MAF 172 - MATEMÁTICA COMPUTACIONAL

Prof. Gérson R. Santos - gerson.santos@ufv.br

MAT - IEF - CAF

Aula 6 - RLS: Regressão Linear Simples

Sumário

Ensino Médio:

Questionamentos:

1. Se por 2 pontos só passa uma reta, que equação expressaria isso?
2. Por exemplo, use $A(2, 1)$ e $B(3, 2)$
3. Para que serviria a equação obtida?

Graduação:

Questionamentos:

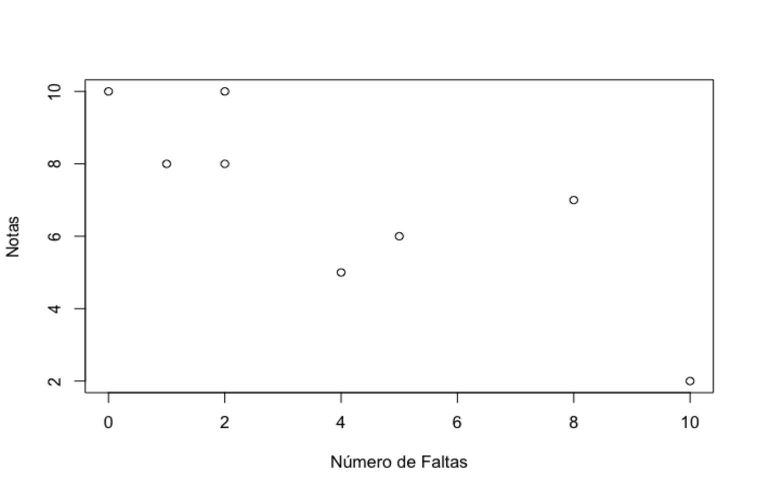
1. Se nós tivéssemos mais pontos, ainda expressaríamos a equação por uma reta?
2. Por exemplo, vamos usar os valores da tabela abaixo entre Faltas e Notas.
3. As notas sofrem influência das faltas?
4. De que forma?

Graduação:

Uma amostra de 8 alunos, escolhidos aleatoriamente, apresenta os dados abaixo sobre as faltas e as notas obtidas durante um certo período. Desejamos saber se as Notas sofrem influência das Faltas; e se Sim, de que forma?

Aluno (i)	1	2	3	4	5	6	7	8
Faltas (X_i)	8	2	5	0	1	4	10	2
Notas (Y_i)	7	10	6	10	8	5	2	8

Diagrama



Medida de Correlação Linear:

IMPORTANTE:

1. Mede APENAS a relação linear entre 2 variáveis
2. As variáveis devem ser amostradas de forma bivariada
3. A amostragem deve ser ALEATÓRIA
4. Há pressupostos teóricos que devem ser atendidos
5. Cuidado com a interpretação CAUSA e EFEITO
6. Os mais utilizados são: Pearson, Spearman e Kendall

Coeficiente:

Fórmulas:

$$r_{XY} = \frac{C\hat{O}V\,(X,Y)}{\sqrt{\hat{V}\,(X)\cdot\hat{V}\,(Y)}} = \frac{\frac{SPD_{XY}}{n-1}}{\sqrt{\frac{SQD_X}{n-1}\cdot\frac{SQD_Y}{n-1}}} = \frac{SPD_{XY}}{\sqrt{SQD_X.SQD_Y}}$$

$$SPD_{XY} = \sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n}$$

$$SQD_X = \sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}$$

Resolvendo:

Resolvendo o exemplo:

»
$$SPD_{XY} = \sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n} = 170 - \frac{32 \times 56}{8} = \boxed{-\ 54}$$

» $$SQD_X = \sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n} = 214 - \frac{32^2}{8} = \boxed{86}$$

» $$SQD_Y = \sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} = 442 - \frac{56^2}{8} = \boxed{50}$$

» $$r_{XY} = \frac{SPD_{XY}}{\sqrt{SQD_X . SQD_Y}} = \frac{-54}{\sqrt{86 \times 50}} = \boxed{-0{,}8235}$$

Pausa para mais teoria:

IMPORTANTE:

1. $-1 \leq r_{XY} \leq +1$;
2. Qdo r_{XY} é negativo: X e Y são inversamente proporcionais;
3. Qdo r_{XY} é positivo: X e Y são diretamente proporcionais;
4. Qdo r_{XY} é zero ou "próximo": X e Y não são correlacionados LINEARMENTE;
5. Qdo r_{XY} é -1 ou "próximo": X e Y são correlacionados LINEARMENTE;
6. Qdo r_{XY} é $+1$ ou "próximo": X e Y são correlacionados LINEARMENTE;

Interpretando:

Voltando ao Exemplo:

1. Como r_{XY} é negativo: X e Y são inversamente proporcionais;
2. Significado: Qto maior o número de faltas menores serão as notas;
3. A associação inversa é na ordem de $82,35\%$
4. **Observação**: Este valor NÃO está dizendo que X explica Y em $82,35\%$.

Regressão - Significado:

Pág. 135:

1. **RLS:** Regressão Linear Simples
2. **Regressão:** Voltar para entender
3. **Linear:** As derivadas parciais do modelo em função dos parâmetros não dependem dos parâmetros
4. **Simples:** Apenas uma variável X envolvida na modelagem

Significado:

Pág. 136:

1. Modelo: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$
2. Parâmetros: β_0 e β_1
3. Erros: ϵ_i
4. As variáveis devem ser amostradas de forma bivariada
5. A amostragem deve ser ALEATÓRIA
6. Há pressupostos teóricos que devem ser atendidos
7. Através do MMQ é possível estimar os parâmetros de $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$

Fórmulas:

Pág. 137:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

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

Voltando ao Exemplo:

Uma amostra de 8 alunos, escolhidos aleatoriamente, apresenta os dados abaixo sobre as faltas e as notas obtidas durante um certo período. Desejamos saber se as Notas sofrem influência das Faltas; e se Sim, de que forma?

Aluno (i)	1	2	3	4	5	6	7	8
Faltas (X_i)	8	2	5	0	1	4	10	2
Notas (Y_i)	7	10	6	10	8	5	2	8

Coeficiente:

Fórmulas:

$$r_{XY} = \frac{C\hat{O}V\,(X,Y)}{\sqrt{\hat{V}\,(X)\cdot\hat{V}\,(Y)}} = \frac{\frac{SPD_{XY}}{n-1}}{\sqrt{\frac{SQD_X}{n-1}\cdot\frac{SQD_Y}{n-1}}} = \frac{SPD_{XY}}{\sqrt{SQD_X.SQD_Y}}$$

$$SPD_{XY} = \sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n}$$

$$SQD_X = \sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}$$

Resolvendo:

Resolvendo o exemplo:

» $$SPD_{XY} = \sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n} = 170 - \frac{32 \times 56}{8} = \boxed{\text{- 54}}$$

» $$SQD_X = \sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n} = 214 - \frac{32^2}{8} = \boxed{86}$$

» $$SQD_Y = \sum_{i=1}^{n} Y_i^2 - \frac{\left(\sum_{i=1}^{n} Y_i\right)^2}{n} = 442 - \frac{56^2}{8} = \boxed{50}$$

» $$r_{XY} = \frac{SPD_{XY}}{\sqrt{SQD_X . SQD_Y}} = \frac{-54}{\sqrt{86 \times 50}} = \boxed{\text{-0,8235}}$$

Interpretando:

Voltando ao Exemplo:

1. Como r_{XY} é negativo: X e Y são inversamente proporcionais;
2. Significado: Qto maior o número de faltas menores serão as notas;
3. A associação inversa é na ordem de $82,35\%$
4. **Observação**: Este valor NÃO está dizendo que X explica Y em $82,35\%$.

Estimando a RLS:

Resolvendo o exemplo:

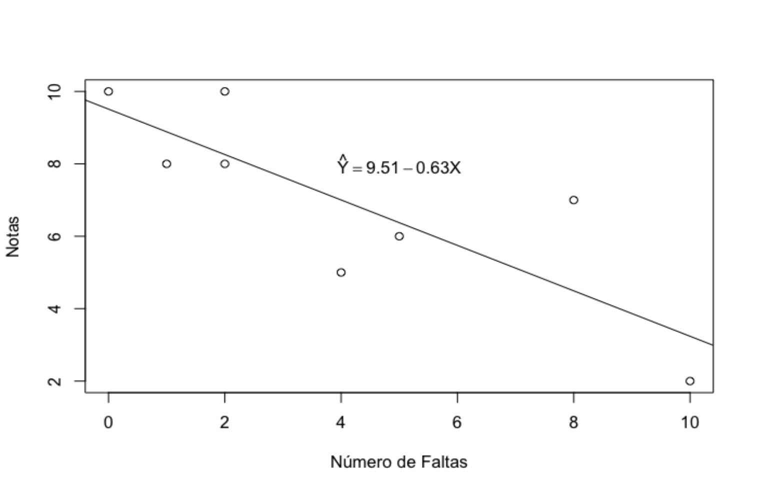
» $$SPD_{XY} = \sum_{i=1}^{n} X_i Y_i - \frac{\left(\sum_{i=1}^{n} X_i\right)\left(\sum_{i=1}^{n} Y_i\right)}{n} = 170 - \frac{32 \times 56}{8} = \boxed{\text{- 54}}$$

» $$SQD_X = \sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n} = 214 - \frac{32^2}{8} = \boxed{86}$$

» $$\hat{\beta}_1 = \frac{SPD_{XY}}{SQD_X} = \frac{-54}{86} = \boxed{\text{-0,63}}$$

» $$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} = 7 - (-0,63) \times 4 = \boxed{9{,}52}$$

Perspectivas



Resultados:

1. $r_{XY} = -0,8235$, ou seja, há uma relação linear inversa entre FALTAS e NOTAS, e a intensidade dessa relação é na ordem de $82,35\%$;
2. $\hat{Y} = 9,52 - 0,63X$, ou seja, essa é a equação do comportamento médio entre as variáveis FALTAS e NOTAS.

Interpretação:

1. Com base em $\hat{\beta}_0 = 9,52$ podemos dizer que, se os alunos não faltarem, a nota MÉDIA será 9,52;
2. Este valor é chamado também de INTERCEPTO, ou coeficiente linear, ou seja, o ponto em Y em que a reta intercepta o eixo;
3. A cada falta que os alunos tiverem, EM MÉDIA, a nota decairá 0,63 pontos;
4. Este valor é chamado também de COEFICIENTE DE REGRESSÃO, ou coeficiente angular, e indica quão inclinada é a reta;
5. Como temos uma equação podemos PREDIZER algum valor, por exemplo, $\hat{Y} = 9,52 - 0,62X = 9,52 - 0,63 \times 3 = 7,63$;
6. Esta operação é chamada também de INTERPOLAÇÃO;

Interpretação:

1. Quando uma interpolação é feita para um valor observado, podemos calcular o ERRO, dado por $\hat{\epsilon}_i = Y_i - \hat{Y}_i$;
2. Por exemplo, $\hat{Y} = 9,52 - 0,63X = 9,52 - 0,63 \times 5 = \boxed{6{,}37}$;
3. Assim, $\hat{\epsilon}_i = 6 - 6,37 = \boxed{-\ 0{,}37}$, ou seja, houve uma SUPERESTIMAÇÃO;
4. CUIDADO!!! $\hat{Y} = 9,52 - 0,63X = 9,52 - 0,63 \times 12 = \boxed{1{,}96}$ pode ser uma operação comum, contudo, trata-se de uma EXTRAPOLAÇÃO e não é permitida, nem garantida, pela Estatística.

Coeficiente:

Pág. 162:

$$R^2 = r_{XY}^2$$

ou

$$R^2 = \frac{\hat{\beta}_1 SPD_{XY}}{SQD_Y}$$

Interpretação:

1. Neste exemplo, $R^2 = (-0,8235)^2 = 0,6782$;
2. Este coeficiente indica que cerca de 68% da VARIAÇÃO de Y (Notas) é explicada pela VARIAÇÃO de X (Faltas);
3. Todas as outras variáveis juntas explicam os 32% restantes. **Obs.:** É possível encontrar variáveis mais expressivas que essa (FALTAS), assim, todo o processo deve ser repetido. Além disso, estatisticamente, sempre devemos testar a SIGNIFICÂNCIA dos valores obtidos, mas não o faremos nessa disciplina.

Um Novo Exemplo:

Uma amostra de 9 alunos, escolhidos aleatoriamente, apresenta os dados abaixo sobre os dias de estudo antes das provas e as notas obtidas naquela prova específica. Desejamos saber se as Notas sofrem influência do número de Dias de Estudo; e se Sim, de que forma?

Aluno (i)	1	2	3	4	5	6	7	8	9
Dias (X_i)	0	2	3	5	6	7	9	10	12
Notas (Y_i)	2	4	4	5	7	8	8	9	10

Resultados: $r_{XY} = \boxed{0{,}9787}$, $\hat{\beta}_1 = \boxed{0{,}67}$, $\hat{\beta}_0 = \boxed{2{,}32}$ e $R^2 = \boxed{0{,}9579}$