# Regressão Linear Múltipla

## Modelo Estatístico - Notação Matricial

Tem-se uma regressão linear múltipla quando se admite que a variável resposta $Y$ é a função de duas ou mais variáveis explicativas (regressoras). O modelo estatístico de uma regressão linear múltipla com $k$ variáveis regressoras $(X_1, X_2, \ldots, X_k)$ é:

$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \varepsilon_i$

ou na forma parametrizada com variéveis centradas:

$Y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i$

Em notação matricial, o modelo de regressão linear múltipla fica:

$Y = X\theta + \varepsilon = \mu + \varepsilon$

em que $Y$ é o vetor de dimensões $n \times 1$ da variével aleatória Y, $X$ é a matriz de dimensões $n \times p$, temos $\theta$ como sendo o vertor de dimensões $p \times 1$ , de parâmetros desconhecidos, $\varepsilon$ é o vetor de dimensões $n \times 1$ das variáveis aleatórias não observáveis.

De forma semelhante a regressão linear simples, têm-se as suposições

1. A variável resposta $Y$ é função linear das variáveis explicativas $X_j$ para $j = 1, 2, \ldots, k$
2. As variáveis explicativas $X_j$ são fixas
3. $E(\varepsilon_i) = 0$ , ou seja, $E(\varepsilon) = \mathbf{0}$, sendo $\mathbf{0}$ o vetor nulo de dimensões $n \times 1$
4. Os erros são *homocedásticos*, isto é, $Var(\varepsilon_i) = E(\varepsilon_i^2) = \sigma^2$
5. Os erros são *independentes*, isto é, $Cov(\varepsilon_1, \varepsilon_j) = 0$ se $i \neq j$
6. Os erros têm distribuição normal

**OBS:** A suposição da normalidade dos erros se dás necessária para a aelabioração de testes de hipóteses e obtenção de intervalos de confiança.

## Estimação dos parâmetros - Método dos mínimos quadrados

Podemos calcular a soma dos quadrados dos desvios $L$ , sendo esta dada pela formula:

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = (Y - X\theta)^T (Y - X\theta)$$

O Estimador por minimos quadrados será $\hat{\theta}$ , sendo este solução para $\theta$ nas equações

$$\frac{\partial L}{\partial \theta} = 0$$

Utilizando propriedades de matrizes podemos concluir que:

$X^T X\hat{\theta} = X^T y \Rightarrow [X^T X]\hat{\theta} = X^T y \Rightarrow \hat{\theta} = (X^T X)^{-1} X^T Y$

Resolvendo assim a última igualdade, temos os parametros para o modelo que minimiozam a função soma dos desvios $L$.

### Exemplo - Satisfação de Pacienrtes em um Hospital

Primeiro vamos importar o conjunto de dados para o calculo do modelo de regressão, este conjunto de dados é sobre o nivel de satisfaçção de visistantes a um dado hospital, com base em alguns parametros como ansiedade e idade do paciente por exemplo.

```
library(readxl)
exemplo_dados_reg <- read_excel("exemplo_dados_reg.xlsx")
View(exemplo_dados_reg)
```

Neste exemplo a variavel $Y$ será os valores de satisfação dos pacientes **Satisfaction**, sendo que vamos verificar o quando os parametros $X_1$ **Age**, $X_2$ **Saverti**, $X_3$ **Surg-Med** e $X_4$ **Anxiety** explicam sobre $Y$ .

Feito isso, vamos calcular a matrix $X$ inserindo uma linha extra somente com valores 1's no lugar da primeira linha, aumentando assim em uma linha a matriz com os nossos dados, feito isso será calculada $X^T$ .

```
X <- matrix(1, nrow = 25, ncol = 5)
M <- exemplo_dados_reg
N <- as.matrix(M)

for (j in 1:25){
  for (k in rev(1:4)){
    X[j, k+1] = N[j, k]
  }
}

X
```

```
##       [,1] [,2] [,3] [,4] [,5]
##  [1,]    1   55   50    0  2.1
##  [2,]    1   46   24    1  2.8
##  [3,]    1   30   46    1  3.3
##  [4,]    1   35   48    1  4.5
##  [5,]    1   59   58    0  2.0
##  [6,]    1   61   60    0  5.1
##  [7,]    1   74   65    1  5.5
##  [8,]    1   38   42    1  3.2
##  [9,]    1   27   42    0  3.1
## [10,]    1   51   50    1  2.4
## [11,]    1   53   38    1  2.2
## [12,]    1   41   30    0  2.1
## [13,]    1   37   31    0  1.9
## [14,]    1   24   34    0  3.1
## [15,]    1   42   30    0  3.0
## [16,]    1   50   48    1  4.2
## [17,]    1   58   61    1  4.6
## [18,]    1   60   71    1  5.3
## [19,]    1   62   62    0  7.2
## [20,]    1   68   38    0  7.8
## [21,]    1   70   41    1  7.0
## [22,]    1   79   66    1  6.2
## [23,]    1   63   31    1  4.1
## [24,]    1   39   42    0  3.5
## [25,]    1   49   40    1  2.1
```

Calculando agora $X^T X$

```
XtX <- t(X)%*%X
XtX
```

```
##        [,1]    [,2]    [,3]  [,4]    [,5]
## [1,]   25.0  1271.0  1148.0  14.0   98.30
## [2,] 1271.0 69881.0 60814.0 756.0 5387.10
## [3,] 1148.0 60814.0 56790.0 671.0 4760.60
## [4,]   14.0   756.0   671.0  14.0   57.40
## [5,]   98.3  5387.1  4760.6  57.4  461.21
```

Calculando a inversa $(X^T X)^{-1}$

```
in_XtX <- solve(XtX)
in_XtX
```

```
##              [,1]          [,2]          [,3]          [,4]          [,5]
## [1,]  0.703140769 -0.0064927611 -0.0077141793 -0.0117614119  0.0070632525
## [2,] -0.006492761  0.0003684202 -0.0001219604 -0.0015277456 -0.0014704381
## [3,] -0.007714179 -0.0001219604  0.0003537337 -0.0005422198 -0.0005150462
## [4,] -0.011761412 -0.0015277456 -0.0005422198  0.1741529911  0.0042739211
## [5,]  0.007063252 -0.0014704381 -0.0005150462  0.0042739211  0.0226224174
```

Multiplicando agora $X^T$ por $Y$ temos:

```
Y <- matrix(0, nrow = 25, ncol = 1)
for (i in 1:25){
  Y[i,1] = N[i,5]
}

Xty <- t(X)%*%Y
Xty
```

```
##         [,1]
## [1,]  1638.0
## [2,] 76487.0
## [3,] 70426.0
## [4,]   857.0
## [5,]  5959.4
```

Deste modo, podemos finalmente calcular $\hat{\theta}$ como sendo $\hat{\theta} = (X^T X)^{-1} X^T Y$.

```
theta <- in_XtX %*% Xty
theta
```

```
##             [,1]
## [1,] 143.8671879
## [2,]  -1.1171771
## [3,]  -0.5862110
## [4,]   0.4148747
## [5,]   1.3063510
```

## Propriedades Estatística dos Estimadores por Mínimos Quadrados

O modelo ajustado é dado em sua forma matricila por $\hat{Y} = X\hat{\theta}$ . Com isso podemos realizar uma análise dos desvios padrão dos residuos $\varepsilon$, sendo estes dados como $\varepsilon = y - \hat{y}$.

Para calcular o desvio padrão dos residuos, utilizamos a fórmula

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \varepsilon_i^2}{n-p} = \frac{SS_E}{n-p}$$

Sendo que $n$ é o número de observaçoes e $p$ é o número de parâmetros em nosso modelo.

Para determinarmos o desvio padrão dos residuos dos estimadores $\hat{\theta}_0, \hat{\theta}_1, \ldots \hat{\theta}_{p-1}$, precisamos calcular primeiro a matrix $C$:

$$C = (X^T X)^{-1} = \begin{bmatrix} C_{00} & C_{01} & \cdots & C_{0(p-1)} \\ C_{10} & C_{11} & \cdots & C_{1(p-1)} \\ \vdots & \vdots & \ddots & \vdots \\ C_{(p-1)0} & C_{(p-1)1} & \cdots & C_{(p-1)(p-1)} \end{bmatrix}$$

Podemos obter a matriz de covariancia $Cov(\hat{\theta})$ multiplicando a matriz $C$ pela estimativa do desvio padrão dos residuos $\hat{\sigma}^2$:

$$Cov(\hat{\theta}) = \hat{\sigma}^2 (X^T X)^{-1} = \hat{\sigma}^2 C$$

O *desvio padrão dos estimadores por minimos quadrados* do estimador $\hat{\theta}_j$ denotado por $Se(\hat{\theta}_j)$, com $j = 0, 1, \ldots, p-1$ é determinado por tomarmos a raiz quadrada do produto de $\hat{\sigma}^2$ e o *j-ésimo elemento da diagonal pricipla de* $C$.

$$Se(\hat{\theta}_j) = \sqrt{\hat{\sigma}^2 C_{jj}}$$

### Voltando ao Exemplo dos Pacientes

Note que já temos calculado o valor da matriz $C$, sendo esta `in_XtX` . Assim vamos calcular as predições feitas pelo modelo fazendo $\hat{Y} = X\hat{\theta}$.

```
Y_chap = X%*%theta
Y_chap
```

```
##            [,1]
##  [1,]  55.85524
##  [2,]  82.48064
##  [3,]  88.11200
##  [4,]  82.92132
##  [5,]  46.56621
##  [6,]  47.20912
##  [7,]  30.69218
##  [8,]  81.38880
##  [9,]  93.13223
## [10,]  61.13073
## [11,]  65.66963
## [12,]  83.21994
## [13,]  86.84116
## [14,] 101.17345
## [15,]  83.27848
## [16,]  65.77176
## [17,]  49.73614
## [18,]  42.55412
## [19,]  47.66286
## [20,]  55.81267
## [21,]  51.18948
## [22,]  25.43453
## [23,]  61.08341
## [24,]  80.24865
## [25,]  68.83528
```

Com isso calculamos os residuos fazendo $\varepsilon = Y - \hat{Y}$.

```
res = Y - Y_chap
res
```

```
##             [,1]
##  [1,]  12.1447623
##  [2,]  -5.4806372
##  [3,]   7.8879960
##  [4,]  -2.9213180
##  [5,]  -3.5662065
##  [6,]  -3.2091186
##  [7,]  -4.6921771
##  [8,]   6.6112036
##  [9,] -18.1322342
## [10,]  -4.1307260
## [11,]  -9.6696335
## [12,]   4.7800638
## [13,]   1.1588368
## [14,]   0.8265468
## [15,]   4.7215250
## [16,]   4.2282432
## [17,]   2.2638620
## [18,]   0.4458802
## [19,]  -1.6628567
## [20,]   0.1873314
## [21,]   7.8105246
## [22,]   0.5654735
## [23,]  -9.0834067
## [24,]   2.7513501
## [25,]   6.1647154
```

Logo, agora podemos calcular $SS_E$ (soma dos quadros dos residuos), bem como dividir isso por $n - p$ (número de observações menos o número de parâmetros), sendo que neste exemplo $n = 25$ e $p = 5$ .

```
quad = res^2
SSe = sum(quad)
sigma_chap = sqrt(SSe/(25 - 5))
sigma_chap
```

```
## [1] 7.20745
```

Deste modo podemos construir um vetor cujo cada entrada correspode ao desvio padrão dos nossos estimadores $\hat{\theta}_k$, com $k = 0, 1, \ldots, 4$.

```
Se_theta = matrix(0, nrow = 5, ncol = 1)
for (i in 1:5){
  Se_theta[i,1] = sqrt(sigma_chap^2*in_XtX[i,i])
}
```

Logo temos o vetor $\hat{\theta}$ com os parâmetros do nosso modelo conseguidos por meio do método dos minimos quadrados:

```
theta
```

```
##             [,1]
## [1,] 143.8671879
## [2,]  -1.1171771
## [3,]  -0.5862110
## [4,]   0.4148747
## [5,]   1.3063510
```

E o vetor $Se(\hat{\theta})$ com os respoectivos desvios padrôes de cada parâmetros:

```
Se_theta
```

```
##           [,1]
## [1,] 6.0436981
## [2,] 0.1383418
## [3,] 0.1355563
## [4,] 3.0077871
## [5,] 1.0840545
```

## Testes de Hipoteses no Modelo de Regressão Multipla

Existem dois testes de hipóteses que podemos realizar em um modelo linear de regressão múltipla, sendo estes:

1. **Teste da Significancia do Modelo:** Testa se o modelo como um todo é apropriado para descrever os dados em estudo, verificadno assim se existe alguma correlação nos dados que pode sex explicada por meio de um modelo linear.
2. **Teste Individual dos Coeficientes de Regressão**: Verifica se os coeficientes são significaticvos de forma individual. Caso não sejam, podemos remover estes do modelo.

Vamos utilizar a análise de variância ANOVA par5a avlaiarmos a variabilidade dos dados, para tal vamos dividir a variabilidade total dos dados $SS_T = \sum_{i=1}^{n} (y_i - \overline{y})^2$ em duas partes:

- $SS_R = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$ *soma dos quadrados da regressão*, com $p - 1$ graus de liberdade. É o quadrado da diferença dos dados preditos pelo modelo para com a média.
- $SS_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ a *soma dos quadrados dos residuos*, calcula a soma dos quadrados dos dados observados para com as prediçoes feitas pelo modelo, possuindo $n - p$ graus de liberdade.

Sendo que temos $SS_T = SS_R + SS_E$. Para o modelo de regressão ser significativo, precisamos que a maior parte da variabilidade dos dados seja explicada pelo modelo, caso contrario, podem haver mais fatores que influenciam nos dados e que não estamos levando em conta em nosso modelo.

Assim, para um modelo significativo, precisamos que $\frac{SS_R}{SS_T}$ seja *consideravelmente grande*, no geral nos normalizamos tanto $SS_R$ quanto $SS_E$ pelos seus graus de liberdade, de modo que:

$$\frac{MS_R}{MS_E} = \frac{\frac{SS_R}{p-1}}{\frac{SS_E}{n-p}} = \frac{SS_R(n-p)}{SS_E(p-1)}$$

Sendo $MS_R$ e $MS_E$ as respectivas normalizações pelos graus de liberdade de $SS_R$ e $SS_E$ . Lembrando que:

- $n$ : Número de observações.
- $p$ : Quantidade de parâmetros do modelo.

Note agora que $SS_T = \sum_{i=1}^{n} (y_i - \overline{y})^2 = s^2(n-1)$, sendo $s^2$ a variancia amostral dos dados.

# 1. Teste da Significancia do Modelo

Realizaremos o teste com base nas seguintes hipóteses:

- $H_0 : \theta = 0$
- $H_1 : \theta_j \neq 0$, para pelo menos um $j$

A hipótese $H_0$ implica o *modelo nulo*, onde o vetor dos parâ,etros é o vetor nulo.

Co mo estamos calculandop a divisão de duas grandezas ao quadrado, isto é, $\frac{MS_R}{MS_E}$, vamios realizar o teste de hipóteses com, base na distribuição $F$.

Assim:

$$f_0 = \frac{MS_R}{MS_E} = \frac{SS_R(n-p)}{SS_E(p-1)}$$

e vamos verificar a condição $f_0 > f_{\alpha,(p-1),(n-p)}$, que caso verdadeira, rejeitamos $H_0$ e aceitamos $H_1$ a um cero nível de significancia.

**(Duvida: Eu vi no video que em boa parte de experimento reais esse texte de hipóteses não é muito útil, por que?)**

### Criando a tabela ANOVA

Para calcular os respectivos parâmetros para o teste de hipóteses, como estamos lidando com vários parâmetros, utilizamos uma tabela *ANOVA* da análise das variâncias da regressão e do residuo.

tabela ANOVA

| Fonte da Variação | Soma dos Quadrados | Graus de Liberdade | Médias dos Quadrados | $F_0$ |
|---|---|---|---|---|
| **Regressão** | $SS_R$ | $p-1$ | $MS_R = SS_R/(p-1)$ | $MS_R/MS_E$ |
| **Erro (Residuo)** | $SS_E$ | $n-p$ | $MS_E = SS_E/(n-p)$ | |
| **Total** | $SS_T$ | $n-1$ | | |

### Realizando o teste

Primeiro, vamos calcular os respectivos valores presentes na tabela da *ANOVA*.

```
SSE_vec = matrix(0, nrow = 25, ncol = 1)
for (i in 1:25){
  SSE_vec[i,1] = (res[i,1])^2
}
SSE = sum(SSE_vec)

SST = (sd(Y))^2*24

SSR = SST - SSE

MSR = SSR/4
MSE = SSE/(25-5)

f0 = MSR/MSE
f0
```

```
## [1] 46.871
```

Feito isso, podemos calcular nosso *p-valor* segundo a distribuição $F$, calculando assim a probabilidade de $f_0 > f_{\alpha,(p-1),(n-p)}$.

```
p_valor = pf(f0, 4, 20, lower.tail = FALSE)
p_valor
```

```
## [1] 6.950524e-10
```

Sendo assim, considerando um nivel de significância de **95%**, temos $\alpha = 0,05$. Como $p \approx 6.95 * 10^{-10} << 0,05$ , rejeitamos $H_0$ e aceitamos $H1$ .

Portanto, $\theta_j \neq 0$, para pelo menos um $j$ , a um nivel de significância de **95%**.

# 2. Teste Individual dos Coeficientes de Regressão

Agora as hipóteses que vamos considerar no teste são:

- $H_0 : \theta_j = \theta_{j0}$
- $H_1 : \theta_j \neq \theta_{j0}$ , sendo tipicamente $\theta_{j0} = 0$

A estatística do teste, caso $\theta_{j0}$ agora é dada por:

$$T_{j0} = \frac{\hat{\theta}_j - \theta_{j0}}{\sqrt{\sigma^2 C_{jj}}} = \frac{\hat{\theta}_j}{se(\hat{\theta}_j)}$$

em suma, estamos verificando se algum dos coeficeintes de nosso vetor dos parâmetros $\hat{\theta}$ é estatisticamente iguala a $0$.

Sendo este um teste de calda dupla, vamos verificar se $|t_0| > t_{(\alpha/2),(n-p)}$, caso isso ocorra, rejeitamos $H_0$ e aceitamos $H_1$.

### Realizando o teste

Calculando primeiro o vetor $T_{j0}$ utilizando a fórmula ascima:

```
T0 = matrix(0, nrow = 5, ncol = 1)
for (i in 1:5){
  T0[i,1] = theta[i,1]/Se_theta[i,1]
}
T0
```

```
##            [,1]
## [1,] 23.8044962
## [2,] -8.0754857
## [3,] -4.3244820
## [4,]  0.1379335
## [5,]  1.2050603
```

Agora vgamos realizar o teste de calda dupla para cada entrada do vetor $T_0$ utilizando a distribuição *t-student*, obtendo assim o vetor $P_0$ dos *p-valores*.

```
P0 = matrix(0, nrow = 5, ncol = 1)
for (i in 1:5){
  P0[i,1] = 2*(pt(abs(T0[i,1]), 20, lower.tail = FALSE))
}
P0
```

```
##              [,1]
## [1,] 3.795908e-16
## [2,] 1.007923e-07
## [3,] 3.294732e-04
## [4,] 8.916722e-01
## [5,] 2.422458e-01
```

Assumindo agora um nivel de significância de **95%**, obtendo assim $\alpha = 0,05$, vamos ver quais *p-valores* são maiores que $\alpha$.

```
verificador = matrix(0, nrow = 5, ncol = 1)
for (i in 1:5){
  if (P0[i,1] > 0.05){
    verificador[i,1] = 0
  } else {
    verificador[i,1] = 1
  }
}
verificador
```

```
##      [,1]
## [1,]    1
## [2,]    1
## [3,]    1
## [4,]    0
## [5,]    0
```

Note que para os parâmetros $\hat{\theta}_3$ e $\hat{\theta}_4$ não rejeitamos $H_0$, logo estes são estatísticamentes iguais a zero, para um nível de significância de **95%**, podendo assim serem removidos do modelo.

Interpretando os dados, temos que os únicos parâmetros que possuem alguma influencia na satisfação dos clientes, são aqueles correspondentes a $\hat{\theta}_1$ e a $\hat{\theta}_2$, isto é, a **Idade do paciente** e a **gravidade da sua condição médica.**