REGRESSION ANALYSIS

Regression analysis is a form of predictive modelling technique which investigates the relationship between a **dependent** (target) and **independent variable(s)** (predictor). This technique is used for forecasting, time series modelling and finding the causal effect relationship between the variables. For example, relationship between rash driving and number of road accidents by a driver is best studied through regression.

One typical example is the relation between Nusselt number and Prandtl and Reynolds numbers, $\mathrm{Nu} = f\left(\mathrm{Re}\,,\, \mathrm{Pr}\right)$

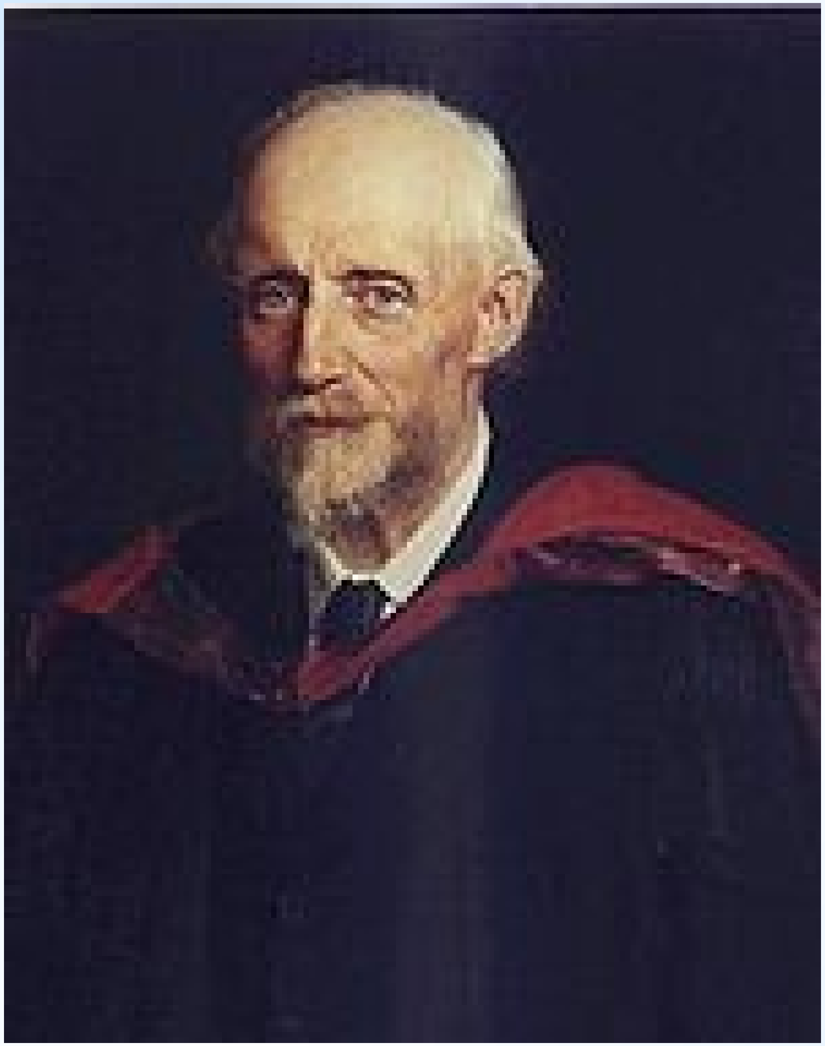
$$\mathrm{Nu} = \frac{\text{Conductive HT}}{\text{Convective HT}} \qquad \mathrm{Re} = \frac{\text{Inertia force}}{\text{Viscous force}} \qquad \mathrm{Pr} = \frac{\text{Momentum diffusivity}}{\text{Thermal diffusivity}}$$

$$\mathrm{Nu} = C\,\mathrm{Re}^{m}\ \mathrm{Pr}^{n} \qquad \mathrm{Nu} = 0.023\,\mathrm{Re}^{0.8}\ \mathrm{Pr}^{0.3}$$

Wilhelm Nusselt

German Engineer

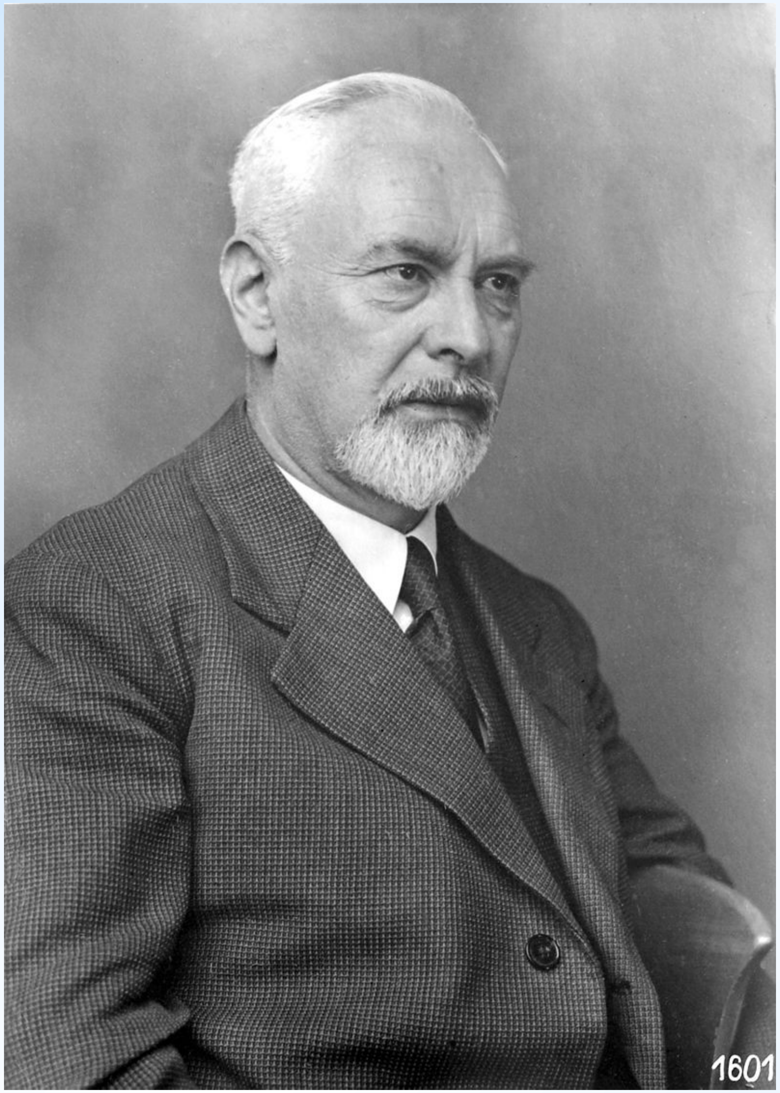
1882 -1957



Osborne Reynolds

British Engineer

1842 -1912



Ludwig Prandtl

German Engineer

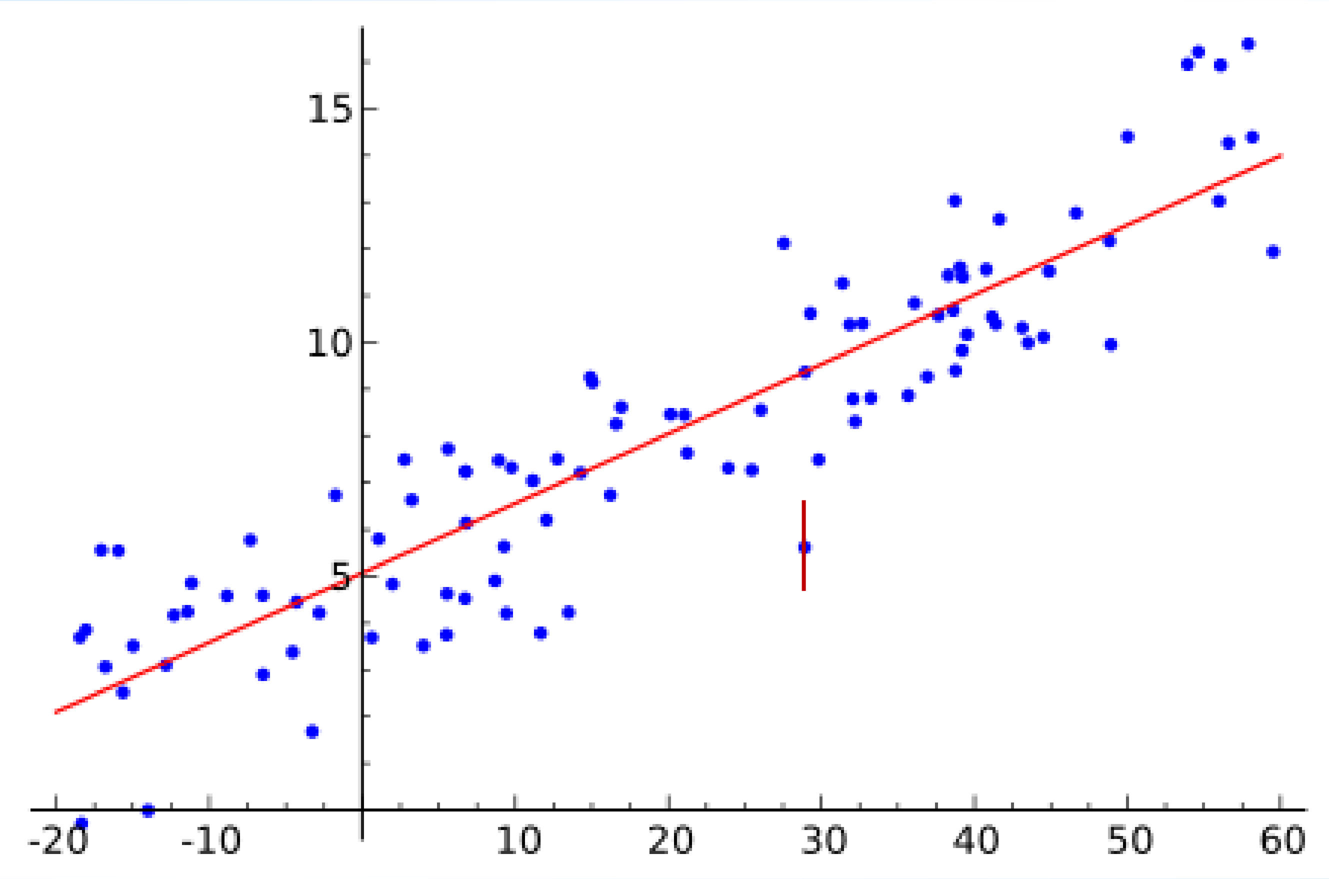
1875 -1953

LEAST-SQUARES APPROXIMATION

Given:	1. Experimental data subject to **random** statistical errors
	2. A function with unknown parameters known (or believed) to represent the data
Problem:	Find the parameters such that the function is the "**best**" representation of the given data

Example	Given Data		Given function with unknown parameters
	x_i	y_i	$y(x) = C_1 x + C_2$
	....	....	$y(x) = C_1 x^2 + C_2 x + C_3$
	....	....	$y(x) = C_1 \exp(C_2 x)$
	...	....	etc.

Question: How to find the unknown parameters, C_i 's , such that $y = f(x)$ is the **best-fitting** function.


15
10
5
-20
-10
10
20
30
40
50
60

Norms

Given: A function, f(x), or tabulated data points in a certain range, [a , b]

Problem: Find one, simple function, g(x), that represents or fits the given f(x) or the data points in [a , b] according to a certain "goodness" or "closeness" is required. How do we define "goodness" of fitting?

Answer: There are basically three ways of defining "goodness" which are called "Norms"

Least Squares or L_2 Norm: Minimize sum of the squares of the error, i.e.

$$\int_a^b [f(x) - g(x)]^2 \, dx$$

Least Deviation or L_1 Norm: Minimize average error, i.e.,

$$\int_a^b |f(x) - g(x)| \, dx$$

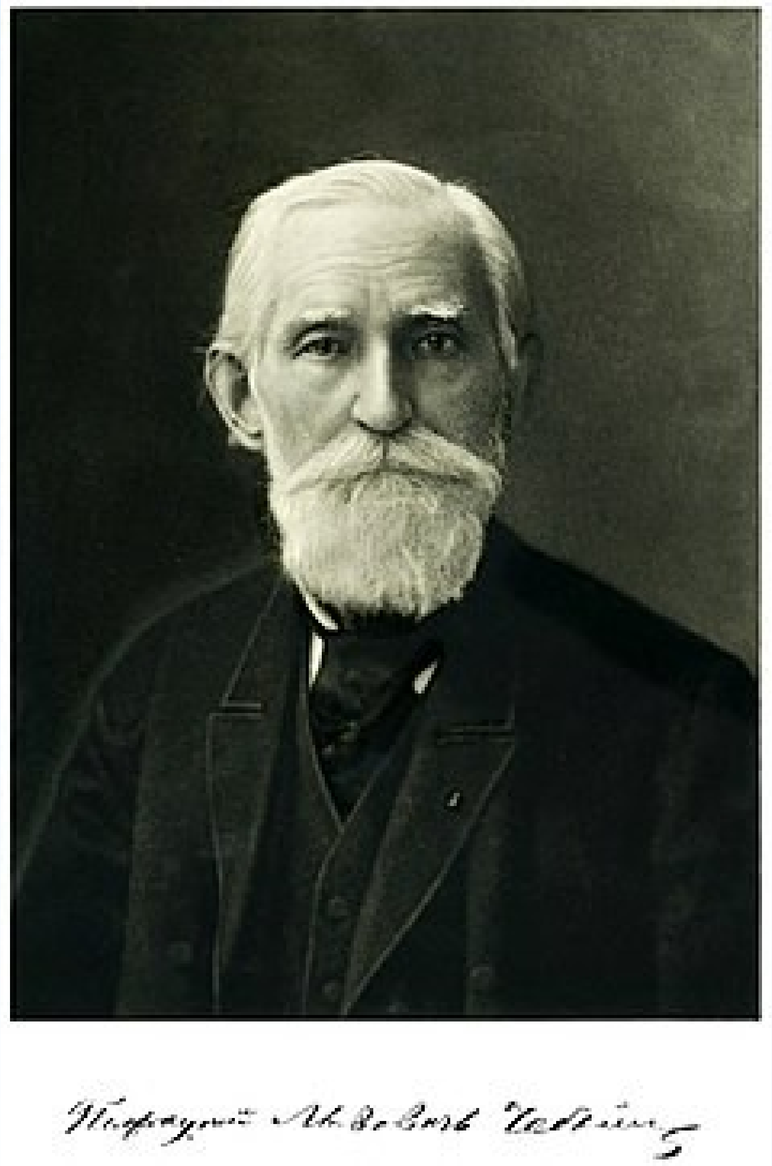
Chebyshev or L_∞ Norm: Minimize maximum error, i.e.,

Max. $|f(x) - g(x)|$ in $[a, b]$

The meaning of the **Least-Squares or L_2 Norm** is intuitively not as clear as the others, but it is very widely used in many problems because of its simplicity in implementation and analysis, even when the other norms are more natural and meaningful.

L_1 Norm is more difficult to use and analyze. Its main attraction is in data analysis and smoothing as it is remarkably insensitive to random errors, wild points, and other uncertainties.

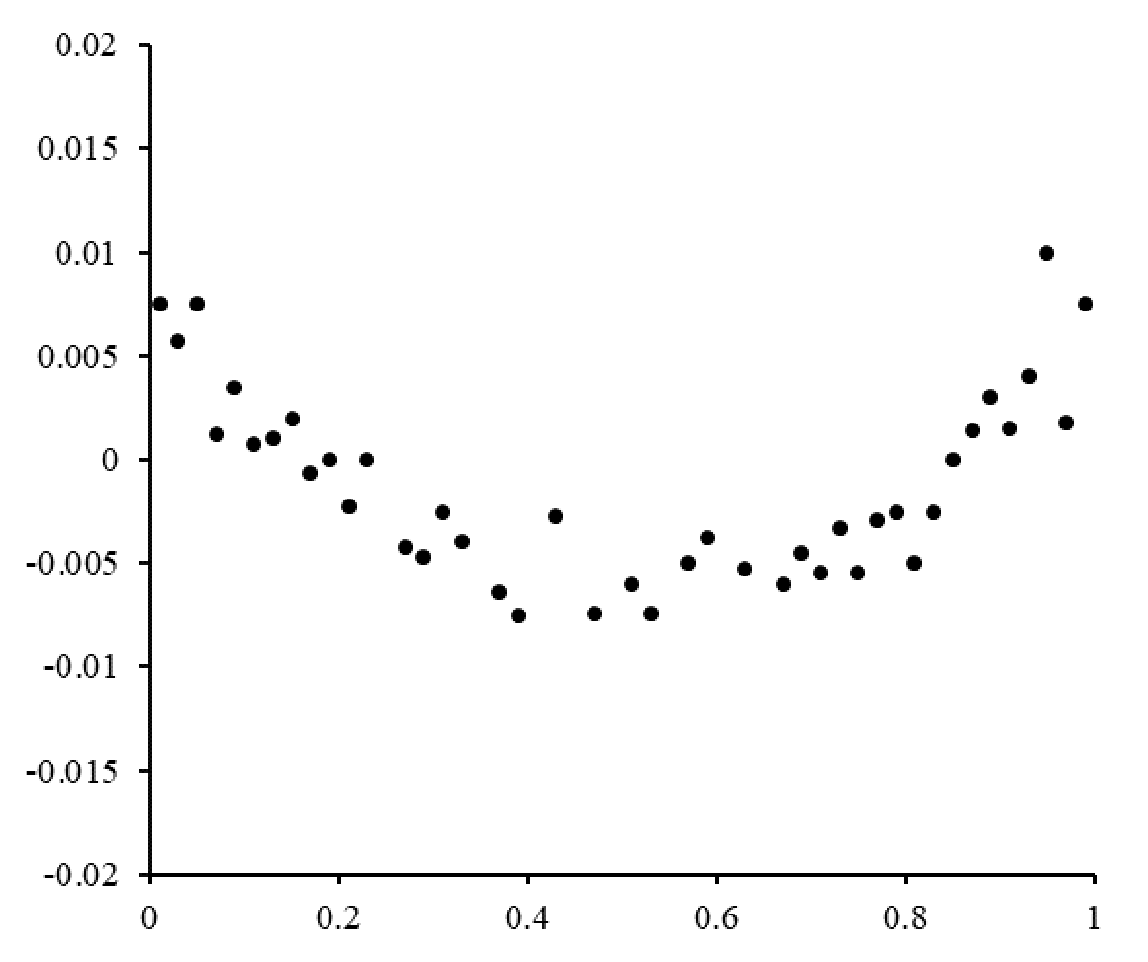
Chebyshev or L_∞ Norm is primarily used for approximating standard mathematical functions rather than data points, because it is unsuitable when uncertainty is present in the data. It often magnifies the effect of a single uncertainty or random error.



Pafnuty Lvovich Chebyshev

Russian Mathematician

1821 - 1894



Given: data with random statistical errors on y_i

Given: the function, y(x) with unkown parameters, C_1, C_2, .

Find the paramters, C_1, C_2, . such the there is best fitting in the sense of LSQ norm

Best Fit: Maximum likelihood of occurence

Theorem in statistics: Maximum likelihood occurs if the sum of the squares of the errors is minimized.

Sum of the squares of the errors:

$$S = \sum_{i=1}^{N} E_i^2 = \sum_{i=1}^{N} [y_i - y(x_i)]^2$$

If $y(x) = C_1 x + C_2$

$$S = \sum_{i=1}^{N} [y_i - (C_1 x_i + C_2)]^2$$

If $y(x) = C_1 \sin(C_2 x)$

$$S = \sum_{i=1}^{N} [y_i - C_1 \sin(C_2 x_i)]^2$$

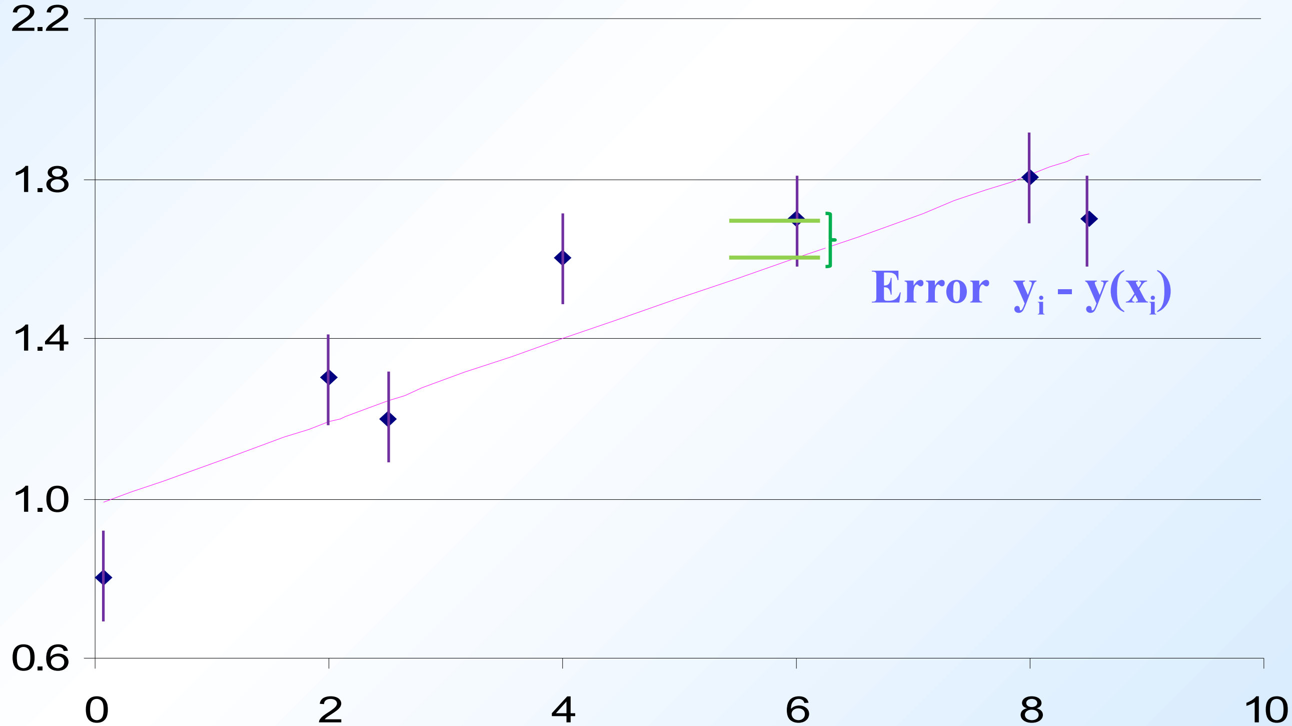
Minimize S:	Set the first partial derivatives of S with respect to the unknown parameters, C_i's, equal to zero.	$\frac{\partial S}{\partial C_j} = 0 \quad \text{for all} \quad j = 1, \ldots, n$

This gives n equations, linear or non-linear, for n unkown parameters. Solve simultaneously.

Example

x	0.07	2.0	2.5	4.0	6.0	8.0	8.5
y	0.8	1.3	1.2	1.6	1.7	1.8	1.7

$$y(x) = C_1 x + C_2$$



$$y(x) = C_1\, x + C_2$$

$$S = \sum_{i=1}^{N=7} E_i^2 = \sum_{i=1}^{N=7} \left[y_i - y(x_i)\right]^2 = \sum_{i=1}^{N=7} \left[y_i - C_1\, x_i - C_2\right]^2$$

$$\frac{\partial S}{\partial C_1} = 0 \quad => \quad \frac{\partial S}{\partial C_1} = \sum_{i=1}^{N=7} 2\,\left[y_i - C_1\, x_i - C_2\right]\,\left[-\, x_i\right] = 0$$

$$\frac{\partial S}{\partial C_2} = 0 \quad => \quad \frac{\partial S}{\partial C_2} = \sum_{i=1}^{N=7} 2\,\left[y_i - C_1\, x_i - C_2\right]\,\left[-\, 1\right] = 0$$

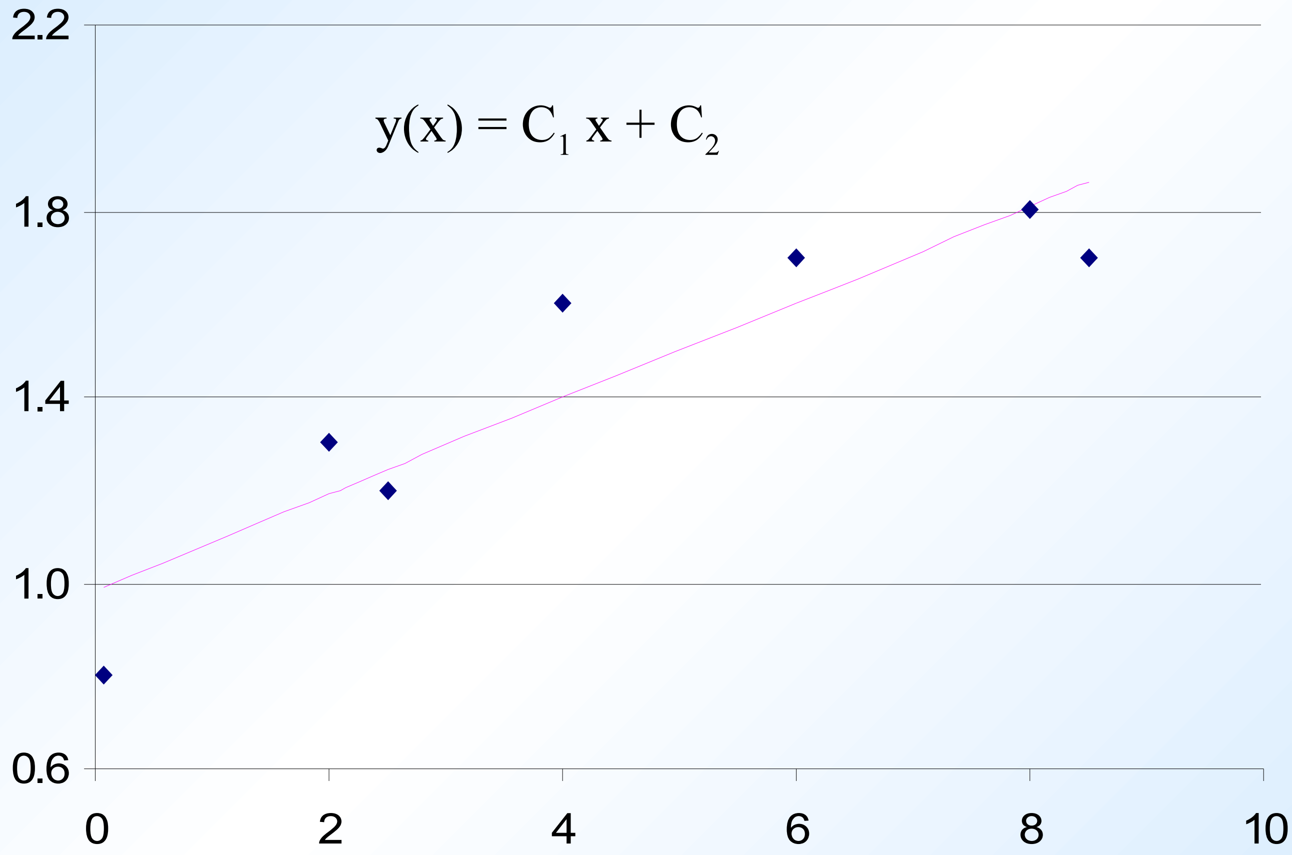
$$\left.\begin{aligned} \sum_{i=1}^{N=7} y_i\, x_i &= C_1 \sum_{i=1}^{N=7} x_i^2 + C_2 \sum_{i=1}^{N=7} x_i \\ \sum_{i=1}^{N=7} y_i &= C_1 \sum_{i=1}^{N=7} x_i + C_2\, N \end{aligned}\right\} \quad \begin{pmatrix} \sum_{i=1}^{N=7} x_i^2 & \sum_{i=1}^{N=7} x_i \\ \sum_{i=1}^{N=7} x_i & 7 \end{pmatrix} \begin{pmatrix} C_1 \\ C_2 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N=7} y_i\, x_i \\ \sum_{i=1}^{N=7} y_i \end{pmatrix}$$

i	x_i	y_i	x_i^2	$x_i\,y_i$
1	0.07	0.80	0.0049	0.056
2	2.00	1.30	4.0000	2.600
3	2.50	1.20	6.2500	3.000
4	4.00	1.60	16.0000	6.400
5	6.00	1.70	36.0000	10.200
6	8.00	1.80	64.0000	14.400
7	8.50	1.70	72.2500	14.450
SUM	31.070	10.100	198.5049	51.106

$$\begin{pmatrix} \sum_{i=1}^{N=7} x_i^2 & \sum_{i=1}^{N=7} x_i \\ \sum_{i=1}^{N=7} x_i & N \end{pmatrix} \begin{pmatrix} C_1 \\ C_2 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N=7} y_i\, x_i \\ \sum_{i=1}^{N=7} y_i \end{pmatrix} \Bigg\} \quad \begin{pmatrix} 198.5049 & 31.070 \\ 31.070 & 7 \end{pmatrix} \begin{pmatrix} C_1 \\ C_2 \end{pmatrix} = \begin{pmatrix} 51.106 \\ 10.100 \end{pmatrix}$$

Example

x	0.07	2.0	2.5	4.0	6.0	8.0	8.5
y	0.8	1.3	1.2	1.6	1.7	1.8	1.7



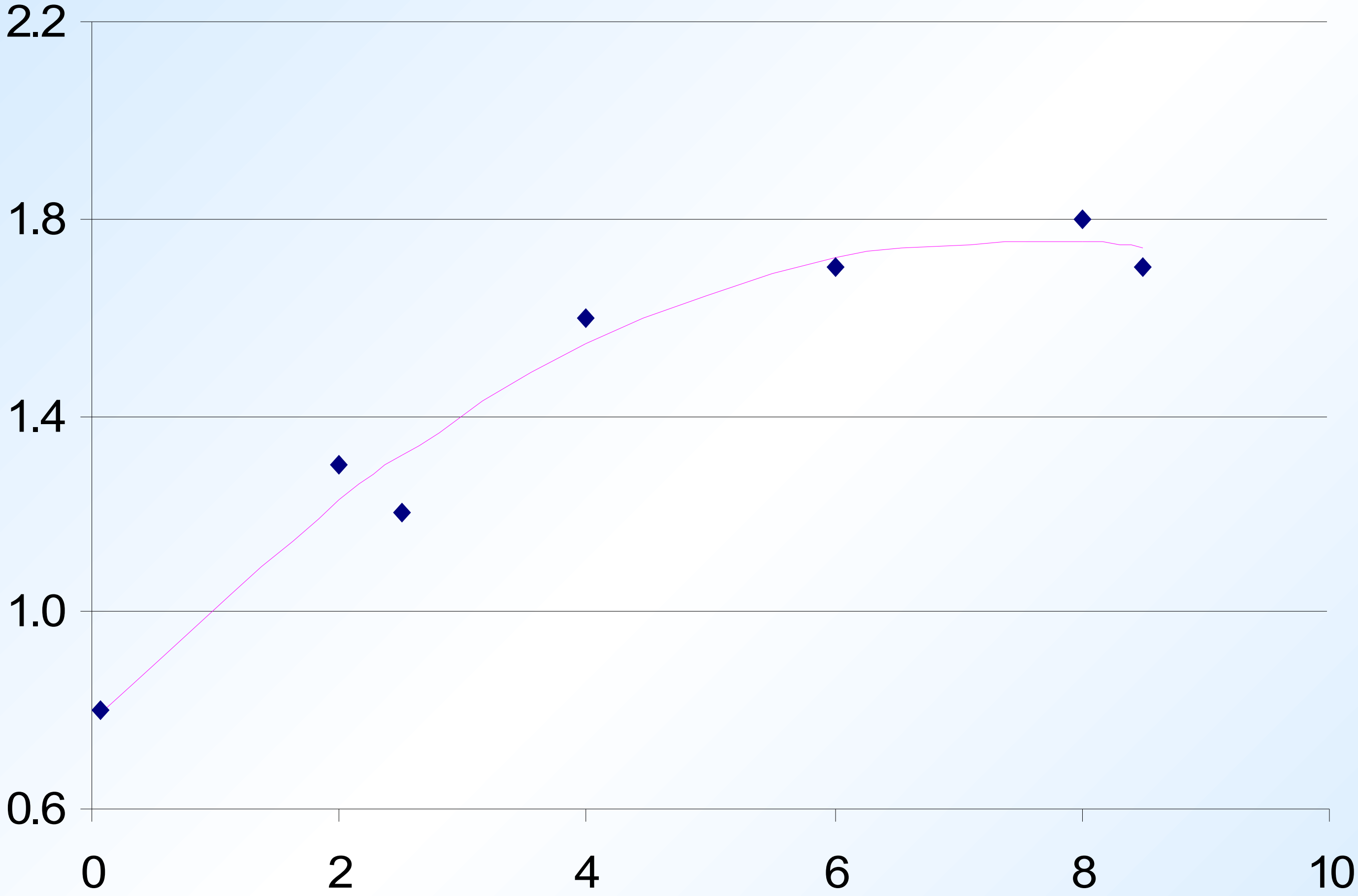
Linear LSQ

$C_1 = 0.10357402$

$C_2 = 0.98313647$

$S = 0.12706794$

$$y(x) = C_1 x^2 + C_2 x + C_3$$



Find

C_1, C_2, and C_3

using LSQ

$y(x) = C_1 x^2 + C_2 x + C_3$

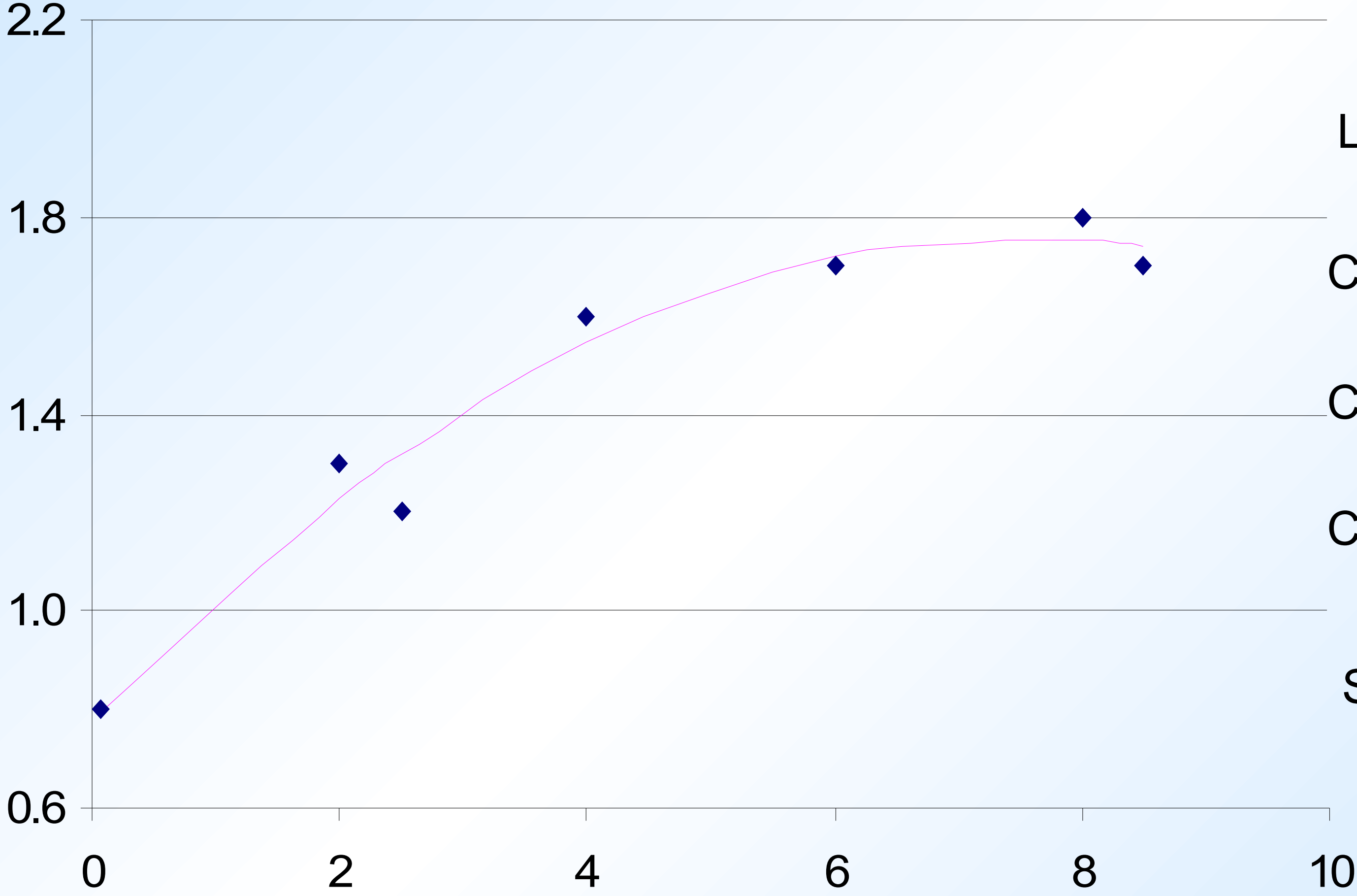
Find C_1, C_2, and C_3 using LSQ

$$\begin{pmatrix} \sum_{i=1}^{N=7} x_i^4 & \sum_{i=1}^{N=7} x_i^3 & \sum_{i=1}^{N=7} x_i^2 \\ \sum_{i=1}^{N=7} x_i^3 & \sum_{i=1}^{N=7} x_i^2 & \sum_{i=1}^{N=7} x_i \\ \sum_{i=1}^{N=7} x_i^2 & \sum_{i=1}^{N=7} x_i & 7 \end{pmatrix} \begin{pmatrix} C_1 \\ C_2 \\ C_3 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N=7} y_i\, x_i^2 \\ \sum_{i=1}^{N=7} y_i\, x_i \\ \sum_{i=1}^{N=7} y_i \end{pmatrix}$$

Linear set of equations in C_1, C_2, and C_3

Solve for C_1, C_2, and C_3

$$y(x) = C_1 x^2 + C_2 x + C_3$$



Linear LSQ

$C_1 = -0.01751769$

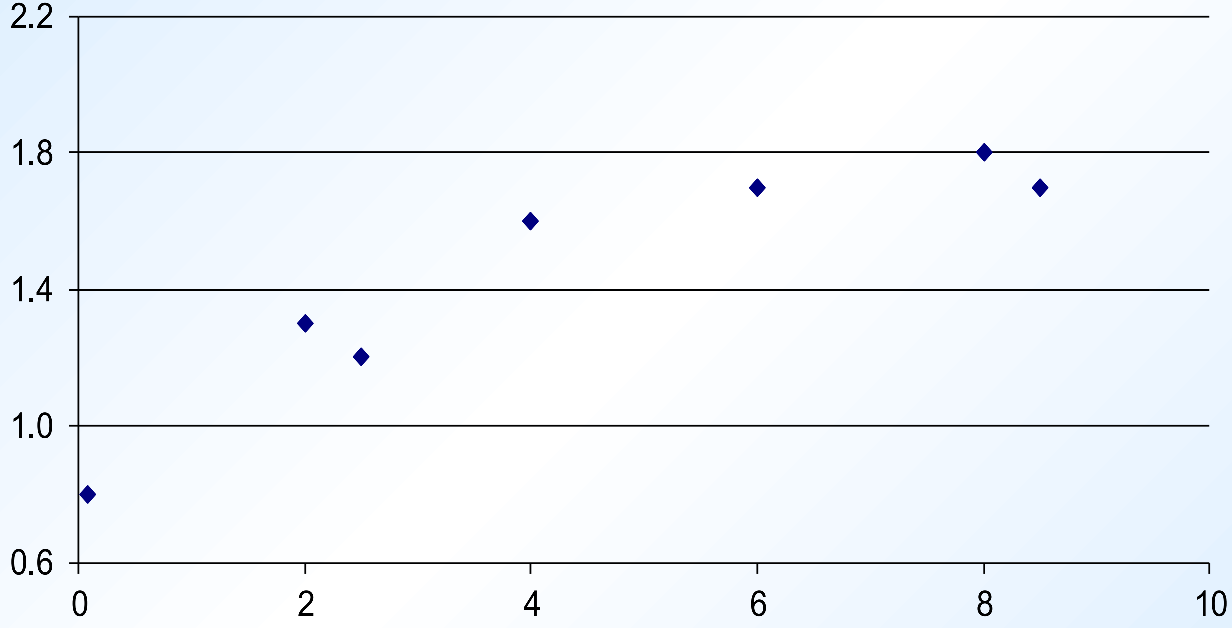
$C_2 = 0.26218308$

$C_3 = 0.775902724$

$S = 0.02698109$

Example

x	0.07	2.0	2.5	4.0	6.0	8.0	8.5
y	0.8	1.3	1.2	1.6	1.7	1.8	1.7



$y(x) = C_1 \exp(C_2 x)$ Find C_1 and C_2 by minimizing S

$$S = \sum_{i=1}^{N=7} E_i^2 = \sum_{i=1}^{N=7} \left[y_i - y(x_i) \right]^2 = \sum_{i=1}^{N=7} \left[y_i - C_1 \exp\left(C_2 x_i\right) \right]^2$$

$$\frac{\partial S}{\partial C_1} = 0 \quad => \quad \frac{\partial S}{\partial C_1} = \sum_{i=1}^{N=7} 2 \left[y_i - C_1 \exp\left(C_2 x_i\right) \right] \left[- \exp\left(C_2 x_i\right) \right] = 0$$

$$\frac{\partial S}{\partial C_2} = 0 \quad => \quad \frac{\partial S}{\partial C_2} = \sum_{i=1}^{N=7} 2 \left[y_i - C_1 \exp\left(C_2 x_i\right) \right] \left[- C_1 \, x_i \exp\left(C_2 x_i\right) \right] = 0$$

These equations are non-linear. How do you solve for C_1 and C_2?

How about simplifying the original form, $y(x) = C_1 \exp(C_2 x)$

$y(x) = C_1 \exp(C_2 x)$

$\ln[y(x)] = \ln(C_1) + C_2\, x$

Define:

$z(x) = \ln[y(x)]$

$A = \ln(C_1)$

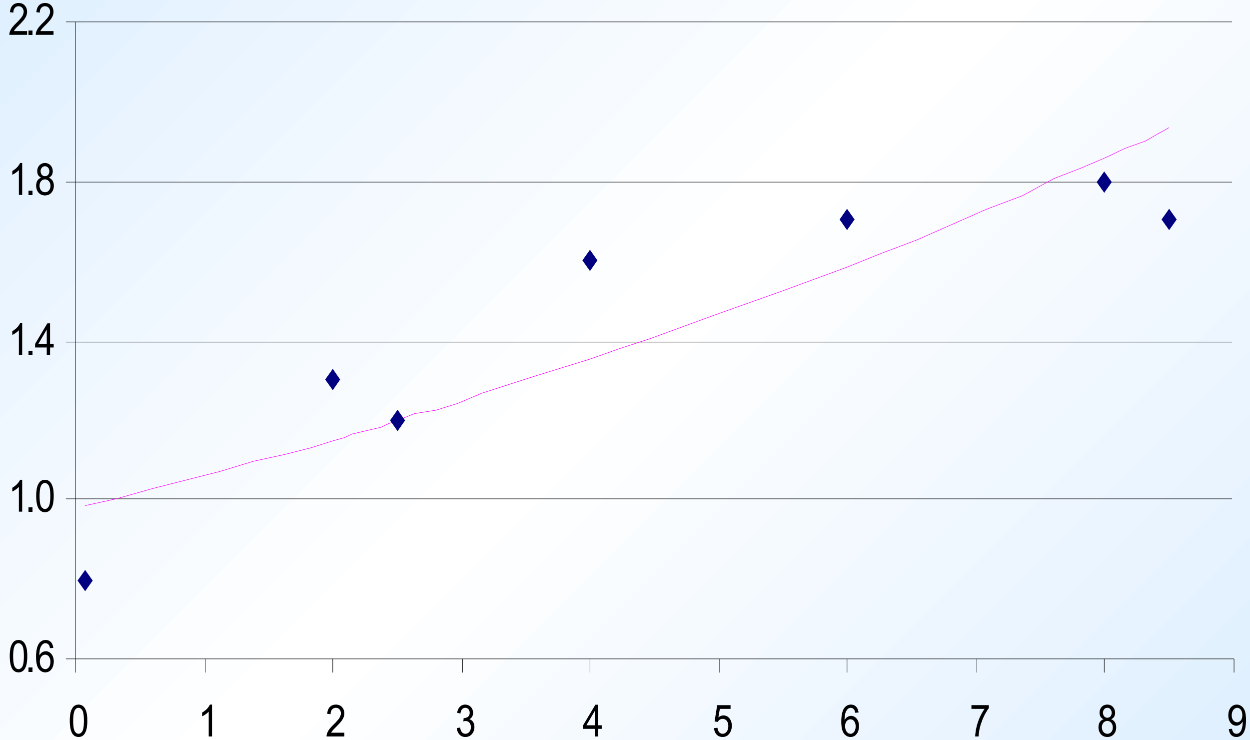
$B = C_2$

$x = x$

$z(x) = A + B\, x$

Find A and B

Then change to C_1 and C_2



Linear LSQ:

$C_1 = 0.97916877$

$C_2 = 0.08007557$

$S = 0.19171593$

LINEARIZATION WITH LEAST-SQUARES

Given Function	Linearized Form
$y = \dfrac{A}{x+B}$	$\left(\dfrac{1}{y}\right) = \dfrac{1}{A}(x) + \dfrac{B}{A}$
$y = \dfrac{A}{x+B}$	$(y) = \dfrac{-1}{B}(x\,y) + \dfrac{A}{B}$
$y = A \ln x + B$	$(y) = A\,(\ln x) + B$
$y = A\,e^{B\,x}$	$(\ln y) = B\,(x) + \ln A$
$y = A\,x^{B}$	$(\ln y) = B\,(\ln x) + \ln A$
$y = \dfrac{A}{(x+B)^2}$	$\left(\sqrt{\dfrac{1}{y}}\right) = \dfrac{1}{\sqrt{A}}(x) + \dfrac{B}{\sqrt{A}}$

Example: Following data and function with two parameters are given

x_i	1.0	2.0	3.0	4.0	5.0
y_i	2.0	5.0	10.0	17.0	26.0

$$y(x) = \frac{A}{x + B}$$

Fit this function to the given data

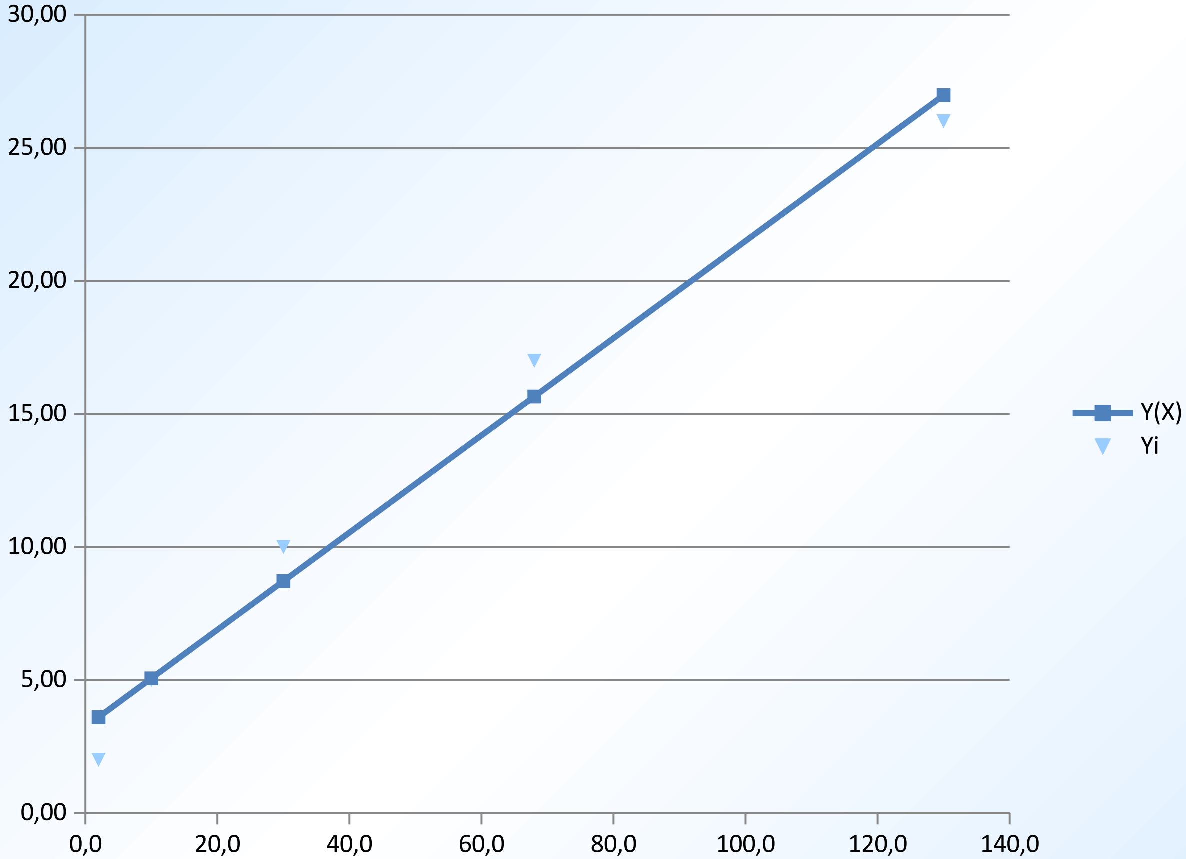
(a) Define variables $Y = y$ and $X = x\,y$

$$Y(X) = \frac{A}{B} - \frac{1}{B} X \qquad => \qquad Y(X) = C_1 + C_2 X$$

(b) Define variables $Y = 1/y$ and $X = x$

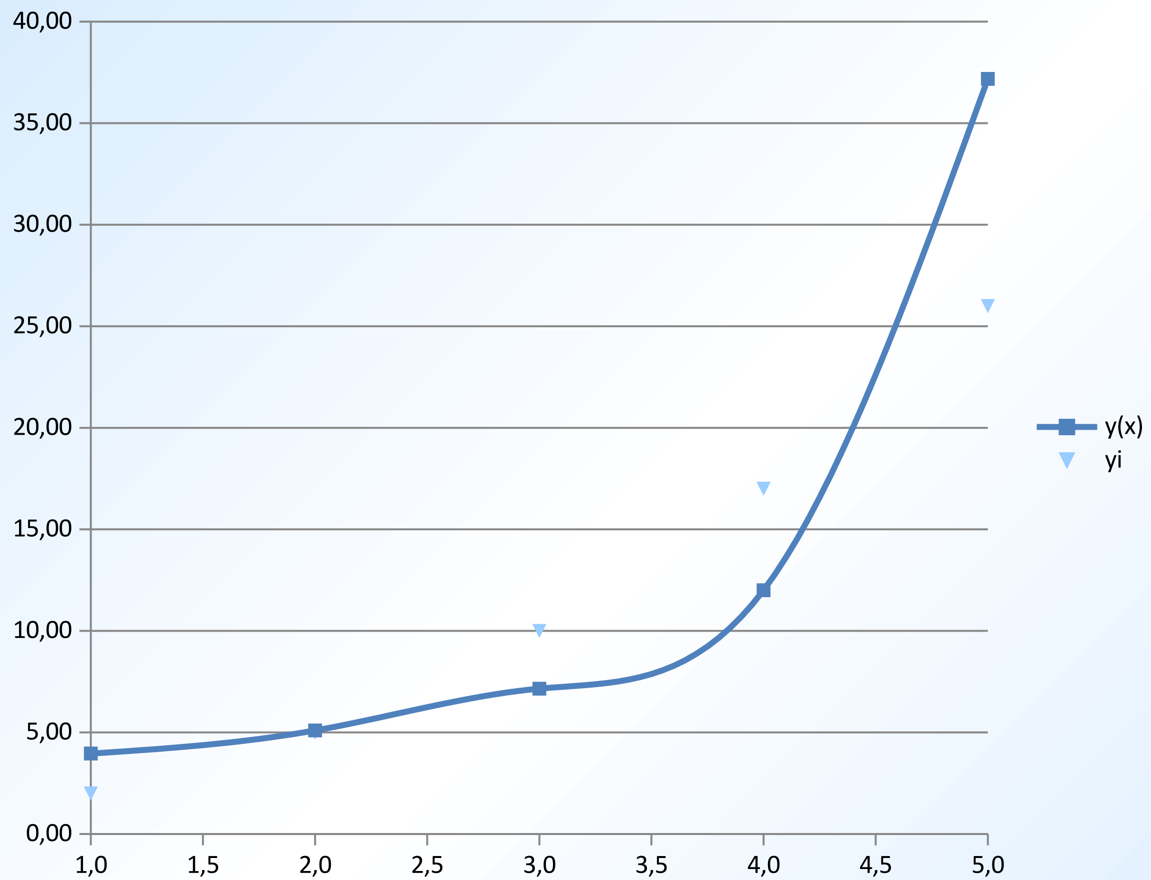
$$Y(X) = \frac{B}{A} + \frac{1}{A} X \qquad => \qquad Y(X) = C_1 + C_2 X$$

$Y = y$ and $X = x\,y$ $\qquad Y(X) = \frac{A}{B} - \frac{1}{B}X \quad \Rightarrow \quad Y(X) = C_1 + C_2\,X$



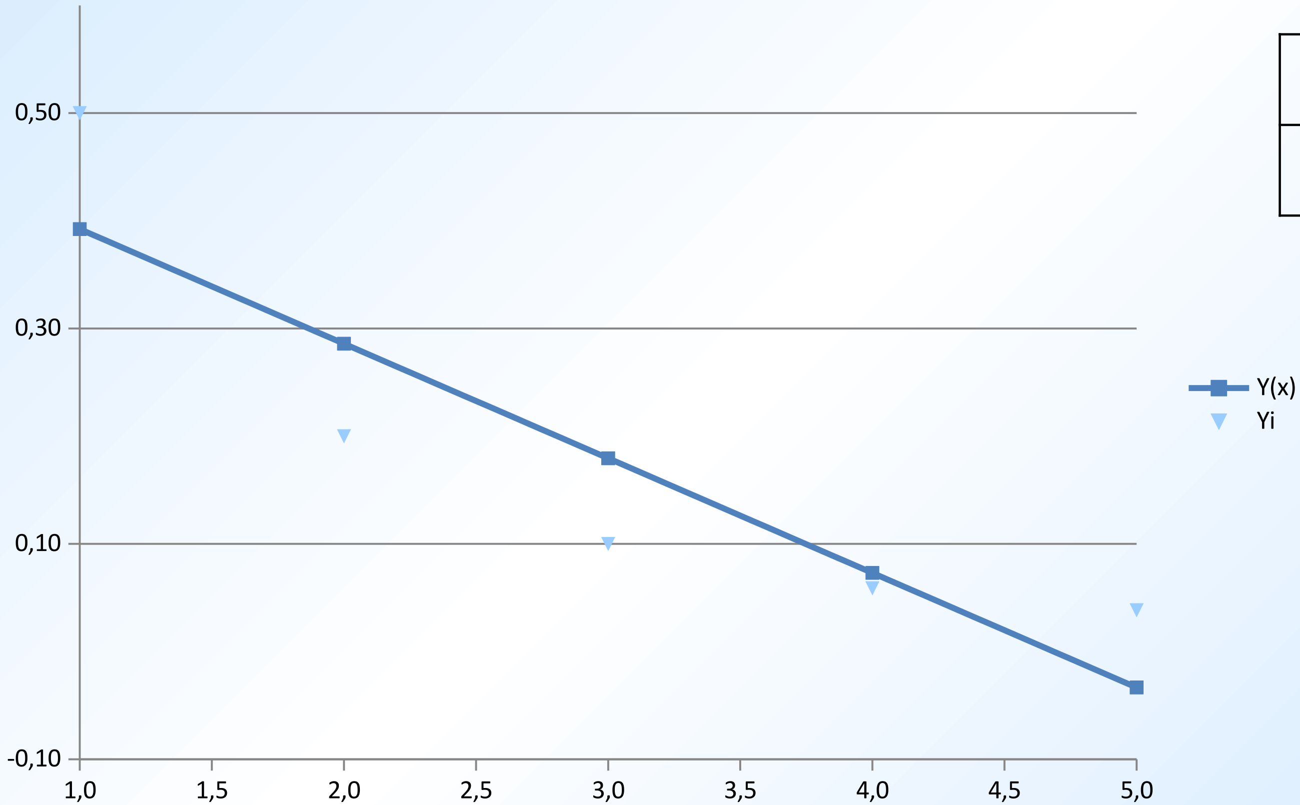
C_1 =	3.2354651
C_2 =	0.1825945

$Y = y$ and $X = x\,y$ $\quad Y(X) = \dfrac{A}{B} - \dfrac{1}{B} X \quad \Rightarrow \quad Y(X) = C_1 + C_2 X$



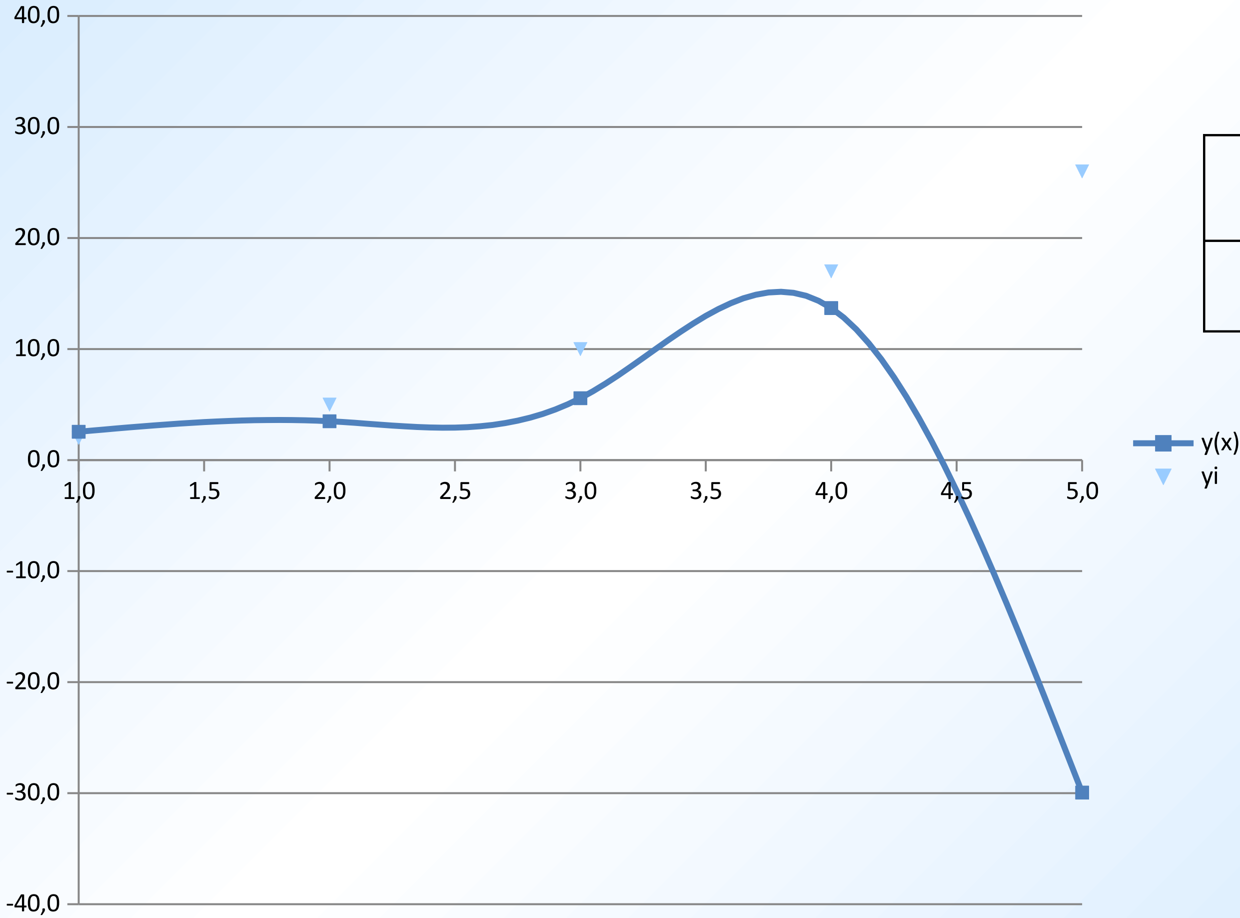
A =	-17.7194030
B =	-5.4766169

$$Y = \frac{1}{y} \text{ and } X = x \qquad Y(X) = \frac{B}{A} + \frac{1}{A} X \qquad \Rightarrow \qquad Y(X) = C_1 + C_2 X$$



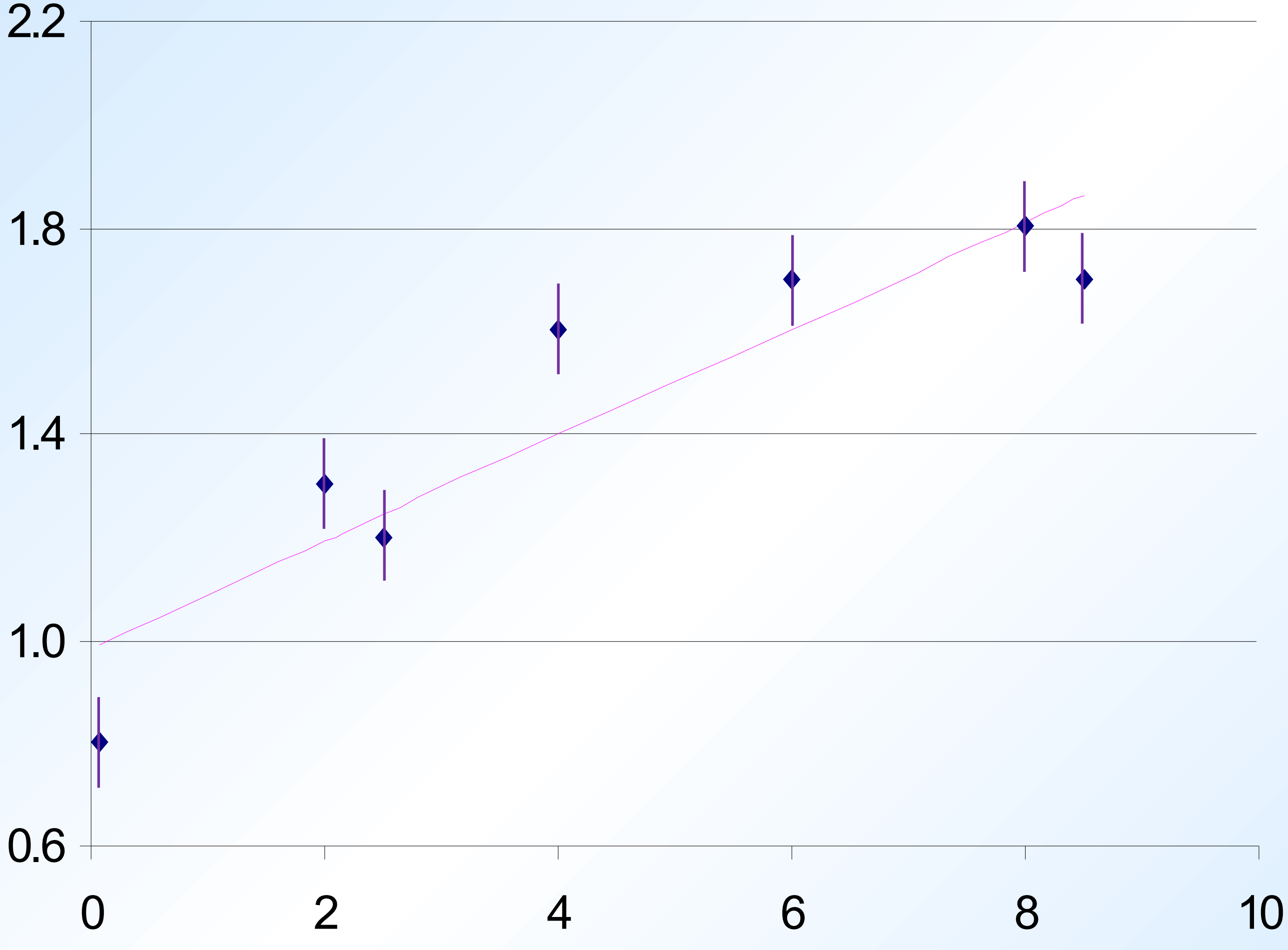
C_1 =	0.4987330
C_2 =	-0.1064253

$$Y = \frac{1}{y} \quad \text{and} \quad X = x \qquad Y(X) = \frac{B}{A} + \frac{1}{A}X \quad => \quad Y(X) = C_1 + C_2 X$$



A =	-9.3962585
B =	-4.6862245

Standard deviations at every point, σ_i, are the same.



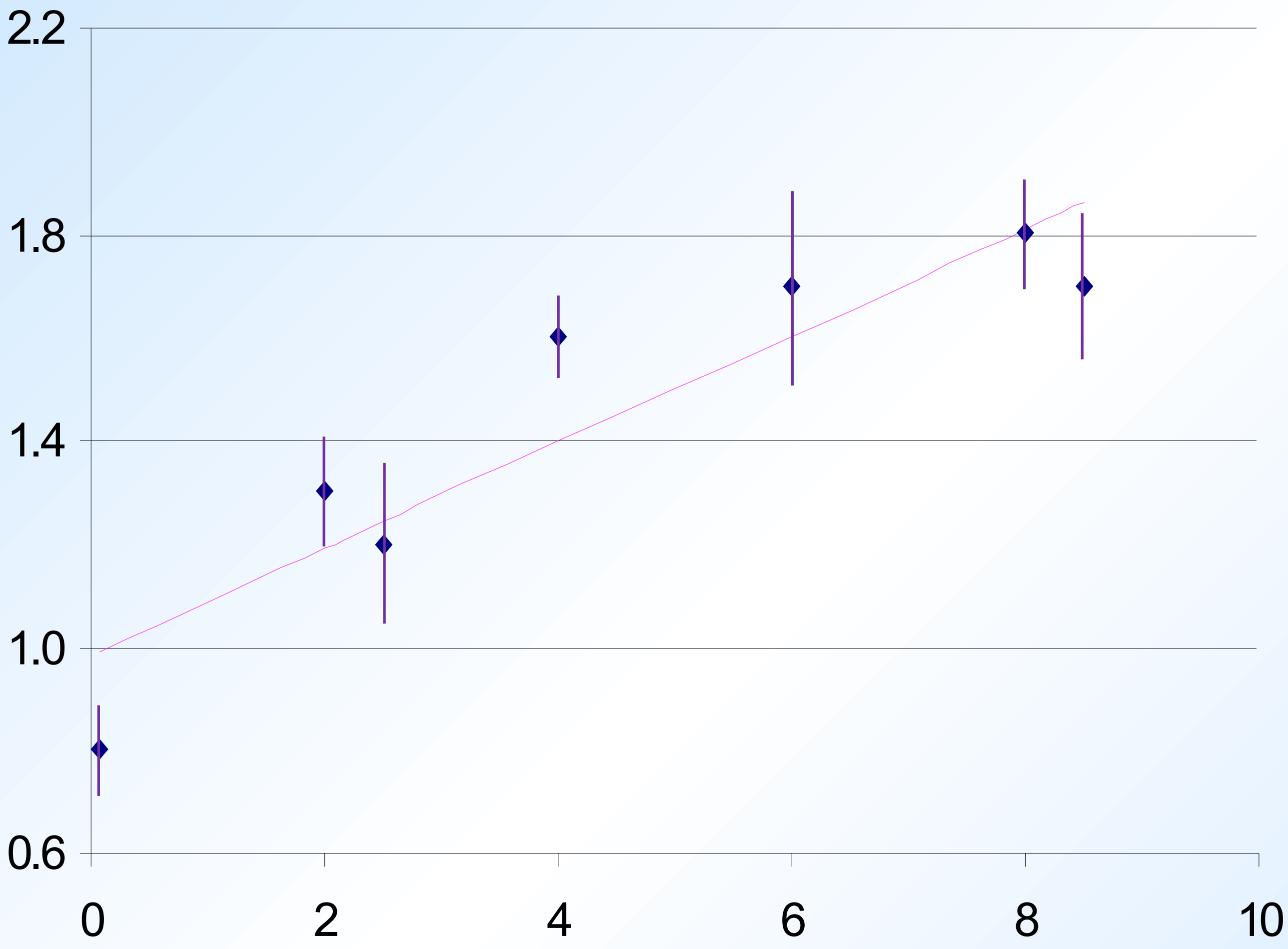
$y(x) = C_1 x + C_2$

Use LSQ

$C_1 = 0.10357402$

$C_2 = 0.98313647$

$S = 0.12706794$



What if Standard deviation at every point, σ_i, is NOT the same.

Minimize what?

Minimize chi square

Minimize the sum of the squares of the errors, each **weighted** by its σ_i

WEIGHTED REGRESSION (LSQ)

Definition of chi square:

$$\chi^2 = \sum_{i=1}^{N} \left[\frac{y_i - y(x_i)}{\sigma_i} \right]^2$$

where σ_i is the standard deviation in y_i is like a weighting factor, w_i.

$$\chi^2 = \sum_{i=1}^{N} \left[w_i \left(y_i - y(x_i) \right) \right]^2$$

Standard deviation in y_i:

$$\sigma_i = \sqrt{\frac{\sum_{i=1}^{n} \left(y_i - \bar{y}_i \right)^2}{n - 1}}$$

COMMENTS

The method, LSQ, is both over-determined and under-determined at the same time.

- Over-determined because there are more data than the number of unknowns.
- Under-determined because ambiguous combinations of parameters may exist, in which case the data may fit equally well to more than one combination of parameters, leading to ill-conditioned set of equations.

WEIGHTED LEAST SQURES

$$y(x) = C_1\, x + C_2$$

$$\chi^2 = \sum_{i=1}^{N} w_i\, E_i^2 = \sum_{i=1}^{N} \frac{1}{\sigma_i^2}\left[y_i - y(x_i)\right]^2 = \sum_{i=1}^{N} \frac{1}{\sigma_i^2}\left[y_i - C_1\, x_i - C_2\right]^2$$

$$\frac{\partial \chi^2}{\partial C_1} = 0 \quad => \quad \frac{\partial \chi^2}{\partial C_1} = \sum_{i=1}^{N} \frac{2}{\sigma_i^2}\left[y_i - C_1\, x_i - C_2\right]\left[- x_i\right] = 0$$

$$\frac{\partial \chi^2}{\partial C_2} = 0 \quad => \quad \frac{\partial \chi^2}{\partial C_2} = \sum_{i=1}^{N} \frac{2}{\sigma_i^2}\left[y_i - C_1\, x_i - C_2\right]\left[- 1\right] = 0$$

$$\sum_{i=1}^{N}\frac{\mathrm{y_i\ x_i}}{\sigma_i^2} = \mathrm{C_1}\sum_{i=1}^{N}\frac{x_i^2}{\sigma_i^2} + \mathrm{C_2}\sum_{i=1}^{N}\frac{\mathrm{x_i}}{\sigma_i^2}$$

$$\sum_{i=1}^{N}\frac{\mathrm{y_i}}{\sigma_i^2} = \mathrm{C_1}\sum_{i=1}^{N}\frac{\mathrm{x_i}}{\sigma_i^2} + \mathrm{C_2}\sum_{i=1}^{N}\frac{1}{\sigma_i^2}$$

$$\begin{pmatrix} \sum_{i=1}^{N}\frac{x_i^2}{\sigma_i^2} & \sum_{i=1}^{N}\frac{\mathrm{x_i}}{\sigma_i^2} \\ \sum_{i=1}^{N}\frac{\mathrm{x_i}}{\sigma_i^2} & \sum_{i=1}^{N}\frac{1}{\sigma_i^2} \end{pmatrix} \begin{pmatrix} \mathrm{C_1} \\ \mathrm{C_2} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N}\frac{\mathrm{y_i\ x_i}}{\sigma_i^2} \\ \sum_{i=1}^{N}\frac{\mathrm{y_i}}{\sigma_i^2} \end{pmatrix}$$

Sometimes, $\sigma_i = \sqrt{\dfrac{\sum_{i=1}^{n}(\mathrm{y_i} - \overline{\mathrm{y}}_\mathrm{i})^2}{\mathrm{n-1}}} \cong \sqrt{\mathrm{y_i}}$

NON-LINEAR REGRESSION

Given $y(x) = y(x, C_1, C_2)$ and data pairs (x_i, y_i), we form

$$\chi^2 = \sum_{i=1}^{N} [y - y(x_i)]^2$$

Assuming $\sigma_i = \sigma$ = constant

$$\frac{\partial \chi^2}{\partial C_j} = 0 \quad \text{for all } j$$

We may end up with a set of non-linear equations

Gauss-Newton Method:

$$\frac{\partial \chi^2}{\partial C_j} = 0 = \sum_{i=1}^{N} 2\,\big(y_i - y(x_i)\big)\left(-\frac{\partial y}{\partial C_j}\right)$$

Expand $y(x, C_1, C_2)$ around $C_{1,0}$ and $C_{2,0}$ in Taylor series:

$$y(x, C_1, C_2) = y(x, C_{1,0}, C_{2,0}) + \Delta C_1 \frac{\partial y}{\partial C_1} + \Delta C_2 \frac{\partial y}{\partial C_2} + \ldots$$

NON-LINEAR REGRESSION

Substitute into χ^2 :

$$\frac{\partial \chi^2}{\partial C_1} = \sum_{i=1}^{N} 2 \left(y_i - y(x, C_{1,0}, C_{2,0}) - \Delta C_1 \frac{\partial y}{\partial C_1} - \Delta C_2 \frac{\partial y}{\partial C_2} \right) \left(- \frac{\partial y}{\partial C_1} \right) = 0$$

$$\frac{\partial \chi^2}{\partial C_2} = \sum_{i=1}^{N} 2 \left(y_i - y(x, C_{1,0}, C_{2,0}) - \Delta C_1 \frac{\partial y}{\partial C_1} - \Delta C_2 \frac{\partial y}{\partial C_2} \right) \left(- \frac{\partial y}{\partial C_2} \right) = 0$$

$$\sum_{i=1}^{N} y_i \frac{\partial y}{\partial C_1} - \sum_{i=1}^{N} y(x_i) \frac{\partial y}{\partial C_1} = \Delta C_1 \sum_{i=1}^{N} \left(\frac{\partial y}{\partial C_1} \right)^2 + \Delta C_2 \sum_{i=1}^{N} \left(\frac{\partial y}{\partial C_1} \right) \left(\frac{\partial y}{\partial C_2} \right)$$

$$\sum_{i=1}^{N} y_i \frac{\partial y}{\partial C_1} - \sum_{i=1}^{N} y(x_i) \frac{\partial y}{\partial C_1} = \Delta C_1 \sum_{i=1}^{N} \left(\frac{\partial y}{\partial C_1} \right) \left(\frac{\partial y}{\partial C_2} \right) + \Delta C_2 \sum_{i=1}^{N} \left(\frac{\partial y}{\partial C_2} \right)^2$$

NON-LINEAR REGRESSION

In matrix notation:

$$\begin{pmatrix} \sum_{i=1}^{N}\left(\frac{\partial y}{\partial C_1}\right)^2 & \sum_{i=1}^{N}\left(\frac{\partial y}{\partial C_1}\right)\left(\frac{\partial y}{\partial C_2}\right) \\ \sum_{i=1}^{N}\left(\frac{\partial y}{\partial C_1}\right)\left(\frac{\partial y}{\partial C_2}\right) & \sum_{i=1}^{N}\left(\frac{\partial y}{\partial C_2}\right)^2 \end{pmatrix} \begin{pmatrix} \Delta C_1 \\ \Delta C_2 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} y_i \frac{\partial y}{\partial C_1} - \sum_{i=1}^{N} y(x_i) \frac{\partial y}{\partial C_1} \\ \sum_{i=1}^{N} y_i \frac{\partial y}{\partial C_2} - \sum_{i=1}^{N} y(x_i) \frac{\partial y}{\partial C_2} \end{pmatrix}$$

$$\Delta C_1 = C_{1,1} - C_{1,0} \quad => \quad C_{1,1} = C_{1,0} + \Delta C_1$$

$$\Delta C_2 = C_{2,1} - C_{2,0} \quad => \quad C_{2,1} = C_{2,0} + \Delta C_2$$

NON-LINEAR REGRESSION

Example: $y(x) = C_1 \, (1 - e^{-C_2 x})$

x	0.25	0.75	1.25	1.75	2.25
y	0.28	0.57	0.68	0.74	0.79

$$\frac{\partial y}{\partial C_1} = 1 - e^{-C_2 x} \qquad \frac{\partial y}{\partial C_2} = C_1 \, x \, e^{-C_2 x}$$

Choose $C_{1,0} = 1$, $C_{2,0} = 1$ =>

$$\begin{pmatrix} 2.31936 & 0.94893 \\ 0.94893 & 0.440395 \end{pmatrix} \begin{pmatrix} \Delta C_1 \\ \Delta C_2 \end{pmatrix} = \begin{pmatrix} -0.1533 \\ -0.0365 \end{pmatrix}$$

Find $C_{1,1} = C_{1,0} + \Delta C_1$, $C_{2,1} = C_{2,0} + \Delta C_2$ => Find the new matrix equation

Find the new ΔC_1 and ΔC_2

Final answer: $C_{1,n} = 0.79186$ $C_{2,n} = 1.67510$

$y(x) = C_1 \exp(C_2 x)$

Linear LSQ:

$C_1 = 0.97916877$

$C_2 = 0.08007557$

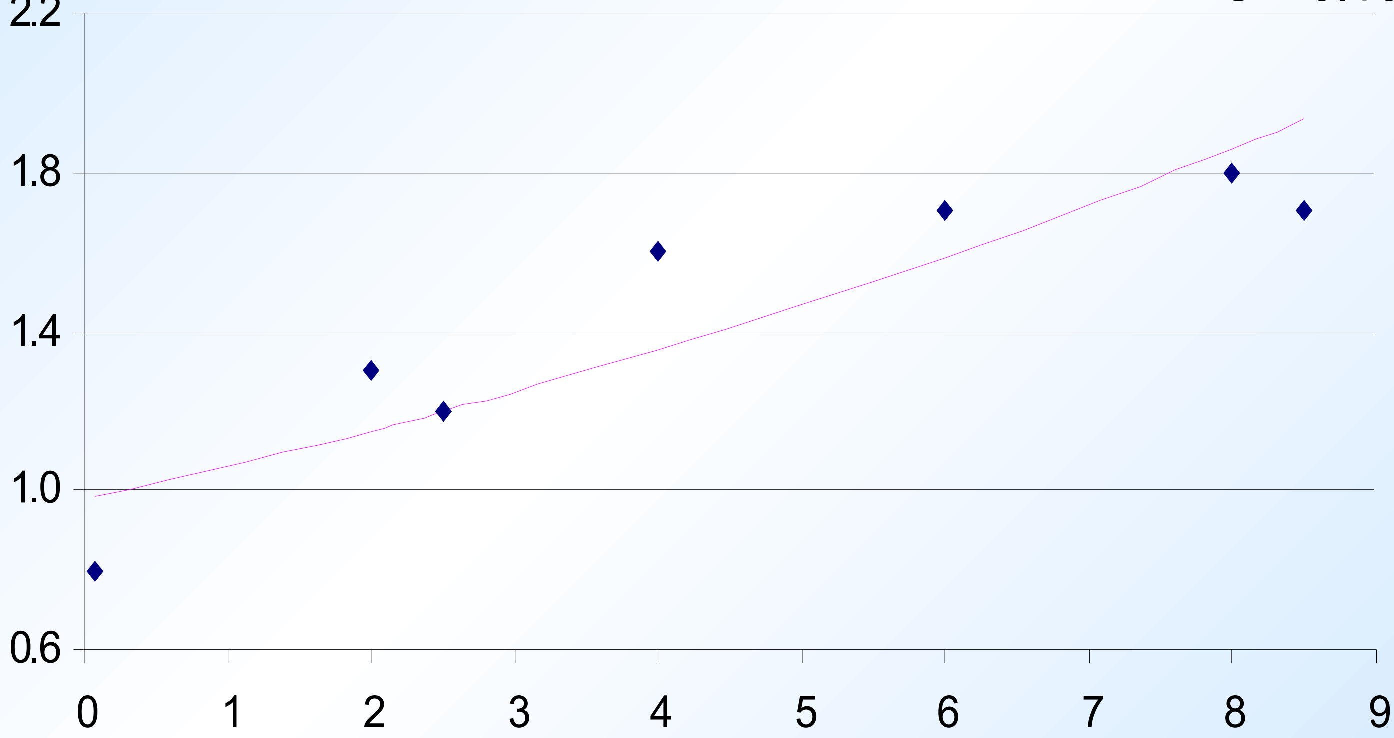
$S = 0.19171593$

Non-linear LSQ

$C_1 = 1.053163159$

$C_2 = 0.066891263$

$S = 0.16947360$



$y(x) = C_1 \ln(x) + C_2$

Linear LSQ:	Non-linear LSQ
$C_1 = 0.20151129$	$C_1 =$
$C_2 = 1.26012269$	$C_2 =$
$S = 0.10016157$	$S =$