MAT4MDS — Practice 5

Linear least squares regression

The linear least squares solution to

$$AX = b$$

is given by the solution to

$$A^T AX = A^T b, \text{ that is, } X = (A^T A)^{-1} A^T b.$$

Given n data points $(x_1 y_1), \ldots, (x_n y_n)$, the linear least squares line of best fit is

$$y = \alpha x + \beta$$

where α and β are found by solving

$$\begin{bmatrix} \sum_{i=1}^{n} x_i^2 & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & n \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} y_i \end{bmatrix}.$$

You will need a calculator to solve most of the numerical questions.

Some questions require a brief explanation.

Question 1. Scientists in the Soviet Antarctic Expeditions at Vostok analysed ice core samples from various depths down to 1000 m to estimate the concentration of carbon dioxide (CO_2) in the atmosphere during the last 60 000 years. Shown below are some of their data with CO_2 levels given in parts per million (ppm) and depth in metres.

Depth (m)	127	303	376	474	602	748	853	976
CO_2 (ppm)	275	259	245	195	223	179	201	201

(a) Determine the linear least-squares line of best fit for estimating CO_2 level from depth.
(b) Use your answer to (a) to estimate the CO_2 concentration at a depth of 1050 m.
(c) Interpret the value of the slope of the regression line in this context.

Question 2. Use the same data given in Question 1.

(a) Estimate the depth at which an ice core sample was obtained if it is known that its CO_2 concentration is 200 ppm.

[**Hint:** which is the independent variable and which is the dependent variable?]

(b) Explain carefully why you needed a different line of best fit in Question 2a from the one derived in Question 1.

Question 3. Consider the following data:

x	0	1	2	3
y	1	1	3	6

(a) From the data above:

(i) Determine the least squares line of best fit for these data.
[For practice, you should do this without the aid of a calculator — it is fairly simple arithmetic.]

(ii) Calculate the residual for each of the data points.

(iii) Calculate the sum of the squares of the residuals for the line of best fit.

(b) Calculate the sum of the squares of the residuals for the line $y = 1.6x + 0.3$.

(c) Compare the two sums of squares of residuals you found in parts (a) and (b).

(d) Explain briefly and carefully the significance of the sum of the squares of the residuals found in part (a).

Question 4. Global warming has many indirect effects on climate. For example, the summer monsoon winds in the Arabian Sea bring rain to India and are critical for agriculture. As the climate warms and winter snow cover in the vast landmass of Asia and Europe decreases, the land heats more rapidly in the summer. This may increase the strength of the monsoon. Here are data on snow cover (in millions of square kilometres) and summer wind stress (in newtons per square meter, N/m^2; 1 N is the force required to accelerate a 1 kg mass at the rate of 1 m/s^2). The data have been taken from a much larger dataset.

Snow cover (10^{6}km^2)	6.6	7.7	8.1	18.2	26.6	29.4
Wind stress (N/m^2)	0.125	0.155	0.196	0.106	0.062	0.024

(a) Determine the least squares line of best fit used to estimate wind stress from snow cover.

(b) What does this model reveal about the nature and strength of the effect of decreasing snow cover on wind stress?

Question 5. The following data gives the population of India in the census years since 1951.

Year	1951	1961	1971	1981	1991	2001	2011
Population (millions)	361	439	548	683	846	1029	1211

Source: Wikipedia. (n.d.). *Census of India*. Wikipedia. Retrieved 8 December, 2021, from https://en.wikipedia.org/wiki/Census_of_India

(a) Assuming that the population growth is exponential over this period, determine the exponential growth rate model

$$\text{population} = P_0\, e^{\alpha x}$$

by first finding the least squares line of best fit $y = \alpha x + \beta$, where x is the independent variable

$$x = \text{year} - 1951$$

and y is the dependent variable

$$y = \ln(\text{population})$$

(b) According to this model, what is the estimate of the population in 2011? Compare that value to the actual population. [In this case, the discrepancy is due to the lower growth rate from 2001 to 2011 than in the period 1951 to 2001.]