
UFV

EST 105

INICIAÇÃO À ESTATÍSTICA

CORRELAÇÃO E REGRESSÃO

Departamento de Estatística – UFV
Av. Peter Henry Rolfs, s/n
Campus Universitário
36570.977 – Viçosa, MG
http://www.det.ufv.br/

UFV

Coeficiente de correlação amostral

Motivação:

- Geralmente existe o interesse em se investigar a relação entre duas ou mais variáveis que foram medidas em uma pesquisa.

- Por exemplo, a quantidade vendida de um produto pode estar relacionada ao preço deste produto. Ou, a quantidade de grãos produzida por uma variedade de arroz, pode estar associada à quantidade de adubo utilizada, etc.

- Neste sentido, uma medida usada para avaliar o grau de **associação linear** entre duas variáveis aleatórias X e Y é chamada **coeficiente de correlação** (ρ).

Coeficiente de correlação amostral

- Considere agora uma amostra de n pares de valores (X_i, Y_i) relativos às variáveis X e Y, para as quais temos o interesse em investigar se existe associação linear.

X		...
Y		...

O coeficiente de correlação linear entre as variáveis X e Y ρ pode ser estimado por:

$$r_{XY} = \frac{SPD_{XY}}{\sqrt{SQD_X \times SQD_Y}}, -1 \leq r_{XY} \leq 1$$

Em que

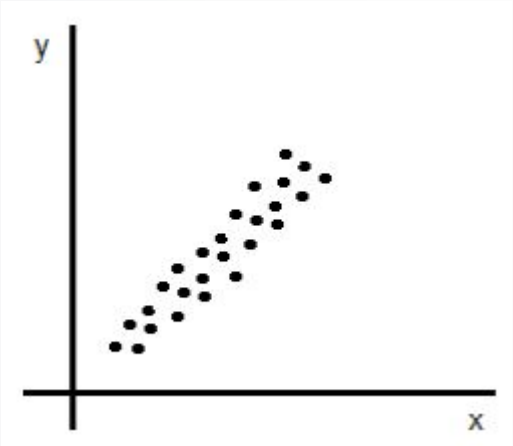
$$SQD_X = \sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n} \qquad SQD_Y = \sum_{i=1}^{n} Y_i^2 - \frac{(\sum_{i=1}^{n} Y_i)^2}{n} \qquad SPD_{XY} = \sum_{i=1}^{n} X_i Y_i - \frac{(\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{n}$$

Coeficiente de correlação amostral

- Se representarmos os pares de valores (X_i, Y_i) num sistema cartesiano, temos um **diagrama de dispersão**. A construção de um gráfico deste tipo pode nos auxiliar a identificar o tipo da associação entre as variáveis aleatórias X e Y.

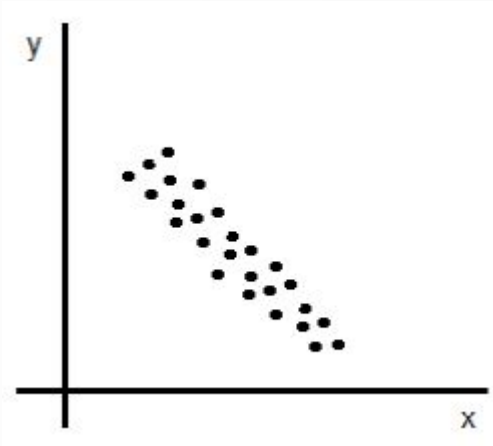
Vejamos algumas possíveis configurações:

(a) Correlação positiva



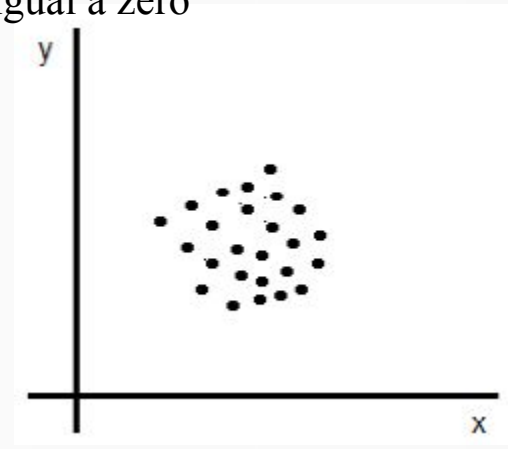
X = altura e Y= peso

(b) Correlação negativa



X = preço e Y= número de itens vendidos

(c) Correlação aproximadamente igual a zero

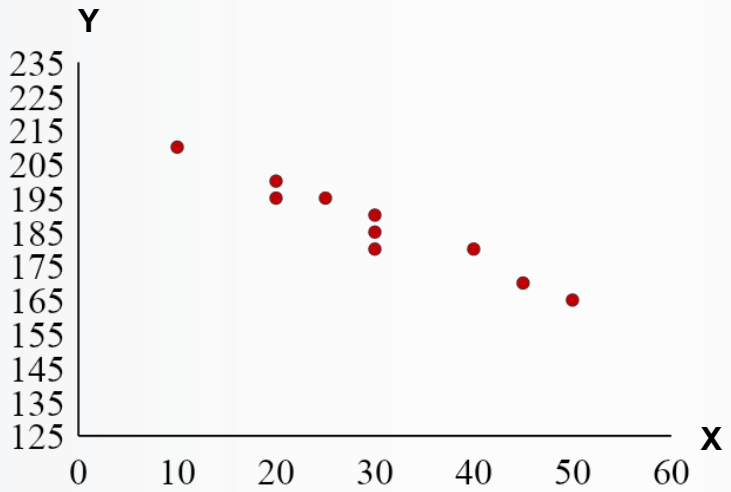


X = altura e Y= renda

Exemplo 1

A tabela a seguir apresenta informações sobre a idade (X, em anos) e o número máximo de batimentos cardíacos (Y, em minutos) de 10 pacientes amostrados em um estudo médico. Calcule o coeficiente de correlação amostral entre as variáveis aleatórias X e Y.

X	10	20	20	25	30	30	30	40	45	50
Y	210	200	195	195	190	180	185	180	170	165



Somatórios

Sabe-se que,

$$r_{XY} = \frac{SPD_{XY}}{\sqrt{SQD_X \times SQD_Y}}$$

em que

$$SPD_{XY} = \sum_{i=1}^{n} X_i Y_i - \frac{(\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{n} = 54625 - \frac{300 \times 1870}{10} = -1475$$

$$SQD_X = \sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n} = 10350 - \frac{(300)^2}{10} = 1350$$

$$SQD_Y = \sum_{i=1}^{n} Y_i^2 - \frac{(\sum_{i=1}^{n} Y_i)^2}{n} = 351400 - \frac{(1870)^2}{10} = 1710$$

Então,

$$r_{XY} = \frac{SPD_{XY}}{\sqrt{SQD_X \times SQD_Y}} = \frac{-1475}{\sqrt{1350 \times 1710}} = -0{,}9708.$$

Observou-se, portanto, uma **associação negativa** entre a idade (X) e o número máximo de batimentos cardíacos (Y), com uma correlação de -0,9708. Assim, a tendência é de que aumentando-se a idade, o número máximo de batimentos cardíacos diminua.

Regressão Linear Simples (RLS)

- Tem por objetivo estabelecer uma relação funcional entre uma variável aleatória e dependente (Y) e uma variável fixa e independente (X).

1. O modelo de RLS

- **Modelo estatístico:**

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

Em que

X_i é i-ésimo valor da variável explicativa ou independente (X);

Y_i é i-ésimo valor da variável dependente ou resposta (Y);

β_0 é a constante da regressão ou intercepto (parâmetro);

β_1 é o coeficiente de regressão ou coeficiente angular (parâmetro);

ε_i é o i-ésimo erro aleatório (não observável).

UFV

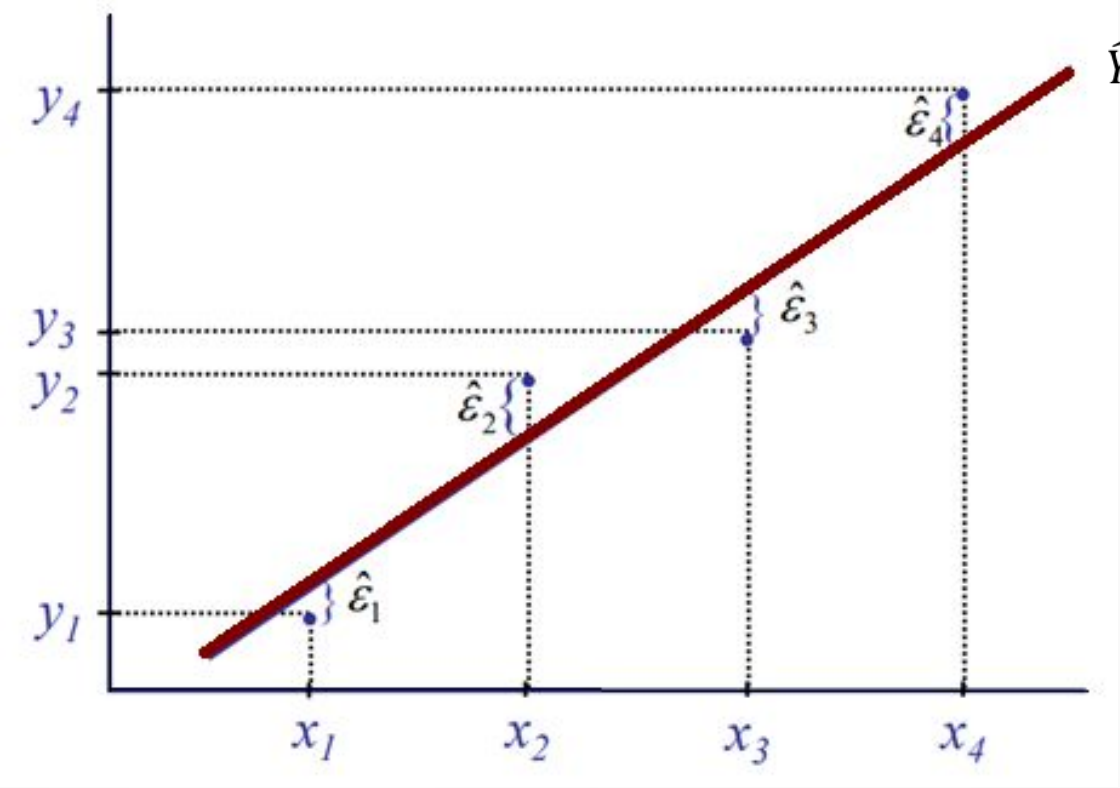
Regressão Linear Simples

2. Método de Estimação

- Apenas uma amostra de pares (x, y) é observada, logo, a verdadeira relação linear entre X e Y não será conhecida e sim estimada pela análise de regressão linear simples.
- Pelo **Método dos Mínimos Quadrados** (MMQ) é possível se ajustar o modelo. O objetivo deste método é obter as estimativas dos parâmetros que minimizam o valor da soma de quadrados dos erros aleatórios.
- Definido o modelo $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, então, $\varepsilon_i = Y_i - \beta_0 - \beta_1 X_i$. O MMQ define $\min Z = \min \sum_{i=1}^{n} \varepsilon_i^2 = \min \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$.
- Os estimadores (fórmulas) que produzem estimativas dos parâmetros (valores) que minimizam Z, são obtidos pela derivação parcial de Z em relação aos parâmetros (β_0 e β_1) do modelo. Isto é, $\frac{\partial Z}{\partial \beta_0}$ e $\frac{\partial Z}{\partial \beta_1}$.

Regressão Linear Simples

$$(b_0, b_1) = (\hat{\beta}_0, \hat{\beta}_1) = \arg\min \left(\sum_{i=1}^{n} \varepsilon_i^2\right)$$



$$\widehat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

$$\widehat{\varepsilon}_i = Y_i - \widehat{Y}_i$$

Regressão Linear Simples

- $\hat{\beta}_1$ ou b_1 é o estimador (fórmula) do parâmetro β_1.

$$\hat{\beta}_1 = b_1 = \frac{SPD_{XY}}{SQD_X} = \frac{\left[\sum_{i=1}^{n} X_i Y_i - \frac{(\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{n}\right]}{\left[\sum_{i=1}^{n} {X_i}^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}\right]}$$

- $\hat{\beta}_0$ ou b_0 é o estimador (fórmula) do parâmetro β_0.

$$\hat{\beta}_0 = b_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = \frac{\sum_{i=1}^{n} Y_i}{n} - \hat{\beta}_1 \frac{\sum_{i=1}^{n} X_i}{n}$$

Equação estimada (ou modelo ajustado):

$$\widehat{Y_i} = \hat{\beta}_0 + \hat{\beta}_1 X_i = b_0 + b_1 X_i$$

Exemplo 2

- Considere os dados do exemplo inicial referentes à idade (X, em anos) e o número máximo de batimentos cardíacos por minuto (Y) de $n = 10$ pacientes amostrados.

X	10	20	20	25	30	30	30	40	45	50
Y	210	200	195	195	190	180	185	180	170	165

Diante da correlação $r_{XY} = -0{,}9708$, admite-se que as variáveis estão relacionadas de acordo com o modelo de RLS: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$. Pede-se:

a) Apresente a equação ajustada.

- A estimativa do coeficiente de regressão:

$$\hat{\beta}_1 = b_1 = \frac{SPD_{XY}}{SQD_X} = \frac{\left[\sum_{i=1}^{n} X_i Y_i - \frac{(\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{n}\right]}{\left[\sum_{i=1}^{n} {X_i}^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}\right]} = \frac{\left(54625 - \frac{300 \times 1870}{10}\right)}{\left[10350 - \frac{(300)^2}{10}\right]} = -1{,}093$$

- A estimativa da constante de regressão:

$$\hat{\beta}_0 = b_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = \frac{\sum_{i=1}^{n} Y_i}{n} - \hat{\beta}_1 \frac{\sum_{i=1}^{n} X_i}{n} = \frac{1870}{10} - (-1{,}093)\frac{300}{10} = 219{,}78$$

- A equação ajustada ou o modelo ajustado:

$$\widehat{Y}_i = 219{,}78 - 1{,}093X_i$$

3. Interpretação

- $\hat{\beta}_0$ representa o valor estimado de Y $(\hat{Y})$ quando X é igual a zero. Algumas vezes, quando o valor $0 \notin (x_{mín\_obs}, x_{máx\_obs})$, essa estimativa não possuirá uma interpretação prática.
- $\hat{\beta}_1$ representa o aumento $(\hat{\beta}_1 > 0)$ ou a redução $(\hat{\beta}_1 < 0)$ média(o) estimada em Y para cada aumento unitário em X.

Exemplo 2

b) Interprete o coeficiente de regressão.

$$\widehat{Y_i} = 219{,}78 - 1{,}093X_i$$

O coeficiente de regressão é b_1 ou $\hat{\beta}_1$, que neste caso, assume o valor -1,093.

Assim, tem-se que, para cada aumento de 1 ano na idade, espera-se uma redução média de 1,093 no número máximo de batimentos cardíacos por minuto.

c) Nesse caso, a interpretação prática da constante de regressão b_0 ou $\hat{\beta}_0$ não deve ser realizada, pois o valor $0 \notin (10, 50)$. Ao final dessa aula detalharemos um pouco mais esse tópico.

Regressão Linear Simples

4. Desvios da regressão (ou resíduos)

São estimativas para os erros aleatórios. Em um modelo bem ajustado, isto é, aquele no qual a variável X é útil para explicar as variações na variável resposta Y, espera-se que os desvios sejam pequenos.

Os resíduos/desvios podem ser calculados como:

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i$$

$\hat{Y}_i$ → **Valor estimado**

Y_i → **Valor observado**

Exemplo 2

X	10	20	20	25	30	30	30	40	45	**50**
Y	210	200	195	195	190	180	185	180	170	**165**

d) Qual é a estimativa do número máximo de batimentos cardíacos para um indivíduo de 50 anos?

$$\hat{Y}_X = 219{,}78 - 1{,}093X$$

$$\hat{Y}_{50} = 219{,}78 - 1{,}093 \times 50 = 165{,}14$$

e) Calcule o desvio da regressão para a observação X=50.

$$\hat{\varepsilon}_X = Y_X - \hat{Y}_X = 165 - 165{,}14 = -0{,}14$$

Regressão Linear Simples

5. Extrapolação

- É possível obter estimativas para Y usando valores de X que não foram estudados. Entretanto, estes devem estar dentro do intervalo coberto pela amostra.

- Utilizar o modelo ajustado fora da amplitude estudada significa fazer uma **extrapolação**. A equação ajustada é razoável para interpolar dentro do intervalo coberto pela amostra, mas pode ser inapropriada para fazer uma extrapolação.

- **ATENÇÃO:** Por este motivo, no nosso exemplo, como o intervalo observado de idade X não continha $X = 0$, então, interpretar b_0 ou $\hat{\beta}_0$ seria uma EXTRAPOLAÇÃO DO MODELO.

Exemplo 2

f) Estime o número máximo de batimentos cardíacos para um indivíduo de 60 anos. Comente a respeito desta estimativa.

$$\hat{Y}_X = 219{,}78 - 1{,}093X$$
$$\hat{Y}_{50} = 219{,}78 - 1{,}093 \times 60 = 154{,}20$$

Estima-se que, em média, um indivíduo de 60 anos tenha no máximo 154,20 batimentos cardíacos por minuto. Entretanto, esta é uma extrapolação com o modelo, uma vez que o mesmo foi ajustado para valores de X compreendidos no intervalo entre 10 e 50. Logo, como não sabemos se a relação linear se mantém para valores superiores a 50, essa estimativa não é confiável e, portanto, sua interpretação, nesse caso, não é aconselhada.

Regressão Linear Simples

6. Coeficiente de Determinação (r^2)

- O coeficiente de determinação é uma medida da qualidade do ajuste do modelo.
- Indica a proporção da variação na variável dependente Y que está sendo explicada pela variável independente X ou pela regressão nos valores de X.
- O r^2 é expresso em porcentagem e calculado a partir da seguinte expressão:

$$r^2(\%) = \frac{SQ\text{Regressão}}{SQ\text{Total}} 100\%, \qquad 0 \leq r^2 \leq 100\%$$

em que

$$SQ\text{Total} = SQD_Y = \sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} \quad \text{e} \quad SQ\text{Regressão} = \hat{\beta}_1 SPD_{XY}.$$

- Quanto maior for o r^2, melhor é a qualidade do ajuste.

Obs.: No caso da RLS, o coeficiente de determinação é igual ao quadrado coeficiente de correlação amostral entre X e Y, isto é, $r^2(\%) = (r_{XY})^2 \times 100(\%)$.

Exemplo 2

g) Calcule e interprete o coeficiente de determinação.

$$r^2(\%) = \frac{SQ\text{Regressão}}{SQ\text{Total}} 100\% = \frac{1612{,}18}{1710} 100\% = 94{,}28\%$$

em que

$$SQ\text{Total} = SQD_Y = 1710$$
$$SQ\text{Regressão} = \hat{\beta}_1 SPD_{XY} = -1{,}093 \times -1475 = 1612{,}18$$

Aproximadamente 94,28% da variação presente no número máximo de batimentos cardíacos por minuto (Y) é explicada pela regressão linear simples nos valores de idade (X, em anos).

Atividade Proposta

Resolver os exercícios do Roteiro de Aulas abaixo relacionados:

- Exercício 4 – pág. 166
- Exercício 6 – pág. 167
- Exercício 8 – pág. 168
- Exercício 9 – pág. 169
- Exercício 10– pág. 169

Campus Viçosa:
Avenida Peter Henry Rolfs, s/n
CEP 36570-900
Viçosa - MG - Brasil | + 55 31 3899-2200

Campus Florestal:
Rodovia LMG 818, km 6
CEP 35690-000
Florestal - MG - Brasil | + 55 31 3536-3300

Campus Rio Paranaíba:
Rodovia MG-230, Km 8
CEP 38810-000
Rio Paranaíba- MG - Brasil | + 55 34 3855-9300

www.ufv.br

UFV
Universidade Federal de Viçosa