Cálculo Numérico – Prova 2

Projeto B – Internação de pacientes com COVID-19

Contexto: As infecções causadas pelo coronavírus (COVID-19) impactam profundamente a saúde dos pacientes, com efeitos variando de sintomas leves a complicações graves, como insuficiência respiratória e danos permanentes. A evolução da doença é imprevisível, e a identificação de fatores que possam prever desfechos clínicos (outcomes), como a alta ou óbito, é essencial para guiar intervenções médicas precoces e otimizar o tratamento.

Os exames de sangue são uma ferramenta acessível e não invasiva para monitorar a progressão da COVID-19. Marcadores como níveis de linfócitos, monócitos, plaquetas e outros têm sido amplamente utilizados para avaliar o estado inflamatório e a função dos órgãos dos pacientes infectados. Além disso, com o avanço da tecnologia e da análise de dados, o monitoramento remoto por meio de exames laboratoriais torna-se uma possibilidade cada vez mais viável, permitindo a detecção precoce de agravamento da condição clínica.

Neste estudo, você será responsável pela análise de dados de exames de sangue de pacientes com COVID-19, focando em como essas variáveis estão associadas ao desfecho clínico de cada paciente. O objetivo não é prever a infecção em si, mas entender como os diferentes parâmetros sanguíneos podem ser usados para prever a gravidade e a evolução da doença, auxiliando na tomada de decisões clínicas.

Dados: Um conjunto de dados foi obtido durante um estudo com pacientes internados em UTI. Cada coluna na tabela representa informação do paciente, desde gênero até informações do sangue.

Os dados estão no arquivo “COVID-19_CBC_Data_cleaned.csv”. Um total de 13 de dados relacionados ao paciente estão disponíveis. Quando os dados são carregados no arquivo p2-b.m na linha 18, as colunas estão na ordem abaixo, sendo a primeira coluna o “outcome”.

As colunas são:

1 – Outcome – Desfecho médico para 0 sendo “não recuperado” e 1 sendo “recuperado”

2 – Patient age – Idade do paciente

3 - Gender – Gênero do paciente, 0 para feminino e 1 para masculino

4 – Ventilation (Y/N) – Se foi utilizada ventilação, 1 para “sim” e 2 para “não”

5 – Red blood cell distribution width – Distribuição de hemáceas

6 – Monocytes(%) – Quantidade relativa de monócitos

7 – White blood cell count – Quantidade de leucócitos

8 – Platelet count – Quantidade de plaquetas

9 – Lymphocyte Count – Quantidade de linfócitos

10 – Neutrophils Count – Quantidade de neutrófilos

11 – Days Hospitalized – Quantidade de dias internado

Questionamentos: seu trabalho deve conter, obrigatoriamente, as análises que seguem, mas outras avaliações que o grupo julgar pertinentes podem ser consideradas. Utilize o Octave para responder:

Análise 1: Seleção das variáveis

1.1. Com base nos dados fornecidos, escolha dois pares de variáveis que tenham o melhor ajuste para uma regressão linear. Isso significa olhar quais pares de variável possuem um comportamento linear entre si.

Por exemplo, investigue as relações entre a **idade do paciente** (Patient age) e a **contagem de leucócitos** (White blood cell count) ou entre a **contagem de neutrófilos** (Neutrophils Count) e a **contagem de linfócitos** (Lymphocyte Count). Justifique sua seleção com base na análise gráfica.

1.2. Tomando os pares na questão (1), utilize análises numéricas vistas em aula para ajudar a embasar sua decisão sobre qual par de variáveis melhor prevê a outra. Justifique sua resposta comentando os resultados obtidos.

Análise 2: Comparação estatística das métricas

2.1. Essa análise é independente da Análise 1. Divida os pacientes em grupos etários (por exemplo, menores de 40, entre 40 e 60, e maiores de 60 anos). Avalie se há diferenças significativas entre esses grupos nas contagens de células sanguíneas (como linfócitos, neutrófilos e leucócitos) em relação ao desfecho clínico.

2.2. Faça plots que julgar pertinentes para mostrar as diferenças de idades e seus desfechos em relação às células sanguíneas.

Sugestão: gráficos de densidade, histogramas ou violinos podem ser interessantes.

Análise 3: Predição

3.1. Faça dois modelos de regressão linear. Aqui você deve implementar um modelo $y_1 = a_{0,1} + a_{1,1}x_1$ e outro $y_2 = a_{0,2} + a_{1,2}x_2$ onde x_1 e x_2 são variáveis que melhor preveem dias hospitalizados. Calcule as somas dos resíduos S_r, r^2 e $s_{y/x}$ para cada modelo. Qual é melhor? Justifique.

3.2. Implemente um terceiro modelo de regressão linear múltipla do tipo $y_3 = a_{0,3} + a_{1,3}x_1 + a_{2,3}x_2$ que combina as duas métricas utilizadas em (3.1).

Isso significa que você deve encontrar os parâmetros $a_{0,3}$, $a_{1,3}$ e $a_{2,3}$ de tal maneira que a regressão $y_3 = a_1x_1 + a_2x_2 + a_0$ seja o melhor possível. Calcule qual a soma dos resíduos S_r, r^2 e $s_{y/x}$ para o modelo y_3. Compare-o com y_1 e y_2. Qual é o melhor? Justifique.

3.3. Verificando as métricas do modelo obtido em (3.1), você diria que esse é um bom modelo de regressão para prever o paciente com a doença? Justifique sua resposta.

Atenção: essa não é uma prova de aprendizado de máquina, mas de cálculo numérico. Logo, não espero uma análise utilizando métricas rebuscadas, discussões sobre modelos mais robustos ou que seu modelo performe bem. Quero que seu grupo foque na utilização do modelo construído, usando a ferramenta Octave e a interpretação dos resultados obtidos segundo a teoria dada em aula.

Códigos: o arquivo que deve resolver seu projeto é dado “p2_b.m”. Esse código já carrega os dados e faz o plot de dispersão (scatter) das métricas duas a duas. Você pode criar outros scripts .m caso precise, mas eu só devo precisar rodar “p2_b.m” para verificar os entregáveis do seu projeto.

Funções proibidas:

- polytfit, linsolve, regress, interp1, interp2, interp3, interpn, spline, fitlm, compact, fitrlinear, mvregress, mvregresslike, plsregress.
- Também é proibido resolver sistemas lineares com o método da inversa (função inv()) ou pelo operador barra invertida (\). Caso você precise, utilize a função lu() para obter a decomposição e implemente uma pequena função que resolve o sistema dado que você possui L e U.

ANEXO

Formato dos dados:

	Outcome	Patient Age	Gender	Ventilated (Y/N)	...	Platelet Count	Lymphocyte Count	Neutrophils Count	Days hospitalized
1	0	65	1	Yes	...	180.66	4.39	7.56	12
2	1	32	0	No	...	336.00	3.47	5.34	17
3	1	36	1	No	...	240.10	0.80	8.66	17
4	1	46	0	No	...	236.58	7.93	13.02	18
5	0	17	0	Yes	...	249.00	4.12	8.15	21

Figura 1. Cinco primeiras linhas dos dados em COVID-19_CBC_Data_cleaned.csv

Pairplots dos dados: