta}$ é o vetor das estimativas dos parâmetros do modelo de regressão múltipla. Portanto, a equação de regressão múltipla estimada é:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k .$$

INTERPRETAÇÃO DOS COEFICIENTES DA REGRESSÃO MÚLTIPLA

Na regressão linear simples, interpreta-se $\hat{\beta}_i$ como uma estimativa da alteração em y correspondente à alteração de uma unidade na variável independente. Na regressão linear múltipla, a interpreta-se $\hat{\beta}_i$ como uma estimativa da alteração em y correspondente à alteração de uma unidade na variável x_i quando as outras variáveis independentes se mantêm constantes.

Exemplo: Para ilustrar os conceitos, utilizaremos um estudo realizado em uma academia de ginástica, durante 15 meses, cujo propósito foi identificar quais fatores afetavam seus gastos gerais. Foram identificados dois fatores potenciais: gastos com energia expressos em kilowatts consumidos e gastos com pessoal expressos em horas de Mão-de-Obra Direta (MOD). O resultado encontrado na pesquisa está apresentado na Tabela 1.

Tabela 1 – Resultados da pesquisa sobre os gastos da academia.

Gastos reais da Academia (y)	Consumo Kilowatts ($x1$)	Horas de MOD ($x2$)
350	6	10
400	8	14
470	12	16
550	10	26
620	15	24
380	7	12
290	6	13
490	9	21
580	11	20
610	13	24
560	12	23
420	14	12
450	11	19
510	12	19
380	9	11

Pretende-se com essas informações ajustar um modelo de regressão linear múltipla. Usando os dados fornecidos, obtemos:

$$X = \begin{bmatrix} 1 & 6 & 10 \\ 1 & 8 & 14 \\ 1 & 12 & 16 \\ 1 & 10 & 26 \\ 1 & 15 & 24 \\ 1 & 7 & 12 \\ 1 & 6 & 13 \\ 1 & 9 & 21 \\ 1 & 11 & 20 \\ 1 & 13 & 24 \\ 1 & 12 & 23 \\ 1 & 14 & 12 \\ 1 & 11 & 19 \\ 1 & 12 & 19 \\ 1 & 9 & 11 \end{bmatrix} \quad \text{e} \quad Y = \begin{bmatrix} 350 \\ 400 \\ 470 \\ 550 \\ 620 \\ 380 \\ 290 \\ 490 \\ 580 \\ 610 \\ 560 \\ 420 \\ 450 \\ 510 \\ 380 \end{bmatrix}$$

A partir dessas matrizes conseguimos definir as seguintes matrizes:

$$X'X = \begin{bmatrix} 15 & 155 & 264 \\ 155 & 1711 & 2847 \\ 264 & 2847 & 5050 \end{bmatrix} \text{ e } X'Y = \begin{bmatrix} 7060 \\ 75950 \\ 131000 \end{bmatrix}.$$

Usando as regras de inversão de matrizes determinou-se a matriz $(X'X)^{-1}$:

$$(X'X)^{-1} = \begin{bmatrix} 1{,}19055 & -0{,}06928 & -0{,}02318 \\ -0{,}06928 & 0{,}01347 & -0{,}00397 \\ -0{,}02318 & -0{,}00397 & 0{,}00365 \end{bmatrix}. \text{ Logo, } \hat{\beta} = (X'X)^{-1} X'Y = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 106{,}7247 \\ 13{,}5794 \\ 12{,}7058 \end{bmatrix}.$$

Portanto, o modelo de regressão linear múltipla estimado foi $\hat{y} = 106{,}7247 + 13{,}5794x_1 + 12{,}7058x_2$.

Interpretação:

O modelo de regressão linear múltipla estimado foi $\hat{y} = 106{,}7247 + 13{,}5794x_1 + 12{,}7058x_2$, então, $\hat{\beta}_1 \cong 13{,}58$ e $\hat{\beta}_2 \cong 12{,}71$. Assim, $\hat{\beta}_1 \cong 13{,}58$, é a estimativa do aumento esperado no gasto mensal da academia correspondente ao aumento de 1 (um) kilowatt no consumo de energia quando os gastos com pessoal expressos em horas de mão-de-obra direta é mantido constante.

Da mesma forma, $\hat{\beta}_2 \cong 12{,}71$ é a estimativa do aumento esperado no gasto mensal da academia correspondente ao aumento de uma hora de mão-de-obra direta quando o consumo de energia (kilowatts) é mantido constante.

TESTE DE SIGNIFICÂNCIA DA REGRESSÃO MÚLTIPLA
ANÁLISE DE VARIANCIA – ANAVA

O teste de hipótese para a significância da regressão é um teste para determinar se existe uma relação linear entre a variável de resposta y e o conjunto de regressores x_1, x_2, ..., x_k. As hipóteses apropriadas são:

$$H_0: \beta_1 = \beta_2 = ... = \beta_k = 0$$
$$H_1: \beta j \neq 0 \text{ para no mínimo um } j.$$

Se H_0 for rejeitada, o teste nos dá suficientes evidências estatísticas para concluirmos que um ou mais dos parâmetros não são iguais a zero e que a relação global entre y e o conjunto de variáveis independentes x_1, x_2, ..., x_k é significativa. Entretanto, se H_0 não puder ser rejeitada, não teremos evidências suficientes para concluir que uma relação significativa está presente.

Portanto, utiliza-se o teste F para concluir se existe ou não uma relação global significativa entre y e o conjunto de variáveis independentes x_1, x_2, ..., x_k . Uma análise de variância geral fornece os resultados do teste F de um modelo de regressão múltipla. Na Tabela 2 é apresentado o procedimento para a ANAVA e consequentemente do teste F, sendo que p é o número de parâmetros do modelo e n é o número de observações da variável resposta (variável dependente).

Tabela 2 – Análise de Variância de um modelo de regressão linear múltipla com *k* variáveis independentes.

FV	***GL***	***SQ***	***QM***	***F***
Regressão	*p - 1*	*SQReg*	*SQReg/(p - 1)*	*QMReg/QMErro*
Erro	*n - p*	*SQErro*	*SQErro/(n - p)*	
Total	*n - 1*	*SQTotal*		

Também neste caso, existe uma relação entre a *SQTotal*, *SQReg* e *SQErro*:

SQTotal = SQReg + SQErro

em que, $SQTotal = Y'Y - \frac{\left(\sum y_i\right)^2}{n}$; $SQ\,\mathrm{Re}\,g = \hat{\beta}'X'Y - \frac{\left(\sum y_i\right)^2}{n}$ e $SQErro = Y'Y - \hat{\beta}'X'Y$.

A regra de decisão é:
Critério do valor p: Rejeita-se H_0 se o valor p ≤ α , em que α é o nível de significância.
Critério do valor crítico: Rejeita-se H_0 se $F \geq F_\alpha$ em que F_α se baseia em uma distribuição F com p – 1 graus de liberdade no numerador e n – p graus de liberdade no denominador.

Exemplo: O procedimento de construção da ANAVA é ilustrado para o exemplo da academia. O modelo de regressão linear múltipla estimado foi $\hat{y} = 106,7247 + 13,5794x_1 + 12,7058x_2$, ou seja, o modelo possui 3 parâmetros (p = 3). Como a coleta dos dados foi realizada durante 15 meses segue-se que n é igual a 15 (quinze) observações.

Soma de Quadrados Total:

$$SQTotal = Y'Y - \frac{\left(\sum y_i\right)^2}{n} = 3.464.400 - \frac{(7060)^2}{15} = 141.493,3333.$$

Soma de Quadrados de Regressão:

$$SQ\,\mathrm{Re}\,g = \hat{\beta}'X'Y - \frac{\left(\sum y_i\right)^2}{n} = 3.449.287,3343 - \frac{(7060)^2}{15} = 126.380,6676.$$

Soma de Quadrados do Erro:

$$SQErro = Y'Y - \hat{\beta}'X'Y = 3.464.400 - 3.449.287,3343 = 15.112,6657.$$

Tabela 3 – Análise de Variância do modelo de regressão linear múltipla para o exemplo de Gastos da Academia.

FV	***GL***	***SQ***	***QM***	***F***
Regressão	3 – 1 = 2	126.380,6676	63.190,3338	50,1754
Resíduo	15 – 3 = 12	15.112,6657	1.259,3888	
Total	15 – 1 = 14	141.493,3333		

Como 50,1754 = $F \geq F_{0,05}(2, 12)$ = 3,89 então rejeita-se H_0 e concluímos que um ou mais dos parâmetros não são iguais a zero e que a relação global entre gastos reais da academia e o conjunto de variáveis independentes é significativa.

TESTE DE HIPÓTESE PARA A SIGNIFICÂNCIA DA REGRESSÃO MÚLTIPLA

Se no teste *F* exibir uma significância global, o teste *t* é usado para determinar se cada uma das variáveis independentes individuais é significativa, ou seja, um teste *t* separado é realizado para cada uma das variáveis independentes do modelo; referimo-nos a cada um desses testes *t* como teste de significância individual.

Hipóteses: $\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_j \neq 0 \end{cases}$; **Estatística do teste:** $t_c = \frac{\hat{\beta}_j}{\sqrt{S^2 C_{jj}}}$

com $v = n - p$ graus de liberdade e $S_{\hat{\beta}_j} = \sqrt{S^2 C_{jj}} = \sqrt{\hat{V}\left(\hat{\beta}_j\right)}$ é a estimativa do erro padrão do parâmetro β_j.

Regra de Decisão: Rejeita-se H_0 se $t \leq -t_{(\alpha/2;n-p)}$ ou $t \geq t_{(\alpha/2;n-p)}$.

Tais testes são úteis na determinação do valor potencial de cada um dos regressores no modelo de regressão. Por exemplo, o modelo pode ser mais eficiente com a inclusão de variáveis adicionais ou talvez com a retirada de um ou mais regressores do modelo. A adição de uma variável ao modelo sempre aumenta a *SQRegressão* e sempre diminui a *SQErro*. Temos que decidir se o aumento na soma quadrática da regressão é grande o suficiente para justificar o uso de uma variável adicional no modelo. Além disso, a adição de uma variável não importante no modelo pode, na verdade, aumentar a soma quadrática do erro, indicando que a adição de tal variável fez realmente o modelo apresentar um ajuste mais pobre aos dados.
A estimativa da matriz de variâncias e covariâncias para os parâmetros do modelo de regressão múltipla, usada para o teste t, é calculada por:

$$\hat{V}\left(\hat{\beta}\right) = \left(X'X\right)^{-1}.S^2 = \left(X'X\right)^{-1}.QMErro = \begin{bmatrix} \hat{V}\left(\hat{\beta}_0\right) & \hat{C}ov\left(\hat{\beta}_1,\hat{\beta}_0\right) & \cdots & \hat{C}ov\left(\hat{\beta}_j,\hat{\beta}_0\right) \\ \hat{C}ov\left(\hat{\beta}_0,\hat{\beta}_1\right) & \hat{V}\left(\hat{\beta}_1\right) & \cdots & \hat{C}ov\left(\hat{\beta}_j,\hat{\beta}_1\right) \\ \vdots & \vdots & \ddots & \vdots \\ \hat{C}ov\left(\hat{\beta}_0,\hat{\beta}_j\right) & \hat{C}ov\left(\hat{\beta}_1,\hat{\beta}_j\right) & \cdots & \hat{V}\left(\hat{\beta}_j\right) \end{bmatrix}$$

com, $j = 1,2,\ldots,p$ e p é igual ao número de parâmetros.

Exemplo: Considerando os dados do exemplo de gastos da academia, um teste t será realizado para determinar a significância de cada um dos parâmetros individuais.
A matriz de variâncias e covariâncias é:

$$\left(X'X\right)^{-1} QMErro = \begin{bmatrix} 1{,}19055 & -0{,}06928 & -0{,}02318 \\ -0{,}06928 & 0{,}01347 & -0{,}00397 \\ -0{,}02318 & -0{,}00397 & 0{,}00365 \end{bmatrix} 1.259{,}3888$$

As estimativas de $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, S_{\hat{\beta}_0}, S_{\hat{\beta}_1}$ e $S_{\hat{\beta}_2}$ são:

$$\hat{\beta}_0 = 106{,}7247 \qquad S_{\hat{\beta}_0} = \sqrt{1{,}19055 \text{x} 1.259{,}3888} = 38{,}7216$$

$$\hat{\beta}_1 = 13{,}5794 \qquad S_{\hat{\beta}_1} = \sqrt{0{,}01347 \text{x} 1.259{,}3888} = 4{,}1187$$

$$\hat{\beta}_2 = 12{,}7058 \qquad S_{\hat{\beta}_2} = \sqrt{0{,}00365 \text{x} 1.259{,}3888} = 2{,}1440$$

As estatísticas de testes de hipóteses para cada um dos parâmetros β_0, β_1 e β_2 são:

	Erro padrão	*Estatística t*
β_0	38,7216	2,7562
β_1	4,1187	3,2970
β_2	2,1440	5,9262

A 5% de significância, podemos rejeitar $H_0 : \beta_0 = 0$, $H_0 : \beta_1 = 0$ e $H_0 : \beta_2 = 0$, pois as estatísticas de teste são maiores que o valor tabelado $t_{(\alpha/2;n-p)} = t_{(0{,}025;\ 12)} = 2{,}179$ em todos os casos, logo concluímos que os gastos da academia estão relacionados com consumo de kilowatts (x_1) e horas de MOD (x_2).

COEFICIENTE DE DETERMINAÇÃO MÚLTIPLO

O coeficiente de determinação múltiplo mede a eficiência de ajuste da equação de regressão múltipla estimada em relação ao conjunto de dados. O coeficiente de determinação múltiplo pode ser interpretado como a proporção da variabilidade da variável dependente que pode ser explicada pela equação de regressão múltipla estimada.

O coeficiente de determinação múltiplo é:

$$R^2 = \frac{SQ\,\mathrm{Re}\,g}{SQTotal}.$$

COEFICIENTE DE DETERMINAÇÃO MÚLTIPLO AJUSTADO

O coeficiente de determinação múltiplo ajustado também mede a eficiência de ajuste da equação de regressão múltipla estimada em relação ao conjunto de dados, porém, ponderando o número de variações (n) e o número de variáveis (k) para evitar a superestimação do R^2.

O coeficiente de determinação múltiplo ajustado é:

$$R^2_{ajust} = 1-\left(1-R^2\right)\frac{\left(n-1\right)}{\left(n-k-1\right)} = 1-\frac{QMErro}{QMTotal}.$$

Exemplo: Vamos calcular o R^2_{ajust} para discutir a eficiência do ajuste do modelo estimado para o exemplo da academia.

O coeficiente de determinação múltiplo é: $R^2 = \frac{SQ\,\mathrm{Re}\,g}{SQTotal} = \frac{126.380,6676}{141.493,3333} = 0,8932$.

O coeficiente de determinação múltiplo ajustado é: $R^2_{ajust} = 1-\left(1-0,8932\right)\frac{\left(15-1\right)}{\left(15-2-1\right)} = 0,8754$.

Interpretação: calcula-se que 87,54% da variabilidade nos gastos da academia (y) pode ser explicada pela presença dos dois regressores, gastos com energia (x_1) e gastos com pessoal (x_2).

INTERVALO DE CONFIANÇA PARA O COEFICIENTE DE REGRESSÃO

A partir da matriz de variâncias e covariâncias dos parâmetros do modelo de regressão múltipla pode-se definir o intervalo de confiança para β_j como:

$$\underset{100(1-\alpha)\%}{IC\left(\beta_j\right)}: \hat{\beta}_j \pm t_{(\alpha/2,n-p)}\cdot\sqrt{S^2C_{jj}}$$

em que, $\sqrt{S^2C_{jj}} = \sqrt{\hat{V}\left(\hat{\beta}_j\right)} = S_{\hat{\beta}_j}$ é o erro padrão do parâmetro β_j.

Exemplo: Considerando os dados do exemplo de gastos da academia, construiremos um intervalo de confiança de 95% para o parâmetro β_1.

A estimativa de β_1 é $\hat{\beta}_1 = 13,5794$. A estimativa de seu erro padrão é $S_{\hat{\beta}_1} = \sqrt{S^2 C_{11}} = \sqrt{\hat{V}\left(\hat{\beta}_1\right)} = \sqrt{1.259,3888 \text{x} 0,01347} = 4,1187$. Consequentemente, o intervalo de confiança de 95% para β_1 é calculado a partir da expressão (39).

$$IC_{95\%}(\beta_1): \hat{\beta}_1 \pm t_{(0,05/2;\ 12)} \cdot \sqrt{S^2 C_{11}}$$

$$IC_{95\%}(\beta_1): 13,5794 \pm 2,179 \text{x} \sqrt{1.259,3888 \text{x} 0,01347}$$

$$IC_{95\%}(\beta_1): 13,5794 \pm 2,179 \text{x} 4,1187$$

$$IC_{95\%}(\beta_1): 13,5794 \pm 8,9746$$

$$IC_{95\%}(\beta_1): [4,6048;\ 22,5540]$$

Pode-se afirmar com 95% de confiança que o verdadeiro valor do parâmetro β_1 está contido no intervalo [4,6048; 22,5540].

INTERVALO DE CONFIANÇA PARA A RESPOSTA MÉDIA

Podemos obter também um intervalo de confiança para a resposta média em um determinado ponto, x_{01}, x_{02}, ..., x_{0k}. Para estimar a resposta média nesse ponto, defina o vetor X_0 como:

$$X_0 = \begin{bmatrix} 1 \\ x_{01} \\ x_{02} \\ \vdots \\ x_{0k} \end{bmatrix}$$

A resposta média nesse ponto é estimada por $\hat{\mu}_{Y|X_0} = X_0' \hat{\beta}$. Portanto, o intervalo de confiança de 100(1- α)% para a resposta no ponto x_{01}, x_{02}, ..., x_{0k} é:

$$IC_{100(1-\alpha)\%}\left(\mu_{Y|X_0}\right): X_0' \hat{\beta} \pm t_{(\alpha/2, n-p)} \cdot \sqrt{S^2 X_0' \left(X'X\right)^{-1} X_0}$$

Exemplo: Considerando os dados do exemplo de gastos da academia, determinaremos um intervalo de confiança de 95% para o gasto médio da academia quando se consome 10 kilowatts $(x_1 = 10)$ e 20 horas de mão-de-obra direta $(x_2 = 20)$.

Usando a equação de regressão estimada $\hat{y} = 106,7247 + 13,5794 x_1 + 12,7058 x_2$, sendo $x_1 = 10$ e $x_2 = 20$, obtemos o seguinte valor de $\hat{y}$:

$$\hat{y} = 106,7247 + 13,5794 \text{x} 10 + 12,7058 \text{x} 20 = R\$496,64$$

Nota: Observe que

$$\hat{y} = X_0' \hat{\beta} = \begin{bmatrix} 1 & 10 & 20 \end{bmatrix} \begin{bmatrix} 106,7247 \\ 13,5794 \\ 12,7058 \end{bmatrix} = 106,7247 + 13,5794 \text{x} 10 + 12,7058 \text{x} 20 = R\$496,64.$$

em que $X_0 = \begin{bmatrix} 1 \\ 10 \\ 20 \end{bmatrix}$ e $\hat{\beta} = \begin{bmatrix} 106,7247 \\ 13,5794 \\ 12,7058 \end{bmatrix}$.

A estimação por ponto do gasto médio da academia quando se consome 10 kilowatts e 20 horas de mão-de-obra direta é de R$ 496,64. A estimação por intervalo para o gasto médio da academia foi obtida a partir da expressão (41):

$$IC\ \underset{95\%}{\left(\mu_{Y|X_0}\right)}: X_0'\hat{\beta}\ \pm\ t_{(0,025;\ 12)}\sqrt{S^2 X_0'\left(X'X\right)^{-1} X_0}$$

$$IC\ \underset{95\%}{\left(\mu_{Y|X_0}\right)}: 496,64\ \pm\ 2,179\sqrt{1.259,3888 \text{x} 0,09675}$$

$$IC\ \underset{95\%}{\left(\mu_{Y|X_0}\right)}: 496,64\ \pm 24,05$$

$$IC\ \underset{95\%}{\left(\mu_{Y|X_0}\right)}: [472,59;\ 520,69]$$

em que,

$$X_0'\left(X'X\right)^{-1} X_0 = \begin{bmatrix} 1 & 10 & 20 \end{bmatrix} \begin{bmatrix} 1,19055 & -0,06928 & -0,02318 \\ -0,06928 & 0,01347 & -0,00397 \\ -0,02318 & -0,00397 & 0,00365 \end{bmatrix} \begin{bmatrix} 1 \\ 10 \\ 20 \end{bmatrix} = 0,09675$$

Pode-se afirmar com 95% de confiança que o gasto médio da academia quando se consome 10 kilowatts e 20 horas de mão-de-obra direta por mês está contido no intervalo [472,59; 520,69].

3. REGRESSÃO LOGÍSTICA

A regressão linear geralmente é adequada quando a variável resposta é quantitativa. No entanto, se a variável resposta é binária (0 ou 1, sucesso ou fracasso) existem apenas dois possíveis valores para o erro (resíduo do modelo) caso se assuma o modelo linear simples. Nesse caso, os erros não podem ser considerados com distribuição normal, e, além disso, a variância dos erros não é constante, pois essa é função da média. Assim, a forma da função de resposta ($E(Y_i)$) deve ser não linear.

Uma função de resposta monotonicamente crescente (ou decrescente) em forma de S, chamada de função logit, geralmente é empregada para modelar a variável resposta. O modelo de regressão é então chamado de regressão logística, e a função logit, considerando uma variável resposta y e apenas uma covariável, tem a forma a seguir.

$$E(Y) = \frac{\exp\left(\beta_o + \beta_1 x\right)}{1 + \exp\left(\beta_o + \beta_1 x\right)} = \frac{1}{1 + \exp\left(-\left(\beta_o + \beta_1 x\right)\right)}$$

A função logit tem a forma dos gráficos a seguir.

O termo $\exp(\beta_o + \beta_1 x)$ é chamado de razão de chances (odds ratio). A razão de chances pode ser interpretada como o número de vezes que o sucesso é mais provável do que o fracasso, considerando um valor particular de x. Assim, se a razão de chances é igual a 2 por exemplo, significa que o sucesso é 2 vezes mais provável que o fracasso, para aquele valor de x. A razão de chances varia de e^{β_1} quando x aumenta em uma unidade.

O log da razão de chances, $\beta_o + \beta_1 x$, é uma função linear da variável regressora. Esse termo é chamado de preditor linear, e a inclinação β_1 é a variação no log das chances que resulta do aumento de 1 unidade em x.

A estimação dos parâmetros é feita pelo método da máxima verossimilhança. No entanto, as funções são não lineares, e assim, são necessários métodos numéricos para se encontrar as estimativas de máxima verossimilhança.

Exemplo. A regressão logística é ilustrada usando os dados de temperatura de lançamento e falha de O-ring para 24 lançamentos de ônibus espaciais antes do desastre do Challenger em janeiro de 1986. O número 1 significa que houve pelo menos uma falha e o zero significa que não houve falha.

Dados:

Temperatura	53	56	57	63	66	67	67	67	68	69	70	70
Falha	1	1	1	0	0	0	0	0	0	0	0	1
Temperatura	70	70	72	73	75	75	76	76	78	79	80	81
Falha	1	1	0	0	0	1	0	0	0	0	0	0

O modelo ajustado é $\hat{y} = \dfrac{1}{1+\exp(-(10{,}875 - 0{,}17132x))}$. O termo e^{β_1} é 0,84; logo, o aumento de um grau na temperatura reduz as chances de falha por 0,84.

No gráfico de probabilidade, fica evidente o aumento na probabilidade de falha, para pequenas temperaturas. A temperatura de lançamento do Challenger foi de 31º, e este valor está fora do intervalo considerado, não sendo aconselhável fazer previsões para esta temperatura. No entanto, analisando-se o gráfico, pode-se verificar forte indício de que uma temperatura tão baixa resultaria quase certamente em uma falha. Esse exemplo indica dramaticamente que todos os cientistas devem ter conhecimentos estatísticos.