CPNA Lecture 21 - Curve Fitting

Mridul Sankar Barik

Jadavpur University

2023

Introduction I

- Problem - to fit a unique curve to data points which are subject to error
- Techniques of Interpolation assume error free data
- Method of least squares is most of common technique

Method of Least Squared Error I

- Let the set of data points be (x_i, y_i), $i = 1, 2, \ldots, n$
- Let the curve $Y = f(x)$ be fitted to this data
- If e_i is the error of approximation at $x = x_i$, then we have

$$e_i = y_i - f(x_i)$$

- Sum of the squares of the errors

$$S = e_1^2 + e_2^2 + \ldots + e_m^2$$

or,

$$S = [y_1 - f(x_1)]^2 + [y_2 - f(x_2)]^2 + \ldots + [y_m - f(x_m)]^2$$

- Method of least squares tries to minimize S

Linear Regression I

- Assume $y = a_1x + a_0$ is the equation of the line
- We need to choose values of a_1 and a_2 that gives the best straight line
- Sum of squared errors

$$\sum_{i=1}^{n}\{y_i - (a_1x_i + a_0)\}^2$$

- In order to minimize S, we take partial derivatives of S with respect to a_0 and a_1 and set them to zero

$$\frac{\partial S}{\partial a_0} = \sum_{i=1}^{n} 2(y_i - a_1x_i - a_0)(-1) = 0$$

$$\frac{\partial S}{\partial a_1} = \sum_{i=1}^{n} 2(y_i - a_1x_i - a_0)(-x_i) = 0$$

Linear Regression II

- Rearranging, we get two linear simultaneous equations for a_0 and a_1 ($\sum$ is used as abbreviation for $\sum_{i=1}^{n}$):

$$na_0 + \left(\sum x_i\right) a_1 = \sum y_i$$

$$\left(\sum x_i\right) a_0 + \left(\sum x_i^2\right) a_1 = \sum x_i y_i$$

- The solutions are:

$$a_0 = \frac{\sum y_i \sum x_i^2 - \sum x_i \sum x_i y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

$$a_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

- a_0 and a_1 are called regression coefficients

Linear Regression III

- **Exercise**: Fit a straight line to the following x and y values

x	y
1	2
2	5
4	7
5	10
6	12
8	15
9	19

- **Solution**: The following quantities are computed

$$n = 7 \qquad \sum x_i y_i = 453 \qquad \sum x_i^2 = 227$$

$$\sum x_i = 35$$

$$\sum y_i = 70$$

Linear Regression IV

$$a_0 = \frac{70 \times 227 - 35 \times 453}{7 \times 227 - 35^2} = \frac{35}{364} = 0.096$$

$$a_1 = \frac{7 \times 453 - 35 \times 70}{7 \times 227 - 35^2} = \frac{721}{364} = 1.98$$

Thus, $y = 1.98x + 0.096$ is the equation of the straight line that is a least squares linear approximation to the given data points

Polynomial Regression I

- In general it may be necessary to fit a higher degree polynomial
- To fit a second degree polynomial, let the equation of the curve be $y = a_2 x^2 + a_1 x + a_0$
- The sum of squares of the errors

$$S = \sum (y_i - a_2 x_i^2 - a_1 x_i - a_0)^2$$

- Differentiating S with respect to a_0, a_1, a_2 respectively and setting each to zero we get

$$na_0 + a_1 \sum x_i + a_2 \sum x_i^2 = \sum y_i$$

$$a_0 \sum x_i + a_1 \sum x_i^2 + a_2 \sum x_i^3 = \sum x_i y_i$$

$$a_0 \sum x_i^2 + a_1 \sum x_i^3 + a_2 \sum x_i^4 = \sum x_i^2 y_i$$

Polynomial Regression II

- These may be solved by Gauss Elimination
- In general, to fit n^{th} degree polynomial there will be $(n+1)$ simultaneous equations in $(n+1)$ unknowns
- **Exercise**: Fit a second degree polynomial to the following x and y values

x	y
0	2.1
1	7.7
2	13.6
3	27.2
4	40.9
5	61.1

Polynomial Regression III

- **Solution**: From the given data

$$n = 6 \qquad \sum x_i = 15 \qquad \sum x_i^4 = 979$$

$$\sum y_i = 152.6 \qquad \sum x_i y_i = 585.6$$

$$\bar{x} = 2.5 \qquad \sum x_i^2 = 55 \qquad \sum x_i^2 y_i = 2488.8$$

$$\bar{y} = 25.433 \qquad \sum x_i^3 = 225$$

We obtain three simultaneous equations as

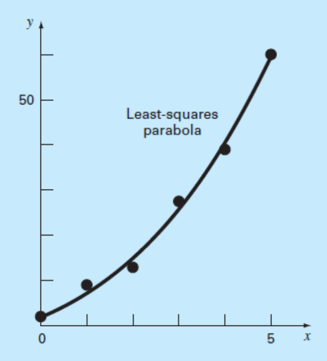
$$\begin{aligned} 6a_0 + 15a_1 + 55a_2 &= 152.6 \\ 15a_0 + 55a_1 + 225a_2 &= 585.6 \\ 55a_0 + 225a_1 + 979a_2 &= 2488.8 \end{aligned}$$

Polynomial Regression IV

Solving these equations using Gauss elimination gives
$a_0 = 2.47857$, $a_1 = 2.35929$, and $a_2 = 1.86071$.

Therefore, the least-squares quadratic equation for this case is

$$y = 2.47857 + 2.35929x + 1.86071x^2$$



Fitting Other Non-Linear Functins I

- Many practical problems generate data from experiments that have forms other than linear or polynomial curve
- Exponential curve
- Power curve
- Saturation growth rate curve
- Trigonometric curve

Fitting an Exponential Curve I

- Let $y = \alpha_1 e^{\beta_1 x}$ be the curve to be fitted, where α_1 and β_1 are constants
- This model is used in many fields, i.e., population growth or radioactive decay etc.
- We take the transformation $z = \log y$
- So, $z = \log y = \log \alpha_1 + \beta_1 x$
- Let, $a_0 = \log \alpha_1$ and $a_1 = \beta_1$
- So, we have $z = a_0 + a_1 x$, which is a linear equation and we can use equations for linear regression to have

$$a_0 = \frac{\sum \log y_i \sum x_i^2 - \sum x_i \sum x_i \log y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

$$a_1 = \frac{n \sum x_i \log y_i - \sum x_i \sum \log y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

- From a_0 and a_1 we obtain the value of α_1 and β_1 as $\alpha_1 = e^{a_0}$ and $\beta_1 = a_1$

Fitting a Power Curve I

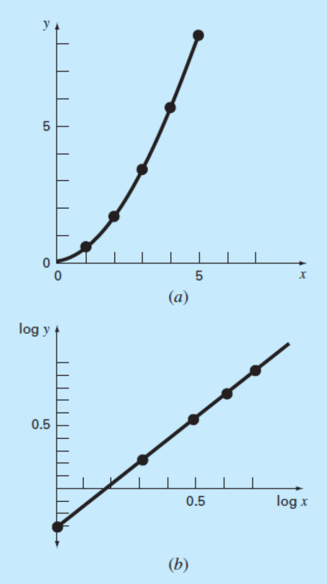
- Let the curve to be fitted be $y = \alpha_2 x^{\beta_2}$
- Taking logarithm on both sides, we get $z = \log(y) = \log \alpha_2 x^{\beta_2} = \log \alpha_2 + \beta_2 \log x$ or $z = a_0 + a_1 t$, where $a_0 = \log \alpha_2$, $a_1 = \beta_2$ and $t = \log x$
- The normal equations for the above are

$$n \log \alpha_2 + \left(\sum \log x_i\right) \beta_2 = \sum \log y_i$$

$$\left(\sum \log x_i\right) \log \alpha_2 + \left(\sum (\log x_i)^2\right) \beta_2 = \sum \log x_i \log y_i$$

- Solving the two equations we get solutions for β_2 and $\log \alpha_2$

Fitting a Power Curve II



(a)

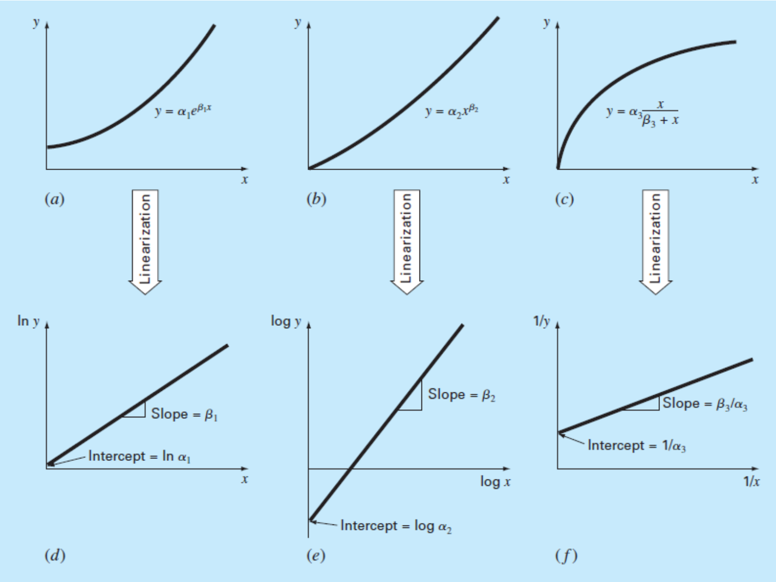
(b)

Fitting a Saturation Growth Rate Curve I

- The assumed equation is $y = \alpha_3 \dfrac{x}{\beta_3 + x}$
- Taking $z = \dfrac{1}{y}$ and $t = \dfrac{1}{x}$,

  we get $z = a + bt$, where $a = \dfrac{1}{\alpha_3}$ and $b = \dfrac{\beta_3}{\alpha_3}$
- This is a linear equation and linear regression approach can be used

Linearization of Nonlinear Relationship



Fitting a Trigonometric Function I

- Assume the equation of the curve be $y = A\ sin\ (\omega x + \varphi)$, where ω is known
- $y = A\ cos\ \varphi\ sin\ \omega x + A\ sin\ \varphi\ cos\ \omega x = a_1\ sin\ \omega x + a_2\ cos\ \omega x$
- We try to minimize the sum of squares of the errors, i.e., minimize $S = \sum(y_i - a_1\ sin\ \omega x_i - a_2\ cos\ \omega x_i)$
- Taking partial derivatives of S with respect to a_1 and a_2, and setting them equal to zero, we get

$$a_1 \sum sin^2\ \omega x_i + a_2 \sum sin\ \omega x_i\ cos\ \omega x_i = \sum y_i\ sin\ \omega x_i$$

$$a_1 \sum sin\ \omega x_i\ cos\ \omega x_i + a_2 \sum cos^2\ \omega x_i = \sum y_i\ cos\ \omega x_i$$

- We solve this two simultaneous linear equations for a_1 and a_2, from which we obtain $A = \sqrt{a_1^2 + a_2^2}$, and $\varphi = tan^{-1}\left(\frac{a_2}{a_1}\right)$