UNIVERSIDADE FEDERAL DO CARIRI
CENTRO DE CIÊNCIAS E TECNOLOGIA
BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

Trabalho de Correlação e Regressão

JUAZEIRO DO NORTE
2022

INTEGRANTES
LEONARDO PEREIRA SILVA
DAVI SANTOS ALEXANDRINO

Trabalho de Correlação e Regressão

JUAZEIRO DO NORTE
2022

SUMÁRIO

1. Introdução

O dataset selecionado baseia-se em uma pesquisa feita por determinado aplicativo que contava as horas de uso dos clientes quando estes usavam as redes sociais. A pedido do aplicativo, os usuários também forneceram alguns dados, entre eles a idade. A partir do conhecimento dessas variáveis, sendo a idade uma variável independente e o tempo de uso diário uma variável dependente, podemos buscar correlacioná-las a fim de obter conclusões.

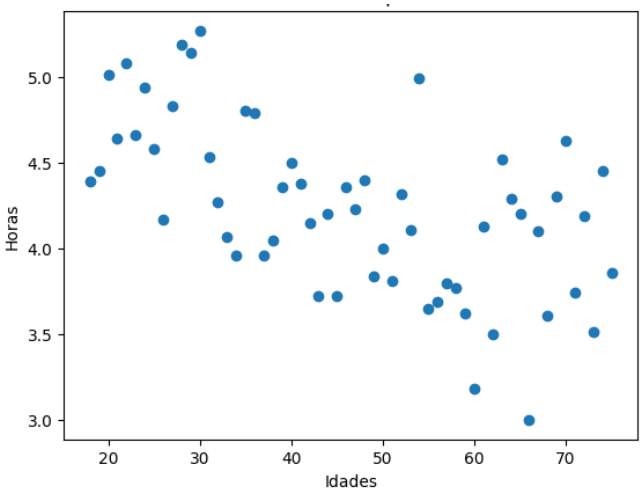
Vale ressaltar que a coleta da quantidade de horas online de cada usuário foi realizada durante 7 dias seguidos e, a partir destas coletas, realizamos o cálculo da média de uso diário dos usuários. Além disso, em ordem de favorecer a exibição intuitiva dos resultados, calculamos, a partir da média supracitada, a média de uso diário dos usuários de mesma idade (a média diária de uso de redes sociais das pessoas de 18 anos que foram entrevistadas, por exemplo). Outra ressalva é para a idade dos usuários, que varia de 18 a 75 anos.

2. Tabela e Diagrama de Dispersão

A partir do dataset foi feita a média de horas diárias para cada idade, tendo esse valor fora feito o gráfico de dispersão abaixo.

Gráfico 1

Gráfico de dispersão entre a idade de usuários de redes sociais e as horas utilizadas diárias



fonte: Kaggle
elaboração dos autores

É possível observar pelo gráfico que, de maneira geral, quanto maior a idade do usuário há uma tendência de que ele use menos redes sociais. Dizemos que há uma tendência pois pode-se observar alguns pontos dispersantes.

3. Verificando a Correlação entre as Variáveis em Estudo

Para a verificação da correlação, estabeleceremos duas hipóteses:

H_0: $\rho = 0$ (As variáveis x e y não são correlacionadas)
H_0: $\rho \neq 0$ (As variáveis x e y são correlacionadas)

Realizamos então o teste de inferência t, calculando o t_{obs} e o $t_{tabelado}$, a partir da tabela *t de student* (com $\alpha = 1\%$) para $n - 2$ graus de liberdade, com n sendo 58 (como mostra a *Tabela para construção do gráfico de dispersão* [1]). Para isso calcularemos r (coeficiente de correlação) com a seguinte fórmula:

$$r = \frac{n\sum_{i}^{n}(xi \cdot yi) - \sum_{i}^{n} yi \left(\sum_{i}^{n} xi\right.}{\sqrt{n \cdot \sum_{i}^{n}(xi^2) - \left(\sum_{i}^{n} xi\right)^2} \cdot \sqrt{n \cdot \sum_{i}^{n}(yi^2) - \left(\sum_{i}^{n} yi\right)^2}}$$

O valor do coeficiente de relação é $r = -0.578$. Interpretando esse valor podemos afirmar que r tem correlação moderada negativa uma vez que está dentro do intervalo [-0,40:-0,69].

Calculando o valor t_{obs} da seguinte forma:

$$t = \frac{r \cdot \sqrt{n-2}}{\sqrt{1-r^2}}$$

O valor t_{obs} é igual a -5.302. Ao observar os valores de $t_{tabelado}$ para graus de liberdade de 50 e 60 respectivamente, achamos $\pm$2,678 e $\pm$2.660. Para ambos os graus de liberdade o tobs está na zona de rejeição, e como o valor $t_{tabelado}$ para 56 graus de liberdade está entre os valores de $t_{tabelado}$ 50 e $t_{tabelado}$ 60, concluímos que para $t_{tabelado}$ para 56 graus de liberdade a nossa hipótese H_0 é rejeitada. Portanto é provável que haja uma correlação entre as variáveis em estudo.

4. Equação de Regressão Linear

Para construir o modelo de regressão usaremos o método de mínimos quadrados, que consiste em obter os valores dos estimadores abaixo, para então construir a equação da reta:

$$b = \frac{n\sum_{i}^{n}(xi \cdot yi) - \left(\sum_{i}^{n} yi\right)\left(\sum_{i}^{n} xi\right)}{n \cdot \sum_{i}^{n}(xi^2) - \left(\sum_{i}^{n} xi\right)^2}$$

[1] Ver o apêndice.

$$a = \frac{\sum_{i}^{n} yi - b\sum_{i}^{n} xi}{n}$$

$$\hat{y} = a + bx$$

Tendo calculado as estimativas para os valores de a e b, obtemos o seguinte modelo de Regressão Linear:

$$\hat{y} = 5.032 - 0.017x$$

Tem-se ainda que o **desvio padrão dos resíduos(S_e)**, calculado por:

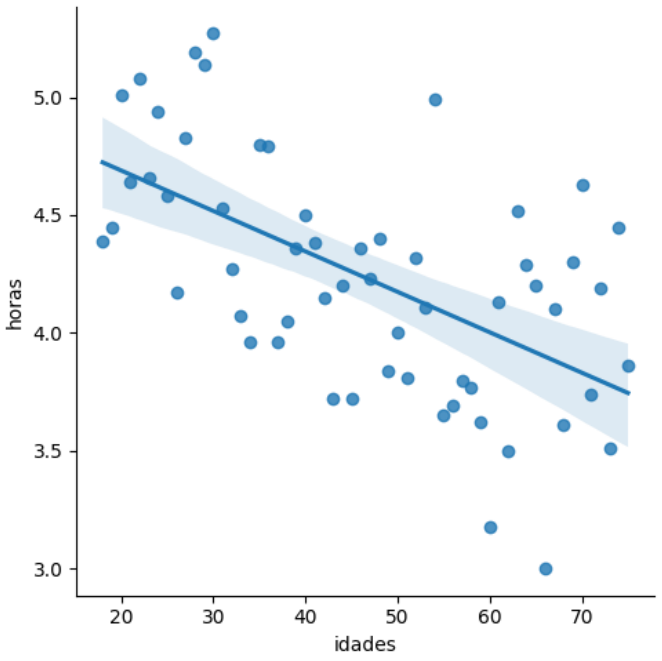
$$S_e = \sqrt{\frac{\sum_{1}^{n} (e_i - \bar{e})^2}{n-2}}$$

É igual a 0.412, indicando que o ajuste da reta de regressão está significativamente próximo da completude.

Dessa maneira, o gráfico de dispersão com o modelo de regressão calculado:

Gráfico 2

Gráfico de dispersão entre a idade de usuários de redes sociais e as horas utilizadas diárias com Reta de Regressão Linear



fonte: Kaggle
elaboração dos autores

5. Cálculo de uma Média de Adequação

Para conferir se o modelo em questão está adequado para descrever o fenômeno usa-se o coeficiente de determinação, dado pela equação a seguir:

$$R^2 = \frac{\sum_{i}^{n}(\widehat{y}-\bar{y})^2}{\sum_{i}^{n}(yi-\bar{y})^2} = \frac{variação\ explicada}{variação\ total};\ 0 \leq R^2 \leq 1$$

Como exemplificado, o coeficiente varia no intervalo [0:1] onde quanto mais próximo de 1, mais adequado é o modelo. Temos que, para o modelo apresentado, $R^2 = 0.333$. Assim, podemos elucidar que a variação das horas diárias usadas de rede social é explicada, em parte, pela variação de idade com 33,3% de explicação e ($1 - R^2 = 0.667$) 66.7% devido a outros fatores.

6. Determinado se a variável independente é significante

Para verificar se a variável independente é significante, temos as seguintes hipóteses:

H_0: $\beta = 0$ (a variável independente não influencia)
H_1: $\beta = 0$ (A variável independente influencia)

Seguindo o cálculo mostrado abaixo:

Fonte de variação	SQ	gl	QM	F
Regressão	$SQR = \sum_{1}^{n}(\widehat{y_i} - \bar{y})^2$	1	$QMR = \frac{SQR}{1}$	$f_{obs} = \frac{QMR}{QME}$
Erro	$SQE = \sum_{1}^{n}(y_i - \widehat{y_i})^2$	$n - 2$	$QME = \frac{SQE}{n-2}$	

Chegamos aos valores de:

Fonte de variação	SQ	gl	QM	F
Regressão	$SQR = 4.785$	1	QMR =4.785	$f_{obs} = 28.066$
Erro	$SQE = 9.548$	56	$QME = 0.170$	

O $f_{tabelado}$ observado (com $\alpha = 1\%$) para os graus de liberdade do numerador 1 e do denominador 50 e 60 respectivamente é 7,17 e 7,08, assim, podemos afirmar que o $f_{tabelado}$ para o numerador 1 e o denominador 56 está entre os valores apresentador, que são menores que que o f_{obs}, que, por sua vez, significa que se rejeita H_0. Logo, há uma provável influência da variável independente.

7. Apêndice

Tabela para construção do Gráfico de Dispersão

index	idade	horas	idades^2	horas^2	(idade*horas)	(idade*horas)^2
0	49	3.84	2401	14.75	188.16	35404.19
1	26	4.17	676	17.39	108.42	11754.9
2	25	4.58	625	20.98	114.5	13110.25
3	44	4.2	1936	17.64	184.8	34151.04
4	55	3.65	3025	13.32	200.75	40300.56
5	43	3.72	1849	13.84	159.96	25587.2
6	31	4.53	961	20.52	140.43	19720.58
7	28	5.19	784	26.94	145.32	21117.9
8	18	4.39	324	19.27	79.02	6244.16
9	73	3.51	5329	12.32	256.23	65653.81
10	64	4.29	4096	18.4	274.56	75383.19
11	30	5.27	900	27.77	158.1	24995.61
12	51	3.81	2601	14.52	194.31	37756.38
13	46	4.36	2116	19.01	200.56	40224.31
14	48	4.4	2304	19.36	211.2	44605.44
15	32	4.27	1024	18.23	136.64	18670.49
16	69	4.3	4761	18.49	296.7	88030.89
17	20	05.01	400	25.1	100.2	10040.04
18	42	4.15	1764	17.22	174.3	30380.49
19	21	4.64	441	21.53	97.44	9494.55
20	29	5.14	841	26.42	149.06	22218.88
21	24	4.94	576	24.4	118.56	14056.47
22	41	4.38	1681	19.18	179.58	32248.98
23	23	4.66	529	21.72	107.18	11487.55
24	65	4.2	4225	17.64	273.0	74529.0
25	19	4.45	361	19.8	84.55	7148.7
26	22	05.08	484	25.81	111.76	12490.3
27	34	3.96	1156	15.68	134.64	18127.93

28	38	04.05	1444	16.4	153.9	23685.21
29	35	4.8	1225	23.04	168.0	28224.0
30	40	4.5	1600	20.25	180.0	32400.0
31	71	3.74	5041	13.99	265.54	70511.49
32	50	4.0	2500	16.0	200.0	40000.0
33	45	3.72	2025	13.84	167.4	28022.76
34	27	4.83	729	23.33	130.41	17006.77
35	33	04.07	1089	16.56	134.31	18039.18
36	39	4.36	1521	19.01	170.04	28913.6
37	63	4.52	3969	20.43	284.76	81088.26
38	61	4.13	3721	17.06	251.93	63468.72
39	54	4.99	2916	24.9	269.46	72608.69
40	58	3.77	3364	14.21	218.66	47812.2
41	75	3.86	5625	14.9	289.5	83810.25
42	72	4.19	5184	17.56	301.68	91010.82
43	68	3.61	4624	13.03	245.48	60260.43
44	37	3.96	1369	15.68	146.52	21468.11
45	47	4.23	2209	17.89	198.81	39525.42
46	36	4.79	1296	22.94	172.44	29735.55
47	53	4.11	2809	16.89	217.83	47449.91
48	62	3.5	3844	12.25	217.0	47089.0
49	59	3.62	3481	13.1	213.58	45616.42
50	70	4.63	4900	21.44	324.1	105040.81
51	67	4.1	4489	16.81	274.7	75460.09
52	74	4.45	5476	19.8	329.3	108438.49
53	52	4.32	2704	18.66	224.64	50463.13
54	66	3.0	4356	9.0	198.0	39204.0
55	57	3.8	3249	14.44	216.6	46915.56
56	60	3.18	3600	10.11	190.8	36404.64
57	56	3.69	3136	13.62	206.64	42700.09

8. Referências

CHENNOJU, Bhuvan. Users Active Time Prediction: Its a multiple timeseries prediction with multiple users. **Kaggle,** 2021. Disponível em: <https://www.kaggle.com/datasets/bhuvanchennoju/mobile-usage-time-prediction>. Acesso em: 01 jun. 2023.