NUMERICAL METHODS
Week-8,9
02.04.2013
09.04.2013

Regression Analysis

Asst. Prof. Dr. Berk Canberk

Regression Analysis

- **What** is a regression?
- **Why** do we use regression analysis?
- **How** do we implement regression?

What is a Regression?

- It is a statistical technique to
  - MODEL
  - ESTIMATE

  the relationships among the variables.

What is a Regression?

Given *n* data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ best fit $y = f(x)$ to the data. The best fit is generally based on minimizing the sum of the square of the residuals, S_r.

Residual at a point is

$$\varepsilon_i = y_i - f(x_i)$$

Sum of the square of the residuals

$$S_r = \sum_{i=1}^{n} (y_i - f(x_i))^2$$

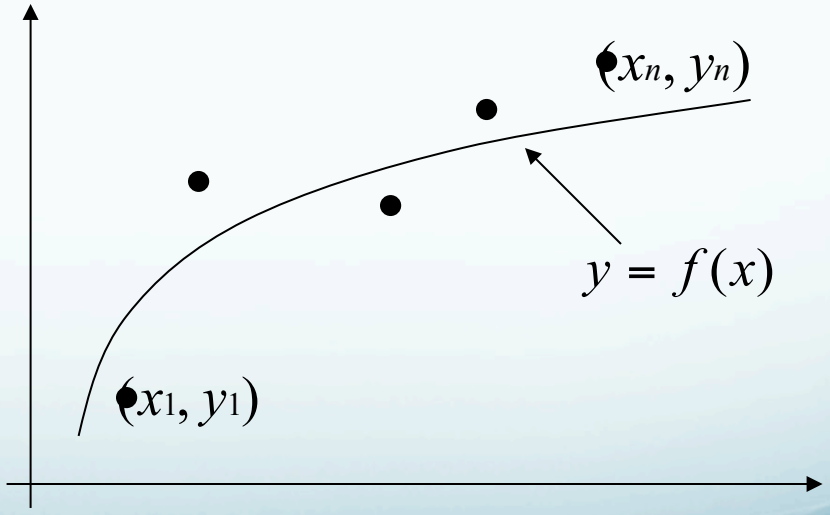


Figure. Basic model for regression

Why to Use Regression Analysis?

- Two Main purposes:
  - To understand the relationship between variables.
  - To predict the value of one based on the other.

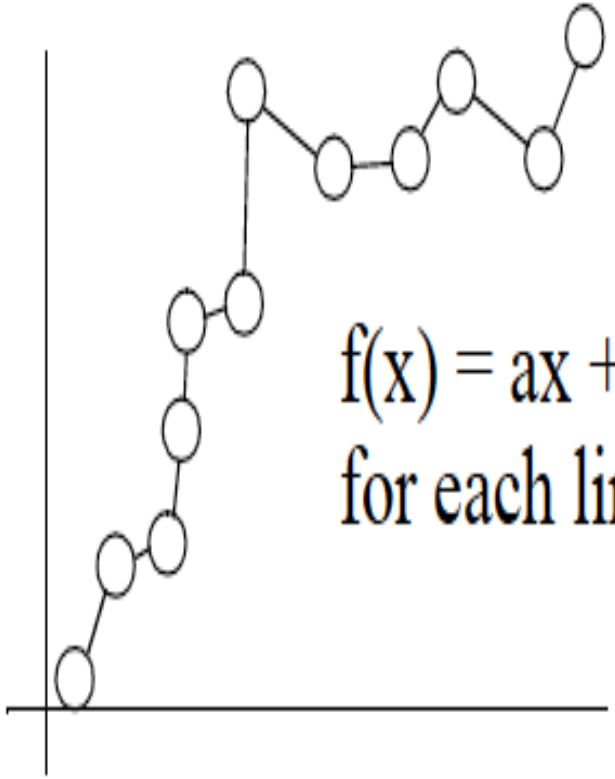
- **Analytical models of phenomena..**
- **Create an equation from the observed data..**
- **Predict the behavior of phenomena..**

Why to Use Regression Analysis?

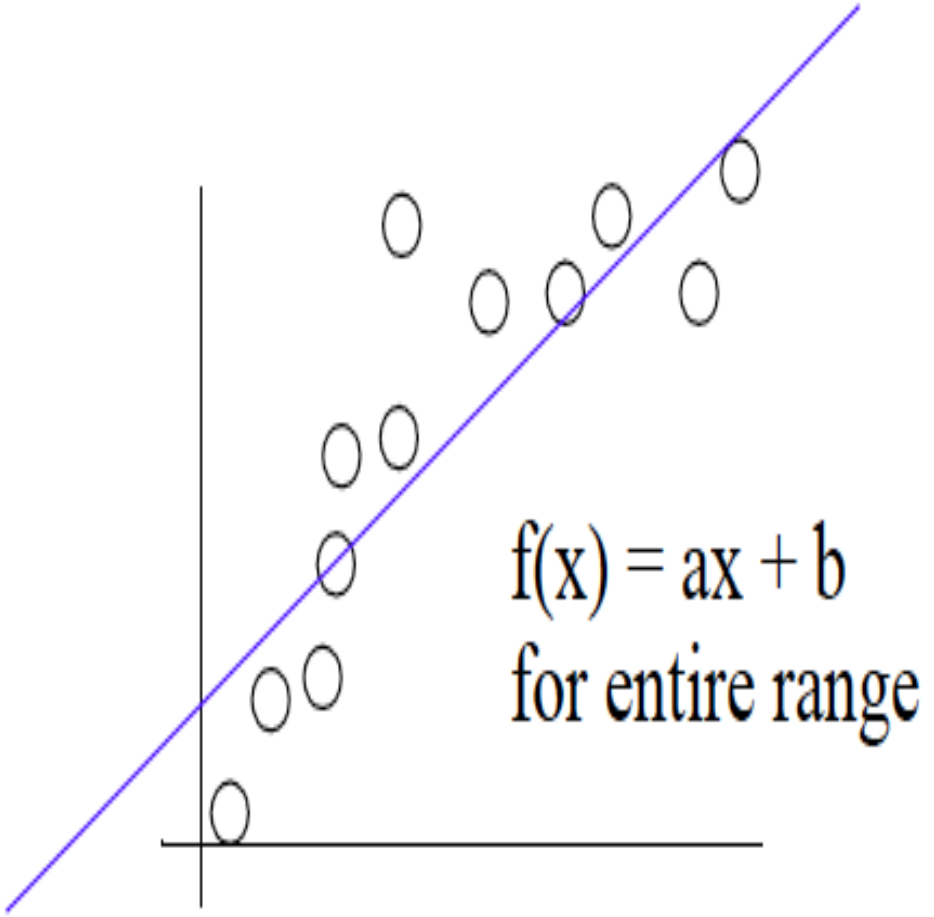
- Some examples:

To model and estimate

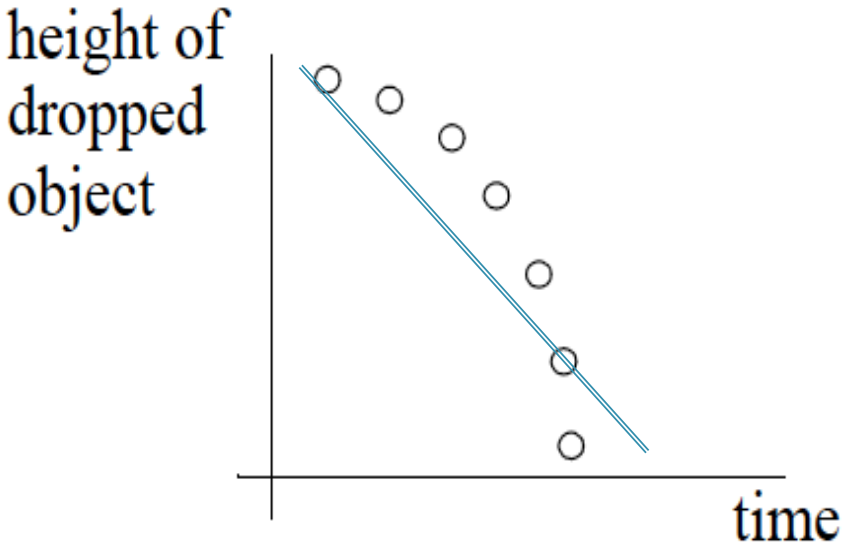
- The relationship between the level of education and the incomes in a population..
- The price of a house and its square footage..
- The sales volumes of a company relative to the money spent on advertising..
- The collected data related to the database (memory, cache etc) size..
- The age vs height of people in a country..
- The internet usage time relative to downloaded data..
- ...

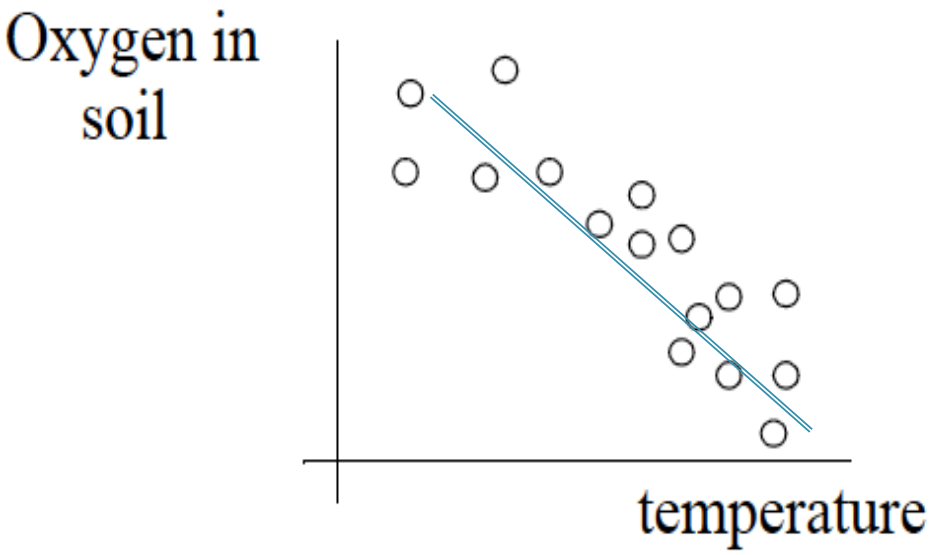


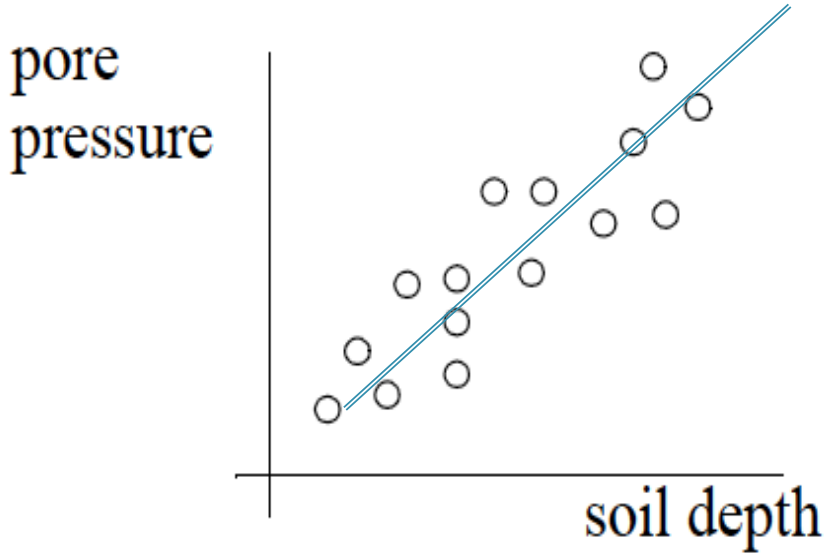
Interpolation

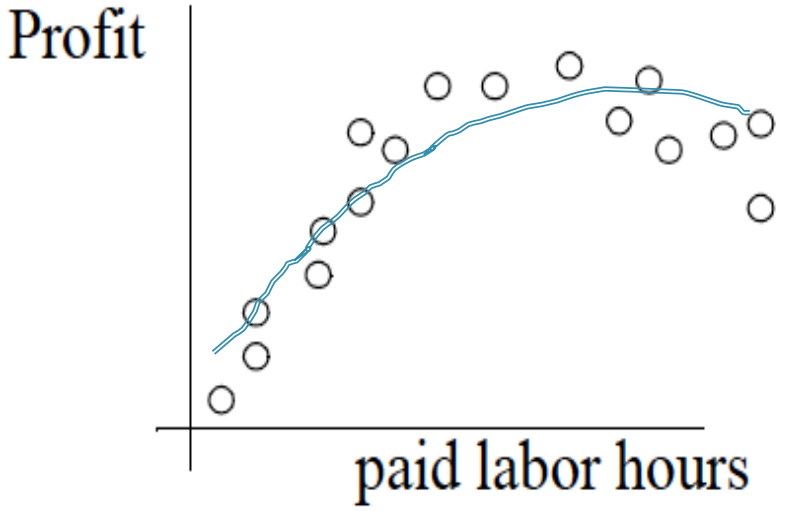


Curve Fitting


height of dropped object
time


Oxygen in soil
temperature


pore pressure
soil depth


Profit
paid labor hours

How to implement regression?

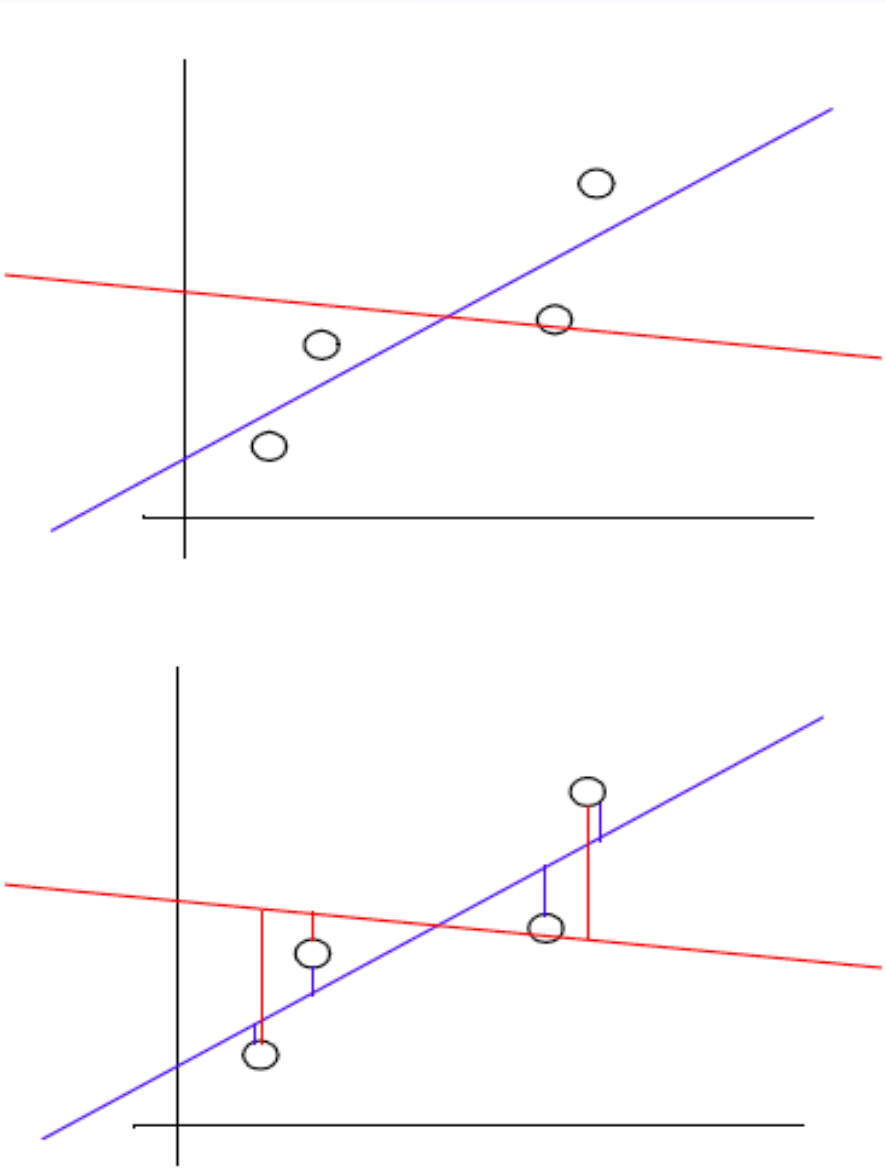
- Linear Regression

Ex: $y=ax+b$

- Nonlinear Regression

Ex: $y=ae^{bx}$

Linear Regression



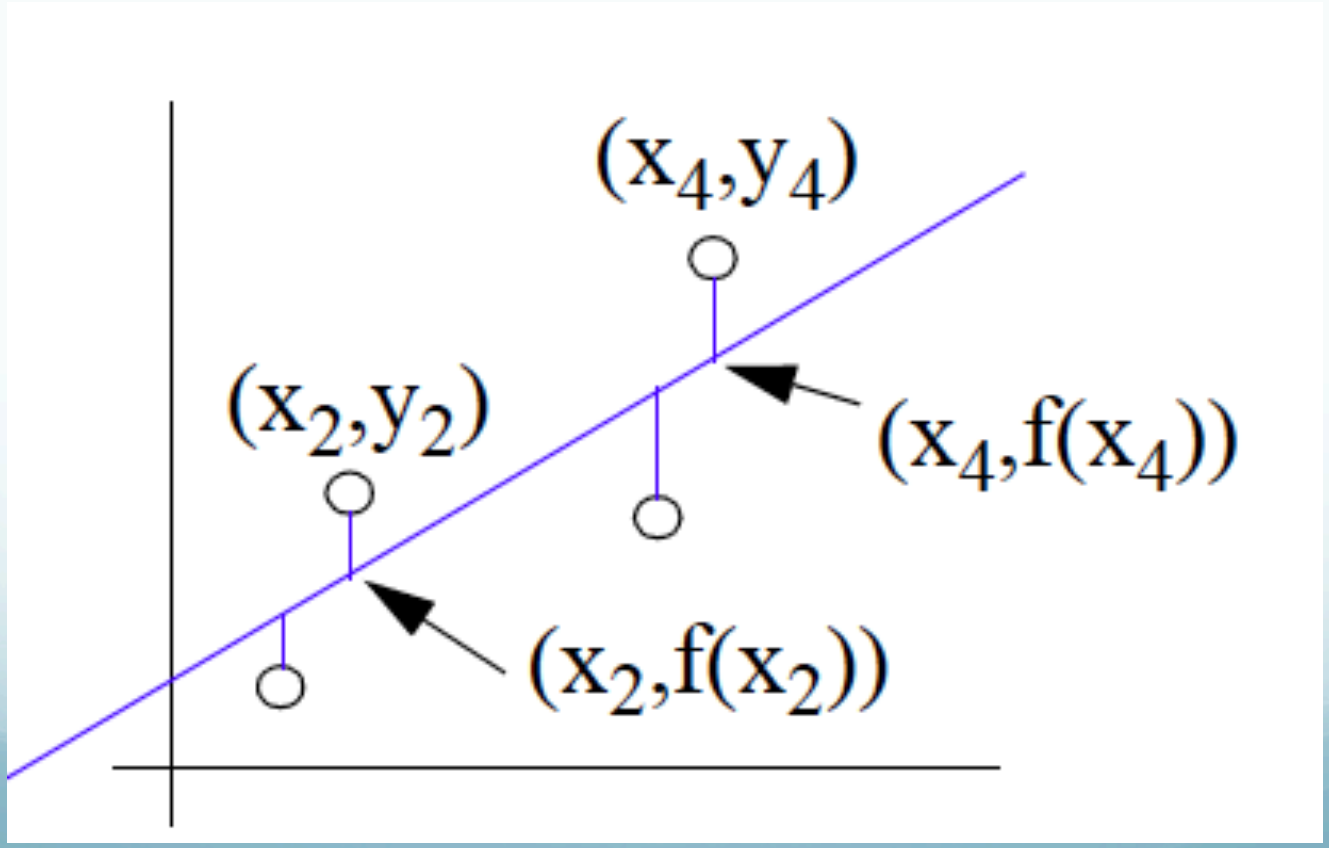
General Form of a Straight line:

f(x)=ax+b

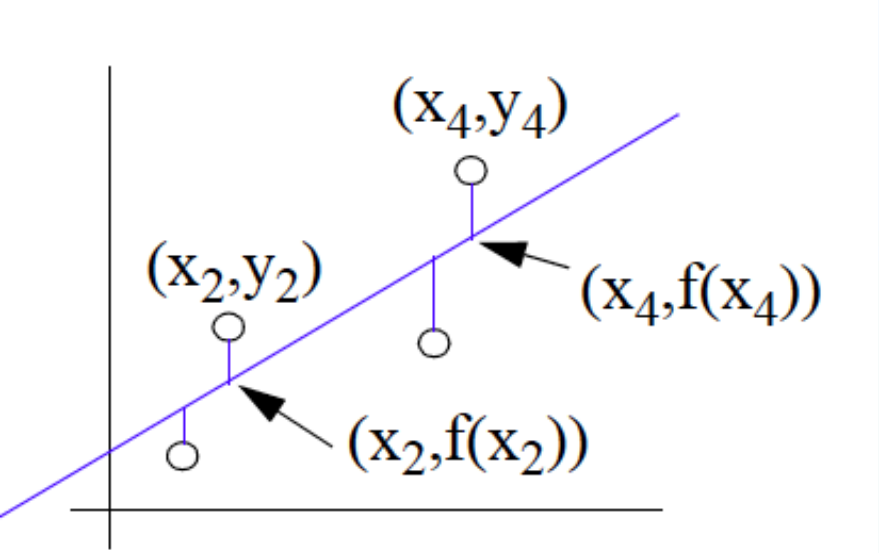
How can we pick the coefficients that best fits the line to the data??

How to Determine the Coefficients?

- Using an estimator:
  - **Least Square Estimator..**



Least Square Approach



$$err = \sum (d_i)^2 = (y_1 - f(x_1))^2 + (y_2 - f(x_2))^2 + (y_3 - f(x_3))^2 + (y_4 - f(x_4))^2$$

Least Square Approach

Our fit is f(x)=ax+b, so we substitute..

$$err = \sum_{i=1}^{\#\text{ data points}} (y_i - f(x_i))^2 = \sum_{i=1}^{\#\text{ data points}} (y_i - (ax_i + b))^2$$

The **'best'** line has minimum error between line and data points.

This is called **the least squares approach, since we minimize the square of the error.**

Least Square Approach

$$\text{minimize} \quad err = \sum_{i=1}^{\#\text{ data points} = n} (y_i - (ax_i + b))^2$$

How to find the minimum of a function??

1) derivative describes the slope

2) slope = zero is a minimum

Least Square Approach

Take the derivative of the error with respect to a and b, and set each to zero

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

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

Least Square Approach

Solve for the a and b so that the previous two equations both = 0.

Re-write these two equations:

$$a\sum x_i^2 + b\sum x_i = \sum (x_i y_i)$$
$$a\sum x_i + b*n = \sum y_i$$

Put them in MATRIX FORM:

$$\begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum (x_i y_i) \end{bmatrix}$$

Least Square Approach

$$A = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}, \quad X = \begin{bmatrix} b \\ a \end{bmatrix}, \quad B = \begin{bmatrix} \sum y_i \\ \sum (x_i y_i) \end{bmatrix}$$

So, $AX = B$ ⟹ $X = (A^T A)^{-1} A^T B$

For linear regression, the coefficients are:

$$a = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

$$b = \bar{y} - a\bar{x}$$

Example 1: Without Noise..

Linear Regression (First Order) Model of the data?

i	1	2	3	4	5	6
x	0	0.5	1.0	1.5	2.0	2.5
y	0	1.5	3.0	4.5	6.0	7.5

$$n = 6$$

$$\sum x_i = 7.5, \quad \sum y_i = 22.5, \quad \sum x_i^2 = 13.75, \quad \sum x_i y_i = 41.25$$

$$\begin{bmatrix} 6 & 7.5 \\ 7.5 & 13.75 \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} 22.5 \\ 41.25 \end{bmatrix}$$

Note: we are using $\sum x_i^2$, NOT $\left(\sum x_i\right)^2$

$$\begin{bmatrix} b \\ a \end{bmatrix} = inv \begin{bmatrix} 6 & 7.5 \\ 7.5 & 13.75 \end{bmatrix} * \begin{bmatrix} 22.5 \\ 41.25 \end{bmatrix}$$ or use Gaussian elimination...

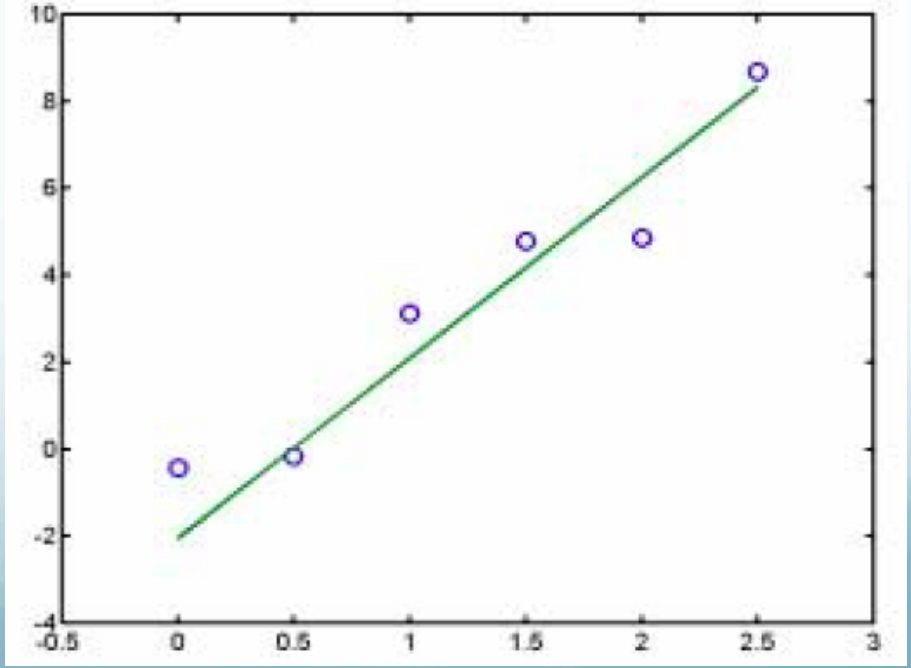
The solution is $\begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} 0 \\ 3 \end{bmatrix} \Longrightarrow f(x) = 3x + 0$

Example 2: With Noise..

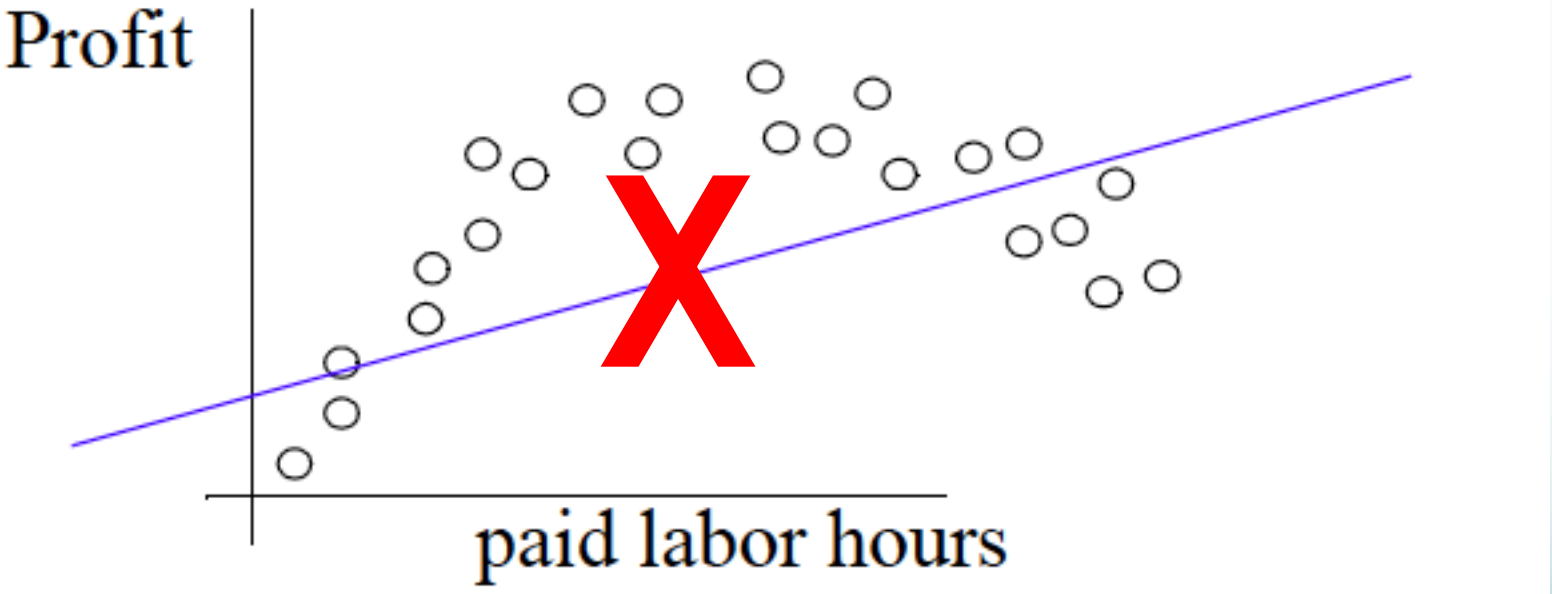
x = [0 .5 1 1.5 2 2.5], y = [-0.4326 -0.1656 3.1253 4.7877 4.8535 8.6909]

$$\begin{bmatrix} 6 & 7.5 \\ 7.5 & 13.75 \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} 20.8593 \\ 41.6584 \end{bmatrix}, \quad \begin{bmatrix} b \\ a \end{bmatrix} = inv \begin{bmatrix} 6 & 7.5 \\ 7.5 & 13.75 \end{bmatrix} * \begin{bmatrix} 20.8593 \\ 41.6584 \end{bmatrix}, \quad \begin{bmatrix} b \\ a \end{bmatrix} = \begin{bmatrix} -0.975 \\ 3.561 \end{bmatrix}$$

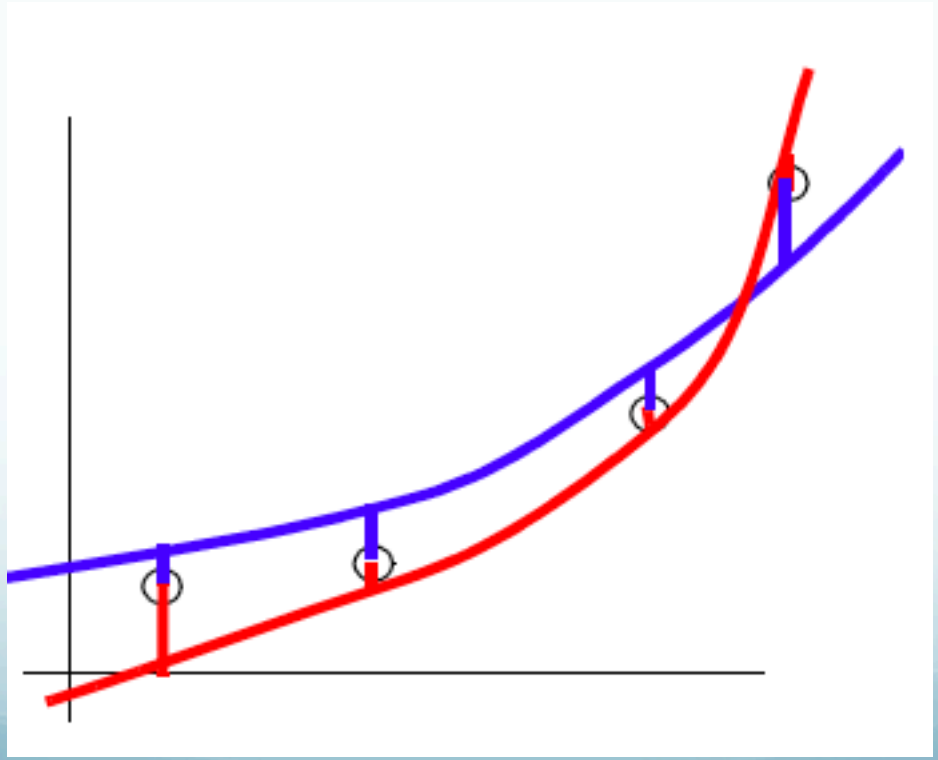
so our fit is $f(x) = 3.561\,x - 0.975$



Higher Order Polynomial Regression



$$f(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \ldots + a_j x^j = a_0 + \sum_{k=1}^{j} a_k x^k$$



Least Square Error Approach

$$err = \sum (d_i)^2 = (y_1 - f(x_1))^2 + (y_2 - f(x_2))^2 + (y_3 - f(x_3))^2 + (y_4 - f(x_4))^2$$

$$err = \sum_{i=1}^{n} \left(y_i - \left(a_0 + a_1 x_i + a_2 x_i^2 + a_3 x_i^3 + \ldots + a_j x_i^j \right) \right)^2$$

$$err = \sum_{i=1}^{n} \left(y_i - \left(a_0 + \sum_{k=1}^{j} a_k x^k \right) \right)^2$$

$$\frac{\partial err}{\partial a_0} = -2 \sum_{i=1}^{n} \left(y_i - \left(a_0 + \sum_{k=1}^{j} a_k x^k \right) \right) = 0$$

$$\frac{\partial err}{\partial a_1} = -2 \sum_{i=1}^{n} \left(y_i - \left(a_0 + \sum_{k=1}^{j} a_k x^k \right) \right) x = 0$$

$$\frac{\partial err}{\partial a_2} = -2 \sum_{i=1}^{n} \left(y_i - \left(a_0 + \sum_{k=1}^{j} a_k x^k \right) \right) x^2 = 0$$

$$\vdots$$

$$\vdots$$

$$\frac{\partial err}{\partial a_j} = -2 \sum_{i=1}^{n} \left(y_i - \left(a_0 + \sum_{k=1}^{j} a_k x^k \right) \right) x^j = 0$$

$$\begin{bmatrix} n & \sum x_i & \sum x_i^2 & \cdots & \sum x_i^j \\ \sum x_i & \sum x_i^2 & \sum x_i^3 & \cdots & \sum x_i^{j+1} \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 & \cdots & \sum x_i^{j+2} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum x_i^j & \sum x_i^{j+1} & \sum x_i^{j+2} & \cdots & \sum x_i^{j+j} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_j \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum (x_i y_i) \\ \sum \left(x_i^2 y_i \right) \\ \vdots \\ \sum \left(x_i^j y_i \right) \end{bmatrix}$$

$$A = \begin{bmatrix} n & \sum x_i & \sum x_i^2 & \dots & \sum x_i^j \\ \sum x_i & \sum x_i^2 & \sum x_i^3 & \dots & \sum x_i^{j+1} \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 & \dots & \sum x_i^{j+2} \\ \vdots & \vdots & \vdots & & \vdots \\ \sum x_i^j & \sum x_i^{j+1} & \sum x_i^{j+2} & \dots & \sum x_i^{j+j} \end{bmatrix}, \quad X = \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_j \end{bmatrix}, \quad B = \begin{bmatrix} \sum y_i \\ \sum (x_i y_i) \\ \sum \left(x_i^2 y_i \right) \\ \vdots \\ \sum \left(x_i^j y_i \right) \end{bmatrix}$$

$$AX = B \quad \Rightarrow \quad \mathtt{X = A^{-1}*B}$$

Example 1: without noise..

Second order regression model of the data?

i	1	2	3	4	5	6
x	0	0.5	1.0	1.5	2.0	2.5
y	0	0.25	1.0	2.25	4.0	6.25

j=2;

$$\begin{bmatrix} n & \sum x_i & \sum x_i^2 \\ \sum x_i & \sum x_i^2 & \sum x_i^3 \\ \sum x_i^2 & \sum x_i^3 & \sum x_i^4 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \\ \sum x_i^2 y_i \end{bmatrix}$$

$n = 6$

$\sum x_i = 7.5,$ $\sum y_i = 13.75$

$\sum x_i^2 = 13.75,$ $\sum x_i y_i = 28.125$

$\sum x_i^3 = 28.125$ $\sum x_i^2 y_i = 61.1875$

$\sum x_i^4 = 61.1875$

$$\begin{bmatrix} 6 & 7.5 & 13.75 \\ 7.5 & 13.75 & 28.125 \\ 13.75 & 28.125 & 61.1875 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 13.75 \\ 28.125 \\ 61.1875 \end{bmatrix}$$

Note: we are using $\sum x_i^2$, NOT $(\sum x_i)^2$. There's a big difference

$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = inv \begin{bmatrix} 6 & 7.5 & 13.75 \\ 7.5 & 13.75 & 28.125 \\ 13.75 & 28.125 & 61.1875 \end{bmatrix} * \begin{bmatrix} 13.75 \\ 28.125 \\ 61.1857 \end{bmatrix}$$

$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \Longrightarrow \quad f(x) = 0 + 0*x + 1*x^2$$

Example 2: with noise..

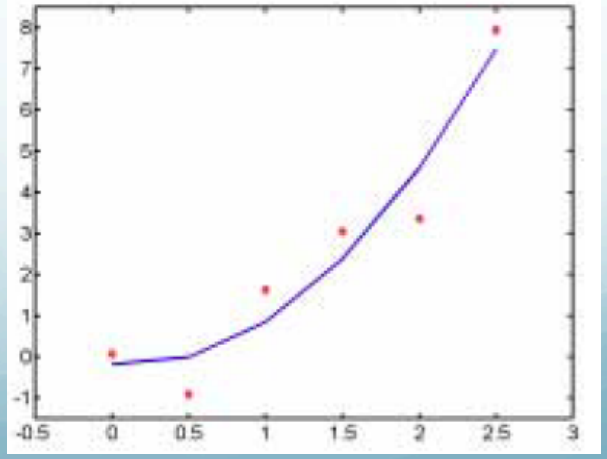
Second order regression model of the data?

$$x = [0 \quad .0 \quad 1 \quad 1.5 \quad 2 \quad 2.5]$$
$$y = [0.0674 \quad -0.9156 \quad 1.6253 \quad 3.0377 \quad 3.3535 \quad 7.9409]$$

$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = inv \begin{bmatrix} 6 & 7.5 & 13.75 \\ 7.5 & 13.75 & 28.125 \\ 13.75 & 28.125 & 61.1875 \end{bmatrix} * \begin{bmatrix} 15.1093 \\ 32.2834 \\ 71.276 \end{bmatrix}$$

$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} -0.1812 \\ -0.3221 \\ 1.3537 \end{bmatrix}$$

$$f(x) = -0.1812 - 0.3221x^* + 1.3537 * x^2$$



Regression Diagnostics

- After determining the parameters of the model, We have to analyze how good and accurate our model is, how does our regression equation fit the data?
- This parameterization is called “goodness of fit”.
- **The coefficient of Determination, R^2,** is one of the common goodness of fit for regression models.
- **R^2 is between 0-1:**

  R^2=1 means that the regression model represents %100 the data.

  R^2=0 means that the regression model represents %0 the data.

Coefficient of Determination

- Some Definitions:

Sum of Squared Errors: $SS_{\mathrm{err}} = \sum (y_i - f_i)^2$

Sum of Squares Total: $SS_{\mathrm{tot}} = \sum_i (y_i - \bar{y})^2,$

$$R^2 = 1 - \frac{SSerr}{SStot} = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2}$$

Correlation Coefficient

$$r = \pm\sqrt{R^2} = \pm\sqrt{1 - \frac{SSerr}{SStot}} = \pm\sqrt{1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2}}$$

The sign of the correlation coefficient is determined by the sign of the slope, i.e. the coefficient of "x".

Some Nonlinear Regression Models

Some popular nonlinear regression models:

1. Exponential model: $(y = ae^{bx})$
2. Power model: $(y = ax^{b})$
3. Saturation growth model: $\left(y = \frac{ax}{b + x}\right)$
4. Polynomial model: $(y = a_0 + a_1x + \ldots + a_mx^{m})$

Nonlinear Regression

Given *n* data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ best fit $y = f(x)$ to the data, where $f(x)$ is a nonlinear function of x .

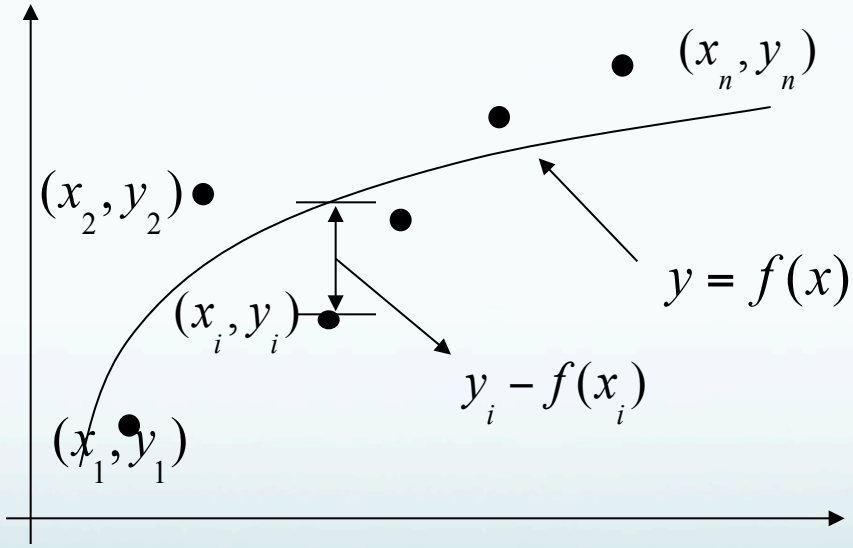


Figure. Nonlinear regression model for discrete y vs. x data

Exponential Model

Given $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ best fit $y = ae^{bx}$ to the data.

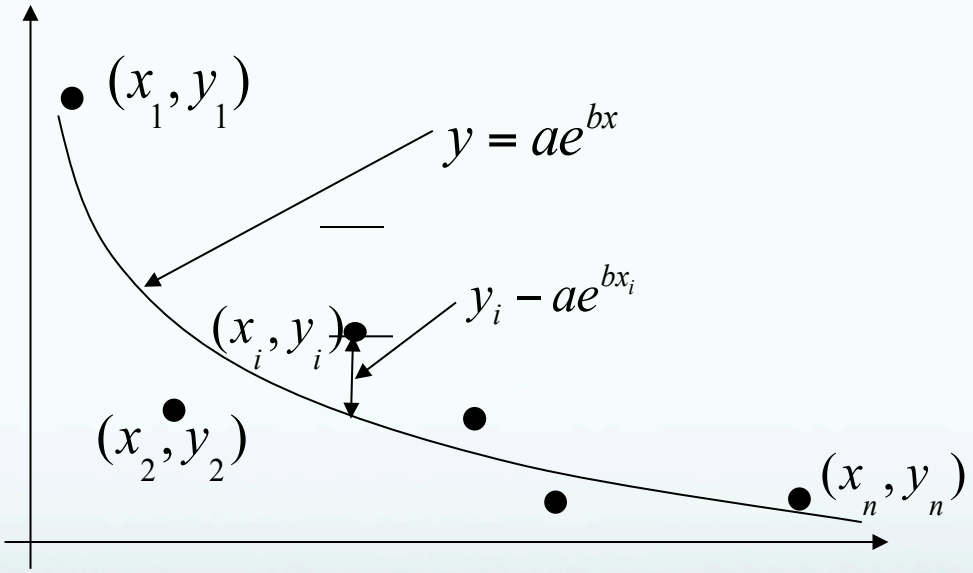


Figure. Exponential model of nonlinear regression for y vs. x data

Finding Constants of Exponential Model

The sum of the square of the residuals is defined as

$$S_r = \sum_{i=1}^{n} \left(y_i - ae^{bx_i} \right)^2$$

Differentiate with respect to *a* and *b*

$$\frac{\partial S_r}{\partial a} = \sum_{i=1}^{n} 2\left(y_i - ae^{bx_i} \right)\left(- e^{bx_i} \right) = 0$$

$$\frac{\partial S_r}{\partial b} = \sum_{i=1}^{n} 2\left(y_i - ae^{bx_i} \right)\left(- ax_i e^{bx_i} \right) = 0$$

Finding Constants of Exponential Model

Rewriting the equations, we obtain

$$-\sum_{i=1}^{n} y_i e^{bx_i} + a\sum_{i=1}^{n} e^{2bx_i} = 0$$

$$\sum_{i=1}^{n} y_i x_i e^{bx_i} - a\sum_{i=1}^{n} x_i e^{2bx_i} = 0$$

Finding constants of Exponential Model

Solving the first equation for a yields

$$a = \frac{\sum_{i=1}^{n} y_i e^{bx_i}}{\sum_{i=1}^{n} e^{2bx_i}}$$

Substituting a back into the previous equation

$$\sum_{i=1}^{n} y_i x_i e^{bx_i} - \frac{\sum_{i=1}^{n} y_i e^{bx_i}}{\sum_{i=1}^{n} e^{2bx_i}} \sum_{i=1}^{n} x_i e^{2bx_i} = 0$$

The constant b can be found through numerical methods such as bisection method.

Example-Exponential Model

Modeling and estimation of the dangerous radiation for a Radioactive Material

Table. Relative intensity of radiation as a function of time.

t(hrs)	0	1	3	5	7	9
γ	1.000	0.891	0.708	0.562	0.447	0.355

Example-Exponential Model.

The relative intensity is related to time by the equation

$$\gamma = Ae^{\lambda t}$$

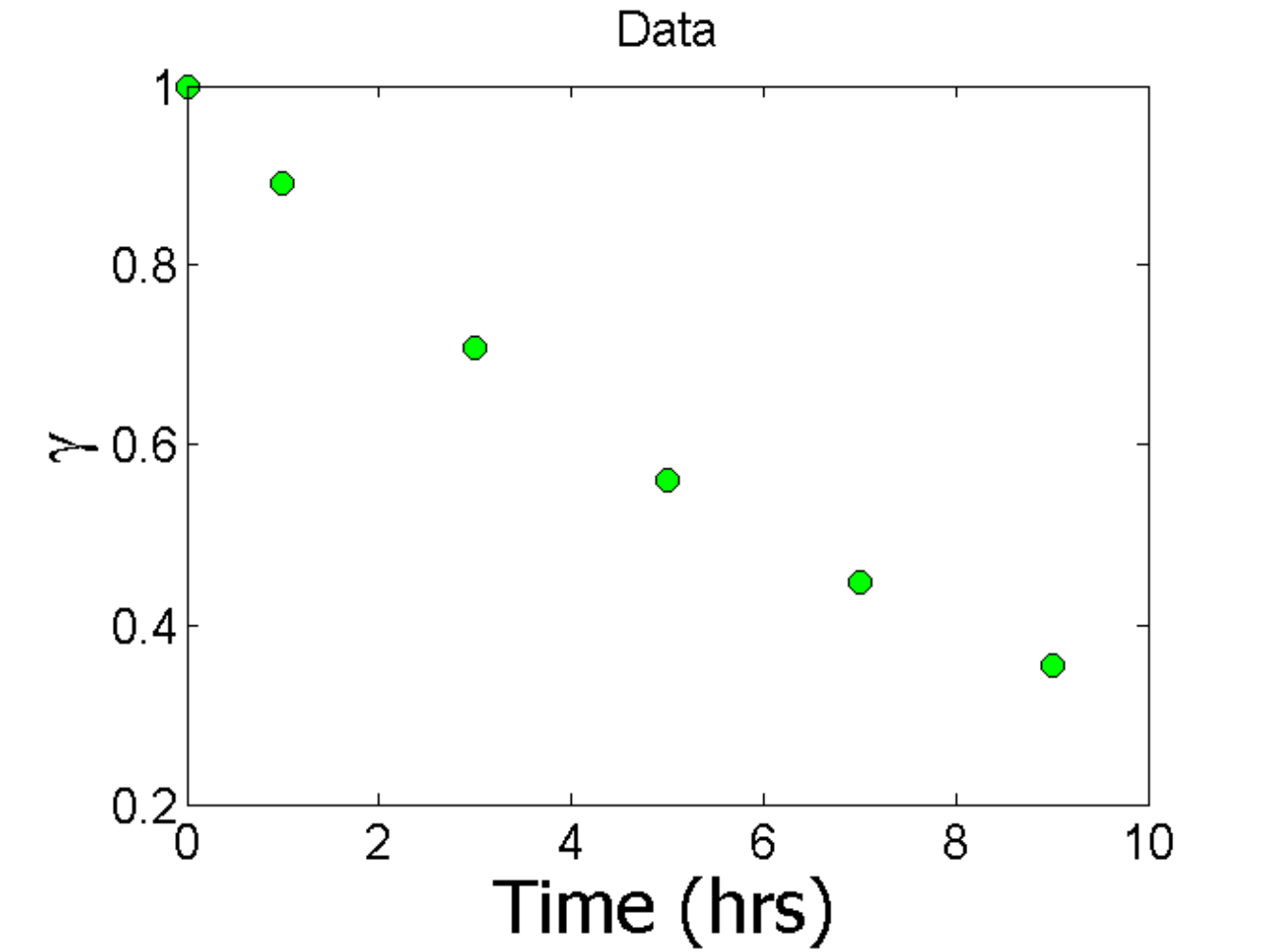
Find:

a) The value of the regression constants A and λ

b) The half-life of this material

c) Radiation intensity after 24 hours

Plot of data



Constants of the Model

$$\gamma = Ae^{\lambda t}$$

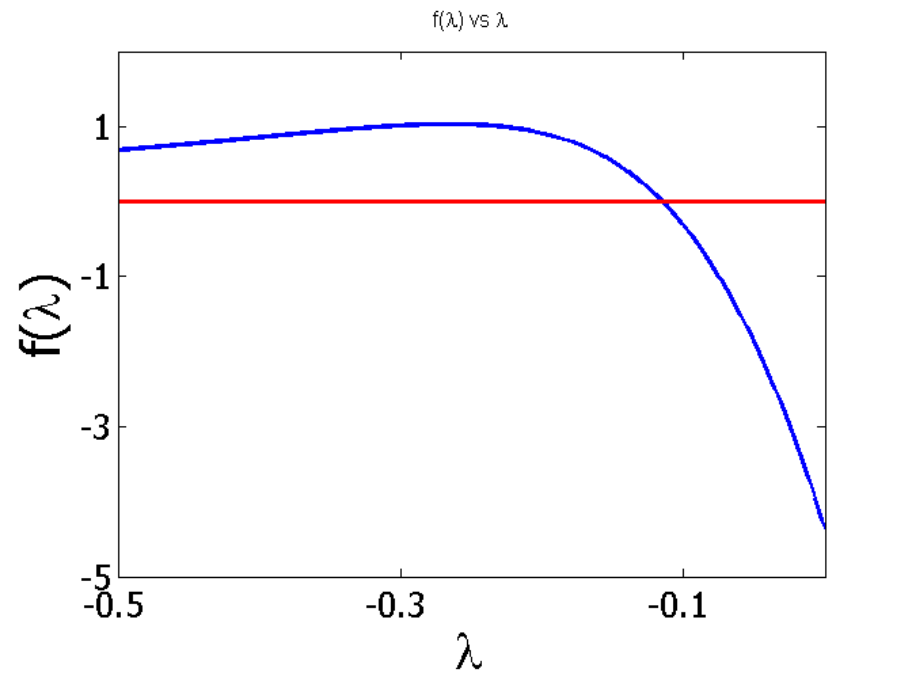
The value of λ is found by solving the nonlinear equation

$$f(\lambda) = \sum_{i=1}^{n} \gamma_i t_i e^{\lambda t_i} - \frac{\sum_{i=1}^{n} \gamma_i e^{\lambda t_i}}{\sum_{i=1}^{n} e^{2\lambda t_i}} \sum_{i=1}^{n} t_i e^{2\lambda t_i} = 0$$

$$A = \frac{\sum_{i=1}^{n} \gamma_i e^{\lambda t_i}}{\sum_{i=1}^{n} e^{2\lambda t_i}}$$

Setting up the Equation in MATLAB

$$f(\lambda) = \sum_{i=1}^{n} \gamma_i t_i e^{\lambda t_i} - \frac{\sum_{i=1}^{n} \gamma_i e^{\lambda t_i}}{\sum_{i=1}^{n} e^{2\lambda t_i}} \sum_{i=1}^{n} t_i e^{2\lambda t_i} = 0$$



t (hrs)	0	1	3	5	7	9
γ	1.000	0.891	0.708	0.562	0.447	0.355

Setting up the Equation in MATLAB

$$f(\lambda) = \sum_{i=1}^{n} \gamma_i t_i e^{\lambda t_i} - \frac{\sum_{i=1}^{n} \gamma_i e^{\lambda t_i}}{\sum_{i=1}^{n} e^{2\lambda t_i}} \sum_{i=1}^{n} t_i e^{2\lambda t_i} = 0$$

$$\lambda = -0.1151$$

```
t=[0 1  3  5  7  9]
gamma=[1  0.891  0.708  0.562  0.447  0.355]
syms lamda
sum1=sum(gamma.*t.*exp(lamda*t));
sum2=sum(gamma.*exp(lamda*t));
sum3=sum(exp(2*lamda*t));
sum4=sum(t.*exp(2*lamda*t));
f=sum1-sum2/sum3*sum4;
```

Calculating the Other Constant

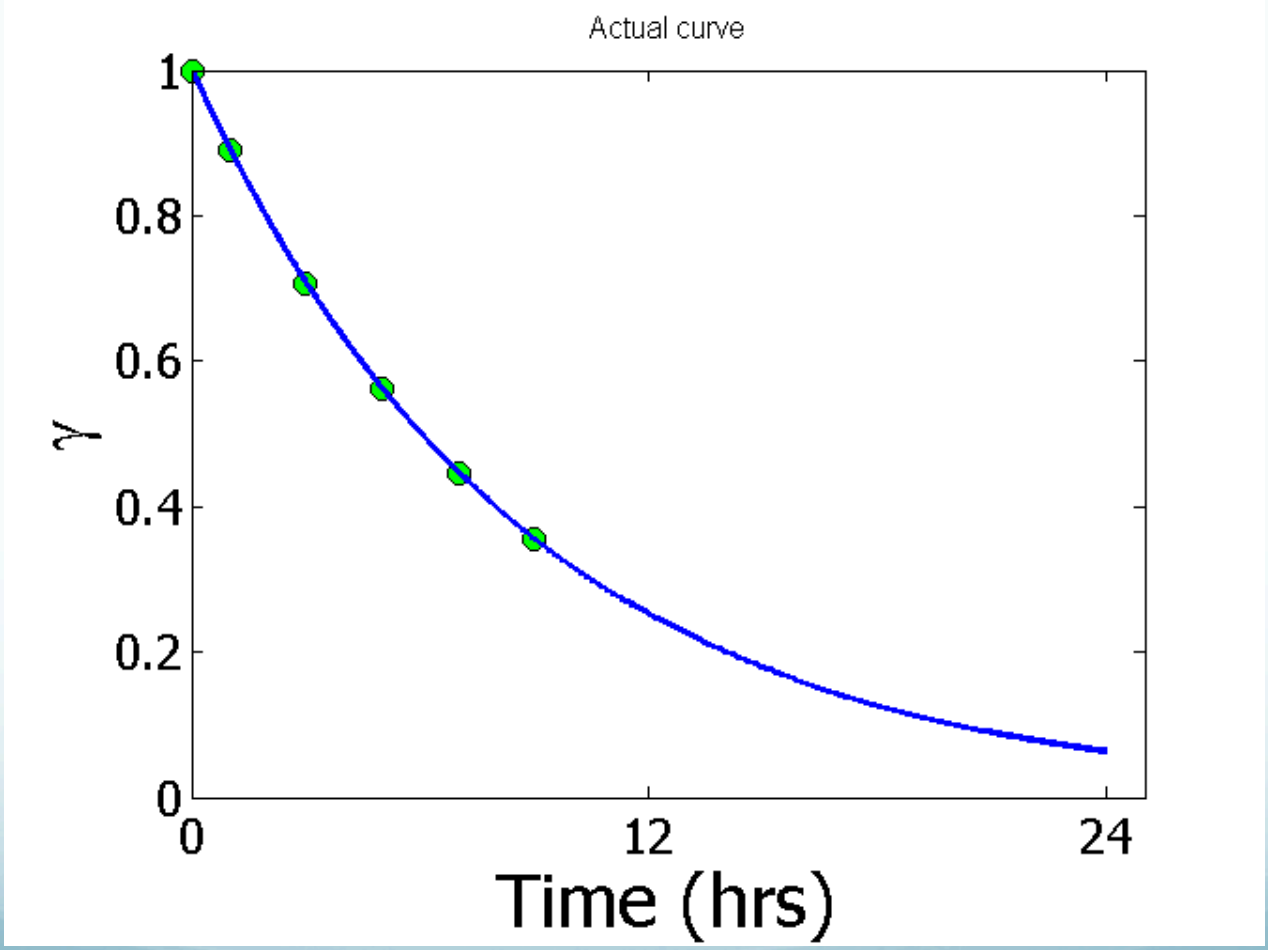
The value of *A* can now be calculated

$$A = \frac{\sum_{i=1}^{6} \gamma_i e^{\lambda t_i}}{\sum_{i=1}^{6} e^{2\lambda t_i}} = 0.9998$$

The exponential regression model then is

$$\gamma = 0.9998\, e^{-0.1151t}$$

Plot of data and regression curve



Relative Intensity After 24 hrs

The relative intensity of radiation after 24 hours

$$\gamma = 0.9998 \times e^{-0.1151(24)}$$

$$= 6.3160 \times 10^{-2}$$

This result implies that only

$$\frac{6.316 \times 10^{-2}}{0.9998} \times 100 = 6.317\%$$

radioactive intensity is left after 24 hours.

Transformation of Data

To find the constants of many nonlinear models, it results in solving simultaneous nonlinear equations. For mathematical convenience, some of the data for such models can be transformed**. For example, the data for an exponential model can be transformed.**

For the nonlinear regression equation, $y = ae^{bx}$

Taking the natural log of both sides yields,

$$\ln y = \ln a + bx$$

Let $z = \ln y$ and $a_0 = \ln a$

We now have a linear regression model where $z = a_0 + a_1 x$

(implying) $a = e^{a_o}$ with $a_1 = b$

Linearization of data

Using linear model regression methods,

$$a_1 = \frac{n\sum_{i=1}^{n} x_i z_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} z_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$a_0 = \bar{z} - a_1 \bar{x}$$

Once a_o, a_1 are found, the original constants of the model are found as

$$b = a_1$$

$$a = e^{a_0}$$

Example-Linearization of data

Exponential model given as,

$$\gamma = Ae^{\lambda t}$$

$$\ln(\gamma) = \ln(A) + \lambda t$$

Assuming $z = \ln\gamma$, $a_o = \ln(A)$ and $a_1 = \lambda$ we obtain

$$z = a_0 + a_1 t$$

This is a linear relationship between z and t

Example-Linearization of data

Using this linear relationship, we can calculate a_0, a_1 where

$$a_1 = \frac{n\sum_{i=1}^{n} t_i z_i - \sum_{i=1}^{n} t_i \sum_{i=1}^{n} z_i}{n\sum_{i=1}^{n} t_1^2 - \left(\sum_{i=1}^{n} t_i\right)^2}$$

and

$$a_0 = \bar{z} - a_1 \bar{t}$$

$$\lambda = a_1$$

$$A = e^{a_0}$$

Example-Linearization of Data

Summations for data linearization are as follows

Table. Summation data for linearization of data model

i	t_i	γ_i	$z_i = \ln \gamma_i$	$t_i z_i$	t_i^2
1	0	1	0.00000	0.0000	0.0000
2	1	0.891	−0.11541	−0.11541	1.0000
3	3	0.708	−0.34531	−1.0359	9.0000
4	5	0.562	−0.57625	−2.8813	25.000
5	7	0.447	−0.80520	−5.6364	49.000
6	9	0.355	−1.0356	−9.3207	81.000
$\sum$	25.000		−2.8778	−18.990	165.00

With $n = 6$

$$\sum_{i=1}^{6} t_i = 25.000$$

$$\sum_{i=1}^{6} z_i = -2.8778$$

$$\sum_{i=1}^{6} t_i z_i = -18.990$$

$$\sum_{i=1}^{6} t_i^2 = 165.00$$

Example-Linearization of Data

Calculating a_0, a_1

$$a_1 = \frac{6(-18.990)-(25)(-2.8778)}{6(165.00)-(25)^2} = -0.11505$$

$$a_0 = \frac{-2.8778}{6} - (-0.11505)\frac{25}{6} = -2.6150\times10^{-4}$$

Since

$$a_0 = \ln(A)$$

$$A = e^{a_0}$$

$$= e^{-2.6150\times10^{-4}} = 0.99974$$

also

$$\lambda = a_1 = -0.11505$$

Example-Linearization of Data

Resulting model is $\gamma = 0.99974 \times e^{-0.11505t}$