Course
MATH2361 Probability and Statistics
(Common for BSC BTECH, CSE, ECE, BBA)
@GITAM, deemed to be University

UNIT-III Curve fitting & Principle of Least square

Dr Malikarjuna Reddy Doodipala
MSc, M.Phil, PGDCA, Ph D
Department of Mathematics
mdoodipa@gitam.edu
Hyderabad Campus

Learning contents

- Fitting of Curves and Principle of Least squares
- Linear curve fitting
- Exponential curve fitting: (i)$y=ae^{bx}$ (ii) $y=ab^{x}$
- Power Law curve fitting: $y=ax^{b}$
- Polynomial curve fitting
- Real time Problems

Introduction to curve fitting

- Let (x, y) be a bivariate data for a set of ‘n’ points, i=1,2,3,…n.
- X being independent variable and y is dependent variable
- Curve fitting is a technique of determining trend or best expected value of variable Y using an analytic expression of the form y=f(x) suggested by original units of the data
- The Analytic expression of the form y=f(x) may be an algebraic, logarithmic, exponential and polynomial expression etc.

.

Introduction to curve fitting (cont'd..)

- Let us suppose that the bivariate data (x, y) for a set of 'n' points, i=1,2,3,…n.
- The Analytic expression of the form y=f(x) may be here, Straight-line, Second-Degree Parabola, Exponential curves, Power curve etc.
- The best trend values of Y obtained by minimizing the total residual error.
- This error is minimized by principle of least square. i.e the algebraic sum the squares of the deviations from expected value is minimum.

Fitting of a straight line

- To fit a straight line of the form .
- Let an Analytic expression of the form,

$$y = f(x) = a + bx \qquad (*)$$

for all (x, y) to a set of 'n' points, i=1,2,3,…n.

- The best trend values of Y obtained by minimizing the total residual error.
  i.e The 'best' line has minimum error between line and data points
- This error is minimized by principle of least square.
- This is called the least squares approach, since square of the error is minimized.

$$Minimize\left\{Error\ E = \sum_{i=1}^{n}[y_i - (Y')]^2\right\}$$

Where Y'= Expected or predicted value

Fitting of a straight line

$$Minimize\left\{Error = \sum_{i=1}^{n}[y_i - (a + bx_i)]^2\right\}$$

Take the derivative of the error with respect to a and b, set each to zero

$$\frac{\partial(Error)}{\partial a} = \frac{\partial\left(\sum_{i=1}^{n}\left[y_i - \left((a + bx_i)\right)\right]^2\right)}{\partial a} = 0$$

$$\frac{\partial(Error)}{\partial b} = \frac{\partial\left(\sum_{i=1}^{n}\left[y_i - \left((a + bx_i)\right)\right]^2\right)}{\partial b} = 0$$

Fitting of a straight line

$$\frac{\partial(Error)}{\partial a} = -2\sum_{i=1}^{n}[y_i - ((a + bx_i))] = 0$$

$$na + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \qquad \ldots.(1)$$

$$\frac{\partial(Error)}{\partial b} = -2\sum_{i=1}^{N} x_i[y_i - ((a + bx_i))] = 0$$

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{N} x_i y_i \qquad \ldots.(2)$$

Eqn. (1), (2) are called normal (Legendre) eqns. On solving two eqns we get values of a, b.

Fitting of a straight line

Above two normal equations can also represent in matrix notation

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix}$$

$$(or)\ a = \frac{N\sum_{i=1}^{N} x_i\, y_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{N\sum_{i=1}^{N} x_i^2 - \left[\sum_{i=1}^{N} x_i\right]^2} \qquad b = \frac{\sum_{i=1}^{N} y_i \sum_{i=1}^{N} x_i^2 - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} x_i\, y_i}{N\sum_{i=1}^{N} x_i^2 - \left[\sum_{i=1}^{N} x_i\right]^2}$$

Example 1: Fit a straight line to the following data.

x	1	6	11	16	20	26
y	13	16	17	23	24	31

Solution: Let the straight line be $Y = a + bX$ and to obtain a and b for this straight line the normal equations are

$$\sum y = na + b\sum x \quad \text{and}$$

$$\sum xy = a\sum x + b\sum x^2$$

Here, there is need of $\sum y$, $\sum x$, $\sum xy$ and $\sum x^2$ which are obtained by the following table

x	y	x^2	xy
1	13	1	13
6	16	36	96
11	17	121	187
16	23	256	368
20	24	400	480
26	31	676	806
$\sum x = 80$	$\sum y = 124$	$\sum x^2 = 1490$	$\sum xy = 1950$

Substituting the values of $\sum y, \sum x, \sum xy$ and $\sum x^2$ in the normal equations we get

$$124 = 6a + 80b \quad \ldots (12)$$

$$1950 = 80a + 1490b \quad \ldots (13)$$

Now we solve equations (12) and (13).

Multiplying equation (12) by 80 and equation (13) by 6, i.e.

$$124 = 6a + 80b \quad] \times 80$$

and

$$1950 = 80a + 1490b \quad] \times 6$$

we get,

$$9920 = 480\,a + 6400\,b \quad \ldots (14)$$

$$11700 = 480\,a + 8940\,b \quad \ldots (15)$$

Subtracting (14) from (15), we obtain

$$1780 = 2540\,b$$

$$\Rightarrow b = 1780/2540 = 0.7008$$

Substituting the value of b in equation (12), we get

$$124 = 6\,a + 80 \times 0.7008$$

$$124 = 6\,a + 56.064$$

$$67.936 = 6a$$

$$\Rightarrow a = 11.3227$$

with these values of a and b the line of best fit is $Y = 11.3227 + 0.7008\,X$.

Now let us do one problem for fitting of straight line.

E 1) Fit a straight line to the following data:

x	6	7	8	9	11
y	5	4	3	2	1

2. Second Degree Parabola

To fit the second-degree Parabola of the form . Let us consider a Second-degree parabola eqn of the form

$$y = a + bx + cx^2 --- (*)$$

to approximate the given set of data, (x_1,y_1), (x_2,y_2)…… (x_n,y_n),, the best fitting curve has the least square error, by principle of least squares

i.e.,

$$Minimize\left\{Error = \sum_{i=1}^{n}\{y_i - (y_i')\}^2\right\}$$

$$Minimize\left\{Error = \sum_{i=1}^{n}\{y_i - [a + bx_i + cx_i^2]\}^2\right\}$$

Take the derivative of the error with respect to a and b, set each to zero

$$\frac{\partial(Error)}{\partial a} = \frac{\partial\left(\sum_{i=1}^{n}\left[y_i - \left((a + bx_i + cx_i{}^2)\right)\right]^2\right)}{\partial a} = 0$$

$$\frac{\partial(Error)}{\partial b} = \frac{\partial\left(\sum_{i=1}^{n}\left[y_i - \left((a + bx_i + cx_i{}^2)\right)\right]^2\right)}{\partial b} = 0$$

$$\frac{\partial(Error)}{\partial c} = \frac{\partial\left(\sum_{i=1}^{n}\left[y_i - \left((a + bx_i + cx_i{}^2)\right)\right]^2\right)}{\partial c} = 0$$

On simplifications we get the following three normal eqns.

$$\sum y = na + b\sum x + c\sum x^2 \text{ ---------- } (1)$$

$$\sum y = a\sum x + b\sum x^2 + c\sum x_i{}^3 \text{ ------ } (2)$$

$$\sum y = a\sum x^2 + b\sum x^3 + c\sum x_i{}^4 \text{ ----- } (3)$$

On solving (1), (2) & (3) we get values of a, b and c.

Now substitute these values in (*)

we get required second degree parabola which is the best fit

Example 2: Fit a second degree parabola for the following data:

x	0	1	2	3	4
y	1	3	4	5	6

Solution: Let $Y = a + bX + cX^2$ be the second degree parabola and we have to determine a, b and c. Normal equations for second degree parabola are

$$\sum y = na + b\sum x + c\sum x^2,$$

$$\sum xy = a\sum x + b\sum x^2 + c\sum x^3, \text{ and}$$

$$\sum x^2y = a\sum x^2 + b\sum x^3 + c\sum x^4$$

To solve above normal equations, we need $\sum y$, $\sum x$, $\sum xy$, $\sum x^2y$, $\sum x^2$, $\sum x^3$ and $\sum x^4$ which are obtained from following table:

x	y	xy	x^2	x^2y	x^3	x^4
0	1	0	0	0	0	0
1	3	3	1	3	1	1
2	4	8	4	16	8	16
3	5	15	9	45	27	81
4	6	24	16	96	64	256
$\sum x = 10$	$\sum y = 19$	$\sum xy = 50$	$\sum x^2 = 30$	$\sum x^2y = 160$	$\sum x^3 = 100$	$\sum x^4 = 354$

Substituting the values of $\sum y, \sum x, \sum xy, \sum x^2y, \sum x^2, \sum x^3$ and $\sum x^4$ in above normal equations, we have

$$19 = 5a + 10b + 30c \quad \ldots (23)$$

$$50 = 10a + 30b + 100c \quad \ldots (24)$$

$$160 = 30a + 100b + 354c \quad \ldots (25)$$

Now, we solve equations (23), (24) and (25).

Multiplying equation (23) by 2, we get

$$38 = 10a + 20b + 60c \quad \ldots (26)$$

Subtracting equation (26) from equation (24)

$$50 = 10a + 30b + 100c$$

$$38 = 10a + 20b + 60c$$

$$12 = 10b + 40c \quad \ldots (27)$$

Multiplying equation (24) by 3, we get

$$150 = 30a + 90b + 300c \quad \ldots (28)$$

Subtracting equation (28) from equation (25), we get

$$160 = 30a + 100b + 354c$$

$$150 = 30a + 90b + 300c$$

$$10 = 10b + 54c \quad \ldots (29)$$

Now, we solve equations (23), (24) and (25).

Multiplying equation (23) by 2, we get

$$38 = 10a + 20b + 60c \quad \ldots (26)$$

Subtracting equation (26) from equation (24)

$$50 = 10a + 30b + 100c$$

$$38 = 10a + 20b + 60c$$

$$12 = 10b + 40c \quad \ldots (27)$$

Multiplying equation (24) by 3, we get

$$150 = 30a + 90b + 300c \quad \ldots (28)$$

Subtracting equation (28) from equation (25), we get

$$160 = 30a + 100b + 354c$$

$$150 = 30a + 90b + 300c$$

$$10 = 10b + 54c \quad \ldots (29)$$

Now we solve equation (27) and (29)

Subtracting equation (27) from equation (29), we get

$$10 = 10b + 54c$$

$$12 = 10b + 40c$$

$$-2 = 14c$$

$$c = -2/14$$

$$c = -0.1429$$

Substituting the value of c in equation (29), we get

$$10 = 10b + 54 \times (-0.1429)$$

$$10 = 10b - 7.7166$$

$$17.7166 = 10b$$

$$b = 1.7717$$

Substituting the value of b and c in equation (23), we get

$$19 = 5a + 10 \times (1.7717) + (-0.1429 \times 30)$$

$$19 = 5a + 17.717 - 4.287$$

$$a = 1.114$$

Thus, the second degree of parabola of best fit is

$$Y = 1.114 + 1.7717X - 0.1429X^2$$

Fitting of a straight line by change of origin & scale(U) (step deviation Method)

Business yr	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Profit (inLakhs)	100	120	125	136	150	226	230	235	245	250	276

Y=A+BU

Where U=X- Middle term/length of the interval(h) for odd n

= X -Mean of mid terms/05*h for even n

Business yr	2010	2011	2012	2013	2014	2015	2016	2017	2018	2019	2020
Profit (in Lakhs)	100	120	125	136	150	226	230	235	245	250	276

Business yr	Profit (in Lakhs)	U=X-2015	UY	U^2
2010	100	-5	-500	25
2011	120	-4	-480	16
2012	125	-3	-375	9
2013	136	-2	-272	4
2014	150	-1	-150	1
2015	226	0	0	0
2016	230	1	230	1
2017	235	2	470	4
2018	245	3	735	9
2019	250	4	1000	16
2020	276	5	1380	25
Sum	2093	0	2038	110

A=190.272727

B=18.52727

Y=A+BU

Where U=X-2015

Exponential curve fitting:
(i) $y=ae^{bx}$

Let $Y = ae^{bX}$ … (47)

be an exponential curve and we have a set of n points (x_i, y_i) $i = 1, 2, ..., n$.

Here problem is to determine a and b such that equation (47) is the curve of best fit.

Taking log of both side of equation (47)

$$\log Y = \log a + X\, b \log e$$

Let $\log Y = U$, $\log a = A$ and $b \log e = B$

Now, equation (47) can be written as

$$U = A + BX \quad \text{… (48)}$$

(which is the equation of straight line)

Normal equations for equation (48) can be obtained as

$$\sum u = nA + B\sum x \quad \text{… (49)}$$

$$\sum ux = A\sum x + B\sum x^2 \quad \text{… (50)}$$

We can get A and B from these normal equations. Then

$$a = \text{antilog } A \text{ and } b = \frac{B}{\log e}$$

With these a and b, the exponential curve $Y = ae^{bX}$ is the best fit equation of the curve for the given set of data.

Exponential curve fitting:
(ii) $y=ab^x$

Let $Y = ab^X$ … (40)

be an exponential curve and we have a set of n points (x_i, y_i) $i = 1, 2, ..., n$.

We have to determine a and b such that equation (40) is the curve of best fit.

Taking log both sides of equation (40)

$$\log Y = \log a + \log b^X$$

$$\log Y = \log a + X \log b$$

Let, $\log Y = U$, $\log a = A$ and $\log b = B$

Now, equation (40) comes in the linear form as

$$U = A + BX \quad \text{… (41)}$$

which is the equation of straight line. Normal equations for equation (41) can be obtained as

$$\sum u = nA + B\sum x \quad \text{… (42)}$$

$$\sum ux = A\sum x + B\sum x^2 \quad \text{… (43)}$$

By solving equation (42) and equation (43), we obtain A and B and finally

$a = \text{antilog } A$ and $b = \text{antilog } B$.

With these a and b, the exponential curve $Y = ab^X$ is the curve of best fit for the given set of data.

Power curve fitting:

$y=aX^b$

Let $Y = aX^b$ … (31)

be a power curve where a and b are constants. We have a set of n points (x_i, y_i) $i = 1, 2, ..., n$. Here, the problem is to determine a and b such that the power curve $Y = aX^b$ is the curve of best fit.

Taking log both sides of equation (31), we get

$$\log Y = \log(aX^b)$$

$$\log Y = \log a + \log X^b$$

$$\Rightarrow \log Y = \log a + b \log X \quad \text{… (32)}$$

Let $\log Y = U$, $\log a = A$ and $\log X = V$

Then equation (32) becomes

$$U = A + bV \quad \text{… (33)}$$

Now equation (33) is the linear form of the power curve (31).

Adopting the procedure of fitting of straight line, the normal equations for straight line equation (33) are

$$\sum u = nA + b\sum v \quad \text{… (34)}$$

$$\sum uv = A\sum v + b\sum v^2 \quad \text{… (35)}$$

equations (34) and (35) can be solved for A and b.

After getting A, we get a = antilog (A)

With these a and b, power curve equation (31) is the best fit equation of the curve to the given set of points.

Note: Here we are using log at the base 10.

The Least-Squares m^{th} Degree Polynomials

When using an m^{th} degree polynomial

$$y = a_0 + a_1 x + a_2 x^2 + \ldots\ldots\ldots a_m x^m$$

to approximate the given set of data, (x_1,y_1), $(x_2,y_2)\ldots\ldots$ (x_n,y_n), where $n \geq m$, the best fitting curve has the least square error, i.e.,

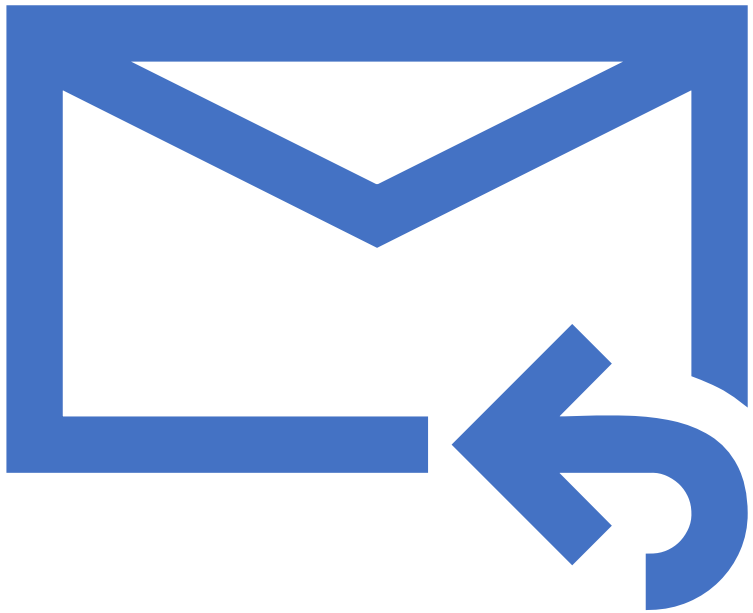
$$Minimize\left\{Error = \sum_{i=1}^{n}\{y_i - f(x_i)\}^2\right\}$$

$$Minimize\left\{Error = \sum_{i=1}^{n}\{y_i - [a_0 + a_1 x_i + a_2 x_i^2 + \ldots\ldots a_m x_i^m]\}^2\right\}$$

The Least-Squares m^{th} Degree Polynomials

$$Minimize\left\{Error = \sum_{i=1}^{n}\{y_i - [a_0 + a_1x_i + a_2x_i^2 + \ldots\ldots a_mx_i^m]\}^2\right\}$$

Thank you



Feedback to mdoodipa@gitam.edu