

Introdução à Probabilidade e Estatística

Ficha N°7: Regressão Linear Simples

Para as licenciaturas em: Eng. Civil, Eng. das Energias Renováveis, Eng. Geológica, Eng. Informática e Eng. Mecatrónica

2º semestre 2014/15 — 2h Teóricas + 2h Práticas

Docentes: Patrícia Filipe e Ana Isabel Santos

1. **(Exame Época Normal — 11 de Junho de 2013)** Pretende-se modelar a resistência de um determinado tipo de plástico em função do tempo que decorre a partir da conclusão do processo de moldagem até ao momento de medição de resistência (horas). Foram registadas as observações correspondentes a 12 peças construídas com este plástico, escolhidas aleatoriamente. Com base nos resultados apresentados abaixo, responda às seguintes questões:

Statistics		
	Resistência	Tempo até medição da resistência
Mean	269,92	48,00
Variance	2125,356	349,091

Coefficients ^a					
Model	Unstandardized Coefficients		Sig.	95,0% Confidence Interval for B	
	B	Std. Error		Lower Bound	Upper Bound
(Constant)	153,917	8,067		135,943	171,890
1 Tempo até medição da resistência	2,417	,157	,000		

a. Dependent Variable: Resistência

- Indique a recta de regressão estimada e interprete os seus coeficientes.
 - Determine e interprete os coeficientes de correlação e de determinação.
 - Ao nível de significância de 10% pode concluir que a recta de regressão não passa pela origem?
 - Diga se existe evidência de que o tempo desde a conclusão do processo de moldagem influencia linearmente de forma significativa a resistência.
 - Decorridas 35 horas desde a conclusão do processo de moldagem, qual o valor da resistência previsto?
2. **(Exame Época Normal — 14 de Junho de 2012)** Uma empresa que avalia a qualidade de água das ETAR, analisa o nível de substâncias nocivas detectadas na água (mg/100ml) de modo a classificar a água nas duas categorias: água rejeitada

e água em boas condições. Para tal, tenta analisar a possível relação linear entre o nível de substâncias nocivas detectadas na água (mg/100ml) e o nível de substâncias químicas utilizadas no tratamento da mesma água (mg/100ml). Admita que existe relação linear entre as variáveis.

Descriptive Statistics			
	N	Mean	Variance
Nível de substâncias químicas utilizadas no tratamento da água	100	4,20	2,800
Nível de substâncias nocivas detectadas na água	100	3,80	1,900

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.7 ^a	,530	,520	,240

a. Predictors: (Constant), Nível de substâncias químicas utilizadas no tratamento da água

Coefficients ^a						
Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	?	,188		?	,000
	Nível de substâncias químicas utilizadas no tratamento da água	?	,240	?	?	?

a. Dependent Variable: Nível de substâncias nocivas detectadas na água

- Qual a equação da recta de regressão estimada? Interprete os coeficientes de regressão estimados.
- Determine os coeficientes de regressão e de determinação e interprete-os.
- Ensaie, ao nível de significância de 5%, a hipótese da recta de regressão passar na origem.
- Ao nível de significância de 5%, teste se o nível de substâncias químicas utilizadas no tratamento da água influencia linearmente o nível de substâncias nocivas detectadas na água.
- Sabendo que um dos nível de substâncias nocivas detectadas na água foi de 3,21 mg/100ml e que o nível de substâncias químicas utilizadas no tratamento da água foi de 2,64 mg/100ml, calcule o resíduo de estimação.
- Por cada mg/100ml a mais no nível de substâncias químicas utilizadas no tratamento da água, quanto aumenta (ou diminui) o nível de substâncias nocivas detectadas na água?

Page 1

3. (Exame Época Recurso — 27 de Junho de 2012) O Presidente da Junta de Freguesia da pacata vila alentejana (e dono da única gasolinera) pretende averiguar se a quantidade de gasolina vendida por dia depende, ou não, do seu preço/l. Assim, com base nos últimos 25 meses (em que em cada mês se seleccionou um dia ao acaso para a recolha das amostras) e admitindo que existe relação linear entre as variáveis, colocaram-se as seguintes questões:

- Escreva a equação da recta de regressão estimada e interprete os coeficientes de regressão estimados.

- (b) Qual a percentagem de variação da quantidade de gasolina vendida por dia em função do seu preço?
- (c) Com base nos dados que conclusão transmitiria ao Presidente da Junta relativamente à sua questão? Ou seja, poderá transmitir-lhe que a associação entre o preço da gasolina e a quantidade vendida estão relacionados de modo significativo? Justifique adequadamente a sua resposta.
- (d) Estime a quantidade de gasolina vendida por dia quando o seu preço por litro atingir os 2 euros.

Descriptive Statistics			
	N	Mean	Variance
Preço da gasolina (por litro)	25	1,65	1,500
Quantidade de gasolina vendida (por dia)	25	37,00	3,500

Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	0,96 ^a	?	,520	,240

a. Predictors: (Constant), Preço da gasolina, por litro

Coefficients ^a					
Model		Unstandardized Coefficients		Standardized Coefficients	
		B	Std. Error	Beta	
1	(Constant)	?	,232		?
	Preço da Gasolina (por litro)	?	,354	?	?

a. Dependent Variable: Quantidade de gasolina vendida por dia

4. Pretende-se obter um modelo que possa explicar a área foliar (em cm^2) através do comprimento da nervura principal (em cm) em folhas de videiras da casta Fernão Pires. Para tal recolheu-se uma amostra aleatória de 12 folhas, tendo-se obtido para cada uma delas os seguintes valores das 2 variáveis acima referidas:

Descriptive Statistics			
	Mean	Std. Deviation	N
Área Foliar (cm^2)	175,42	30,776	12
Nervura principal (cm)	11,617	1,2298	12

Model Summary			
Model	R Square	Adjusted R Square	Std. Error of the Estimate
1	0,940	0,934	7,887

Admita que existe relação linear entre as variáveis. Considere ainda que $Var[\hat{\beta}_0] = 509,77$ e que $Var[\hat{\beta}_1] = 3,74$.

- (a) Qual a equação da recta de regressão estimada? Interprete os coeficientes de regressão estimados.
- (b) Qual o valor do coeficiente de determinação e interprete-o.
- (c) Por cada 2 cm a mais no comprimento da nervura principal, quanto aumenta a área foliar da folha de videira?
- (d) Ao nível de significância de 5%, teste se o comprimento da nervura principal influencia linearmente a área foliar.

- (e) Obtenha o intervalo de confiança a 99% para o parâmetro que representa a ordenada na origem. (**Nota:** Caso não tenha determinado a equação da recta, considere valores hipotéticos para β_0 e β_1 .)
- (f) Sabendo que numa folha a nervura principal media 9,1 cm de comprimento e a área foliar era 126 cm^2 , calcule o resíduo de estimação.
5. Para alguns países da Europa, foram registados alguns indicadores económicos, nomeadamente o PIBA (produto interno bruto, originado pela agricultura) e o PURB (percentagem de população urbana). Os valores apresentam-se na tabela seguinte:

Países	PIBA	PURB
Alemanha	2	85
Áustria	4	54
Bélgica	2	72
Dinamarca	4	84
Espanha	8	74
Finlândia	0	62
França	4	78
Grécia	16	62
Holanda	4	76
Itália	6	69
Noruega	5	53
Portugal	13	31
Reino Unido	2	91
Suécia	3	87

Considere que $\sum_{i=1}^{14} y_i = 73$, $\sum_{i=1}^{14} x_i = 978$, $\sum_{i=1}^{14} y_i^2 = 635$, $\sum_{i=1}^{14} x_i^2 = 71786$ e $\sum_{i=1}^{14} x_i y_i = 4591$.

- (a) Estime da recta de regressão entre as variáveis PIBA e PURB.
- (b) Ensaie a hipótese do declive da recta de regressão ser igual a -0,2 (ao nível de significância de 5%).
- (c) Qual a percentagem de variabilidade da variável Y que é explicada pela variável X ?
6. No gráfico de dispersão abaixo encontram-se registados os tempos (em segundos) do vencedor na final da corrida de 400 metros masculinos, em cada Olimpíada entre 1896 e 2004 (i.e., dos 25 anos em que as Olimpíada da era moderna se realizaram): Considere os seguintes outputs que foram obtidos recorrendo ao software estatístico SPSS.

Vencedor 400m masculino

N	Valid	25
	Missing	0
	Mean	46,306
	Median	45,15
	Std. Deviation	2,66223
Percentiles	25	44,13
	50	45,15
	75	48

Data dos Jogos Olímpicos

N	Valid	25
	Missing	0
	Mean	1952
	Median	1956
	Std. Deviation	34,059
Percentiles	25	1922
	50	1956
	75	1982

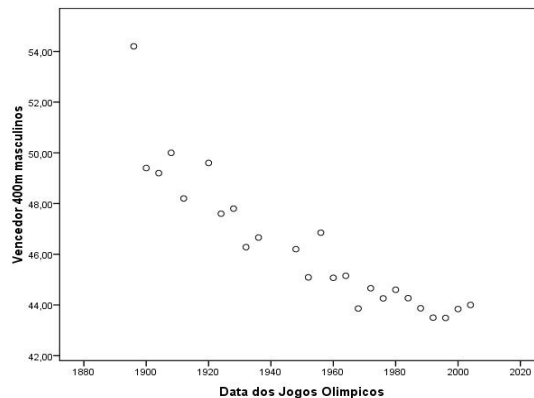


Figure 1: Gráfico de Dispersão (Scatter plot).

Considere ainda que $\sum_{i=1}^{25} x_i y_i = 2257753$.

- Diga, destas duas variáveis, qual a variável dependente e qual a variável explicativa?
 - A relação entre as duas variáveis será positiva ou negativa? Porquê?
 - Determine o coeficiente de determinação e interprete-o.
 - Construa o modelo de regressão linear simples e interprete as estimativas dos parâmetros do modelo.
 - Obtenha uma previsão para o tempo do vencedor, na referida prova, para a Olimpíada de 2008.
7. Considerando a idade (em anos) e a capacidade pulmonar (em L) de 9 crianças, observou-se a seguinte tabela:

Idade (X)	4	5	6	7	8	9	10	11	12
Cap. Pulmonar (Y)	0,7	0,9	1,2	1,3	1,3	1,5	1,7	1,9	2,1

Com o auxílio do software estatístico SPSS obtiveram-se os seguintes resultados:

	Unstandardized	Coefficients	
Model	B	Std. Error	t
1 (constant)	0,093	0,081	1,147
Idade, em anos	0,163	0,010	16,878

- Calcule os coeficientes de correlação e determinação e interprete-os.
- Apresente a recta de regressão dos mínimos quadrados. Interprete o declive estimado.
- Determine o intervalo de 95% de confiança para o declive da recta.
- Ensaie a hipótese da recta de regressão passar na origem, ao nível de significância de 1%.

8. Pretende-se, se possível, modelar através de uma recta de regressão linear simples a quantidade de vidro, Y , produzido num Ecoponto (Kg), usando como variável independente, X , o número de dias sem despejar o mesmo. Para tal, registaram-se os seguintes dados:

x_i	2	3	4	5	10	15	20	25
y_i	100	150	250	320	650	810	1040	1480

Dos quais resulta $\sum_{i=1}^8 x_i y_i = 79700$.

Nº dias sem despejar o Ecoponto			Quantidade de vidro num Ecoponto (Kg)		
N	Valid	8	N	Valid	8
	Missing	0		Missing	0
	Mean	10,5		Mean	600
	Median	7,5		Median	485
	variance	74,571		Variance	238285,7

- (a) Escreva a recta de regressão estimada através do método dos mínimos quadrados.
- (b) Diga, justificando convenientemente, se o modelo encontrado se ajusta bem ou não aos dados reais.
- (c) Teste a hipótese de o declive da recta de regressão ser nulo, ao nível de significância de 10%.
- (d) Qual o valor da quantidade de vidro produzida no Ecoponto que prevê ocorrer em 10 dias sem o despejar. Seria possível calcular o mesmo para um período de 35 dias?
9. Mediu-se o número de pulsações por minuto antes e depois de uma determinada prova de esforço, num grupo de 10 fumadores com mais de 20 anos consecutivos de consumo de tabaco, tendo-se obtido os seguintes resultados:

Pulsação antes (x_i)	49	50	52	53	55	55	61	62	64	70
Pulsação depois (y_i)	98	95	97	91	96	100	110	109	115	120

Com base nos outputs da folha seguinte:

- (a) Calcule os coeficientes de correlação e determinação e interprete-os.
- (b) Apresente a recta de regressão. Interprete o declive estimado.
- (c) Determine o intervalo de 95% de confiança para o declive da recta.
- (d) Ensaie a hipótese da recta de regressão passar na origem, ao nível de significância de 1%.
- (e) O que poderá dizer sobre a significância da regressão?

Descriptive Statistics

	Mean	Std. Deviation	N
Y Pulsação Depois	103,100	9,6891	10
X Pulsação Antes	57,10	6,839	10

Correlations

		Y Pulsação Depois	X Pulsação Antes
Pearson Correlation	Y Pulsação Depois	1,000	,934
	X Pulsação Antes	,934	1,000
Sig. (1-tailed)	Y Pulsação Depois	.	,000
	X Pulsação Antes	,000	.
N	Y Pulsação Depois	10	10
	X Pulsação Antes	10	10

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,934 ^a	,872	,856	3,6752

a. Predictors: (Constant), X Pulsação Antes

b. Dependent Variable: Y Pulsação Depois

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	736,844	1	736,844	54,553	,000 ^a
	Residual	108,056	8	13,507		
	Total	844,900	9			

a. Predictors: (Constant), X Pulsação Antes

b. Dependent Variable: Y Pulsação Depois

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	27,550	10,295		2,676	,028
	X Pulsação Antes	1,323	,179	,934	7,386	,000

a. Dependent Variable: Y Pulsação Depois

Coefficients^a

Model		95,0% Confidence Interval for B	
		Lower Bound	Upper Bound
1	(Constant)	3,811	51,290
	X Pulsação Antes	,910	1,736

a. Dependent Variable: Y Pulsação Depois