Correlação e Regressão – Lista de Exercícios

1) Barbetta (2001, p.275). Considerando os dados da Tabela 1:
   a) Construir um diagrama de dispersão para as variáveis *taxa de alfabetização* e *taxa de mortalidade infantil*. Quais as informações observadas no gráfico?
   b) Calcule o coeficiente de correlação de Pearson entre as variáveis *taxa de alfabetização* e *taxa de mortalidade infantil*. Interprete o resultado obtido.

Tabela 1 - Alguns dados de doze importantes municípios catarinenses

município	população (em 1000 hab.)	pop. urbana	% de pop. urbana	taxa de cresc. demográfico	taxa de mort. infantil	taxa de alfabetização
Itajaí	101	94	93	3,19	37	85
Blumenau	193	181	94	4,60	27	90
Rio do Sul	42	39	94	2,78	38	85
Joinville	304	292	96	6,46	25	87
Curitibanos	42	32	76	1,99	67	75
Lages	152	126	83	1,89	63	78
Canoinhas	55	36	66	2,92	41	81
Chapecó	105	77	73	5,32	13	75
Concórdia	68	25	37	2,71	28	84
Florianópolis	219	186	85	3,11	17	87
Criciúma	129	116	90	3,11	32	85
Laguna	42	33	78	1,21	32	77

2) Barbetta (2001, p.275). Sejam X= nota na prova do vestibular de matemática e Y = nota final na disciplina de cálculo. Estas variáveis foram observadas em 20 alunos, ao final do primeiro período letivo de um curso de engenharia. Os dados são apresentados a seguir:

X	Y	X	Y	X	Y	X	Y	X	Y
39	65	43	78	21	52	64	82	65	88
57	92	47	89	28	73	75	98	47	71
34	56	52	75	35	50	30	50	28	52
40	70	70	50	80	90	32	58	67	88

   a) Construa um diagrama de dispersão e verifique se existe correlação entre os dados observados destas duas variáveis.
   b) Existe algum aluno que *foge* ao comportamento geral dos demais (ponto discrepante)?
   c) Calcule a correlação entre a nota no vestibular de matemática e a nota na disciplina de cálculo.
   d) Retire o valor discrepante detectado e calcule novamente o coeficiente r. Interprete.

3) Barbetta (2001, p.275). Sejam os dados do conjunto de dados do anexo do Capítulo 4 (fazer download no site). Considerando apenas a localidade da Encosta do Morro, faça um diagrama de dispersão com os dados de: *X= renda familiar* e *Y= número de moradores no domicílio*. Interprete.

4) Barbetta (2001, p.286). Com o objetivo de verificar se existe correlação positiva entre *aptidão em matemática* e *aptidão em música,* foi selecionado um grupo de crianças de 8 a 10 anos de idade, que foram submetidas a dois testes de aptidão: um de matemática e outro de música. A ordem da aplicação dos testes em cada criança foi aleatória. Os dados estão relacionados na Tabela 2

Tabela 2 - Testes de aptidão em crianças

criança	Valores de aptidão em		criança	Valores de aptidão em	
	matemática	música		matemática	música
1	60	80	7	48	79
2	58	62	8	72	88
3	73	70	9	75	54
4	51	83	10	83	82
5	54	62	11	62	64
6	75	92	12	52	69

Faça o cálculo do coeficiente *r* e confira o resultado encontrado

5) Barbetta (2001, p.286). Com respeito aos 23 alunos de uma turma de estatística, foram observadas as seguintes variáveis: *número de faltas e nota final da disciplina.* Estes dados acusaram a segui8ntes correlação, descrita pelo coeficiente de correlação de Pearson: r = – 0,56. Comente as seguintes frases relativas à turma em estudo e ao coeficiente obtido.
   a) "Como r = – 0,56 (correlação relativa moderada), nenhum aluno com grande número de faltas tirou nota alta."
   b) "Como as duas variáveis são correlacionadas, bastaria usar uma delas como critério de avaliação, pois uma acarreta a outra."
   c) "Os dados observados mostraram uma leve tendência de a nota final se relaciona inversamente com o número de faltas, então os alunos *freqüentadores* tiveram, em geral, melhor desempenho nas avaliações, do que os alunos que faltaram muito."

6) Barbetta (2001, p.286). Numa amostra aleatória de $n = 12$ livros da Biblioteca Central, encontramos $r = 0,207$ entre a *idade da edição* e o *número de páginas* do livro.
   a) O que se pode dizer com base no valor deste coeficiente de correlação?
   b) Esta correlação pode ser explicada meramente por fatores casuais?

7) Barbetta (2001, p.297). Nos últimos anos, em várias regiões, houve um movimento migratório que fez crescer bastante a população urbana nos municípios médios e grandes. Neste contexto, vamos tentar *explicar* o crescimento demográfico de um município em função de sua população urbana, para os municípios da Tabela 1.

   a) Qual deve ser a variável dependente e a independente?
   b) Estabeleça a equação de regressão.
   c) Faça um gráfico com os pontos observados e a reta de regressão estimada.
   d) Qual é a taxa de crescimento demográfico, predita pela equação de regressão, para um município de 300 mil habitantes?
   e) Calcule o coeficiente R^2.
   f) Quais são as principais informações que podem ser obtidas pela presente análise?

8) Barbetta (2001, p.298). Considerando que a satisfação de um aluno com curso universitário (Y) pode ser afetada pelo seu desempenho no curso (X), faça uma análise de regressão, usando os dados do conjunto de dados do anexo do Capítulo 2 (fazer download no site).

9) Dados o tempo de serviço de 10 funcionários de uma companhia de seguros e o número de clientes que cada um possui, verifique se existe uma associação entre estas variáveis:

Anos de serviço (x)	2	3	4	5	4	6	7	8	8	10
Nº de clientes (y)	48	50	56	52	43	60	62	58	64	72

   a) Calcule as medidas descritivas destas duas variáveis;
   b) Construa o diagrama de dispersão e anote os valores mínimo e máximo de **X** e **Y** que aparecem no gráfico;
   c) Trace no diagrama de dispersão as retas $y = \overline{X}$ e $x = \overline{Y}$ e analise o gráfico;
   d) Calcule e interprete o coeficiente de correlação.

10) Numa pesquisa feita com 10 famílias com renda bruta mensal entre 10 e 60 salários mínimos, mediram-se: X: renda bruta mensal (expressa em números de salários mínimos) e Y: a porcentagem da renda bruta anual gasta com assistência médica.

X	12	16	18	20	28	30	40	48	50	54
Y	7,2	7,4	7,0	6,5	6,6	6,7	6,0	5,6	6,0	5,5

   a) Escolha adequadamente X e Y.
   b) Construa o diagrama de dispersão;
   c) Calcule o coeficiente de correlação.

11) Os quatro conjuntos de dados a seguir foram preparados pelo estatístico F. J. Anscombe e são usados com freqüência em aulas sobre correlação.

Conjunto 1		Conjunto 2		Conjunto 3		Conjunto 4	
X	Y	X	Y	X	Y	X	Y
10	8,04	10	9,14	10	7,46	8	6,58
8	6,95	8	8,14	8	6,77	8	5,76
13	7,58	13	8,74	13	12,74	8	7,71
9	8,81	9	8,77	9	7,11	8	8,84
11	8,33	11	9,26	11	7,81	8	8,47

14	9,96	14	8,10	14	8,84	8	7,04
6	7,24	6	6,13	6	6,08	8	5,25
4	4,26	4	3,10	4	5,39	19	12,50
12	10,84	12	9,13	12	8,15	8	5,56
7	4,82	7	7,26	7	6,42	8	7,91
5	5,68	5	4,74	5	5,73	8	6,89

a) Calcule a média e o desvio padrão para cada conjunto de dados.
b) Calcule o coeficiente de correlação para cada conjunto de dados.
c) Construa o diagrama de dispersão para cada conjunto de dados.
d) Analise os resultados.

12) Uma empresa que produz bens de alta tecnologia está preocupada com a produtividade de funcionários que exercem funções repetitivas e procura descobrir como algumas variáveis podem influenciar no rendimento dessas pessoas. Para isso implementa em cada uma de suas três fábricas um programa específico: alimentação especial sugerida pelos nutricionistas; intervalos para exercícios de relaxamento sugerido pelos fisioterapeutas; rodízio de funções sugerido pelos psicólogos. A tabela a seguir mostra o resultado da produtividade para diversos níveis implementados no programa.

Produtividade (menor=100%)	100	102	105	108	112	120
Alimentação (frequência semanal)	4	5	1	3	6	2
Exercícios (frequência semanal)	1	3	2	4	5	6
Rodízio (frequência semanal)	3	1	2	6	4	5

a) Construa o diagrama de dispersão da produtividade contra cada uma das variáveis explicativas. Qual variável parece manter melhor correlação com a produtividade?
b) Calcule o coeficiente de correlação linear de Pearson nos três casos. O coeficiente confirma a impressão visual dos diagramas?

13) Use as observações de poupança agregada e renda (bilhões de reais) em um país X no período de 1990 a 1999 (dados fictícios), para estimar a influência do nível de renda sobre a poupança..
a) Construa o diagrama de dispersão e trace a reta de regressão da poupança em função da renda. Interprete os coeficientes.
b) Diga qual é o acréscimo na poupança agregada para cada bilhão a mais na renda.
c) Estime a poupança para uma renda de R$469 400 000 000,00. Quanto seria o consumo agregado das famílias? OBS.: consumo + poupança = renda
d) Calcule e interprete o coeficiente de correlação.
e) Calcule e interprete o coeficiente de determinação.
f) Construa o diagrama de dispersão considerando o consumo como variável resposta e a renda como variável explicativa (preditora). Estime a reta de regressão e compare o resultado com o item a.

14) Suponha os seguintes dados na tabela

Despesas com Propaganda (1000 000 R$)	**Vendas de Certo Produto (1000 unidades)**
2,5	120
6,5	190
11,0	240
4,0	140
8,5	180
14,0	280
6,0	150
5,0	115
10,0	215
13,5	220
16,0	320

a) Construa o diagrama de dispersão;
b) Ajuste uma reta aos dados e estime as vendas do produto, para um gasto com propaganda de 12 milhões de reais;
c) Qual o acréscimo nas vendas para cada milhão a mais gasto com propaganda?
d) Trace a reta no diagrama de dispersão;
e) Determine o coeficiente de correlação e interprete-o;
f) Calcule e interprete o coeficiente de determinação.

15) (Toledo e Ovalle, 1995) A tabela abaixo fornece os resultados de uma pesquisa com 10 famílias de determinada região.

Famílias	Renda (R$100,00)	Poupança (R$10,00)	Número de Filhos	Média de anos de estudo da família
A	10	4	8	3
B	15	7	6	4
C	12	5	5	5
D	70	20	1	12
E	80	20	2	16
F	100	30	2	18
G	20	8	3	8
H	30	8	2	8
I	10	3	6	4
J	60	15	1	8

a) Calcule o coeficiente de correlação entre as variáveis renda e poupança, renda e número de filhos, poupança e número de filhos, média dos anos de estudo e número de filhos e entre as variáveis renda familiar e média de anos de estudo. Retire conclusões.
b) Ajuste um modelo linear utilizando as variáveis Renda (X) e Poupança (Y). Estime o valor poupado quando a renda for de 2.000 reais.

16) A administração de um banco desejava estabelecer um critério objetivo para avaliar a eficiência de seus gerentes. Para isso, levantou (para cada um dos subdistritos onde possuía agência) dados a respeito do depósito médio mensal por agência e o número de estabelecimentos comerciais existentes nesses subdistritos. Os dados são os seguintes:

Subdistritos	Número de Estabelecimentos Comerciais	Depósito Médio Mensal por Agência (10 000 R$)
Nossa Senhora do Ó	16	14
Casa verde	30	16
Vila Formosa	35	19
Santana	70	30
Barra Funda	90	31
Jardim Paulista	120	33
Santo Amaro	160	35
Lapa	237	43
Pinheiros	378	50

a) Construa o diagrama de dispersão;
b) Ajuste uma reta aos dados e estime depósito médio para um número de estabelecimentos comerciais igual a 350;
c) Qual o acréscimo nos depósitos médios, para cada estabelecimento a mais no subdistrito?
d) Trace a reta no diagrama de dispersão;
e) Determine o coeficiente de correlação e interprete-o;
f) Calcule e interprete o coeficiente de determinação.

17) Barbetta (2001, p.308). Com o objetivo de verificar se numa certa região existe correlação entre o nível de escolaridade médio dos pais e o nível de escolaridade dos filhos, observou-se uma amostra aleatória de 8 indivíduos adultos, verificando o número de anos que estes freqüentaram (e tiveram aprovação) em escolas regulares (Y) e o número médio de anos que os seus pais freqüentaram (e tiveram aprovação) em escolas regulares (X). Os resultados são apresentados na tabela abaixo:

X	0	0	2	3	4	4	5	7
Y	2	3	2	5	9	8	8	15

a) Calcule o coeficiente de correlação de Pearson.
b) Em termos do resultado do item (a), o que se pode dizer sobre a correlação entre o número de anos que os 8 indivíduos freqüentaram escolas regulares (Y) e o número médio de anos que os seus pais freqüentaram escolas regulares?
c) Estabeleça a reta de regressão de y em relação a x.

d) Apresente o diagrama de dispersão acompanhado da reta de regressão.

18) Barbetta (2001, p.308). A tabela a seguir relaciona os pesos (em centenas de kg) e as taxas de consumo de combustível em rodovia (km/litro) numa amostra de 10 carros de passeio novos.

Peso	12	13	14	14	16	18	19	22	24	26
Consumo	16	14	14	13	11	12	9	9	8	6

a) Calcule o coeficiente de correlação de Pearson.
b) Considerando o resultado do item a), como você avalia o relacionamento entre o peso e o consumo, na amostra observada?
c) Para estabelecer uma reta de regressão, qual deve ser a variável dependente e qual deve ser a variável independente? Justifique a sua resposta.
d) Estabeleça a equação de regressão, considerando a resposta do item c).
e) Apresente o diagrama de dispersão e a reta de regressão obtida em d).
f) Você considera adequado o ajuste do modelo de regressão do item d)? Dê uma medida desta adequação interpretando-a.
g) Qual o consumo esperado para um carro de 2000 kg? Lembrete: os dados de consumo na tabela estão em centenas de kg.
h) Você considera seu estudo capaz de predizer o consumo esperado para um carro de 7000 kg? Justifique sua resposta.

19) Barbetta (2001, p.308). Um administrador de uma grande sorveteria anotou por um longo período de tempo a *temperatura média diária*, em ºC (X), e o *volume de vendas diária de sorvete*, em kg (Y). Com os dados, estabeleceu uma equação de regressão, resultando em:

$$y = 0,5 + 1,8x, \text{ com } R^2 = 0,80$$

Pede-se:
a) Qual o consumo esperado de sorvete num dia de 27ºC?
b) Qual o incremento esperado nas vendas de sorvete a cada 1°C de aumento da temperatura?

20) Atkinson et al. (1994) investigaram em que medida partículas de chumbo potencialmente tóxica emitidas por veículos automotores são absorvidas por ciclistas que participam de competições. A tabela abaixo, construída a partir de um gráfico apresentado em seu artigo, fornece níveis de chumbo no sangue e horas de treinamento de 10 ciclistas.

Horas de treinamento	8	10	10	12	15	18	18	21	25	25
Chumbo no sangue (mmol/L)	0,53	0,25	0,34	0,25	0,29	0,3	0,53	0,53	0,53	0,87

Pede-se:
a) Faça um gráfico dos dados. Quais suas impressões?
b) Verifique se há uma relação entre níveis de chumbo no sangue e horas de treinamento.
c) O ciclista 10 tem níveis muito altos. Nossa evidência de uma relação é proveniente quase que inteiramente desta observação? Repita (b) omitindo o ciclista 10.
d) O que fizemos em (c) parece razoável?
e) Está claro a partir do gráfico obtido em (a) que há variação nos dados que não é explicada pelas horas de treinamento. (O que nos dá esta informação?) Talvez o efeito de horas de treinamento não apareça tão fortemente como deveria, porque estamos deixando de levar em consideração outras variáveis importantes. Sugira algumas outras variáveis que poderiam ser importantes.