Atividade Avaliativa Semana 9
Lista de exercícios

Número de Matrícula: 18111679

1. Faça essa questão manualmente. Você pode digitar as contas (faça de uma forma que eu entenda), não é necessário fazer a mão. (Usei o MathType para transformar LaTeX em uma imagem com as fórmulas)

   Um analista mediu o desempenho de um banco de dados com o tempo decorrido em função da complexidade de uma consulta em um banco de dados. A complexidade foi medida pelo número de palavras-chave na consulta. O número de operações de leitura de disco também foi medido, conforme mostrado na Tabela a seguir. Para esses dados, prepare o modelo de regressão para prever o tempo decorrido em função do número de palavras-chave e interpretar os resultados respondendo as questões a seguir:

   a. Determine os valores dos parâmetros do seu modelo.

$$\bar{x} = \frac{1+2+4+8+16}{5} = 6.2$$

$$\bar{y} = \frac{0.75+0.7+0.8+1.28+1.60}{5} = 1.026$$

$$\sum x^2 = 1^2 + 2^2 + 4^2 + 8^2 + 16^2 = 341$$

$$\sum xy = (0.75 + (2 \times 0.7) + (4 \times 0.8) + (8 \times 1.28) + (16 \times 1.60)) = 41.19$$

$$b_1 = \frac{\sum xy - n\bar{x}\bar{y}}{\sum x^2 - n(\bar{x})^2} = \frac{41.19 - 5 \cdot 6.2 \cdot 1.026}{341 - 5 \cdot (6.2)^2} = \frac{9.384}{148.8} = 0.063064516$$

$$b_0 = \bar{y} - b_1\bar{x} = 1.026 - 0.063064516 \cdot 6.2 = 0.635$$

   b. Qual a porcentagem da variação é explicada pela regressão? Quais parâmetros são significativos, com uma confiança de 90%? E de 95% ?

   Tanto b0 como b1 são significativos, pois não incluem o 0 no intervalo de confiança nem para 90% de confiança, nem para 95% de confiança. Fato que será mostrado nos cálculos abaixo.

$$\sum y^2 = 0.75^2 + 0.7^2 + 0.8^2 + 1.28^2 + 1.60^2 = 5.8909$$

$$\sum y = 0.75 + 0.7 + 0.8 + 1.28 + 1.60 = 5.13$$

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum y^2 - b_0 \sum y - b_1 \sum xy}{\left(\sum y^2\right) - n(\bar{y})^2} = 1 - \frac{5.8909 - 0.635 \cdot 5.13 - 0.0631 \cdot 41.19}{5.8909 - 5 \cdot 1.053}$$

$$R^2 = 1 - \frac{0.034261}{0.6259} = 1 - 0.054738776 = 0.945261224 = 94.5\%$$

$$S_{b_0} = S_e \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum x^2 - n\bar{x}^2}\right]^{\frac{1}{2}} \quad e \quad S_{b_1} = \frac{S_e}{\left[\sum x^2 - n\bar{x}^2\right]^{\frac{1}{2}}}$$

$$S_e = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{0.0343}{3}} = 0.107$$

$$S_{b_0} = 0.107 \cdot \left[\frac{1}{5} + \frac{38.44}{341 - 5 \cdot 38.44}\right]^{\frac{1}{2}} = 0.072439; \quad e \quad S_{b_1} = \frac{0.107}{[341 - 5 \cdot 38.44]^{\frac{1}{2}}} = 0.00877$$

Intervalo de Confiança para 90%:

$$IC_{b_0} = 0.635 \pm 2.353 \cdot 0.072439 = (0.464551033,\ 0.805448967)$$

$$IC_{b_1} = 0.063 \pm 2.353 \cdot 0.00877 = (0.04236419,\ 0.08363581)$$

Intervalo de Confiança para 95%:

$$IC_{b_0} = 0.635 \pm 3.182 \cdot 0.072439 = (0.404499102,\ 0.865500898)$$

$$IC_{b_1} = 0.063 \pm 3.182 \cdot 0.00877 = (0.03509386,\ 0.09090614)$$

Number of Keywords	Elapsed Time	Number of Disk Reads
1	0.75	3
2	0.70	6
4	0.80	7
8	1.28	78
16	1.60	92

2. Nessa questão você deve usar uma ferramenta de programação para fazer a regressão linear. Indique bibliotecas e funções que você usou. Anexe o link do repositório do seu código.

Repositório:
https://github.com/MarechalLima/fluffy-potato/tree/main/PE/S9

Bibliotecas usadas: numpy, sklearn, statsmodels, pandas, matplotlib

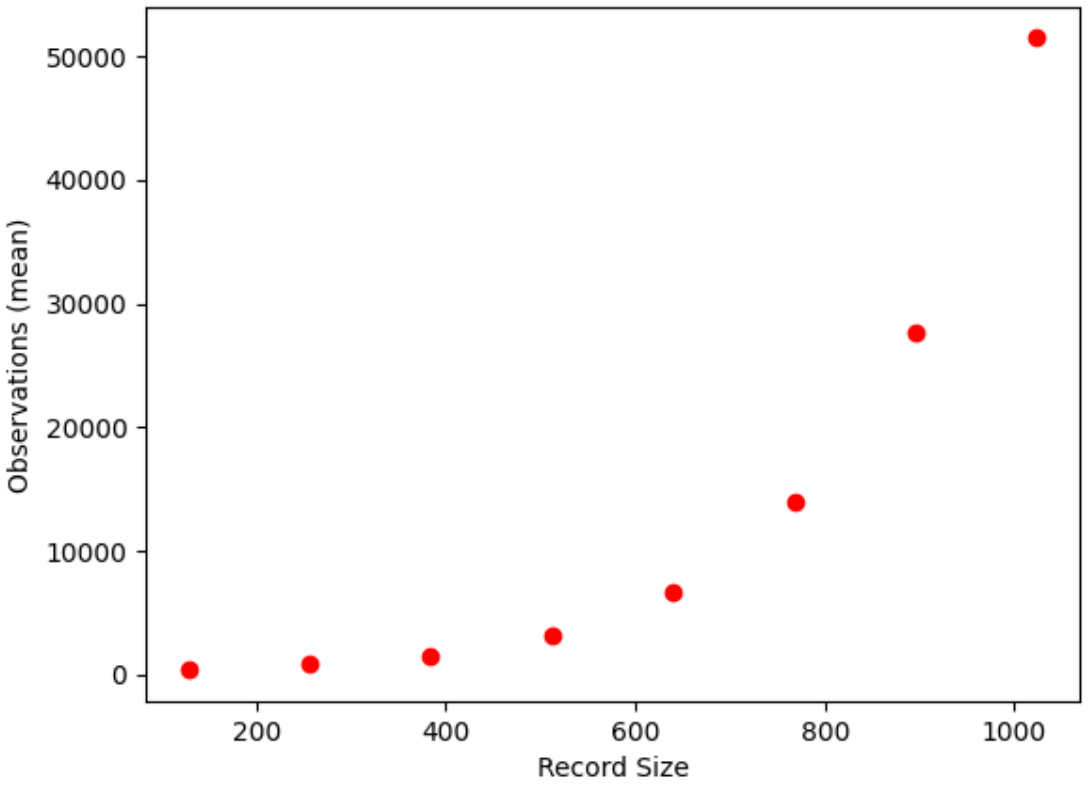
Várias funções numéricas do numpy foram usadas, como np.array(e a opção de reshape para mudar a forma do array), np.mean para calcular a média (no axis=1). Já no sklearn, utilizou-se a classe Linear Regression e suas funções. Statsmodels foi utilizado tanto para criar um modelo de Linear Regression (também), como para calcular o intervalo de confiança (já que o sklearn não tem isso implementado), e foi usada também para o qqplot (quartile-quartile plot). O pandas foi usado tanto por causa da função pd.read_csv, como para a criação de um Dataframe na quarta questão, já que o predict do statsmodels OLS (linear regression) espera um vetor do shape (2,) (duas dimensões) como entrada. E por fim, o matplotlib foi usado para plotar os gráficos, tanto scatter como line plot.

O tempo para criptografar um registro de k-bytes usando uma técnica de criptografia é mostrado na Tabela a seguir. Ajuste um modelo de regressão linear a esses dados respondendo às questões a seguir.

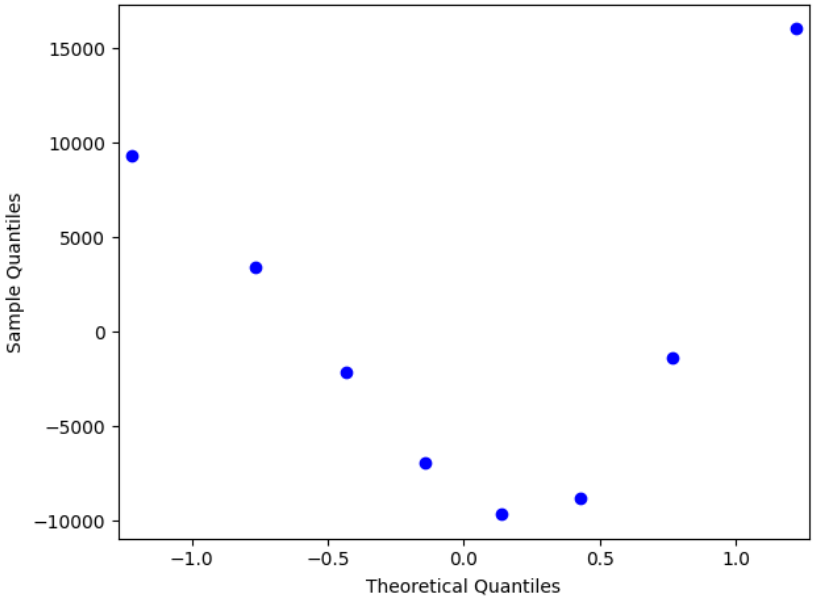
Record Size	Observations		
	1	2	3
128	386	375	393
256	850	805	824
384	1,544	1,644	1,553
512	3,035	3,123	3,235
640	6,650	6,839	6,768
768	13,887	14,567	13,456
896	28,059	27,439	27,659
1,024	50,916	52,129	51,360

a. Faça os testes visuais para verificar se a regressão é adequada a esses dados.

'Record Size' foi usado como variável independente, e 'Observations' foram usadas como a variável dependente, e como as três observações possuem o mesmo significado, sendo apenas várias execuções de uma mesma coisa, optou-se por utilizar a média como um parâmetro para descrever as 3 observações de forma mais simples. E como pode ser observado na figura abaixo que as variáveis são ***não-lineares****.*

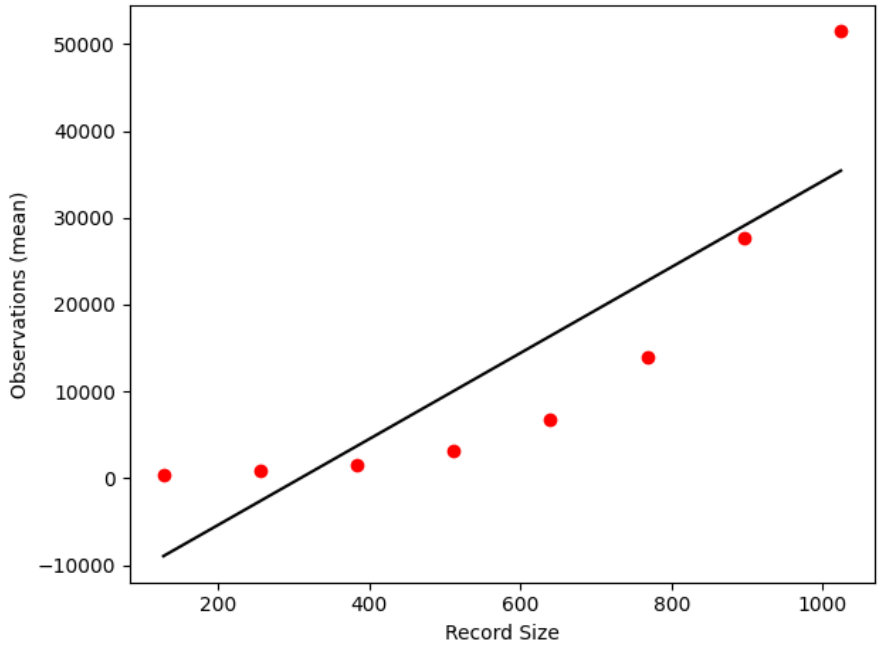


*E se analisarmos o gráfico **Quantil de Erro (gráfico abaixo)**, iremos perceber que além de ser não-normal, o gráfico tem uma clara **Tendência**, corroborando mais uma vez que as variáveis são **não-lineares**.*



b. Ajuste um modelo de regressão linear a esses dados.

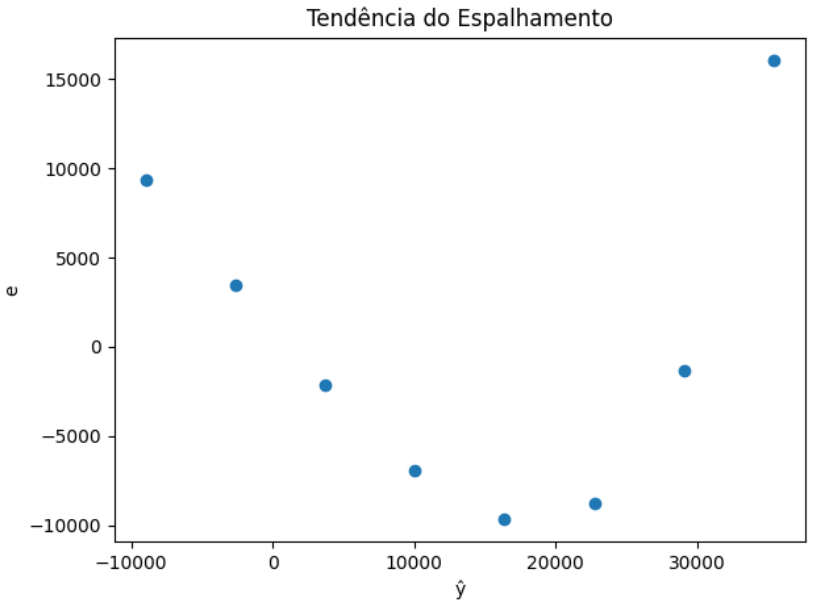
Assim como na questão anterior, a médias das observações (Observations) foi usada como valor de y, e o tamanho do registro foi utilizado como a variável independente x, a linha de regressão foi determinada pela função predict do sklearn, função que calcula os valores de y da reta de regressão com base nos valores de x.



c. Qual a porcentagem da variação é explicada pela regressão? Você está satisfeito com seu modelo? Se não, qual seria o seu próximo passo

R^2 = 0.7441049359691524 ≈ 75%

O próximo passo seria analisar como está o espalhamento do desvio padrão.



Depois desses passos vemos que até o espalhamento tem tendência crescente, nenhum outro teste de regressão linear sobram, a única coisa que sobra seria aplicar uma outra regressão, a curvilínea por exemplo, que serve para relações não-lineares.

d. Quais parâmetros são significativos, com uma confiança de 90%?

$$CI_{b0} = (-30217.75334486,\ -414.17522657)$$

$$CI_{b1} = (26.50258931,\ 72.61187001)$$

*Nenhum dos dois parâmetros inclui o 0, logo com uma confiança de 90%, tanto **b0** como **b1** são **significativos**.*

e. Qual o tempo esperado para criptografar um registro de 2ˆ20kbits? Quais limites você colocaria para esta estimativa se você aceita um erro máximo de 10% para uma única medida futura?

*Utilizando o modelo já criado, podemos utilizar model.predict((2^20)/8) para estipularmos qual poderia ser o valor aproximado. Contudo, é válido lembrar que as relações desse modelo são não-lineares, logo, a estimativa dada pelo programa é provavelmente muito divergente do real valor. Contudo o **IC da estimativa é: (-4467032242.948143, 452044101.23385954)***

3. Explique e exemplifique os seguintes conceitos encontrados na regressão linear múltipla: Analysis of Variance (ANOVA), F-Test e multicolinearidade. O seu exemplo deve ser diferente do livro. Você não precisa fazer os cálculos, apenas mostrar com o exemplo como esses conceitos são aplicados na regressão linear múltipla.

ANOVA: Analysis of variance (ANOVA) é uma ferramenta de análise usada em estatística que divide uma variação encontrada em duas partes, fatores sistemáticos e fatores aleatórios. Os fatores sistemáticos têm influência estatística nos dados, enquanto que fatores aleatórios são apenas ruídos. O teste ANOVA é usado para determinar a influência que uma variável independente tem sobre uma variável dependente em um modelo de regressão. Exemplo: Suponha que uma pessoa está testando uma nova forma de fazer bolo, e cada bolo é realizado com os mesmos ingredientes, mudando apenas o tempo em que o bolo fica no forno e percebeu-se que o resultado de crescimento do bolo não foi linear. Aqui temos que a variável independente é a temperatura do forno e o tempo que o bolo ficou no forno, e como variável dependente tem-se o crescimento do bolo. Porém, existe uma outra variável dependente que não foi levada em consideração e que interferia da variável independente, gerando ruído, que seria o tempo levado entre pôr o fermento na massa e adicionar os líquidos, valor o qual interfere também no resultado final, que seria o crescimento do bolo.

F-test: F-test é um termo geral para descrever teste que usam a F-distribuição. Um F-test irá dizer se um grupo de variáveis são significantes conjuntamente. O F-test em uma análise de variância unilateral é usada para avaliar se os valores esperados de uma variável preditora dentro de um grupo de variáveis diferem uma das outras. Por exemplo, considerando o exemplo anterior é possível avaliar se alguma das versões do bolo é em média superior ou inferior aos outros em relação à hipótese nula de que todos os outros bolos produzem a mesma resposta média.

Multicolinearidade: É um problema comum em regressões lineares, na qual as variáveis independentes possuem relações lineares exatas ou quase exatas. Uma pista de que isso possa estar acontecendo é um valor de R^2 muito alto, mas nenhum dos coeficientes é estatisticamente significativo pelo t-test. O exemplo claro desse evento estão nos dados da quarta questão, que possui um R^2 bastante alto (90,25%), mas que nenhum dos coeficientes apresenta significância ao nível de 90%.

4. Os resultados de uma regressão múltipla baseada em nove observações são mostrados na Tabela a seguir. Com base nesses resultados, responda às seguintes perguntas. Justifique todas as respostas.

a. Qual porcentagem de variância é explicada pela regressão?

   Pela própria questão temos que R = 0.95, e consequentemente R^2 = 0.9025 = 90.25%, dessa forma temos que 90.25% da variância é explicada pela regressão.

b. A regressão é significativa no nível de confiança de 90%?

$$S_e = \sqrt{\frac{SSE}{n-k-1}} \Rightarrow SSE = (n-k-1) \cdot {S_e}^2$$

$$SSE = (9-4-1) \cdot 12^2 = 576$$

$$R^2 = \frac{SSR}{SST} = \frac{SSR}{SSR + SSE} = \frac{SSR}{SSR + 576} = 0.9025$$

$$SSR = 0.9025 \cdot (SSR + 576) = 0.9025(SSR) + 519.84$$

$$SSR = \frac{519.84}{0.0975} = 5331.69923$$

$$MSR = \frac{SSR}{k} = \frac{5331.69923}{4} = 1332.9248$$

$$MSE = {S_e}^2 = 12^2 = 144$$

$$\frac{MSR}{MSE} = 9.2564$$

$$F - value(0.9,\ 4,\ 4) = 4.1073$$

Como MSR/MSE é maior do que o F-value(0.9, 4, 4), podemos dizer que com 90% de confiança, a ***regressão é significativa****.*

c. Qual variável tem o coeficiente mais alto?

A variável que possui o coeficiente mais alto, é a variável que possui bj = 5, que é a que tem j = 4

d. Qual variável é mais significativa?

<u>Valores de P aproximados, extraídos de uma tabela t-student bicaudal</u>

p-value variabel 1 = 2P(t> 1.3/3.6 | t~t_4) = 2P(t> 0.36 | t~t_4) = 0.7

p-value variabel 2 = 2P(t> 2.7/1.8 | t~t_4) = 2P(t> 1.5 | t~t_4) = 0.2

p-value variabel 3 = 2P(t> 0.5/0.6 | t~t_4) = 2P(t> 0.83 | t~t_4) = 0.4

p-value variabel 4 = 2P(t> 5/8.3 | t~t_4) = 2P(t> 0.61 | t~t_4) = 0.6

Logo, podemos dizer que a variável mais significativa é ***variável 2,*** *pois ela possui o menor valor p.*

e. Quais parâmetros não são significativos em 90%?

*IC: bj ±t*Sbj e t(1-0.1/2, 4) = 2.132*

*b1 = 1.3 ± 2.132*3.6 = (−6.3752, 8.9752)*

*b2 = 2.7 ± 2.132*1.8 = (−1.1376, 6.5376)*

*b3 = 0.5 ± 2.132*0.6 = (−0.7792, 1.7792)*

*b4 = 5.0 ± 2.132*8.3 = (−12.6956, 22.6956)*

Logo, nenhum parâmetro é significativo ao nível de 90%, pois todos tiveram 0 incluído em seu intervalo, e por isso não é possível dizer que são significativos.

f. Qual é o problema com essa regressão?

Pode ser o problema da multicolinearidade.

g. O que você tentaria a seguir para resolver o problema?

Calcular a correlação entre as variáveis preditoras e dentre os pares que tiverem uma alta correlação, testar a regressão com cada preditor e separadamente encontrar aquele que obtém o melhor R^2, diminuindo assim o número de variáveis preditoras.

j	b_j	s_{bj}
1	1.3	3.6
2	2.7	1.8
3	0.5	0.6
4	5.0	8.3

Intercept = 75.3
Coefficient of multiple correlation = 0.95
Standard deviation of errors = 12.0
F-value = 14.1