ANÁLISE DE REGRESSÃO MÚLTIPLA

1) Em um estudo foi utilizada, erroneamente, uma amostra de apenas 3 observações para se estimarem os coeficientes de uma equação de regressão. Obteve-se R^2 = 0,96. A título de brincadeira, foi dito ao analista responsável que, se ele quisesse melhorar os resultados, bastaria eliminar uma observação e ficar com apenas n = 2. Faça uma crítica sobre o uso de amostras muito pequenas em regressão linear.

Com n=2, R^2 = 1 pois a reta vai ligar perfeitamente dois pontos. Importante notar que a regressão linear tem outras variáveis a serem analisadas além do R^2 , com poucos pontos a tendência é que esse indicador seja sempre alto. É também conveniente verificar o tamanho mínimo da amostra

2) A tabela a seguir apresenta os dados correspondentes à produção brasileira de automóveis, em milhares, no período de 17 anos. Ajuste os dados, usando os modelos de regressão linear. Analise os resultados e estime a produção para o décimo oitavo ano.

Produção	**Tempo**
30,5	1
61	2
96,1	3
133	4
145,6	5
191,2	6
174,2	7
183,7	8
185,2	9
224,6	10
225,4	11
278,5	12
349,5	13
416	14
516	15
609	16
729,1	17

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                Produção   R-squared:                       0.849
```

```
Model:                            OLS   Adj. R-squared:                  0.839
Method:                 Least Squares   F-statistic:                     84.58
Date:                Mon, 01 Apr 2024   Prob (F-statistic):           1.49e-07
Time:                        07:24:08   Log-Likelihood:                -97.228
No. Observations:                  17   AIC:                             198.5
Df Residuals:                      15   BIC:                             200.1
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -54.0529     39.816     -1.358      0.195    -138.918      30.812
Tempo         35.7353      3.886      9.197      0.000      27.453      44.017
==============================================================================
Omnibus:                        0.910   Durbin-Watson:                   0.255
Prob(Omnibus):                  0.635   Jarque-Bera (JB):                0.526
Skew:                           0.417   Prob(JB):                        0.769
Kurtosis:                       2.783   Cond. No.                         21.6
==============================================================================
```

R-squared: 0,849 - Bom ajuste Coeficiente de tempo > 0 relação positiva entre as variáveis p-valor de tempo <0,05 - variável relevante para o modelo

```
Teste de Shapiro-Wilk para normalidade dos resíduos:
Estatística de teste: 0.944674015045166
Valor-p: 0.3778810501098633
```

Aceitamos a hipótese de igualdade dos resíduos p>0,05

```
Previsão Producao 589.1823529411766
```

3) A companhia Multifator está analisando o comportamento dos Custos Indiretos de Fabricação (CIF) em função das variáveis: horas de mão-de-obra direta (HMOD) e horas - máquina (HM) nos últimos 15 meses. Analise a variável CIF em função de cada uma das variáveis (HMOD e HM) isoladamente e em função das duas simultaneamente. Para facilitar as análises, obtenha também a matriz de correlação de todas as variáveis envolvidas. Após a análise do modelo de regressão com as duas variáveis simultaneamente, refaça o estudo, considerando o modelo de regressão *stepwise*. Compare os resultados das duas modelagens de regressão múltipla.

Período	**CIF**	**HMOD**	**HM**
1,00	350,00	4,00	10,00
2,00	400,00	8,00	14,00
3,00	470,00	12,00	16,00
4,00	550,00	10,00	26,00
5,00	620,00	15,00	31,00
6,00	380,00	7,00	12,00

7,00	290,00	6,00	13,00
8,00	490,00	10,00	21,00
9,00	580,00	11,00	26,00
10,00	610,00	13,00	24,00
11,00	560,00	12,00	23,00
12,00	420,00	8,00	12,00
13,00	450,00	11,00	19,00
14,00	510,00	12,00	19,00
15,00	380,00	5,00	11,00

	CIF	HMOD	HM
CIF	1.000000	0.882914	0.919862
HMOD	0.882914	1.000000	0.845405
HM	0.919862	0.845405	1.000000

AS CORRELAÇÕES SÃO ALTAS E POSITIVAS

OLS Regression Results

Dep. Variable:	CIF	R-squared:	0.780
Model:	OLS	Adj. R-squared:	0.763
Method:	Least Squares	F-statistic:	45.97
Date:	Mon, 01 Apr 2024	Prob (F-statistic):	1.30e-05
Time:	07:50:53	Log-Likelihood:	-78.584
No. Observations:	15	AIC:	161.2
Df Residuals:	13	BIC:	162.6
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	200.8214	41.762	4.809	0.000	110.600	291.043
HMOD	28.1089	4.146	6.780	0.000	19.152	37.066

Omnibus:	0.876	Durbin-Watson:	1.768
Prob(Omnibus):	0.645	Jarque-Bera (JB):	0.698
Skew:	-0.154	Prob(JB):	0.705
Kurtosis:	1.989	Cond. No.	33.6

R-squared: 0,780 - Bom ajuste Coeficiente de HMOD > 0 relação positiva entre as variáveis p-valor de HMOD <0,05 - variável relevante para o modelo

Teste de Shapiro-Wilk para normalidade dos resíduos:
Estatística de teste: 0.9590085744857788

Valor-p: 0.6751689314842224

Valor p acima de 0.05 aceito a normalidade dos resíduos

OLS Regression Results

Dep. Variable:	CIF	R-squared:	0.846
Model:	OLS	Adj. R-squared:	0.834
Method:	Least Squares	F-statistic:	71.50
Date:	Mon, 01 Apr 2024	Prob (F-statistic):	1.21e-06
Time:	07:42:42	Log-Likelihood:	-75.886
No. Observations:	15	AIC:	155.8
Df Residuals:	13	BIC:	157.2
Df Model:	1		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	208.8765	32.714	6.385	0.000	138.202	279.551
HM	14.1764	1.677	8.456	0.000	10.554	17.798

Omnibus:	6.390	Durbin-Watson:	1.782
Prob(Omnibus):	0.041	Jarque-Bera (JB):	3.222
Skew:	-0.944	Prob(JB):	0.200
Kurtosis:	4.262	Cond. No.	60.6

R-squared: 0,846 - Ajuste melhor que o anterior Coeficiente de HM > 0 relação positiva entre as variáveis p-valor de HM <0,05 - variável relevante para o modelo

Teste de Shapiro-Wilk para normalidade dos resíduos:
Estatística de teste: 0.9249430894851685
Valor-p: 0.22901682555675507

Valor p acima de 0.05 aceito a normalidade dos resíduos

OLS Regression Results

Dep. Variable:	CIF	R-squared:	0.885
Model:	OLS	Adj. R-squared:	0.866
Method:	Least Squares	F-statistic:	46.17
Date:	Mon, 01 Apr 2024	Prob (F-statistic):	2.32e-06
Time:	07:42:52	Log-Likelihood:	-73.704
No. Observations:	15	AIC:	153.4
Df Residuals:	12	BIC:	155.5
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]

Intercept	184.8836	31.762	5.821	0.000	115.680	254.087
HMOD	11.7460	5.835	2.013	0.067	-0.968	24.460
HM	9.3694	2.825	3.317	0.006	3.215	15.524

Omnibus:	6.418	Durbin-Watson:	1.721
Prob(Omnibus):	0.040	Jarque-Bera (JB):	3.376
Skew:	-1.062	Prob(JB):	0.185
Kurtosis:	3.943	Cond. No.	73.4

R-squared: 0,885 - Ajuste melhor que o anterior Coeficientes de HM e HMOD > 0 relação positiva entre as variáveis p-valor de HMOD > 0,05 - variável irrelevante para o modelo. É melhor o modelo só com HM

4) Uma rede de lojas de material de construção (CONSTRUCAO) que atua em 52 regiões quer fazer um estudo sobre a quantidade vendida (qt_vend) de determinado tipo de material. Como possíveis informações que poderiam ter alguma influência estão: gasto com propaganda (gast_prop), número de contas ativas (n_cont), número de marcas (n_marc), número de lojas na região (n_loj).

Faça uma regressão entre quantidade vendida e as demais variáveis.

OLS Regression Results

Dep. Variable:	qt_venda	R-squared:	0.989
Model:	OLS	Adj. R-squared:	0.988
Method:	Least Squares	F-statistic:	1075.
Date:	Mon, 01 Apr 2024	Prob (F-statistic):	1.52e-45
Time:	07:59:08	Log-Likelihood:	-185.83
No. Observations:	52	AIC:	381.7
Df Residuals:	47	BIC:	391.4
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	178.4976	8.656	20.621	0.000	161.084	195.912
gast_prop	1.7943	0.722	2.485	0.017	0.342	3.247
n_cont	3.3190	0.109	30.507	0.000	3.100	3.538
n_marc	-21.1989	0.526	-40.282	0.000	-22.258	-20.140
n_loj	0.3218	0.312	1.030	0.308	-0.307	0.950

Omnibus:	9 1.647	Durbin-Watson:	1.548
Prob(Omnibus):	0.439	Jarque-Bera (JB):	1.410
Skew:	-0.245	Prob(JB):	0.494
Kurtosis:	2.360	Cond. No.	383.

R-squared: 0,989 - p-valor de nloj > 0,05 - variável irrelevante para o modelo. Vamos repetir sem essa variável

OLS Regression Results

Dep. Variable:	qt_venda	R-squared:	0.989
Model:	OLS	Adj. R-squared:	0.988
Method:	Least Squares	F-statistic:	1431.
Date:	Mon, 08 Apr 2024	Prob (F-statistic):	6.23e-47
Time:	08:47:53	Log-Likelihood:	-186.41
No. Observations:	52	AIC:	380.8
Df Residuals:	48	BIC:	388.6
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	180.0088	8.536	21.087	0.000	162.845	197.172
gast_prop	1.6656	0.712	2.341	0.023	0.235	3.096
n_cont	3.3701	0.097	34.789	0.000	3.175	3.565
n_marc	-21.2302	0.526	-40.383	0.000	-22.287	-20.173

Omnibus:	1.243	Durbin-Watson:	1.594
Prob(Omnibus):	0.537	Jarque-Bera (JB):	1.265
Skew:	-0.296	Prob(JB):	0.531
Kurtosis:	2.516	Cond. No.	371.

R-squared: 0,989 – todas as variáveis relevantes para o modelo. N_marc tem relação negativa com a qt_venda, todas as demais são positivas.

```
Teste de Shapiro-Wilk para normalidade dos resíduos:
Estatística de teste: 0.9676164984703064
Valor-p: 0.1671878695487976
```

Valor p acima de 0.05 aceito a normalidade dos resíduos

5) Um estudo revelou acentuada correlação entre o consumo de bebidas alcoólicas e a elevação dos salários dos professores. Existe relação de causa e efeito entre essas variáveis que justificaria um modelo de análise de regressão?

   Aparentemente não existe relação de causa e efeito e essa correlação alta deve ser apenas coincidência.

6) Cite:

   a) duas variáveis que podem apresentar alta correlação, mas não têm relação de causa e efeito;

      Vendas de picolé e número de afogamentos

   b) duas variáveis que podem apresentar alta correlação, sendo razoável supor relação de causa e efeito entre elas.

      Número de gols de uma equipe e a pontuação desta equipe no campeonato

7) Considere o peso e o comprimento de alguns cães. Calcule o coeficiente de correlação entre estas duas variáveis:

Peso (Kg)	Comprimento (cm)
14	85
14	90
16	95
17	100
20	95
22	96
22	100
23	109

28	105
28	110

	Peso_(Kg)	Comprimento_(cm)
Peso_(Kg)	1.000000	0.847614
Comprimento_(cm)	0.847614	1.000000

Correlação alta e positiva

8) Calcule, pelo método dos mínimos quadrados, a equação de regressão linear para os dados do exercício anterior

```
OLS Regression Results
==============================================================================
Dep. Variable:                   Peso   R-squared:                       0.718
Model:                            OLS   Adj. R-squared:                  0.683
Method:                 Least Squares   F-statistic:                     20.41
Date:                Mon, 08 Apr 2024   Prob (F-statistic):            0.00195
Time:                        09:09:51   Log-Likelihood:                -23.751
No. Observations:                  10   AIC:                             51.50
Df Residuals:                       8   BIC:                             52.11
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept    -33.6078     11.989     -2.803      0.023     -61.254      -5.962
Comprimento    0.5483      0.121      4.518      0.002       0.268       0.828
==============================================================================
Omnibus:                        1.027   Durbin-Watson:                   1.857
Prob(Omnibus):                  0.598   Jarque-Bera (JB):                0.659
Skew:                          -0.160   Prob(JB):                        0.719
Kurtosis:                       1.783   Cond. No.                     1.29e+03
==============================================================================
```

R-squared: 0,718 - razoável. Relação positiva entre peso e comprimento p-valor menor que 0,05.

```
Teste de Shapiro-Wilk para normalidade dos resíduos:
Estatística de teste: 0.9438753128051758
Valor-p: 0.5968831777572632
```

Valor p acima de 0.05 aceito a normalidade dos resíduos

9) Para o arquivo Biscobis.xlsx, referente a uma amostra de 100 empresas clientes de uma grande empresa que é fornecedora no setor industrial , processe a análise de regressão múltipla *stepwise* e analise os resultados obtidos, sendo:

Variável dependente: X_9 = nível de uso do serviço (quanto do total de produtos da empresa é comprado da Biscobis)
Variáveis independentes: avaliação de 0 a 10 de atributos da Biscobis:
X_1 = rapidez na entrega do produto
X_2 = nível de preço
X_3 = flexibilidade de preço
X_4 = imagem do fornecedor
X_5 = serviço como um todo
X_6 = imagem da força de vendas
X_7 = qualidade do produto
X_8 = Variável nominal – status da compra 1=primeira compra 2=segunda compra 3=comprador frequente

Passo 1 criar variável dummy (está no código em anexo)
Passo 2 executar procedimento stepwise (está no código em anexo)

```
Resultado stepwise: ['x3', 'x5', 'x6', 'x7', 'x8_2', 'x8_3']
```

OLS Regression Results

Dep. Variable:	x9	R-squared:	0.845
Model:	OLS	Adj. R-squared:	0.835
Method:	Least Squares	F-statistic:	84.38
Date:	Mon, 08 Apr 2024	Prob (F-statistic):	1.97e-35
Time:	10:31:06	Log-Likelihood:	-267.83
No. Observations:	100	AIC:	549.7
Df Residuals:	93	BIC:	567.9
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	1.9801	4.331	0.457	0.649	-6.621	10.582
x3	1.9147	0.395	4.848	0.000	1.130	2.699
x5	4.8305	0.722	6.695	0.000	3.398	6.263
x6	1.5569	0.503	3.098	0.003	0.559	2.555
x7	0.8487	0.265	3.205	0.002	0.323	1.374
x8_2	4.3577	1.187	3.671	0.000	2.000	6.715
x8_3	10.2542	1.576	6.506	0.000	7.124	13.384

Omnibus:	22.721	Durbin-Watson:	1.810
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5.563
Skew:	-0.164	Prob(JB):	0.0619
Kurtosis:	1.892	Cond. No.	138.

==========

Todas as variáveis tem relação positiva – e p-valor<0,05 (variáveis relevantes) – r-quadrado 0. 845 bom valor

```
Teste de Shapiro-Wilk para normalidade dos resíduos:
Estatística de teste: 0.9590353965759277
Valor-p: 0.0034342966973781586
```

Os resíduos não tem distribuição normal

10) Para o arquivo Imoveis.xlsx, faça a análise para o consumo de energia, área e idade explicando o valor. Analise os resultados.

```
                  OLS Regression Results
==============================================================================
Dep. Variable:                  Valor   R-squared:                       0.844
Model:                            OLS   Adj. R-squared:                  0.834
Method:                 Least Squares   F-statistic:                     83.11
Date:                Mon, 08 Apr 2024   Prob (F-statistic):           1.35e-18
Time:                        09:16:01   Log-Likelihood:                -249.36
No. Observations:                  50   AIC:                             506.7
Df Residuals:                      46   BIC:                             514.4
Df Model:                           3
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   -128.2598     38.541     -3.328      0.002    -205.838     -50.681
Área           1.4590      0.124     11.776      0.000       1.210       1.708
Idade         -2.7722      0.825     -3.359      0.002      -4.434      -1.111
Energia        0.6805      0.254      2.681      0.010       0.170       1.191
==============================================================================
Omnibus:                       14.503   Durbin-Watson:                   1.781
Prob(Omnibus):                  0.001   Jarque-Bera (JB):               19.605
Skew:                           0.969   Prob(JB):                     5.53e-05
Kurtosis:                       5.378   Cond. No.                     1.43e+03
==============================================================================
```

Idade tem relação negativa com o valor. Todas as demais positivas R-square 0,844 – valor alto. Todas as variáveis significativas (p-valor<0,05)

```
Teste de Shapiro-Wilk para normalidade dos resíduos:
Estatística de teste: 0.9448480606079102
Valor-p: 0.02104487642645836
```

Valor p acima de 0.05 aceito a normalidade dos resíduos