Application of Computer in Chemistry SSC 3533

REGRESSION ANALYSIS

Prof. Mohamed Noor Hasan
Dr. Hasmerya Maarof
Department of Chemistry

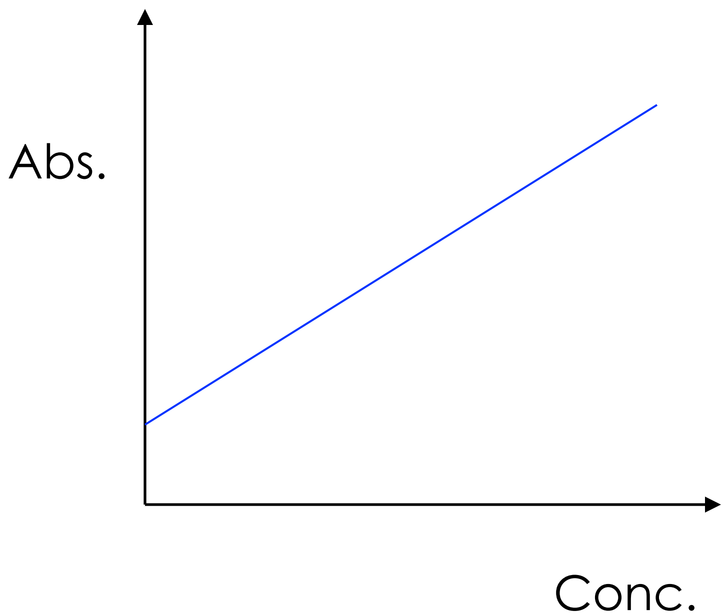
Introduction

Data obtained from experiments are usually plotted to produce a straight line.

Reasons for plotting a straight line curve:

- Calibration
- Extrapolation
- To find gradient
- Prediction

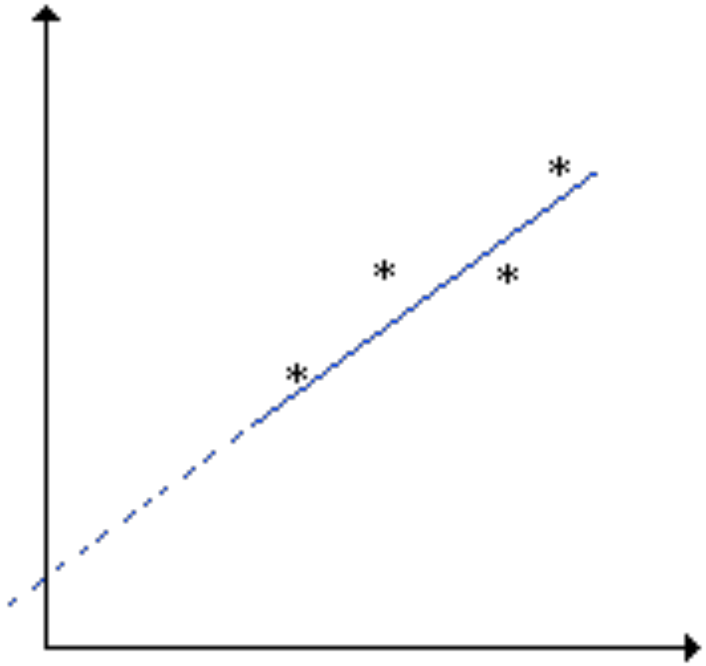
Calibration



A straight line equation $y = mx + c$ with gradient m and intercept c.

Calibration plot is used to determine concentration of unknown sample.

Extrapolation

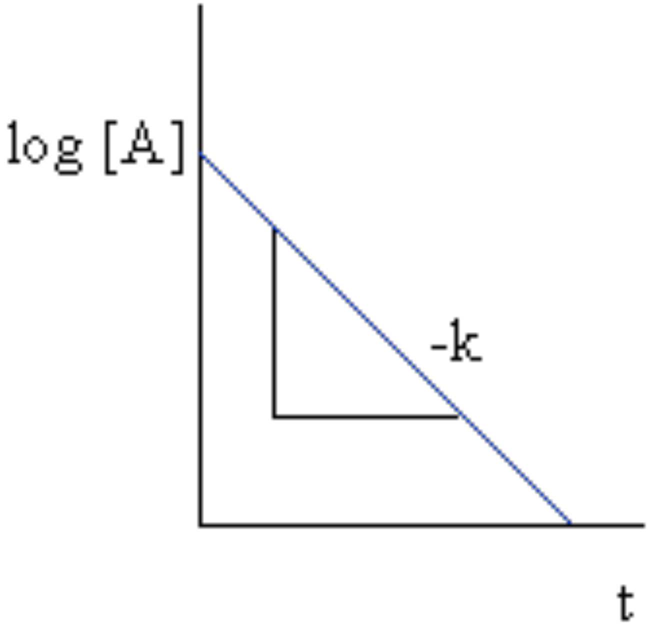


A straight line is used to obtain a value of intercept.

Value of *y* at *x*=0 could not be measured experimentally

Determination of Gradient

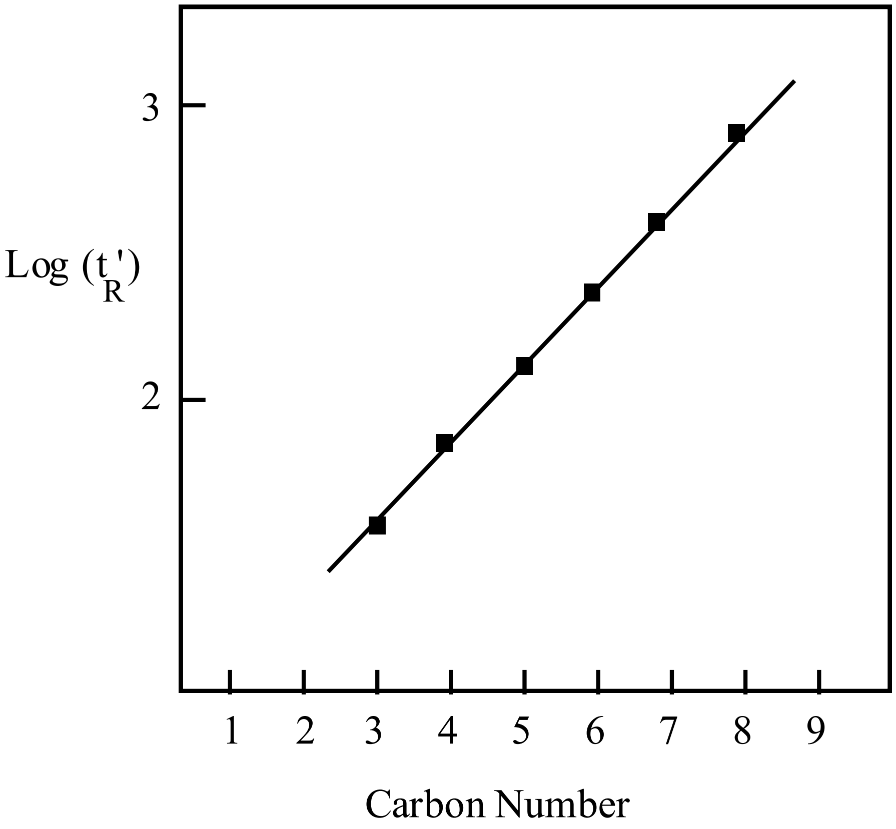
$$\text{Log}\,[A] = \log\,[A]_0 - kt$$



Example:
Determination of rate constant in first order reaction.

Gradient = -k

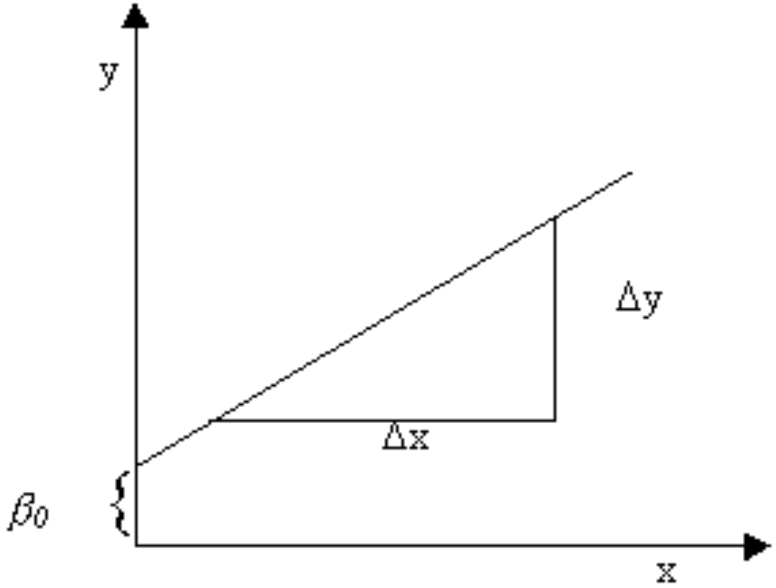
Predicting parameters



In chromatography, plot of retention time against number of carbon can be used to predict number of carbon atoms in the unknown.

Simple Linear Regression

The simplest relation between x and y is the linear relation or straight line relation.



$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

β_0 - intercept
β_1 - gradient
ε_i - error

Estimating Value of the Constants

Since we do not have access to all the population, we can estimate β_0 and β_1 from the sample:

$$\hat{y}_i = b_0 + b_1 x_i$$

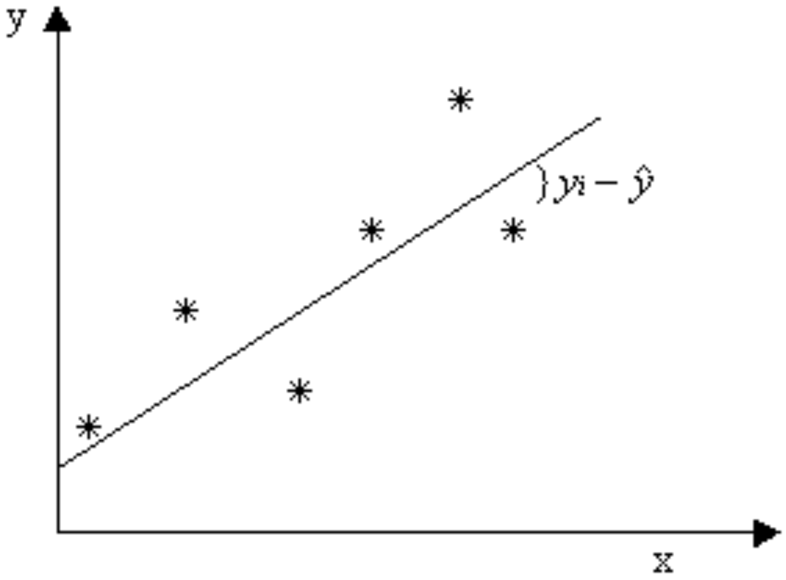
b_0 and b_1 can be calculated using least squares method.

The Least Squares Method

- Minimize deviation (error) between the observed and predicted values
- The method is used to obtain b_0 and b_1
- The set of selected b_0 and b_1 is the one that minimizes the error, Q

$$Q = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

The Deviation, Q



$$Q = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Q Sum of squared error

y_i observed value

$\hat{y}_i$ values obtained from the equation

n number of data

Defining the Equation

$$\hat{y}_i = b_0 + b_1 x_i$$

$$b_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

Another method to get b_o and b_1

b_0 and b_1 can be calculated using a computer
Values that have to be calculated:

$$\sum x_i \qquad \sum y_i \qquad \sum x_i^2 \qquad \sum x_i y_i$$

$$b_1 = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{D}$$

$$D = n\sum x_i^2 - (\sum x_i)^2$$

Example Calculation

In determination of a metal using spectrophotometric method, standard solutions of the metal were measured.

Example Calculation

ppm	Absorbance
10	0.087
20	0.154
30	0.202
40	0.283
50	0.313

Standard error

Standard error, *s*, is a measure how good is the relationship between y and x.

Small *s* means good relationship between *y* and *x*.

$$s = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$$

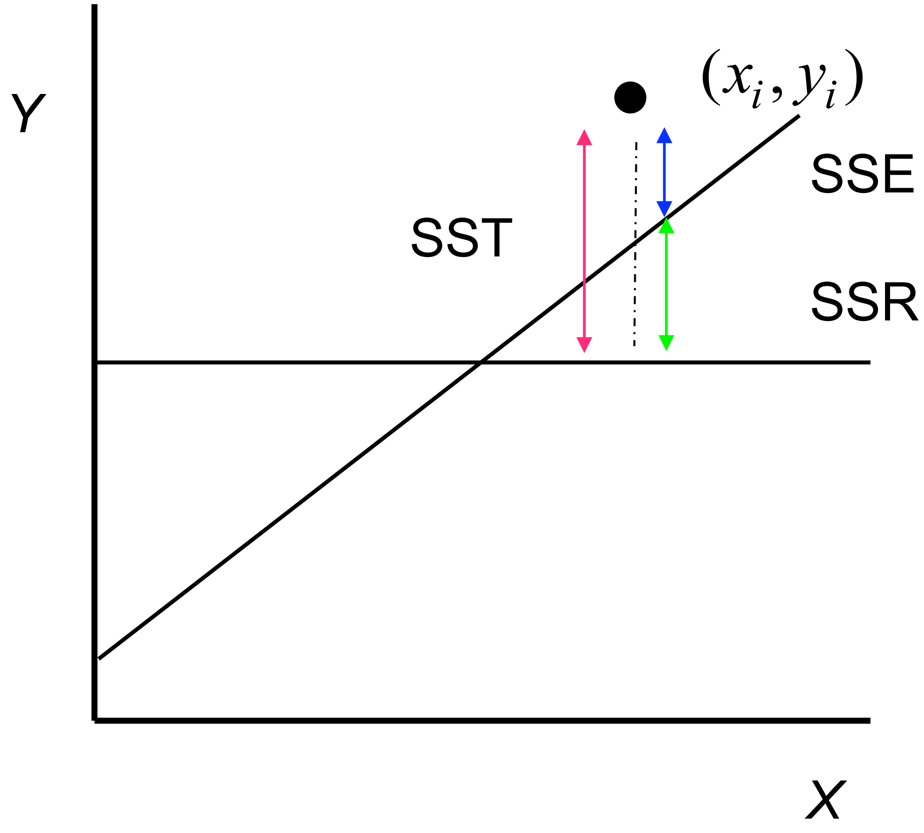
Correlation coefficient, *r*

- Correlation coefficient *r* is a parameter that can be used to show degree of correlation between *y* and *x*
- Value of *r* is between 0 - 1. The higher value of *r*, the better
- If gradient is negative, value of *r* is between 0 to -1

Coefficient of Determination, r^2

- Coefficient of determination is the percentage of variance that can be accounted for (explained) by the equation.
- Value of r^2 : 0 – 1 or 0 – 100%
- In statistics, r^2 is more meaningful than r and must be reported if the equation is to be used for prediction.

Analysis of Variance



Partition of variance

Total variance is a measure of variation around the mean value, can be divided into two components:

SST = SSR + SSE

SST –	Total Sum of Squares
SSR –	Sum of squares due to regression
SSE –	Error sum of squares

Formulae for the variance

$$\text{SST} = \sum (y_i - \bar{y})^2$$

$$\text{SSR} = \sum (\hat{y}_i - \bar{y})^2$$

$$\text{SSE} = \sum (y_i - \hat{y}_i)^2$$

Standard Error

- Standard error for a regression equation is the square root of MSE

$$s = \sqrt{MSE}$$

$$MSE = SSE/(n-2)$$

Coefficient of determination, r^2

$$r^2 = \frac{\text{SSR}}{\text{SST}}$$

- r^2 is the proportion of variance in y that can be explained by the equation
- Value of r^2 is between 0 – 1
- The higher value of r^2 the better is the equation, especially for prediction purposes.

Correlation coefficient, *r*

$$r = \pm\sqrt{r^2}$$

- Value of r : –1 ….. 0 ….. 1
- Higher value of r means higher correlation between y and x.
- Must know when to use r or r^2. Usually r^2 must be reported if the equation is to be used for predicting another value.
- r can only be used to show correlation between y and x.

F Test

$$F = \frac{\text{MSR}}{\text{MSE}}$$

- F test is conducted to determine whether b_1 is significant or not
- Look up F test table with degree of freedom = 1, n-2

Example Output

Weighted least Squares

- In least squares method, all y_i values are assumed to have the same precision. In experimental data, there are y_i values that are less precise.
- Less precise y_i values must be given lower weightage (w_i) – less influence on regression line

$$Q = \sum w_i (y_i - \bar{y}_i)^2$$

$$Q = \sum w_i (y_i - b_0 - b_i x_i)^2$$

Examples of weight factor, w

Value of w	Comment
1	No weight. All values have the same precision
$1/y_i$	Smaller values assumed to be more precise
$1/s^2$	s – standard deviation for the i^{th} value

Linearization by Transformation

- Not all relation between y and x are linear. One way to overcome this is by making it linear.

  Example: Exponential function

$$f(x_i) = y_i = \alpha \, e^{\beta x_i}$$

if $\beta > 0$ – increasing exponent
eg. The change in population with time

if $\beta < 0$ – decreasing exponent
eg. In radioactive decay process

To make an equation linear

The simplest way to make an exponential equation linear is by taking its logarithm

$y' = \ln y = \ln \alpha + \beta x$

This is a linear equation with:

$y' = \ln y$

$bo = \ln \alpha$

$b_1 = \beta$

$y' = bo + b_1 x$