Lecture 6

Data Manipulation; Curve Fitting; Statistics

Import and Export of data files

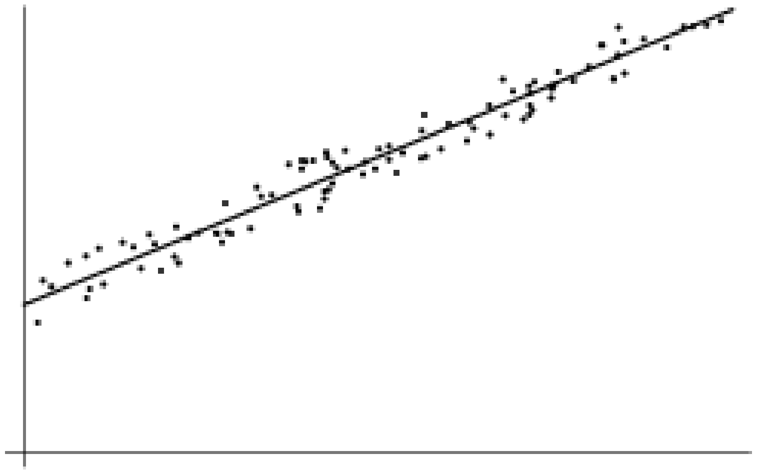
Syntax

- **Directory[ ]**
- **SetDirectory[NotebookDirectory[ ]]**
- **SetDirectory["path − of − my − directory"]**
- **Import[“datasemicircle. dat”];**

Least Squares Fitting

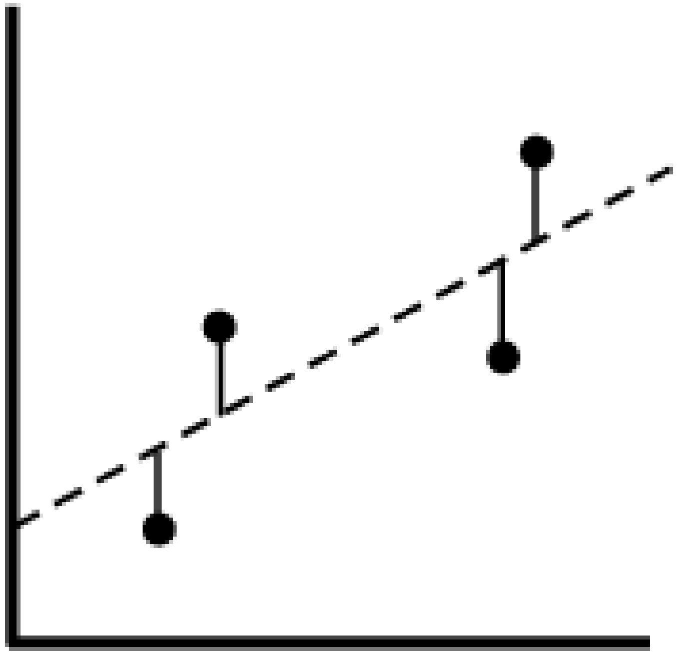
http://mathworld.wolfram.com/LeastSquaresFitting.htm

You have measured a set of data points, $\{x_i, y_i\}, i = 1,2, \ldots, N$; and you know that they should approximately lie on a straight line of the form *y* = *a* x + *b* if the y_i's are plotted against x_i's.



- We wish to know what are the best values for *a* and *b* that make the best fit for the data set. The process is called 'data fitting'. The function to be fit against is in a linear form, $y = a + bx$.

Vertical offset

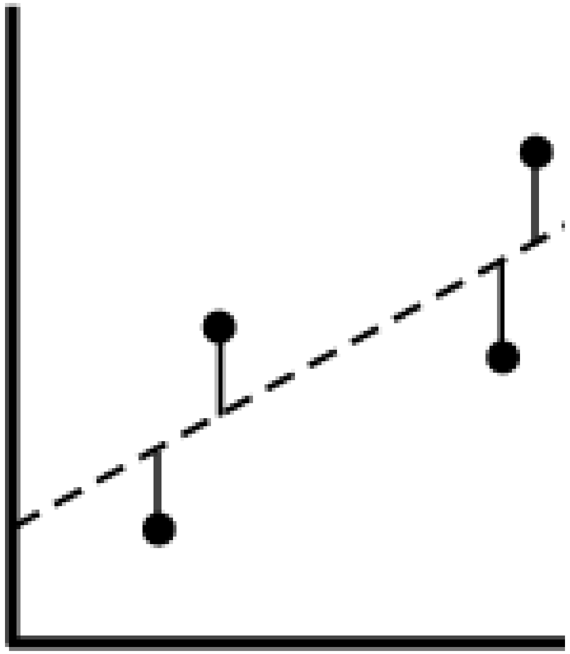


Let $f(x_i, a, b) = b\, x_i + a$, in which we are looking for the best values for *b* and *a*.

Vertical least squares fitting proceeds by minimizing the sum of the *squares* of the *vertical* deviations R^2 of a set of *n* data points

$$R^2\,(a, b) \equiv \sum_{i=1}^{n} [y_i - (a + b\, x_i)]^2$$

Minimisation of R^2



The values of a and b for which R^2 is minimized are the best fit values.

You can find these best fit values by minimize R^2

Least Squared Minimisation

$$R^2(a, b) \equiv \sum_{i=1}^{n} [y_i - (a + b x_i)]^2$$

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

$$\frac{\partial(R^2)}{\partial b} = -2 \sum_{i=1}^{n} [y_i - (a + b x_i)] x_i = 0.$$

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

$$a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i.$$

In matrix form

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

$$\begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i\, y_i \end{bmatrix}. \qquad \text{Eq. (1)}$$

$$\begin{bmatrix} a \\ b \end{bmatrix} = \frac{1}{n \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \begin{bmatrix} \sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i\, y_i \\ n \sum_{i=1}^{n} x_i\, y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i \end{bmatrix},$$

$$a = \frac{\sum_{i=1}^{n} y_i \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i\, y_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

$$= \frac{\bar{y}\left(\sum_{i=1}^{n} x_i^2\right) - \bar{x} \sum_{i=1}^{n} x_i\, y_i}{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2} \qquad \text{Eq. (2)}$$

$$b = \frac{n \sum_{i=1}^{n} x_i\, y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \qquad \text{Eq. (3)}$$

$$= \frac{(\sum_{i=1}^{n} x_i\, y_i) - n\,\bar{x}\,\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2}$$

$$\bar{x} = \frac{1}{N}\sum_{i}^{N} x_i\,, \bar{y} = \frac{1}{N}\sum_{i}^{N} y_i$$

Standard errors in a and b are given by $\mathrm{SE}(a)$ and $\mathrm{SE}(b)$

$$SS_{xx} = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

$$SS_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

$$SS_{yy} = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

$$s = \sqrt{\frac{SS_{yy} - b\,SS_{xy}}{n-2}} = \sqrt{\frac{SS_{yy} - \frac{SS_{xy}^2}{SS_{xx}}}{n-2}}$$

$$\mathrm{SE}(a) = s\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{SS_{xx}}} \qquad \mathrm{SE}(b) = \frac{s}{\sqrt{SS_{xx}}}.$$

Mathematica's built-in functions for data fitting

Syntax:

NonlinearModelFit[ ], **Normal**[ ], **model**["**ParameterTable**"]

These are Mathematica's built in functions to fit a set of data against a linear formula, such as $y = a + b x$, and at the same time automatically provide errors of the best fit parameters – very handy way to fit a set of data against any linear formula.

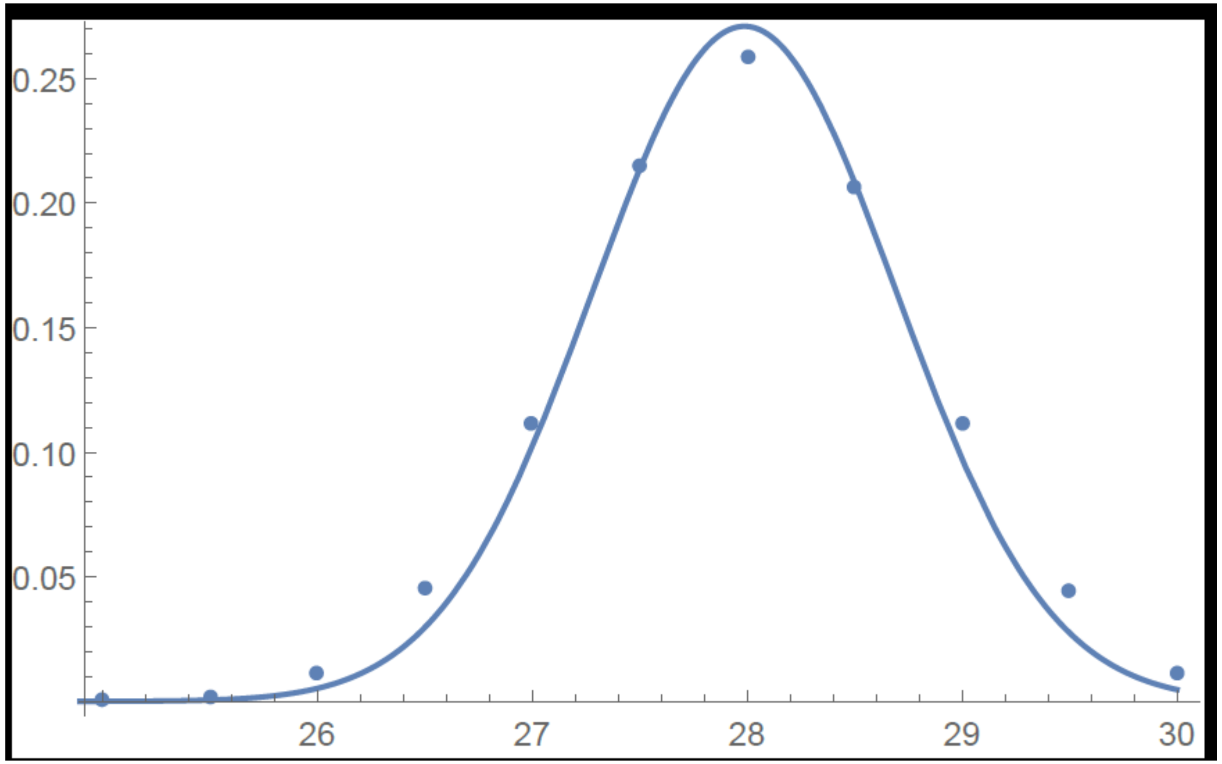
Exercise: Line fitting

- Download the data "data_for_linear_fit.dat" online.
- "linear.dat" is supposed to be a list of measured data of pairs of data points in the form of $\{x_i, y_i\}, i = 1,2, \dots, n$.
- Visualise the data using $\mathbf{ListPlot}[]$. You should realise that this data set lie along a supposed linear function, $y = a + bx$.
- Find the slope b and intersection a that best fit this data set.
- Overlapped the fitted function on the original data to show that you have done a good fit.

Gaussian function

A Gaussian function has the general form: $y = ae^{-(x-b)^2}$

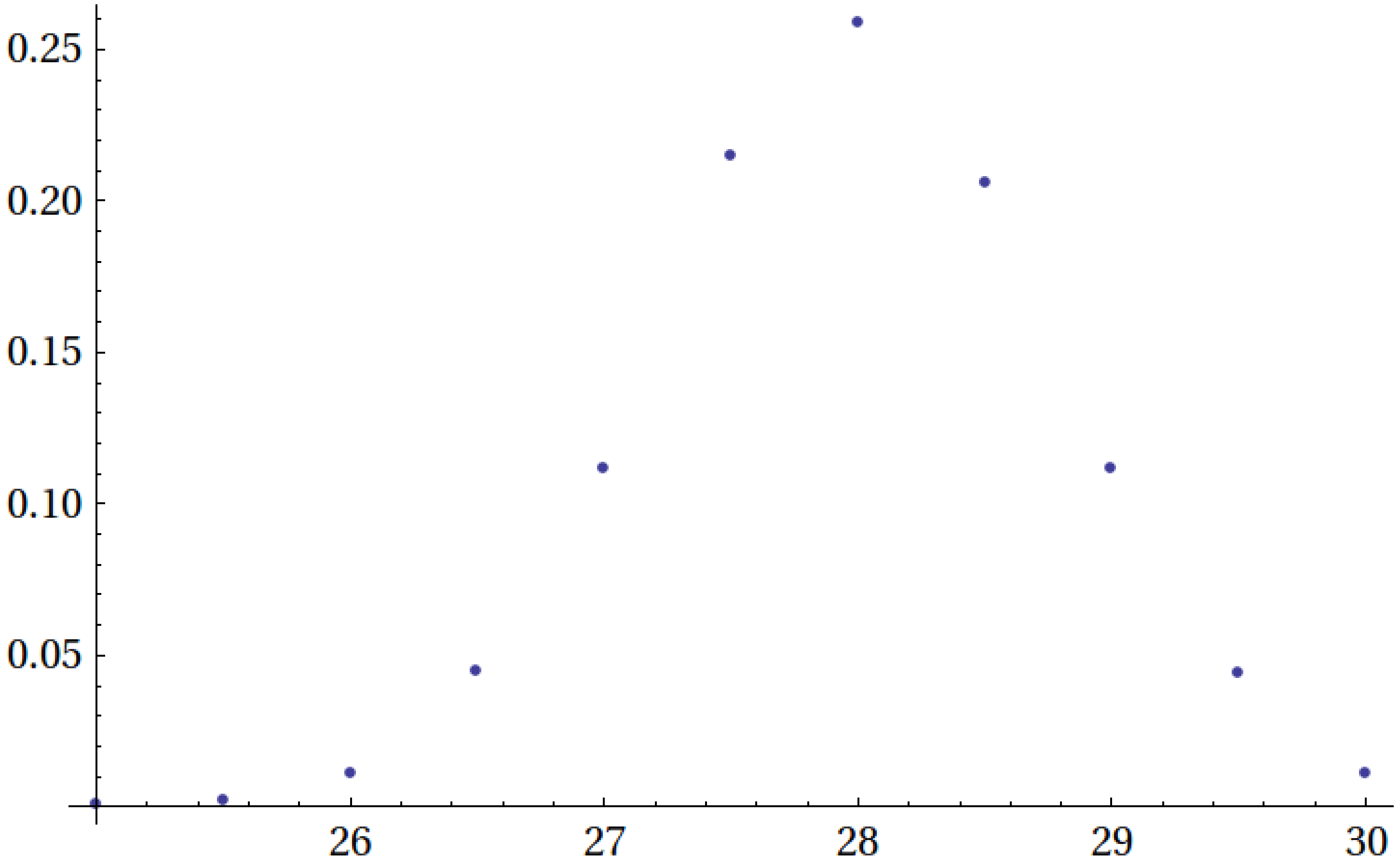
Parametrised by two parameters, a and b.



Exercise: Line fitting

- Download the data "gaussian.dat" online.
- It is supposed to be a list of measured data of pairs of data points in the form of $\{x_i, y_i\}, i = 1,2, \ldots, n$.
- Visualise the data using $\mathbf{ListPlot}[]$.
- This data set lie along a nonlinear curve of the form $y = ae^{-(x-b)^2}$.
- Find b and a that best fit this data set.
- Overlapped the fitted function on the original data to show that you have done a good fit.

gaussian.dat.



Data in XYZ format

See
http://openbabel.org/wiki/XYZ_(format) for data file in XYZ format.
https://reference.wolfram.com/language/ref/format/XYZ.html

Example File

```
12
benzene example
  C        0.00000        1.40272        0.00000
  H        0.00000        2.49029        0.00000
  C       -1.21479        0.70136        0.00000
  H       -2.15666        1.24515        0.00000
  C       -1.21479       -0.70136        0.00000
  H       -2.15666       -1.24515        0.00000
  C        0.00000       -1.40272        0.00000
  H        0.00000       -2.49029        0.00000
  C        1.21479       -0.70136        0.00000
  H        2.15666       -1.24515        0.00000
  C        1.21479        0.70136        0.00000
  H        2.15666        1.24515        0.00000
```

Visualising sample XYZ data

Download and install VMD at either

- https://staffusm-my.sharepoint.com/personal/tlyoon_usm_my/_layouts/15/guestaccess.aspx?docid=04b5757e70c3543038c7502d5c8b5702d&authkey=AdXO5ypu0iE8iXH-bpS5LfE
- or
- http://www.ks.uiuc.edu/Development/Download/download.cgi?PackageName=VMD

- Download the sample XYZ data files N3PD.xyz.
- Use VMD to visualise N3PD.xyz.

Data manipulation

- Import the online data file: http://comsics.usm.my/tlyoon/teaching/ZCE111/1617SEM2/data/atom1.lammpstrj

- Manipulate the data so that it can be converted into a *.xyz format.
- Export the *.xyz formatted file.
- Install vmd so that you can visualize the *.xyz file.

Converting *.lammpstrj into *.xyz format

- To this end, you need to know how to abstract the following information from atom1.lammpstrj

- 1. Total number of atom
- 2. Types of the atoms
- 3. x-,y- and z-coordinates of these atom
- 4. Write these info in a *.XYZ format into a named *.xyz file.

Exercise

- By making use of the **Manipulate**[] command, develop a code to visualize NP3D.xyz using Mathematica automatically without manual intervention.

Exercise: log.lammps

- If you are given a data file with certain format, can you write a code to read in the data, process them and visualise the content according to your need?
- Try this out on the file log.lammps, which is part of an output produced by a Molecular Dynamics simulation software package LAMMPS.
- log.lammps is a formatted file containing assorted information of the LAMPPS output, such as "Step" "Atoms" "Temp" "Press" "PotEng" "KinEng" "TotEng" "Volume" "Enthalpy

Exercise: log.lammps

- Write a Mathematica code to abstract the data of "Step" "Atoms" "Temp" "Press" "PotEng" "KinEng" "TotEng" "Volume" "Enthalpy from log.lammps.

Then plot

- Temp vs. Step
- PotEng vs. Step
- PotEng vs. Temp