Licenciatura em Engenharia Informática – DEI/ISEP

Análise de Dados em Informática 2023/2024

Ficha Teórico-Prática 4

Correlação e Regressão Linear

Objetivos:

- Familiarização com a ferramenta R no suporte a problemas relativos a Correlação e Regressão Linear;
- Análise e discussão de resultados.

Exercícios Propostos

1. Observe a seguinte tabela:

TEMPO DE RESPOSTA DE UM MONITOR (ms)		
Marca e modelo	Anunciado	Medido
A	8	4.8
B	12	5
C	16	17,5
D	8	6,4
E	8	6,7
F	8	4,3
G	27	20,3
H	12	8
I	8	3,5
J	16	6,3

Use o coeficiente de correlação de Kendall para verificar se existe alguma relação entre as duas variáveis (use $\alpha = 5\%$).

2. Numa determinada instituição de ensino as classificações de Estatística são categorizadas em 6 níveis: Excelente, Muito Bom, Bom, Suficiente, Insuficiente e Mau. Já no caso do Cálculo, as classificações podem ser A, B, C, D, E e F, por ordem decrescente de valor. No ficheiro "**`Notas.txt`**", encontram-se os registos das classificações obtidas por 25 alunos. Pretende-se estudar se a nota obtida a Cálculo está associada positivamente com

a classificação obtida em Estatística. Use um nível de 5% e o coeficiente de correlação de Spearman para analisar o problema.

3. Um engenheiro informático, responsável pela gestão de um conjunto de servidores, pretende analisar se existe uma correlação entre as falhas de conetividade e o número de acessos diários. Para tal, gerou um script para registar diariamente e durante o período de 1 mês, o número de falhas e o número de acessos a um determinado servidor. Os valores obtidos encontram-se no ficheiro "**`Servidor.csv`**".

a) Construa um diagrama de dispersão dos dados e verifique a existência de uma associação entre as duas variáveis.

b) Calcule um coeficiente de correlação apropriado e conclua, a um nível de $5% $, se existe uma correlação positiva entre as duas variáveis.

4. O ficheiro "**`fang_data`**" contém as contações diárias na bolsa de valores das empresas tecnológicas: Facebook, Amazon, Netflix e Google, do ano 2013 até 2016. Construa a matriz das correlações entre as cotações das 4 empresas e comente os resultados.

5. Considere os seguintes valores observados das variáveis X e Y :

xi	yi
21	185,79
24	214,47
32	288,03
47	424,84
50	454,58
59	539,03
68	621,55
74	675,06
62	562,03
50	452,93
41	369,95
30	273,98

a) Usando um diagrama de dispersão, verifique se existe uma relação linear entre as duas variáveis.

b) Estime os parâmetros da recta de regressão e $y(40)$.

c) Calcule o coeficiente de correlação linear de Pearson e comente os resultados.

d) Determine se os pressupostos relativos aos resíduos se verificam.

6. Na seguinte tabela apresentam-se os montantes dos seguros de vida e os rendimentos anuais, em milhares de unidades monetárias (u.m.), de 12 agregados familiares de certo país.

Rendimento anual (milhares de u.m.)	Capital seguro (milhares de u.m.)
14	31
19	40
23	49
12	20
9	21
15	34
22	54
25	52
15	28
10	21
12	24
16	34

a) Construa o diagrama de dispersão para estes dados e adicione a recta de regressão. Comente os resultados.

b) Estime o montante do seguro de vida para um agregado familiar com rendimento anual de 20000 u.m.

c) Verifique se os resíduos gozam de homocedasticidade e se são independentes.

d) Teste a normalidade dos resíduos.

7. Um engenheiro mecânico pretende analisar o acabamento da superfície das peças de metal produzidas num torno e suspeita que este está dependente da velocidade (em rotações por minuto) do torno e do tipo de ferramenta de corte usada. O ficheiro "**`ExemploMontegomery-12-11.csv`**" contém os dados da amostra recolhida:

a) Apresente um modelo de regressão linear adequado e interprete os seus coeficientes.

b) Estime os parâmetros da reta de regressão.

c) Calcule o coeficiente de determinação ajustado e comente os resultados.

d) Determine se os pressupostos relativos aos resíduos se verificam.

e) Verifique se existe multicolinearidade.

Exercícios de Consolidação

1. Os seguintes dados mostram o volume de produção de trigo, em milhares de toneladas, numa dada região entre 1986 e 1994.

Ano	Volume de produção
1986	285
1987	270
1988	294
1989	279
1990	260
1991	262
1992	258
1993	272
1994	255

a) Construa o diagrama de dispersão e acrescente a recta de regressão correspondente. Comente os resultados.

b) Calcule os coeficientes de determinação e de correlação. Comente os resultados.

c) Usando o teste de Durbin-Watson, verifique a independência dos resíduos.

2. O ficheiro "**`regressao_exerc9.txt`**" contém dados de 9 variáveis relacionadas com 517 incêndios ocorridos no parque nacional de montesinho. Considere a variável dependente *area* sendo as restantes 8 variáveis independentes (*FFMC*, *DMC*, *DC*, *ISI*, *tem*, *RH*, *wind*, *rain*):

a) Estime o modelo de regressão linear múltipla.

b) Determine se os pressupostos relativos aos resíduos se verificam.

c) Verifique se existe multicolinearidade.

d) Encontre um modelo mais simples com menor coeficiente de informação de Akaike (AIC)

3. O ficheiro "**`Covid19.csv`**" contém os dados relativos ao número total de infetados em Portugal, pelo covid19 no período de 3 a 28 de março de 2020.

a) Faça um diagrama de dispersão e indique se há uma relação linear entre as duas variáveis.

b) Faça a mudança de variáveis, *Z=log(nº de infetados)*. Faça um diagrama de dispersão (entre as variáveis *Dia* e *Z*) e indique se há uma relação linear entre as duas variáveis.

c) Encontre a reta de regressão, $Z = m.Dia + b$, faça um plot dos dados e a reta de regressão. Verifique se os resíduos são normais (com média zero), são independentes e se são homocedásticos. Estime o número de infetados nos dias 29 e 30 de março.