Linear Regression – Linear Least Squares

ME 120 Notes

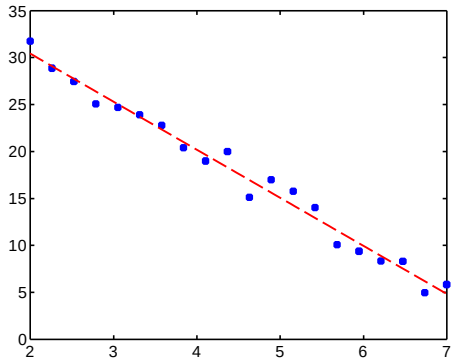
Gerald Recktenwald
Portland State University
Department of Mechanical Engineering
gerry@me.pdx.edu

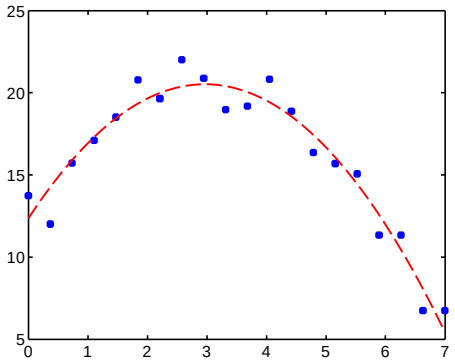
Introduction

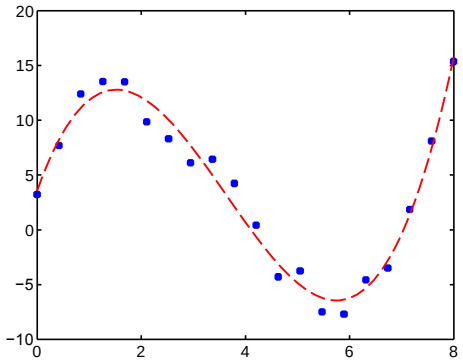
- Engineers and Scientists work with lots of data.
  - Scientists try to undersand the way things behave in the physical world.
  - Engineers try to use the discoveries of scientists to produce useful products or services.

- Given data from a measurement, how can we obtain a simple mathematical model that fits the data? By "fit the data", we mean that the function follows the trend of the data.

Polynomial Curve Fits







Basic Idea

- Given data set (x_i, y_i), $i = 1, \ldots, n$
- Find a function $y = f(x)$ that is *close* to the data

The process avoids guesswork.

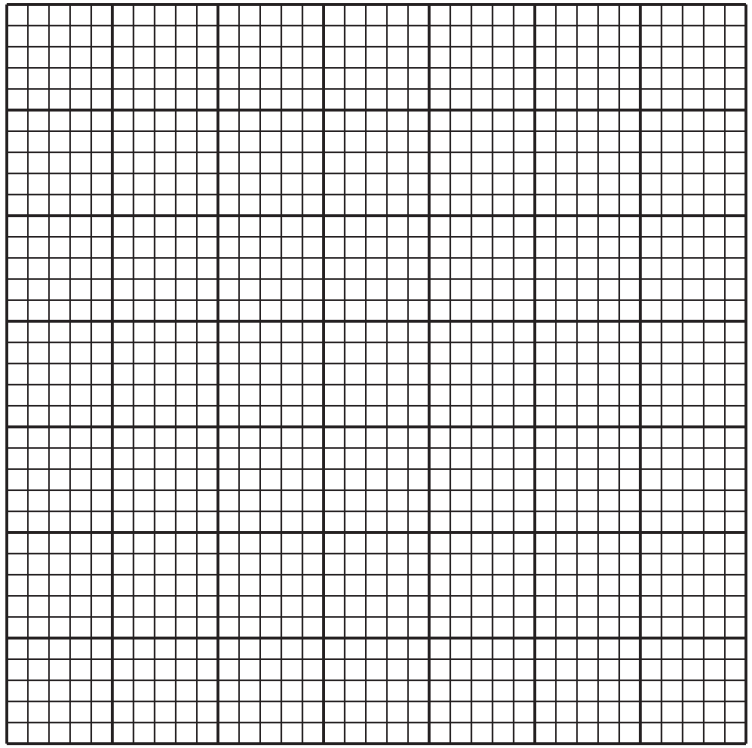
Some sample data

x (time)	*y* (velocity)
1	9
2	21
3	28
4	41
5	47

It is aways important to visualize your data. You should be able to plot this data by hand.

x (time)	y (velocity)
1	9
2	21
3	28
4	41
5	47

Plot the Data



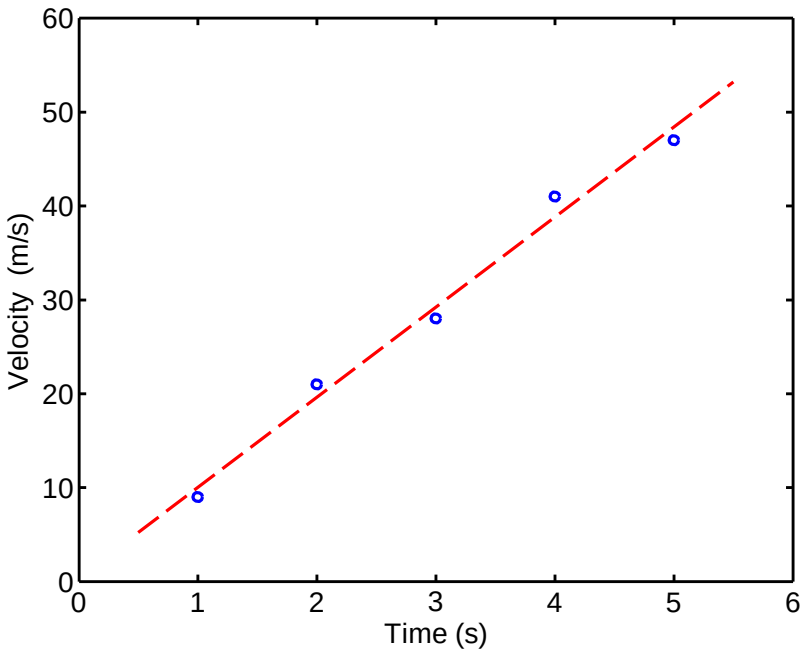
- Plot x on the horizontal axis, y on the vertical axis.
  - What is the range of the data?
  - Use the range to select an appropriate scale so that the data uses all (or most) of the available paper.
- In this case, x is the *independent* variable.
- y is the *dependent* variable.

Suppose that x is a measured value of time, and y is a measured velocity of a ball.

- Label the axes

x (time)	*y* (velocity)
1	9
2	21
3	28
4	41
5	47

Analyze the plot



An equation that represents the data is valuable

- A simple formula is more compact and reusable than a set of points.
- This data looks linear so our "fit function" will be

$$y = mx + b$$

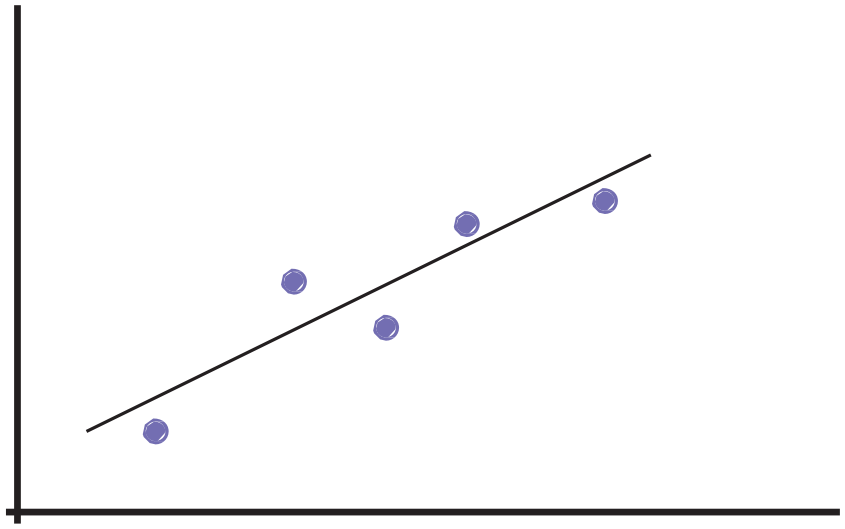
- The value of slope or intercept may have physical significance.

Least Squares Method

- Compute slope and intercept in a way that minimizes an error (to be defined).
- Use calculus or linear algebra to derive equations for m and b.
- There is only one slope and intercept for a given set of data.

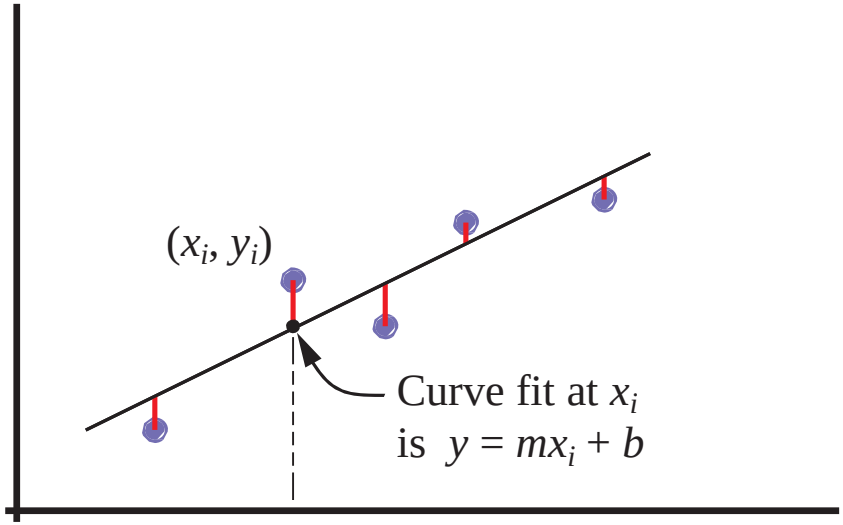
Do not guess m and b. Use least squares.

Least Squares: The Basic Idea



The best fit line goes near the data, but not through them.

Least Squares: The Basic Idea



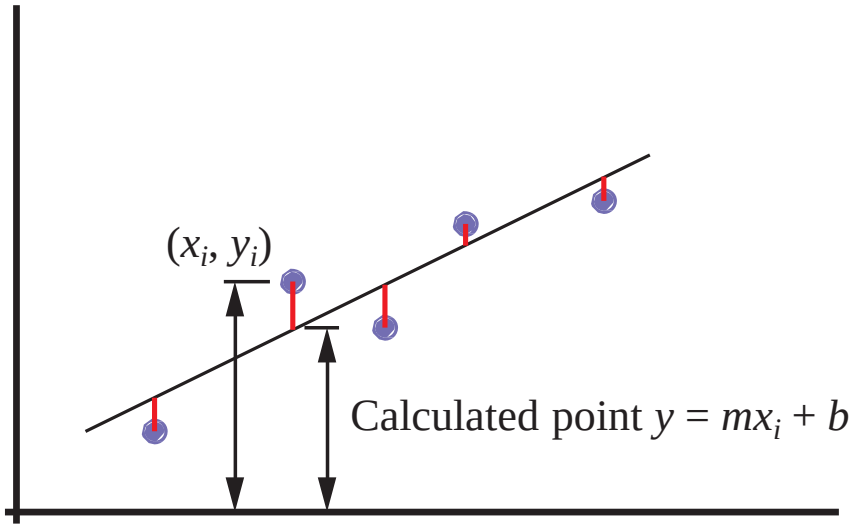
The best fit line goes near the data, but not through them.

The equation of the line is

$$y = mx + b$$

The data (x_i, y_i) are known.
m and b are unknown.

Least Squares: The Basic Idea



The best fit line goes near the data, but not through them.

The discrepancy between the known data and the unknown fit function is taken as the *vertical distance*

$$y_i - (mx_i + b)$$

But the error can be positive or negative, so we use the *square of the error*

$$[y_i - (mx_i + b)]^2$$

Least Squares Computational Formula

Use calculus to *minimize the sum of squares* of the errors

$$\text{Total error in the fit} = \sum_{i=1}^{n} [y_i - (mx_i + b)]^2$$

Minimizing the total error with respect to the two parameters *m* and *b* gives

$$m = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum {x_i}^2 - (\sum x_i)^2} \qquad \text{b} = \frac{\sum y_i - m\sum x_i}{n}$$

Notice that *b* depends on *m*, so solve for *m* first.

Subscript and Summation Notation

Suppose we have a set of numbers

$$x_1, x_2, x_3, x_4$$

We can use a variable for the subscript

$$x_i, \qquad i = 1, \ldots, 4$$

To add the numbers in a set we can write

$$s = x_1 + x_2 + x_3 + x_4$$

as

$$s = \sum_{i=1}^{n} x_i$$

Subscript and Summation Notation

We can put any formula with subscripts inside the summation notation.

Therefore $$\sum_{i=1}^{4} x_i y_i = x_1 y_1 + x_2 y_2 + x_3 y_3 + x_4 y_4$$

$$\sum_{i=1}^{4} {x_i}^2 = {x_1}^2 + {x_2}^2 + {x_3}^2 + {x_4}^2$$

The result of a sum is a number, which we can then use in other computations.

Subscript and Summation Notation

The order of operations matters. Thus,

$$\sum_{i=1}^{4} {x_i}^2 \neq \left(\sum_{i=1}^{4} x_i \right)^2$$

$${x_1}^2 + {x_2}^2 + {x_3}^2 + {x_4}^2 \neq (x_1 + x_2 + x_3 + x_4)^2$$

Subscript and Summation: Lazy Notation

Sometimes we do not bother to write the range of the set of data.

Thus,

$$s = \sum x_i$$

implies

$$s = \sum_{i=1}^{n} x_i$$

where *n* is understood to mean, "however many data are in the set"

Subscript and Summation: Lazy Notation

So, in general

$$x_i$$

implies that there is a set of x values

$$x_1, x_2, \ldots, x_n$$

or

$$x_i, \qquad i = 1, \ldots, n$$

Subscript and Summation Notation

Practice!

Given some data (x_i, y_i), $\quad i = 1, \ldots, n$

1. *Always plot your data first!*
2. Compute the slope and intercept of the least squares line fit.

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

$$\mathrm{b} = \frac{\sum y_i - m\sum x_i}{n}$$

Sample Data:

x (time)	y (velocity)
1	9
2	21
3	28
4	41
5	47