ÁLGEBRA LINEAR NUMÉRICA
AULA PRÁTICA 4
MÉTODO DOS MÍNIMOS QUADRADOS

1) (*Texto do livro Cálculo - Volume 2 – James Stewart)* Em 1928, Charles Cobb e Paul Douglas publicaram um estudo no qual modelaram o crescimento da economia norte-americana durante o período de 1899-1922. Eles consideraram uma visão simplificada da economia em que a saída da produção é determinada pela quantidade de trabalho envolvido e pela quantidade de capital investido. Apesar de existirem muitos outros fatores afetando o desempenho da economia, o modelo mostrou-se bastante preciso. A função utilizada para modelar a produção era da forma

$$P = bL^{\propto}K^{1-\propto}$$

onde *P* é a produção total (valor monetário dos bens produzidos no ano); *L* é a quantidade de trabalho (número total de pessoas-hora trabalhadas no ano); e *K* é a quantidade de capital investido (valor monetário das máquinas, equipamentos e prédios); *b* e α são parâmetros (constantes) a serem determinados.

Cobb e Douglas usaram os dados da tabela a seguir e o Método dos Mínimos Quadrados para obter os valores de *b* e de *α*.

Ano	*P*	*L*	*K*
1899	100	100	100
1900	101	105	107
1901	112	110	114
1902	122	117	122
1903	124	122	131
1904	122	121	138
1905	143	125	149
1906	152	134	163
1907	151	140	176
1908	126	123	185
1909	155	143	198
1910	159	147	208
1911	153	148	216
1912	177	155	226
1913	184	156	236
1914	169	152	244
1915	189	156	266
1916	225	183	298
1917	227	198	335
1918	223	201	366
1919	218	196	387
1920	231	194	407
1921	179	146	417
1922	240	161	431

a) Faça como Cobb e Douglas: use o Método dos Mínimos Quadrados para estimar os valores dos parâmetros *b* e α. Mostre a sua modelagem para o problema ser resolvido pelo Método dos Mínimos Quadrados.
b) Agora, use a função de Cobb-Douglas encontrada no item a) e teste a sua adequação calculando os valores da produção nos anos de 1910 e 1920. Comente!

2) Agora vamos usar o método dos mínimos quadrados para implementar um método rudimentar de "machine learning" para diagnosticar câncer de mama a partir de um conjunto de características fornecidas para cada paciente. São dados dois arquivos: um arquivo para "treinamento" (cancer_train.csv) do modelo e um arquivo para "teste" (cancer_test.csv).
O primeiro arquivo contém 300 registros e o segundo 260 registros, partes do "Wiscosin Diagnostic Breast Cancer dataset". Cada registro de cada arquivo contém 11 valores: os 10 primeiros correspondem a valores reais de 10 características dos núcleos celulares observados em imagens digitalizadas de uma fina camada de massa mamária coletada de cada paciente. O décimo primeiro valor é +1 se a paciente tem câncer de mama e -1, caso contrário.
Sendo **x** o vetor das 10 características de cada paciente (variáveis independentes) e y o valor (+1 ou -1) que indica o diagnóstico (variável dependente), a ideia é, usando o arquivo de treinamento, obter o hiperplano $y = h(\boldsymbol{x})$ ($y = \alpha_0 + \sum_{i=1}^{10} \alpha_i x_i$) que "melhor se ajuste aos dados fornecidos" usando o método dos mínimos quadrados.
Uma vez obtido o hiperplano, o mesmo será usado para classificar cada paciente da seguinte forma: se $h(\boldsymbol{x}) \geq 0$, então o diagnóstico é +1 (tem câncer), caso contrário, o diagnóstico é −1 (não tem câncer).
Use o seu classificador (hiperplano) e calcule a porcentagem de acertos sobre o arquivo de treinamento (de certa forma é uma medida do ajuste do seu modelo aos dados de treinamento) e sobre o arquivo de teste (de certa forma é uma medida da capacidade de generalização do seu modelo).
Construa uma Matriz de Confusão (Confusion Matrix) (pesquise a respeito) com o conjunto de teste e calcule as diversas medidas daí decorrentes, tais como: acurácia, precisão, recall, probabilidade de falso alarme, probabilidade de falsa omissão de alarme. Interprete essas medidas e comente os resultados obtidos.