

## Exercícios

### Ano letivo 2024/25

#### 1. Correlação

1.1. O departamento de qualidade de um determinado hospital, pretende fazer um estudo sobre o tempo que os utentes encaminhados para cirurgia pelos seus médicos de família demoram a ser efetivamente operados. Estes utentes têm que primeiro fazer uma consulta de referência no hospital ( $x$  dias após o encaminhamento) e só posteriormente são operados ( $y$  dias após a consulta de referência). Escolheram-se aleatoriamente 15 utentes nestas condições, tendo sido reportados os seguintes tempos de espera, também disponíveis em *tempos.csv* (fonte: Afonso e Nunes, 2019):

$x$	69	76	51	34	62	13	40	7	64	41	64	26	40	44	48
$y$	28	64	7	26	38	18	40	20	44	32	31	32	36	25	73

- Através da análise gráfica, parece-lhe existir relação linear entre as duas variáveis?
- Calcule e interprete o valor do coeficiente de correlação linear de Pearson.

1.2. (📄) Observou-se uma amostra de 26 modelos de *machine learning* tendo-se registado para cada um deles o tempo de execução (em segundos) e o número de parâmetros, tendo-se obtido os seguintes resultados, que também estão disponíveis em *ML.csv*:

Modelo	N.º parâmetros	Tempo	Modelo	N.º parâmetros	Tempo
1	49	71,66	14	21	34,88
2	15	20,3	15	22	35,23
3	1	1,46	16	39	61,27
4	35	52,2	17	17	21,35
5	7	10,24	18	6	9,14
6	11	18,41	19	43	65,95
7	9	10,37	20	1	4,95
8	43	64,23	21	39	57,82
9	32	49,5	22	45	67,29
10	50	75,15	23	28	44,98
11	49	77,32	24	14	24,88
12	19	28,4	25	25	44,56
13	19	24,87	26	17	23,6

- Considera que o tempo de execução está relacionado com o número de parâmetros? Justifique com auxílio de um gráfico.
- Calcule a covariância entre o tempo de execução e o número de parâmetros.
- Calcule o coeficiente de correlação de Pearson entre o tempo de execução e o número de parâmetros.
- Repita as alíneas anteriores, considerando o tempo em minutos. Pronuncie-se sobre as diferenças obtidas.

1.3. (🖨️) Os dados apresentados na tabela seguinte foram extraídos do Relatório do Desenvolvimento Humano, de 2023/24, publicado pelo Programa das Nações Unidas para o Desenvolvimento (link: [Relatório do Desenvolvimento Humano \(RDH\) 2023-2024 | United Nations Development Programme](#)). Para os 27 países da União Europeia foram calculados: o Índice de Desenvolvimento Humano (IDH), o IDH ajustado às desigualdades (IDHAjD), o Índice de Desenvolvimento do Género (IDG), o Índice de Desigualdade de Género (IDiG), e o IDH ajustado à pressão planetária (IDHAjPP). Os dados estão disponíveis no ficheiro *RDH2324.xlsx*.

País	IDH	IDHAjD	IDG	IDiG	IDHAjPP
Alemanha	0,950	0,881	0,966	0,071	0,833
Áustria	0,926	0,859	0,972	0,048	0,789
Bélgica	0,942	0,878	0,975	0,044	0,803
Bulgária	0,799	0,703	0,995	0,206	0,72
Chéquia	0,895	0,848	0,988	0,113	0,782
Chipre	0,907	0,827	0,977	0,253	0,803
Croácia	0,878	0,817	0,993	0,087	0,807
Dinamarca	0,952	0,898	0,981	0,009	0,839
Eslováquia	0,855	0,808	1,002	0,184	0,776
Eslovénia	0,926	0,882	0,999	0,049	0,832
Espanha	0,911	0,796	0,988	0,059	0,839
Estónia	0,899	0,835	1,022	0,093	0,766
Finlândia	0,942	0,886	0,989	0,032	0,787
França	0,910	0,820	0,986	0,084	0,823
Grécia	0,893	0,801	0,969	0,120	0,809
Hungria	0,851	0,800	0,989	0,230	0,769
Irlanda	0,950	0,886	0,991	0,072	0,814
Itália	0,906	0,802	0,969	0,057	0,825
Letónia	0,879	0,802	1,022	0,142	0,782
Lituânia	0,879	0,795	1,028	0,098	0,748
Luxemburgo	0,927	0,839	0,993	0,043	0,685
Malta	0,915	0,837	0,980	0,117	0,806
Países Baixos	0,946	0,885	0,960	0,025	0,796
Polónia	0,881	0,797	1,009	0,105	0,78
Portugal	0,874	0,774	0,998	0,076	0,807
Roménia	0,827	0,739	0,981	0,230	0,759
Suécia	0,952	0,878	0,983	0,023	0,839

- Quais os índices que lhe parecem estar mais correlacionados? Confirme o seu palpite calculando os respetivos coeficientes de correlação.
- Se recorresse ao coeficiente de Spearman em vez do coeficiente de Pearson, parece-lhe que a avaliação da associação entre as variáveis se iria alterar substancialmente? Justifique.

## 2. Regressão linear simples

2.1. Pretende-se verificar se, nos jogos do campeonato do Mundo de 2018, o número de remates dependia do n.º de cruzamentos efetuados por jogo. Para tal selecionaram-se ao acaso 9 jogos, tendo-se obtido os seguintes resultados.

Número de remates	19	15	12	16	11	28	22	8	16
Número de cruzamentos	45	25	22	32	28	56	38	10	32

- Através da análise gráfica, considera que existe relação linear entre o número de remates à e o número de cruzamentos? Justifique.
- Com base na nuvem de pontos, sugira um valor para o coeficiente de correlação linear entre o número de remates à e o número de cruzamentos. Confirme o seu palpite calculando o valor do coeficiente e comente o valor obtido.
- Qual a equação da reta de regressão ajustada pelo método dos mínimos quadrados?
- Interprete os coeficientes de regressão estimados.
- Represente a reta de regressão ajustada em cima da nuvem de pontos e comente.
- Calcule a Soma dos Quadrados Totais ( $SQT$ ), a partir do cálculo da variância amostral de  $y$ .
- Calcule a Soma dos Quadrados da Regressão ( $SQR$ ), a partir do cálculo da variância amostral dos valores estimados para  $y$ .
- Calcule a Soma dos Quadrados dos Resíduos ( $SQRE$ ), a partir do cálculo da variância amostral dos resíduos.
- Verifique numericamente a relação:  $SQT = SQR + SQRE$ .
- Analise a qualidade estatística do modelo estimado.

2.2. Durante o desenvolvimento de um novo medicamento para alergias, foi realizada uma experiência para estudar o efeito de diferentes dosagens, no período de tempo que os doentes se libertam dos sintomas alérgicos. Foram incluídos na experiência 10 pacientes. A cada um foi dada uma dosagem específica do medicamento, e foi-lhes pedido que comunicassem, de imediato, assim que o efeito desaparecesse.

As dosagens ( $x$ ) foram medidas em miligramas e o tempo de duração do medicamento ( $y$ ) em dias. Os resultados obtidos foram os seguintes (fonte: Afonso e Nunes, 2019):

$$\sum_{i=1}^{10} x_i = 59; \quad \sum_{i=1}^{10} y_i = 151; \quad \sum_{i=1}^{10} x_i^2 = 389; \quad \sum_{i=1}^{10} y_i^2 = 2651; \quad \sum_{i=1}^{10} x_i y_i = 1003.$$

- Construa o modelo de regressão linear simples que explique a duração do efeito do medicamento em função da sua dosagem.
- Determine os coeficientes de correlação e de determinação. Interprete.
- Qual a previsão para o número de dias que um paciente se liberta dos sintomas alérgicos, se lhe forem administrados 6,5 mg do medicamento.

2.3. Suponha que se utiliza um novo método para determinar o montante de magnésio na água do mar. Se o método for bom, haverá uma forte relação entre o montante real de magnésio na água do mar e o montante indicado por este novo método. Foram preparadas 10 amostras de água do mar, cada uma contendo um montante conhecido de magnésio a fim de serem testadas pelo novo método.

Os dados desta experiência são apresentados na forma de estatísticas, onde  $x$  representa o montante real de magnésio presente e  $y$  o montante determinado pelo novo método.

$$\sum_{i=1}^{10} x_i = 311; \quad \sum_{i=1}^{10} y_i = 310,1; \quad \sum_{i=1}^{10} (x_i - \bar{x})^2 = 427,9; \quad \sum_{i=1}^{10} (y_i - \bar{y})^2 = 438,89;$$

$$\sum_{i=1}^{10} (x_i - \bar{x})(y_i - \bar{y}) = 429,89.$$

Determine a equação de regressão e o coeficiente de determinação.

2.4. (📄) O ficheiro taxaabandono.csv contém dados do Eurostat sobre a evolução da taxa de abandono escolar precoce, ou seja, a percentagem de pessoas dos 18 aos 24 anos que deixaram de estudar sem terminar o secundário em Portugal, no período de 2002 a 2022 (Fonte: Pordata).

Ano	Taxa	Ano	Taxa	Ano	Taxa
2002	45	2009	34,9	2016	13,9
2003	41,2	2010	30,9	2017	12,6
2004	39,3	2011	22,8	2018	11,6
2005	38,3	2012	20,5	2019	10,5
2006	38,3	2013	18,7	2020	9,1
2007	38,5	2014	17,3	2021	6,4
2008	36,5	2015	13,5	2022	6,3

- Construa uma nuvem de pontos de taxa de abandono vs. ano e comente.
- Com base no gráfico obtido na alínea anterior, dê um palpite para o valor do coeficiente de correlação entre as duas variáveis. Avalie o palpite calculando o valor do coeficiente de correlação. Comente o valor obtido.
- Ajuste, pelo método dos mínimos quadrados, uma reta de regressão para taxa de abandono baseada nos anos. Interprete o significado dos coeficientes de regressão estimados.
- Avalie a qualidade da reta obtida.
- Represente a reta de regressão ajustada em cima da nuvem de pontos e comente.
- Transforme a variável ano num contador dos anos do estudo, i.e., 0 corresponde ao ano 2002, 1 corresponde ao ano 2003, e por aí fora. Ajuste novamente a regressão, após efetuar esta alteração. O que aconteceu aos parâmetros estimados e ao coeficiente de determinação? Comente.

2.5. (📄) Pensa-se que o número de embalagens vendidas de um determinado medicamento genérico ( $y$ ) depende do seu preço ( $x$ ), em euros. Para o efeito observou-se durante 12 semanas os valores destas variáveis, tendo-se obtido os seguintes resultados apresentados no ficheiro *vendas.xlsx* (fonte: Afonso e Nunes, 2019):

$y$	892	1012	1060	987	680	739	809	1275	946	874	720	1096
$x$	1,23	1,15	1,1	1,2	1,35	1,25	1,28	0,99	1,22	1,25	1,3	1,05

- Estime a reta de regressão linear, pelo método dos mínimos quadrados. Interprete os coeficientes de regressão.
- Represente graficamente a nuvem de pontos e a reta ajustada.
- Obtenha os resíduos de estimação.
- Valide os pressupostos subjacentes ao modelo de regressão linear.
- Determine o coeficiente de correlação e interprete o valor obtido.
- Estime a variância do erro do modelo.
- Construa a tabela ANOVA.
- Construa um intervalo de confiança a 99% para  $\beta_0$ .
- Complete: “Com 95% de confiança o verdadeiro valor de  $\beta_1$  situa-se entre ... e ...”.
- A partir de que nível de significância é rejeitada a hipótese do coeficiente  $\beta_0$  ser nulo?
- Ensaie a hipótese de que o preço não influencia linearmente o número de embalagens vendidas (considere  $\alpha = 1\%$ ).
- Determine e interprete o coeficiente de determinação.
- Estime o número esperado de embalagens vendidas quando o preço de cada embalagem é 1,23 euros. Construa um intervalo de confiança a 95% para esse valor esperado.
- Construa um intervalo de predição (95%) associado ao número de embalagens quando o preço é 1,23 euros. Compare com o intervalo de confiança obtido na alínea anterior e comente.

2.6. (📄) Uma empresa de *e-commerce* pretende verificar se um maior investimento em publicidade *online* está associado a um aumento nas vendas mensais, ambas em milhares de euros. Para tal foi recolhida uma amostra de 50 meses, tendo-se obtido os dados disponíveis no ficheiro *publicidade.csv*.

- Desenhe a nuvem de pontos e comente.
- Ajuste uma regressão linear simples das vendas mensais ( $y$ ) sobre o investimento em publicidade ( $x$ ). Interprete os coeficientes de regressão e comente o coeficiente de determinação obtido.
- Obtenha estimativas das variâncias e dos desvios-padrão dos estimadores dos parâmetros da reta populacional  $\beta_0$  e  $\beta_1$ .
- Admitindo a validade do modelo:
  - Utilize um teste de hipóteses sobre o declive da reta populacional  $\beta_1$  para validar a seguinte afirmação: “não existe uma relação linear significativa entre o investimento e as vendas mensais”.
  - Teste, com um nível de significância de 0.01, a hipótese de que “por milhar de euros investido a mais, as vendas crescem em média 2,05 milhares de euros”. Mantém a sua decisão ao nível de significância de 0.05?

- iii) Teste, com um nível de significância de 0.01, a hipótese de que “por milhar de euros investido a mais, as vendas crescem em média menos de 2,05 milhares de euros”
- e) Numa reunião foram apresentadas 3 propostas de investimento para o próximo mês: 2100, 2150 e 2300 milhares de euros. Para cada uma destas propostas, calcule:
  - i) O valor estimado para a venda mensal.
  - ii) Um intervalo de confiança a 90% para o valor esperado da venda mensal associado às várias propostas de investimento.
  - iii) Um intervalo de predição a 90% para o valor da venda mensal individual.
- f) Valide os pressupostos do modelo de regressão linear ajustado, analisando o comportamento dos resíduos e restantes ferramentas de diagnóstico. Comente.

2.7. A tabela seguinte representa o número  $y$  de bactérias, por unidade de volume, presentes numa cultura ao fim de  $x$  horas.

$x$ (horas)	1	2	3	4	5	6
$y$ (nº bactérias)	48	66	93	132	190	275

- a) Ajuste aos dados uma exponencial de mínimos quadrados da forma  $Y_i = \alpha_0 e^{\alpha_1 x} \varepsilon_i$  e calcule o coeficiente de determinação.
- b) O ajustamento através de um modelo linear também seria bom? Justifique.
- c) Qual o número estimado de bactérias na cultura ao fim de 8 horas?

2.8. (📊) Considere o conjunto de dados *Animals*, disponível no *package* MASS do R, que tem os pesos médios dos cérebros (em g) e dos corpos (em kg) para 28 espécies de animais terrestres. Pretende-se estudar uma relação entre pesos do cérebro ( $y$ ) e pesos do corpo ( $x$ ).

- a) Desenhe a nuvem de pontos e comente.
- b) Calcule o coeficiente de correlação correspondente e comente.
- c) Construa nuvens de pontos com as seguintes transformações de uma ou ambas as variáveis:
  - i)  $\ln(y)$  vs.  $x$ ;
  - ii)  $y$  vs.  $\ln(x)$ ;
  - iii)  $\ln(y)$  vs.  $\ln(x)$ .
- d) Considere uma relação linear entre  $\ln(y)$  e  $\ln(x)$ . Explique a relação de base correspondente entre as variáveis originais (não logaritmizadas). Comente.
- e) Nas próximas alíneas, considere sempre os dados logaritmizados, i.e,  $\ln(y)$  e  $\ln(x)$ .
  - i) Calcule os coeficientes de correlação e de determinação associados à relação entre  $\ln(y)$  e  $\ln(x)$ . Interprete os valores obtidos. Como se explica que o Coeficiente de Determinação não seja particularmente elevado, sendo evidente a partir da nuvem de pontos que existe uma boa relação linear entre log-peso do corpo e log-peso do cérebro para a generalidade das espécies?
- f) Ajuste a reta de regressão de log-peso do cérebro sobre log-peso do corpo (utilizando a totalidade das observações). Interprete o valor obtido para o declive da reta, do ponto de vista biológico, quer na relação entre variáveis logaritmizadas, quer na relação entre as variáveis originais (não logaritmizadas).

- g) Considere a nuvem de pontos das variáveis logaritmizadas. Com auxílio da função `identify` do R, identifique os três pontos que se destacam na parte inferior direita da nuvem.
- h) Ajuste a reta de regressão de log-peso do cérebro sobre log-peso do corpo considerando os dados (logaritmizados) sem as 3 observações identificadas em g). Obtenha o valor do coeficiente de determinação.
- i) Represente na nuvem de pontos das variáveis logaritmizadas os modelos obtidos em f) e h). Comente, sem esquecer de comentar a diferença considerável nos coeficientes de determinação.

2.9. (📄) Considere os dados referentes ao índice de salários ( $IS$ ) e ao rendimento nacional ( $RN$ ) a preços de 2010, para um certo país, no período de 2000 a 2017 os quais se apresentam no quadro seguinte (SalariosRN.xlsx):

Anos	$t$	Rendimento nacional, a preços de 2010 (biliões de euros)	Índice de salários (100 = 2010)
2000	0	4491	84,48
2001	1	4695	88,12
2002	2	4841	95,40
2003	3	4780	97,65
2004	4	4796	98,72
2005	5	4786	93,69
2006	6	4604	98,54
2007	7	4824	95,20
2008	8	5328	95,97
2009	9	5750	97,76
2010	10	6079	100,00
2011	11	6403	159,39
2012	12	6777	174,09
2013	13	6967	196,63
2014	14	7231	212,53
2015	15	7301	215,91
2016	16	7381	208,56
2017	17	7546	218,03

- a) Para estimar a variação absoluta média anual do índice de salários, temos de relacionar  $IS$  com a variável tempo ( $t$ ), de acordo com o modelo  $IS_t = \beta_0 + \beta_1 Ano_t + \varepsilon_t$ .
- i) Ajuste o modelo e interprete os coeficientes obtidos.
- ii) Avalie a significância dos coeficientes.
- iii) Considera que fez um bom ajustamento? Na sua resposta considere também a validação dos pressupostos do modelo de regressão linear ajustado.
- b) Para estimar a taxa média anual de crescimento do índice de salários, o modelo que possibilitará a obtenção direta dessa taxa é  $IS_t = \beta_0 \beta_1^{Ano_t} \varepsilon_t$  (exponencial/crescimento constante).
- i) Ajuste o modelo, interprete os coeficientes obtidos e explicita a relação de base correspondente entre as variáveis originais (não logaritmizadas).
- ii) Avalie a significância dos coeficientes.
- iii) Considera que fez um bom ajustamento? Na sua resposta considere também a validação dos pressupostos do modelo de regressão linear ajustado.

- c) Para estimar a elasticidade<sup>1</sup> de valor constante dos salários em relação ao rendimento nacional é  $IS_t = \beta_0 RN_t^{\beta_1} \varepsilon_t$  (potência/elasticidade constante).
- Ajuste o modelo, interprete os coeficientes obtidos e explicita a relação de base correspondente entre as variáveis originais (não logaritmizadas).
  - Avalie a significância dos coeficientes.
  - Considera que fez um bom ajustamento? Na sua resposta considere também a validação dos pressupostos do modelo de regressão linear ajustado.
- d) Para estimar o efeito da taxa de crescimento do rendimento nacional sobre a variação absoluta do índice de salários, o modelo que possibilitará a obtenção direta dessa taxa é  $IS_t = \beta_0 + \beta_1 \ln(IS_t) + \varepsilon_t$  (exponencial/crescimento constante).
- Ajuste o modelo, interprete os coeficientes obtidos e explicita a relação de base correspondente entre as variáveis originais (não logaritmizadas).
  - Avalie a significância dos coeficientes.
  - Considera que fez um bom ajustamento? Na sua resposta considere também a validação dos pressupostos do modelo de regressão linear ajustado.

2.10. (📊) Num estudo sobre a eficiência de um algoritmo de processamento de dados, analisou-se o tempo necessário para processar diferentes volumes de dados (variável volume), medidos em megabytes (MB). Para cada volume de entrada, registou-se o tempo médio de processamento (variável tempo), medido em segundos. Os dados recolhidos encontram-se na tabela seguinte e estão disponíveis no ficheiro *eficiencia.xlsx*:

volume (MB)	0,02	0,02	0,06	0,06	0,11	0,11	0,22	0,22	0,56	0,56	1,10	1,10
tempo (s)	76	47	97	107	123	139	159	152	191	201	207	200

- Construa a nuvem de pontos do tempo de processamento (tempo, eixo vertical) versus volume de entrada (volume, eixo horizontal).
- Tendo em conta a curvatura observada no gráfico, admite-se que o modelo de Michaelis-Menten é adequado à descrição da relação referida, e decide-se usar este modelo com a seguinte parametrização (onde  $y$  representa o tempo médio de processamento e  $x$  o volume de dados),  $y = \frac{ax}{b+x}$ ,  $a > 0$ ,  $b > 0$  e  $x > 0$ .
  - Mostre que o modelo referido pode ser linearizado, indicando a relação linearizada e as transformações de variáveis necessárias.
  - Ajuste o modelo linearizado que escolheu na alínea anterior, através do comando `lm` do R.
  - Estime os parâmetros  $a$  e  $b$  na relação original no modelo de Michaelis-Menten. Como interpreta o valor estimado do parâmetro  $a$ ? Trace a curva correspondente ao ajustamento na nuvem de pontos original para melhor compreender o seu significado. Comente o resultado.

<sup>1</sup> Elasticidade mede a variação percentual de  $y$  quando  $x$  varia 1%.

### **3. Regressão linear múltipla**