II yr.

Applied statistics   Paper – III

Unit I :–

Curve Fitting :–

curve fitting is the process of constructing a curve or a mathematical function that has been fit to a series of data, that has the best fit to a series of data points, possibly subject to constraints

curve fitting is one of the most powerful and most widely used analysis tools in origin.

curve fitting examines the relationship between one or more predictors (Independent variable) and a response variable (dependent variable) with the goal of defining a "best fit" model of the relationship.

Let (x_i, y_i) ; $i = 1, 2, \ldots n$ be a given set of n pair of values, X being indept. variable and Y being the dependent variable.

The problem of curve fitting is to find an analytic expression of the form

$$y = f(x)$$

or to obtain the line of best fit.

Uses :–

(i) Fitting of curves to a set of numerical data theoretical as well as practical.

(ii) Theoretically it is used in the study of correlation and regression.

eg: lines of regression to fit linear curve to a bivariate distribution.

(iii) Practically it is useful in represent the relationship between two variables by simple

Page No.
Date:

algebric expressions e.g., Polynomials, exponential or logarithmic functions.
(iv) It is used to estimate the value of one variable which would be supposed to the specified values of the other variable.

Method of least Squares :-

The method of least squares determines the coefficient such that the sum of the deviations between the data and the curve fit is minimized.
If the coefficient in the curve fit appear in a linear fashion, then the problem reduces to solving a system of linear equations.

Fitting of Straight Line :-

The graphical method has the drawback in that the straight line drawn may not be unique.
To overcome this drawback we apply method of principle of least square (given by Legender's).
This method provide a unique set of values to the constants and hence suggests a curve of best fit.
The method of least square consists in minimising the sum of squares of the deviations of the actual values ~~of~~ from their estimated values as given by the line of best fit.

Page No.
Date:

Let (x_i, y_i), $i = 1, 2, \ldots, n$ be the n sets of observations and let the related relation by

$$y = a + bx$$

Now we have to select a and b so that the straight line is the best fit to the data.

Let
$$E = \sum_{i=1}^{n} (Y - f(x))^2$$ be the residual or error of estimate.

or $$E = \sum_{i=1}^{n} [Y - (a+bx)]^2$$

According to the principle of least square we have to determine a and b so that

$$E = \sum_{i=1}^{n} [y_i - (a+bx_i)]^2$$
$$= \sum_{i=1}^{n} [y_i - a - bx_i]^2$$ is minimum.

By the principle of least square E is minimum.

Partially derivating E_i with respect to a and b and equating them to zero we obtain normal equations

$\therefore$ solving $$\frac{\partial E}{\partial a} = 0$$

& $$\frac{\partial E}{\partial b} = 0$$

respectively we get

$$-2\sum[y_i - a - bx_i] = 0 \quad \cancel{\& \; 2\sum[y_i - a - bx_i] \times (x_i) = 0}$$

& $-2x_i\sum_{i=1}^{n}[y_i - a - bx_i]x_i = 0$

or $\sum_{i=1}^{n}(y_i - a - bx_i] = 0$

& $\sum_{i=1}^{n}x_i(y_i - a - bx_i) = 0$

or $\sum_{i=1}^{n}y_i = a\sum_{i=1}^{n} - b\sum_{i=1}^{n}x_i$

& $\sum x_i y_i = a\sum x_i - b\sum_{i=1}^{n}x_i^2$

or $\sum y_i = na - b\sum x_i$ —①

$\sum x_i y_i = a\sum x_i - b\sum x_i^2$ —②

Solving eqn ① & ② we obtain the values of a and b. Then by putting the value of a and b in line of best fit we obtain the best fit

$y = a + bx.$

Fitting of Second degree parabola :-

Let

$$Y = a + bx + cx^2$$

be the second degree parabola of best fit to set of points (x_i, y_i); $i = 1, 2, --, n$ applying the principle of least squares we have to determine the constants a, b and c so that

$$E = \sum_{i=1}^{n}[y_i - a - bx_i - cx_i^2]^2$$

is minimum.

Partially derivating E with respect to a, b and c and equating them to zero we get three normal equation. By solving them we get the value of a, b, & c.

i.e.

$$\frac{\partial E}{\partial a} = 0; \quad \frac{\partial E}{\partial b} = 0 \quad ; \quad \frac{\partial E}{\partial c} = 0$$

respectively we get

$$-2\sum_{i=1}^{n}[y_i - a - bx_i - cx_i^2] = 0$$

$$-2\sum_{i=1}^{n}[y_i - a - bx_i - cx_i^2](x_i) = 0$$

$$-2\sum_{i=1}^{n}[y_i - a - bx_i - cx_i^2](x_i^2) = 0$$

or

$$\sum_{i=1}^{n}(y_i - a - bx_i - cx_i^2) = 0$$

$$\sum_{i=1}^{n}x_i(y_i - a - bx_i - cx_i^2) = 0$$

$$\sum_{i=1}^{n}x_i^2(y_i - a - bx_i - cx_i^2) = 0$$

or

$$\sum y_i = na + b\sum x_i + c\sum x_i^2 \quad (1)$$

$$\sum x_i y_i = a\sum x_i + b\sum x_i^2 + c\sum x_i^3 \quad (2)$$

$$\sum x_i^2 y_i = a\sum x_i^2 + b\sum x_i^3 + c\sum x_i^4 \quad (3)$$

Solving equation (1) (2) and (3) we get the value of a, b, c. Putting the values of a, b, & c in (*) we get the best fit of second degree parabola.

Fitting of Polynomial of k^{th} Degree :—

If
$$y = a_0 + a_1 x + a_2 x^2 + \cdots + a_k x^k \quad \text{—}(*)$$
is the k^{th} degree polynomial of best fit to the set of points (x_i, y_i) ; $i = 1, 2, \ldots, n$, the constants $a_0, a_1, a_2, \ldots, a_k$ are to be obtained so that

$$E = \sum_{i=1}^{n} \left[y_i - a_0 - a_1 x_i - a_2 x_i^2 - \cdots - a_k x^k \right]^2 \quad (**)$$

is minimum.

Thus the normal equations for estimating $a_0, a_1, \ldots, a_k$ are obtained on equating to zero the partial derivatives of E w.r.t. $a_0, a_1, \ldots, a_k$ separately i.e.

$$\frac{\partial E}{\partial a_0} = 0 = -2 \sum \left[y_i - a_0 - a_1 x_i - a_2 x_i^2 - \cdots - a_k x_i^k \right]$$

$$\frac{\partial E}{\partial a_1} = 0 = -2 \sum x_i \left[y_i - a_0 - a_1 x_i - a_2 x_i^2 - \cdots - a_k x^k \right]$$

(1) $\cancel{(***)}$

$$\frac{\partial E}{\partial a_2} = 0 = -2 \sum x_i^2 \left[y_i - a_0 - a_1 x_i - a_2 x_i^2 - \cdots - a_k x^k \right]$$

$$\vdots$$

$$\frac{\partial E}{\partial a_k} = 0 = -2 \sum x_i^k \left[y_i - a_0 - a_1 x_i - a_2 x_i^2 - \cdots - a_k x^k \right]$$

$\Rightarrow$

(2)
$$\left.\begin{aligned}
\Sigma y_i &= n a_0 + a_1 \Sigma x_i + a_2 \Sigma x_i^2 + \cdots + a_k \Sigma x_i^k \\
\Sigma x_i y_i &= a_0 \Sigma x_i + a_1 \Sigma x_i^2 + a_2 \Sigma x_i^3 + \cdots + a_k \Sigma x_i^{k+1} \\
\Sigma x_i^2 y_i &= a_0 \Sigma x_i^2 + a_1 \Sigma x_i^3 + a_2 \Sigma x_i^4 + \cdots + a_k \Sigma x_i^{k+2} \\
&\vdots \\
\Sigma x_i^k y_i &= a_0 \Sigma x_i^k + a_1 \Sigma x_i^{k+1} + a_2 \Sigma x_i^{k+2} + \cdots + a_k \Sigma x_i^{2k}
\end{aligned}\right.$$

summation extended over i from 1 to n. These are $(k+1)$ equations in $(k+1)$ unknown $a_0, a_1, a_2, \ldots, a_k$ and can be solved with the help of algebra.

Remark :- By solving ~~all~~ above eqn ~~(xx)~~ at second order derivatives

viz $\frac{\partial^2 E}{\partial a_0^2}, \frac{\partial^2 E}{\partial^2 a_1^2}, \ldots$

we find that the values are positive at the point $a_0, a_1, \ldots, a_k$, the solutions of normal equation. Hence they provide minima of E.

Fitting of Exponential and logarithmic Curves :-

The curve ~~havin~~ having exponential form is known as exponential curve. Following are exponential form

(i) $y = ab^x$

(ii) $y = ae^{bx}$

PREMIUM

(i) $Y = ab^X$

Taking logarithm of both sides, we get

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

Let $U = A + BX$

where $U = \log Y$

$A = \log a$

$B = \log b$

This is linear equation in X and U.

The normal equations for estimating A & B are

$$\Sigma U = nA + B\Sigma X$$

& $$\Sigma XU = A\Sigma X + B\Sigma X^2$$

Solving these equations for A and B, we finally get

$$a = \text{antilog}(A)$$

$$b = \text{antilog}(B)$$

By putting these value we get the curve of best fit to the given set of n points.

(ii) $Y = ae^{bX}$

Taking logarithm of both sides, we get

$$\log Y = \log a + bX \log e$$
$$= \log a + (b \log e) X$$

$\Rightarrow \quad U = A + BX$

where $U = \log y$

$A = \log a$

$B = b \log e$

This is linear equation in X and U

Thus the normal equation are

$$\Sigma U = nA + B\Sigma X$$

$$\Sigma XU = A\Sigma X + B\Sigma X^2$$

From these we find A & B and consequently

$$a = \text{antilog}(A)$$

$$b = \frac{B}{\log e}$$

By putting these value we get the curve of best fit to the given points.

Most Plausible Solution of a System of Linear Equations :

Method of least squares is used to finding the most plausible values of the variables satisfying a system of independent linear equations whose number is more than the number of variables under study.

Consider the following set of m equations in n variables $X, Y, Z, \ldots, T$:

Page No.
Date:

$$\left.\begin{aligned} a_1X + b_1Y + c_1Z + \cdots + k_1T &= l_1 \\ a_2X + b_2Y + c_2Z + \cdots + k_2T &= l_2 \\ &\vdots \\ a_mX + b_mY + c_mZ + \cdots + k_mT &= l_m \end{aligned}\right\} \quad (1)$$

where $a_i, b_i, \cdots k_i, l_i$ are constants ; $i = 1, 2, \cdots, m$.

If $m = n$, the system of equation (1) can be solved uniquely with the help of algebra.

If $m > n$ it is not possible to determine a unique solution $X, Y, Z, \cdots, T$ which will satisfy the system (1), In this case we find the values of $X, Y, Z, \cdots, T$ which will satisfy the system (1) as nearly as possible.

Legender's principle of least square consists in minimising the sum of the squares of the 'residuals' or the "errors".

If

$$E_i = a_iX + b_iY + c_iZ + \cdots + k_iT - l_i$$
$$(i = 1, 2, \cdots, m)$$

is the residual for the i^{th} equation, then we have to determine $X, Y, Z, \cdots, T$ so that

$$U = \sum_{i=1}^{m} E_i^2 = \sum_{i=1}^{m} (a_iX_i + b_iY + c_iZ + \cdots + k_iT - l_i)$$

is minimum.

Using the principle of maxima & minima in differential calculus, the partial derivatives of 'U' w.r.t. X, Y, Z, --, T should vanish seperately.

Thus.

$$\left.\begin{aligned}
\frac{\partial U}{\partial X} &= 0 = \sum_{i=1}^{m} a_i (a_i X + b_i Y + \cdots + k_i Z - l_i) \\
\frac{\partial U}{\partial Y} &= 0 = \sum_{i=1}^{m} b_i (a_i X + b_i Y + \cdots + k_i Z - l_i) \\
\frac{\partial U}{\partial T} &= 0 = \sum_{i=1}^{m} k_i (a_i X + b_i Y + \cdots + k_i Z - l_i)
\end{aligned}\right\} \quad (2)$$

These are known as the normal equations for X, Y, Z, --, T respectively.

Thus we have n - normal equations in n unknown X, Y, Z, --, T and their solution gives the best or most plausible solution of the system (1).