EST 105 - Exercícios de Regressão Linear Simples e Correlação [1]

1 (II/2006). Em regressão linear simples utiliza-se o modelo $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ que após ajustado é representado por $\widehat{Y}_i = b_0 + b_1 X_i$.

a. Explique as diferenças entre **erros aleatórios** e **desvios da regressão**.

b. Mostre que $\sum_{i=1}^{n} \widehat{\varepsilon}_i = 0$.

2 (II/2001). A tabela a seguir apresenta dados de uma amostra de 10 pacientes de um estudo médico conduzido para se pesquisar o relacionamento entre as variáveis idade (X) em anos e o número máximo de batimentos cardíacos por minuto (Y).

Idade	10	20	20	25	30	30	30	40	45	50
N^o de batimentos	210	200	195	195	190	180	185	180	170	165

Dados:

$$SQD_X = 1350 \quad SQD_Y = 1710 \quad SPD_{XY} = -1475 \quad \overline{X} = 30 \quad \overline{Y} = 187$$

Assinale (V) se a afirmativa for totalmente verdadeira ou (F) caso contrário.

a.( ) A equação de regressão linear simples ajustada é: $\widehat{Y}_i = 219,78 + 1,093 X_i$.

b.( ) Aproximadamente 94,2% da variação observada nos valores do número máximo de batimentos cardíacos por minuto é explicada pela regressão linear nos valores da idade.

c.( ) O coeficiente de correlação linear aproximadamente igual a 0.971 (correlação positiva) indica que com o aumento da idade espera-se uma diminuição do número máximo de batimentos cardíacos por minuto e vice-versa .

d.( ) Uma estimativa do valor médio do número máximo de batimentos cardíacos por minuto para um indivíduo com idade igual a 50 anos é $\approx 165,14$.

e.( ) A estimativa do correspondente (item d.) erro ou o desvio da regressão é igual a 0,14.

3 (II/2002). Suponha que se estimou o coeficiente de correlação entre as notas da primeira prova (X) e as notas médias finais (Y) de um curso e obteve-se $r_{X,Y} =$

[1] Exercícios das avaliações dos semestres indicados. Contém 10 exercícios em páginas numeradas de 1 a 7.

0, 73. Os $n = 600$ pares de notas $(X_i, Y_i),\ i = 1, 2, \ldots, n$ apresentaram as seguintes estatísticas:

$$\begin{array}{rll} \text{primeira prova:} & \overline{X} = 72,8 & S_X = 8,1 \\ \text{médias finais:} & \overline{Y} = 76,4 & S_Y = 7,0 \end{array}$$

a. Verifique que a seguinte fórmula é uma alternativa para se estimar β_1: $\widehat{\beta}_1 = r_{X,Y}\frac{S_Y}{S_X}$.

b. Ajuste uma equação de **regressão linear simples** e interprete a estimativa do coeficiente da regressão $b_1 = \widehat{\beta}_1$

c. Sabendo-se que a média final mínima para ser aprovado no curso é $Y = 60$, qual deve ser a decisão de um aluno que obteve $X = 55$ como nota da primeira avaliação? Continuar no curso ou desistir do curso? **justifique sua resposta com base na equação ajustada no item (a.)**.

d. Qual é a proporção da variabilidade nas notas médias finais explicada pela regressão nas notas da primeira prova ?

4 (II/2003). Um economista interessado em estudar a relação entre o valor da renda familiar extra (X) ou disponível para gastos extras (chamada de *disposable income* na literatura em inglês) e o valor dos gastos com alimentação (Y) conduziu um estudo preliminar com 8 famílias, todas compostas por marido, esposa e dois filhos. Os resultados estão na tabela a seguir com valores X em milhares de dólares por ano e Y em centenas de dólares por ano.

X	30	36	27	20	16	24	19	25	SQD_X=291,88	$\overline{X} = 24,63$
Y	55	60	42	40	37	26	39	43	SQD_Y=783,50	$\overline{Y} = 42,75$

a. Ajuste a equação de regressão linear simples.

b. Interprete o valor do coeficiente da regressão (β_1) em termos do problema anunciado.

c. Calcule o coeficiente de determinação e interprete o valor calculado.

5 (proposto por E.B., monitor em 2001). Pode-se determinar o teor de proteínas (mg/ml) no leite de uma forma indireta analisando-se a absorvância de luz, medida em um aparelho denominado fotocolorímetro. A absorvância consiste na fração da

luz incidente que a amostra é capaz de absorver. Por exemplo, uma absorvância de 0,70 indica que a solução absorveu 70% da luz incidente. Por razões históricas, este método é denominado Método do Biureto. A tabela a seguir apresenta os resultados obtidos em um teste com cinco amostras padrão, de concentração previamente conhecida.

Conc. (mg/ml)	1,00	2,00	3,00	4,00	5,00
Absorvância	0,12	0,31	0,49	0,64	0,77

Pede-se:

a. Ajuste a reta de regressão linear simples para estimar a absorvância (Y) em função da concentração de proteínas (X).

b. Interprete o valor da estimativa do coeficiente de regressão (b_1).

c. Para uma absorvância igual a 0,58 estime a concentração média de proteínas, em mg/ml (regressão inversa).

d. Pode-se utilizar o modelo ajustado em **a.** para se estimar Y quando $X = 8,00$? explique.

e. Calcule o coeficiente de determinação e interprete.

6 (II/2005). Exemplo extraído de D.S. Falconer, Introdução à genética quantitativa, 1^a edição. Os dados abaixo ilustram o efeito do gene anão em ratos com 6 semanas de idade, sendo X o número de genes e Y o peso médio dos ratos em gramas. O objetivo é relacionar as duas variáveis com um modelo de regressão linear simples (RLS): $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$.

i	1	2	3
X_i	0	1	2
Y_i	15	12	6

a. Apresente a equação de RLS ajustada.

b. Interprete as estimativas dos parâmetros β_0 e β_1.

c. Apresente os desvios da regressão e mostre que a soma deles é igual a zero.

d. Calcule e interprete o coeficiente de determinação.

7 (I/2006). A eficiência de uma enzima utilizada no processo de fabricação de medicamentos é avaliada pela quantidade do princípio ativo do medicamento que é

produzido na reação química catalizada pela enzima. Considere que a quantidade do princípio ativo (Y, em mg/kg do soluto) produzido em função da concentração do soluto (X, em g/kg do solvente) pode ser explicado por um modelo de regressão linear simples: $Y_i = \beta_1 X_i + \varepsilon_i$, ($\beta_0 = 0$). Os resultados obtidos por uma empresa que conduziu testes com duas enzimas, A e B, estão apresentados na tabela a seguir. Note que não ocorre reação química e portanto nenhum princípio ativo é produzido quando não há soluto.

ENZIMA	MODELO AJUSTADO	$r^2(\%)$
A	$\widehat{Y}_i = 6,5X_i$	91
B	$\widehat{Y}_i = 10,2X_i$	98

a. Qual das duas enzimas foi a mais eficiente nos testes? Justifique sua resposta com base nos modelos ajustados.

b. O modelo ajustado explicou melhor o fenômeno estudado para qual das duas enzimas? Justifique sua resposta.

c. Quando a concentração do soluto for igual a 20g/kg do solvente e a reação for catalizada pela enzime A, qual é a estimativa da quantidade média do princípio ativo produzida (mg/kg do soluto)?

8 (II/2006). O preço de um modelo de motocicleta usada está linearmente relacionado ao ano de fabricação. A tabela a seguir apresenta os valores do preço (em milhares de reais, $R\$ \times 1000$) e o respectivo ano de fabricação (1993 a 1999, exceto 1996) de 6 motocicletas pesquisadas,

Motocicleta (i)	1	2	3	4	5	6
Ano (X_i)	93	94	95	97	98	99
Preço (Y_i)	6,3	7,0	8,2	9,0	10,5	12

Pede-se:

a. A estimativa do acréscimo médio no preço da motocicleta, para cada aumento de um ano (mais nova), é igual a $R\$$.....

b. O percentual do valor da variância observada nos preços, representado pelo valor da variância dos preços estimados, ou *explicado* pela regressão nos valores do ano de fabricação, é igual a .....%.

c. $R\$$.....é o preço mediano das motocicletas pesquisadas.

d. R\$..... é a amplitude total dos preços das motocicletas pesquisadas..

e. Estimar que o preço médio de uma motocicleta ano 1992 seja igual a R\$..... seria uma ..... com o modelo.

f. Os estimadores b_0 e b_1 foram obtidos pelo método..... ....

g. ..... é o desvio da regressão para o ano 1997.

h. O ..... do modelo, indicado por ε_i, é não observável e representa o efeito de todas as variáveis explicativas não incluídas no modelo, além das causas não controláveis de variação.

i. R\$..... é uma estimativa do preço médio de uma motocicleta 1996.

9 (II/2006). Estudou-se o relacionamento entre o tempo de uma reação química, expresso em minutos (Y) e o valor da concentração, expressa em %, de um composto ativador da reação (X). Os valores testados para X variaram de 0% a 51%, tendo este último valor causado reação *instantânea*. O estudo possibilitou o ajuste da seguinte equação de regressão linear simples,

$$\widehat{Y}_i = 10,2 - 0,20X \qquad r^2 = 0,9362$$

a. Interprete a estimativa do coeficiente da regressão.

b. Interprete a estimativa da constante da regressão.

c. Interprete a estimativa do coeficiente de determinação.

10 (I/2007). (Exemplo obtido de http://statmaster.edu.dk) Em um estudo sobre nutrição infantil em países em desenvolvimento, avaliou-se mensalmente as alturas (Y, em cm) de crianças com 18 a 30 meses de idade (X, em meses) da vila de Kalama no Egito. O objetivo do estudo era modelar por regressão linear simples (RLS), o relacionamento entre idade e altura com o propósito de compará-lo com outros países investigados no estudo.

X	18	19	20	21	22	23	24	25	26	27	28	29	30
Y	76,1	77,0	78,1	78,2	78,8	79,7	79,9	81,1	81,2	81,8	82,8	83,5	84,6

Pede-se:

a. Informe os valores das somas a seguir, $\sum X$ $\sum X^2$ $\sum Y$ $\sum Y^2$ e $\sum XY$.

b. Estime o acréscimo médio na altura, para cada aumento de um mês na idade.

c. Calcule o percentual do valor da variância observada nas alturas, representado pelo valor da variância das alturas estimadas pelo modelo de RLS, ou *explicado* pela regressão nos valores das idades.

d. Calcule os desvios da regressão para as idades 18 e 30 meses.

RESPOSTAS

1. a. $\widehat{\varepsilon}_i$ são os desvios da regressão, valores estimados após o ajuste do modelo, ε_i são os erros aleatórios, não observáveis, do modelo e se referem aos efeitos de todas as fontes de variação não consideradas no modelo, essencialmente outras variáveis explicativas e causas aleatórios não controláveis. **b.** trabalhe por propriedades de $\sum$ até obter $\sum \widehat{\varepsilon} = \sum(Y - \overline{Y}) + b_1 \sum(X - \overline{X})$.

2. a.(F) **b.**(V) **c.**(F) **d.**(V) **e.**(F)

3. a. $b_1 = \frac{SPD_{XY}}{SQD_X} = \frac{r_{XY}\sqrt{SQD_X\ SQD_Y}}{SQD_X} = \frac{r_{XY}\sqrt{S_X^2\ S_Y^2\ (n-1)^2}}{S_X^2\ (n-1)} = r_{XY}\frac{S_Y}{S_X}$ **b.** $\widehat{Y}_i = 30,47 + 0,63X_i$. A estimativa $b_1 = 0,63$ significa que para cada ponto obtido na primeira prova estima-se um aumento médio de 0,63 pontos na média final. **c.** $\widehat{Y} = 30,47 + 0,63(55) \approx 65,1$. Deve continuar pois a média final estimada é superior a 60. **d.** $r^2 = 0,73^2 = 0,5329$ ou $53,29\%$. r^2 é o coeficiente de determinação e r é o coeficiente de correlação.

4. a. $\widehat{Y}_i = 12,8 + 1,2X_i$ **b.** Estima-se aumento médio de 120 dólares nos gastos com alimentos para cada 1000 dólares de aumento na renda extra. **c.** $r^2 = 54,88\%$ é o percentual da variabilidade observada nos gastos sendo *explicada* pela RLS nos valores de renda extra.

5. a. $\widehat{Y}_i = -0,023 + 0,163X_i$ **b.** Para cada aumento de 1 mg/ml na conc. de proteína estima-se aumento médio de 0,163 ou 16,3% na absorvância. **c.** $\widehat{X}_i = \frac{0,023}{0,163} + \frac{1}{0,163}Y_i$ portanto $b_0^* = 0,141$ e $b_1^* = 6,135$ fornece $\widehat{X} = 3,699$ mg/ml **d.** Sim, $\widehat{Y}_i = -0,023 + 0,163 \times 8 = 1,281$ ou $128,1\%$ mas além de ser uma extrapolação, o valor estimado supera 100% **d.** $r^2 = 99,4\%$ é o percentual da variabilidade observada nos valores da absorvância *explicado* pela RLS nos valores da conc. de proteínas.

6. a. $\widehat{Y}_i = 15,5 - 4,5X_i$ **b.** $b_0 = 15,5$ gramas é uma estimativa do peso médio dos ratos que não possuem o gene anão e $b_1 = -4,5$ é uma estimativa do decréscimo médio no peso para cada um gene anão de acréscimo. **c.** $\widehat{\varepsilon}_1 =$

$15 - 15,5 = -0,5$, $\widehat{\varepsilon}_2 = 12 - 11 = 1$ e $\widehat{\varepsilon}_3 = 6 - 6,5 = -0,5$, portanto $\sum_{i=1}^{3} \widehat{\varepsilon}_i = 0$ **d.** $r^2 \approx 96,4\%$ é o percentual da variabilidade nos valores de peso sendo *explicados* pela RLS nos valores do número de genes anão.

7. a. Enzima B, por apresentar maior valor b_1, o que significa maior aumento médio estimado do P.A. para cada aumento de uma unidade do soluto **b.** Enzima B, maior r^2 **c.** $\widehat{Y} = 6,5 \times 20 = 130$ mg/Kg do soluto.

8. a. $R\$890,00$ **b.** $96,22\%$ **c.** $R\$8600,00$ **d.** $R\$5700,00$ **e.** $R\$5276,00$ seria uma extrapolação **f.** dos mínimos quadrados **g.** -0,72 **h.** erro aleatório **i.** $R\$8830,00$.

9. a. $b_1 = -0,20$, para cada acréscimo de 1% na conc. do composto, estima-se uma diminuição média de 0,20 minutos no tempo da reação (aumento de velocidade) **b.** $b_0 = 10,2$, estima-se um tempo médio de 10,2 minutos quando nenhum composto (0%) é utilizado **c.** $r^2 = 93,62\%$ é o percentual da variabilidade observada nos valores do tempo de reação que foram explicados pela RLS nos valores da concentração do composto.

10. a. $\sum X = 312$ $\sum X^2 = 7670$ $\sum Y = 1042,8$ $\sum Y^2 = 83727,74$ e $\sum XY = 25146,5$. **b.** $b_1 = 0,6555$ **c.** $r^2 = 98,8\%$ **d.** $\widehat{\varepsilon}_{18} = -0,1824$ e $\widehat{\varepsilon}_{30} = 0,4516$