Curve Fitting

Lecture Strecture:

Topic	Description
1. Introduction to Curve Fitting	Explanation of what curve fitting is, its importance in engineering, and the different types of curve fitting.
2. Least Squares Method	Detailed explanation of the least squares method, how it works, and its application in fitting a line to data points.
3. Example: Linear Curve Fitting	A worked example of using the least squares method to fit a line to a set of data points.
4. Linear Correlation Coefficient	Explanation of the linear correlation coefficient, how to calculate it, and its interpretation with an example and chart.
5. Example of Correlation Coefficient	Example of calculating the linear correlation coefficient using sample data and interpretation of the results.
6. Assignment	Homework assignment based on calculating the linear correlation coefficient and fitting a curve to data using the least squares method.

1. Introduction to Curve Fitting

What is Curve Fitting?

Curve fitting is a process of finding a curve (or mathematical function) that best represents a set of data points. This is especially useful when the relationship between variables is not perfectly linear or when there are uncertainties or errors in the data. By using curve fitting, we can create a model that helps predict or describe the behavior of the data.

In engineering, data is often collected from experiments or field measurements, and curve fitting is used to approximate the relationship between the variables. For example, if you measure the deflection of a beam under varying loads, curve fitting can be used to model the relationship between load and deflection.

Why Do We Use Curve Fitting?

In real-world engineering problems, the data we collect from experiments or observations often contains **noise** or **measurement errors**. Curve fitting allows us to:

- **Summarize Data**: Simplify complex data by finding a mathematical model that represents the general trend.
- **Make Predictions**: Use the model to predict outcomes for values not present in the data.
- **Analyze Relationships**: Understand the relationship between variables, such as stress and strain in materials testing.

2- Least Squares Method

The **least squares method** is one of the most commonly used techniques for curve fitting. It minimizes the **sum of the squared differences** between the observed data points and the values predicted by the fitted curve.

Given a set of data points $(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)$the goal is to find a curve that minimizes the sum of the squared differences between the observed values y_i and the values predicted by the curve $f(x_i)$.

These set of points have the equation :

$$y = a + bx$$

where the constant **a** and **b** are determined by solving the following equation:

$$\sum y = an + b\sum x$$

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

$$\rightarrow a = \frac{(\sum y)\,(\sum x^2) - (\sum x)(\sum xy)}{n\sum x^2 - (\sum x)^2} \qquad b = \frac{n\sum xy - (\sum x)(\sum y)}{n\sum x^2 - (\sum x)^2}$$

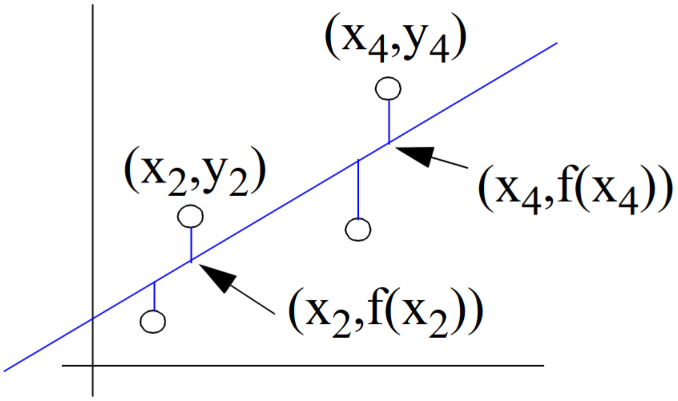
$$\therefore\ b = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sum (x-\bar{x})^2}$$

$$\because\ \bar{x} = \frac{\sum x}{n} \qquad \bar{y} = \frac{\sum y}{n}$$

$$\therefore \bar{y} = a + b\bar{x} \qquad \rightarrow \qquad a = \bar{y} - b\bar{x}$$

$$y - \bar{y} = b(x - \bar{x})$$

The result shows that the constant **b**, which is the slope of the line (**y = a + bx**), is the fundamental constant in determining the line. It is also seen that the least-squares line passes through the point $(\bar{x}, \bar{y})$ which is called the centroid or center of gravity of the data.



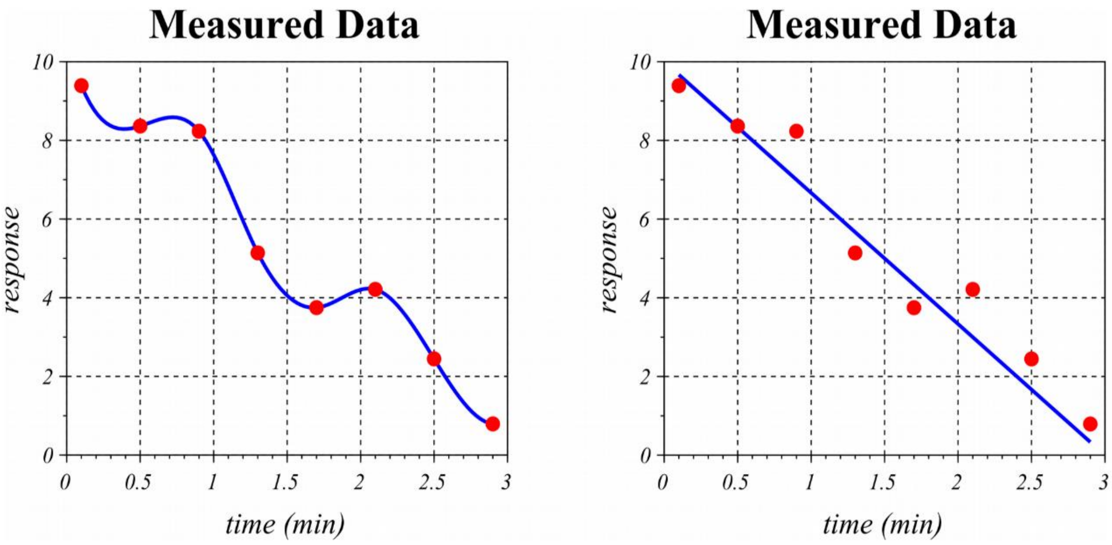


Fig. 1: Measured data with: (left) spline interpolation, (right) line fit.

3- Example

EX. 1: The following two random variables x and y are linearly correlated, Find out the linear regression line y=f(x).

X	Y
50	37
100	48
150	60
200	71
250	80
300	90
350	102
400	109

Solution:

x	y	x^2	xy	y^2
50	37	2500	1850	1369
100	48	10000	4800	2304
150	60	22500	9000	3600
200	71	40000	14200	5041
250	80	62500	20000	6400
300	90	90000	27000	8100
350	102	1222500	35700	10404
400	109	160000	43600	11881
Σx =1800	**Σy = 597**	**Σx^2 = 510000**	**Σxy = 156150**	**Σy^2 = 49099**

$$\mathbf{a} = \frac{(\sum \mathbf{y})\,(\sum \mathbf{x}^2) - (\sum \mathbf{x})(\sum \mathbf{xy})}{\mathbf{n}\sum \mathbf{x}^2 - (\sum \mathbf{x})^2} = \frac{(597)(510000)-(1800)(156150)}{(8)(510000)-(1800)^2} = \frac{195}{7} = 27.857$$

$$\mathbf{b} = \frac{\mathbf{n}\sum \mathbf{xy} - (\sum \mathbf{x})(\sum \mathbf{y})}{\mathbf{n}\sum \mathbf{x}^2 - (\sum \mathbf{x})^2} = \frac{(8)(156150)-(1800)(597)}{(8)(524)-(56)^2} = \frac{291}{1400} = 0.208$$

$$\therefore\ y = 27.857 + 0.28x$$

4- Linear Correlation Coefficient

The **linear correlation coefficient**, often denoted by **rrr**, is a statistical measure that quantifies the strength and direction of a **linear relationship** between two variables. It is also known as **Pearson's correlation coefficient**. The correlation coefficient provides a numerical value between **-1** and **1** that indicates how closely the data points fit a straight line.

- **r = 1**: Perfect positive correlation — as one variable increases, the other variable increases proportionally.
- **r = −1**: Perfect negative correlation — as one variable increases, the other decreases proportionally.
- **r = 0**: No correlation — there is no linear relationship between the variables.

In the context of **curve fitting** and **regression analysis**, the linear correlation coefficient is crucial for determining how well a linear model fits the data. A higher correlation coefficient indicates that the linear model is a good fit for the data, whereas a low correlation coefficient suggests that a linear model may not be appropriate.

The linear correlation coefficient rrr is calculated using the following formula:

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left(n\sum x^2 - (\sum x)^2\right)\left(n\sum y^2 - (\sum y)^2\right)}}$$

Interpreting the Correlation Coefficient

- **0.7 ≤ r ≤ 1** : **Strong positive correlation** — a linear relationship where y tends to increase as x increases.
- **0.3 ≤ r < 0.7** : **Moderate positive correlation** — a positive trend exists, but the data points are more spread out.
- **0 ≤ r < 0.3** : **Weak positive correlation** — a weak relationship where the data points do not form a clear linear pattern.

- **$-0.3 \leq r < 0$** : **Weak negative correlation** — a weak negative relationship where y tends to decrease slightly as x increases.
- **$-0.7 \leq r < -0.3$** : **Moderate negative correlation** — a clear, but not perfect, negative relationship.
- **$-1 \leq r < -0.7$** : **Strong negative correlation** — a strong linear relationship where y decreases as x increases.

5- Example

EX. 2: Specify the correlation type and the degree for the data in example 1.

Solution:

x	y	x^2	xy	y^2
50	37	2500	1850	1369
100	48	10000	4800	2304
150	60	22500	9000	3600
200	71	40000	14200	5041
250	80	62500	20000	6400
300	90	90000	27000	8100
350	102	1222500	35700	10404
400	109	160000	43600	11881
Σx =1800	**Σy = 597**	**Σx^2 = 510000**	**Σxy = 156150**	**Σy^2 = 49099**

$$r = \frac{n\sum xy - \sum x \sum y}{\sqrt{\left(n\sum x^2 - \left(\sum x\right)^2\right)\left(n\sum y^2 - \left(\sum y\right)^2\right)}} = \frac{8(156150) - (1800)(597)}{\sqrt{[8(510000) - (1800)^2][8(49099) - (597)^2]}} = 0.999$$

∵ $r \approx 1 \longrightarrow$ Perfect positive linear correlation.

6- Assignment:

For the raw data shown in Table, three regression lines were suggested:

$y_1 = a_1 – 11.5x$

$y_2 = a_2 – 12x$

$y_3 = a_3 – 12.5x$

1) Find out which one of these 3 equations gives better regression.

2) Calculate the y intersection (a) for the best equation.

3) Specify the correlation type and the degree.

x	y
4.4	588
6.7	564
10.5	516
9.6	534
12.4	480
5.5	558