Sample: Linear Algebra - Approximation Assignment

3) Often we have a set of data from an experiment and we want to fit a straight line (or other curve) to that data. A common approach to this is to use the method of least squares to find the line of the best fit. This approach involves solving a linear system called the normal equation.

An experiment was conducting with the following results

x	0	1	2	3	4
y	-0.18	2.90	7.12	14.91	23.63

We would like to determine if x and y have a linear or quadratic relationship.

(a) Research using the method of least squares to find the line of best fit to find an expression for the normal equation.

Suppose we want to fit the data using the curve

$$y = a_0 f_0(x) + a_1 f_1(x) + a_2 f_2(x)$$

Substituting data entries into this equation we get:

$$y_i = a_0 f_0(x_i) + a_1 f_1(x_i) + a_2 f_2(x_i), i = 1,2, \dots, N$$

Let's find sum of squared differences:

$$F(a_0, a_a, a_2) = \sum_{i=1}^{N} \left(y_i - a_0 f_0(x_i) - a_1 f_1(x_i) - a_2 f_2(x_i)\right)^2$$

Let's minimize this function by parameters a_0, a_1, a_2.

$$\frac{\partial F}{\partial a_0} = 2\sum_{i=1}^{N} \left(y_i - a_0 f_0(x_i) - a_1 f_1(x_i) - a_2 f_2(x_i)\right)\left(-f_0(x_i)\right) = 0$$

$$\frac{\partial F}{\partial a_1} = 2\sum_{i=1}^{N} \left(y_i - a_0 f_0(x_i) - a_1 f_1(x_i) - a_2 f_2(x_i)\right)\left(-f_1(x_i)\right) = 0$$

$$\frac{\partial F}{\partial a_2} = 2\sum_{i=1}^{N} \left(y_i - a_0 f_0(x_i) - a_1 f_1(x_i) - a_2 f_2(x_i)\right)\left(-f_2(x_i)\right) = 0$$

This system of equations determines coefficients a_0, a_1, a_2 of best fit line.

(b) Use the normal equation to fit a straight line $y = a_0 + a_1 x$ to the data.
In case

$$y = a_0 + a_1 x$$

we have:

$$f_0(x) = 1; f_1(x) = x$$

Thus we have system of equations:

$$\begin{cases} \sum_{i=0}^{4} y_i - a_0 - a_1 x_i = 0 \\ \sum_{i=0}^{4} (y_i - a_0 - a_1 x_i) x_i = 0 \end{cases}$$

$$\begin{cases} 5a_0 + a_1 \sum_{i=1}^{4} x_i = \sum_{i=1}^{4} y_i \\ a_0 \sum_{i=1}^{4} x_i + a_1 \sum_{i=1}^{4} x_i^2 = \sum_{i=1}^{4} x_i y_i \end{cases}$$

Substituting values we have:

$$\begin{cases} 5a_0 + 10a_1 = 48.38 \\ 10a_0 + 30a_1 = 156.39 \end{cases}$$

Solution of this system is

$$a_0 = -\frac{9}{4} = -2.25$$

$$a_1 = \frac{5963}{1000} = 5.963$$

Thus best fit line is

$$y = -2.25 + 5.963x$$

(c) Use the normal equation to fit a parabola $y = a_0 + a_1 x + a_2 x^2$ to the data.

In case

$$y = a_0 + a_1 x + a_2 x^2$$

we have:

$$f_0(x) = 1; f_1(x) = x; f_2(x) = x^2$$

Thus system of equations is such:

$$\begin{cases} \sum_{i=0}^{4} y_i - a_0 - a_1 x_i - a_2 x_i^2 = 0 \\ \sum_{i=0}^{4} (y_i - a_0 - a_1 x_i - a_2 x_i^2) x_i = 0 \\ \sum_{i=0}^{4} (y_i - a_0 - a_1 x_i - a_2 x_i^2) x_i^2 = 0 \end{cases}$$

$$\begin{cases} 5a_0 + a_1 \sum_{i=1}^{4} x_i + a_2 \sum_{i=1}^{4} x_i^2 = \sum_{i=1}^{4} y_i \\ a_0 \sum_{i=1}^{4} x_i + a_1 \sum_{i=1}^{4} x_i^2 + a_2 \sum_{i=1}^{4} x_i^3 = \sum_{i=1}^{4} x_i y_i \\ a_0 \sum_{i=1}^{4} x_i^2 + a_1 \sum_{i=1}^{4} x_i^3 + a_2 \sum_{i=1}^{4} x_i^4 = \sum_{i=1}^{4} x_i^2 y_i \end{cases}$$

Substituting all x_i and y_i we have:

$$\begin{cases} 5a_0 + 10a_1 + 30a_2 = 48.38 \\ 10a_0 + 30a_1 + 100a_2 = 156.39 \\ 30a_0 + 100a_1 + 354a_2 = 543.65 \end{cases}$$

Solution of this system is

$$a_0 = -\frac{9}{70} = -0.1286$$

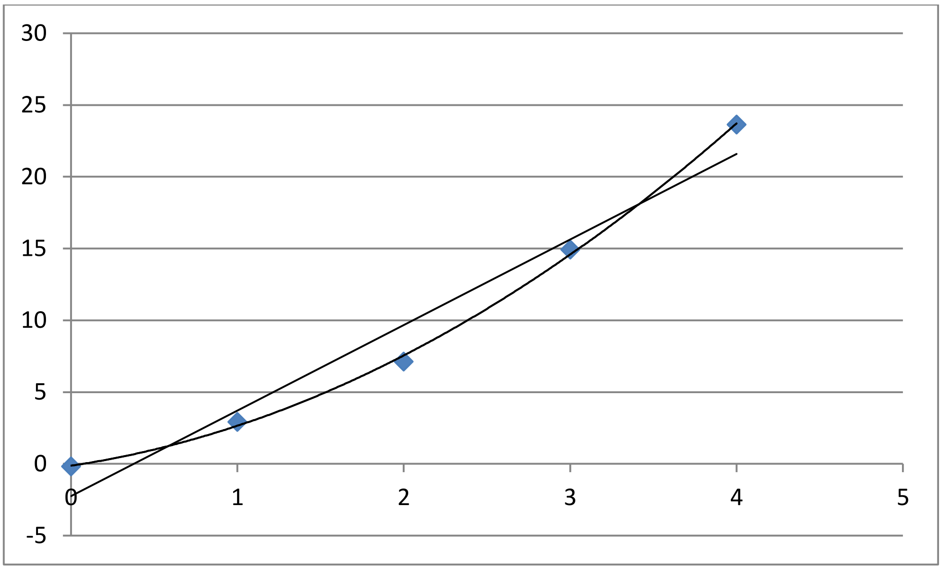
$$a_1 = \frac{12041}{7000} = 1.7201$$

$$a_2 = \frac{297}{280} = 1.0607$$

So quadratic fit function is

$$y = -0.1286 + 1.7201x + 1.0607x^2$$

(d) Graph the data using appropriate software and your fitted curves on the same axis. Which do you think fits the data best?



Quadratic line fits the data better.

(e) Construct the research a mathematical method to determine which line fits best. Describe the method and use it to determine which of the curves gives the best fit.

The line gives better fit if sum of squared differences between data points and points on the line is lesser.

Sum of squares for line fit:

$$S_1^2 = \sum_{i=1}^{N} (y_i - a_0 - a_1 x_i)^2 = 16.1232$$

Sum of squares for parabola fit:

$$S_2^2 = \sum_{i=1}^{N} \left(y_i - a_0 - a_1 x_i - a_2 x_i^2\right)^2 = 0.2716$$

We see that sum of squared differences is much lesser for quadratic model. Thus parabola fits the data better.

4) Consider the matrix

$$A = \begin{bmatrix} 4 & 8 & -2 & 10 & -16 \\ 9 & 18 & -7 & 30 & -41 \\ -3 & -6 & 2 & -9 & 13 \end{bmatrix}$$

(a) Find the basis for the null space of A.

Let's reduce matrix A to row echelon form:

$$A = \begin{bmatrix} 4 & 8 & -2 & 10 & -16 \\ 9 & 18 & -7 & 30 & -41 \\ -3 & -6 & 2 & -9 & 13 \end{bmatrix} \sim \begin{bmatrix} 1 & 2 & -\frac{1}{2} & \frac{5}{2} & -4 \\ 9 & 18 & -7 & 30 & -41 \\ -3 & -6 & 2 & -9 & 13 \end{bmatrix} \sim$$

$$\sim \begin{bmatrix} 1 & 2 & -\frac{1}{2} & \frac{5}{2} & -4 \\ 0 & 0 & -\frac{5}{2} & \frac{15}{2} & -5 \\ 0 & 0 & \frac{1}{2} & -\frac{3}{2} & 1 \end{bmatrix} \sim \begin{bmatrix} 1 & 2 & -\frac{1}{2} & \frac{5}{2} & -4 \\ 0 & 0 & 1 & -3 & 2 \\ 0 & 0 & \frac{1}{2} & -\frac{3}{2} & 1 \end{bmatrix} \sim \begin{bmatrix} 1 & 2 & 0 & 1 & -3 \\ 0 & 0 & 1 & -3 & 2 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Thus null space of the matrix is described by equalities:

$$x_1 = -2x_2 - x_4 + 3x_5$$

x_2 is arbitrary

$$x_3 = 3x_4 - 2x_5$$

x_4, x_5 are arbitrary

So basis for the null space of A is

$$B = \{(-2,1,0,0,0), (-1,0,3,1,0), (3,0,-2,0,1)\}$$

(b)

Let's check conditions:

$$x_1 = -2x_2 - x_4 + 3x_5$$

and

$$x_3 = 3x_4 - 2x_5$$

$-8 = -2 \cdot 2 - 1 + 3 \cdot (-1)$ – holds

$5 = 3 \cdot 1 - 2 \cdot (-1)$ – holds

So vector v is in the nullspace of A.