Practical Numerical Training UKNum

Statistics, data modeling

Dr. C. Mordasini

Max Planck Institute for Astronomy, Heidelberg

Program:

1) Elementary statistics
2) Regression analysis
3) Linear regression
4) Non-linear regression

1 Elementary statistics

Basic statistical quantities I

- Study a statistical sample containing n values e.g. measurements.
- We can characterise the sample by so called moments, i.e. sums of integer powers of the values.

Mean

Estimates the value around which central clustering occurs

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Alternatives are the median or the mode.

Range

$$R = y_{\max} - y_{\min}$$

$y_{\max}$ is the maximum of the values of y_i, $i = 1,2,...,n$,
$y_{\min}$ is the minimum of the values of y_i, $i = 1,2,...,n$..

Problem: outliers

Basic statistical quantities II

Residual

The residual between a data point i and the mean is the residual i

$$e_i = y_i - \bar{y}$$

The residual can be negative or positive and hence if one calculates the sum of such differences to find the overall spread, the differences may simply cancel each other. That is why the sum of the square of the differences is a better measure.

Summed squared error

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

The magnitude of the summed squared error is dependent on the number of data points. Therefore, we want an average.

Basic statistical quantities III

Variance

An average value of the summed squared error is the variance

$$\sigma^2 = \frac{S_t}{n-1} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

The variance is divided by (n –1) and not n because with the use of the mean in calculating the variance, we lose the independence of one of the data points. That is, if you know the mean of n data points, then the value of one of the n data points can be calculated by knowing the other (n–1) values. N should be used if the mean is known externally (not calculated from the sample).

Basic statistical quantities IV

Variance continued

There are different ways of writing the variance numerically like

$$\mathrm{Var}(x_1 \ldots x_N) = \frac{1}{N-1}\left[\left(\sum_{j=1}^{N} x_j^2\right) - N\overline{x}^2\right] \approx \overline{x^2} - \overline{x}^2$$

One which reduces round-off errors (large samples) is the corrected two pass algorithm. First calculate the mean, then the variance as

$$\mathrm{Var}(x_1 \ldots x_N) = \frac{1}{N-1}\left\{\sum_{j=1}^{N}(x_j - \overline{x})^2 - \frac{1}{N}\left[\sum_{j=1}^{N}(x_j - \overline{x})\right]^2\right\}$$

Basic statistical quantities V

Standard deviation

To bring the variation back to the same level of units as the original data, the standard deviation is defined as

$$\sigma = \sqrt{\frac{S_t}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}}$$

Coefficient variation

The ratio of the standard deviation to the mean, known as the coefficient variation is used to normalize the spread of a sample

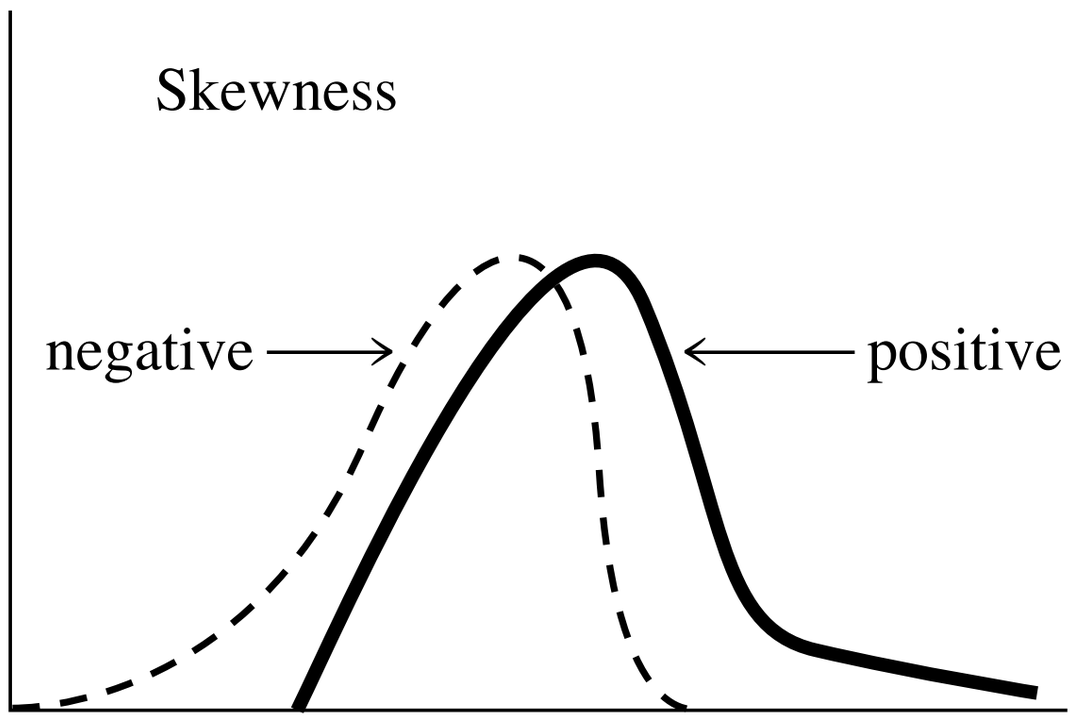
$$c.v = \frac{\sigma}{\bar{y}} \times 100 \qquad [\%]$$

Basic statistical quantities VI

Skewness

Also known as the third moment, skewness characterizes the degree of asymmetry of a distribution around its mean. It is a non-dimensional number.

$$\mathrm{Skew}(x_1 \ldots x_N) = \frac{1}{N} \sum_{j=1}^{N} \left[\frac{x_j - \overline{x}}{\sigma} \right]^3$$



Press et al.

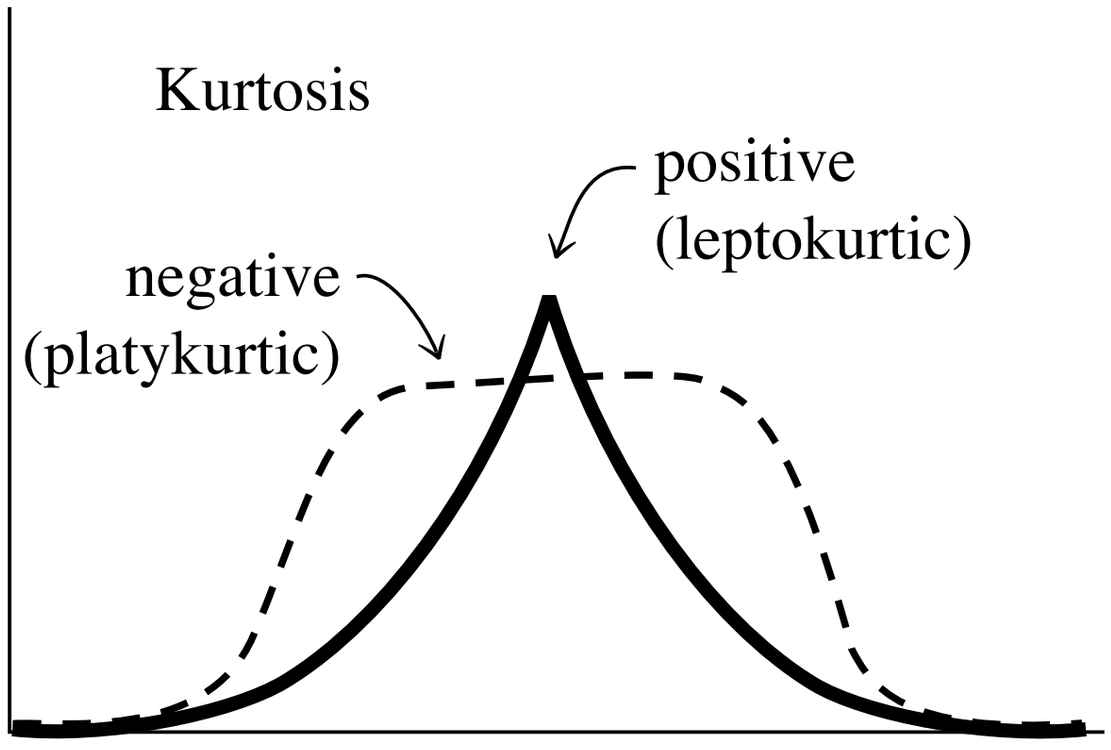
A positive value signifies a distribution with an asymmetric tail extending out towards more positive x; a negative value signifies a distribution whose tail extends out towards more negative x.

Basic statistical quantities VII

Kurtosis

Also known as the forth moment, the kurtosis characterizes the relative peakedness of flatness of a distribution. It is a non-dimensional number.

$$\mathbf{Kurt}(x_1 \ldots x_N) = \left\{ \frac{1}{N} \sum_{j=1}^{N} \left[\frac{x_j - \overline{x}}{\sigma} \right]^4 \right\} - 3$$



Press et al.

A distribution with positive kurtosis is termed leptokurtic. The Matterhorn is an example. A distribution with negative kurtosis is termed platykurtic. An in-between distribution is mesokurtic. The reference is a normal distribution.

High moments (skew, kurt) are less robust than the mean or std. deviation.

Basic statistical quantities VIII

Median

The median of a probability distribution function p(x) is the value x_{med} for which larger and smaller values of x are equally probable:

$$\int_{-\infty}^{x_{\rm med}} p(x)\,dx = \frac{1}{2} = \int_{x_{\rm med}}^{\infty} p(x)\,dx$$

The median of a sample of values $x_1,\ldots, x_N$ is that value x_i which has equal numbers of values above it and below it. Of course, this is not possible when N is even. In that case it is conventional to estimate the median as the mean of the unique two central values. If the values are sorted into ascending order, then the formula for the median is

$$x_{\rm med} = \begin{cases} x_{(N+1)/2}, & N \text{ odd} \\ \frac{1}{2}(x_{N/2} + x_{(N/2)+1}), & N \text{ even} \end{cases}$$

Press et al.

Basic statistical quantities IX

Mode

The mode of a probability distribution function p(x) is the value of x where p takes on a maximum value. The mode is useful primarily when there is a single, sharp maximum, in which case it estimates the central value.

Occasionally, a distribution will be bimodal, with two relative maxima; then one may wish to know the two modes individually. Note that, in such cases, the mean and median are not very useful, since they will give only a “compromise” value between the two peaks.

Press et al.

2 Regression analysis

Regression analysis

What is regression analysis?
Regression analysis gives quantitative information on the relationship between a response (dependent) variable and one or more predictor (independent) variables to the extent that the necessary information is contained in the data.

What is regression analysis used for?

1. prediction
2. model adaption and
3. parameter estimation.

In other words: fitting

Least squares method

This is the most popular method of parameter estimation for coefficients of regression models. It has well known probability distributions and gives unbiased estimators of regression parameters with the smallest variance.

We wish to predict the response to n data points $(x_1,y_1),(x_2,y_2),\ldots\ldots,(x_n,y_n)$ by a regression model given by

$$y = f(x)$$

where the function f (x) has regression constants that need to be estimated. For example

$f(x) = a_0 + a_1x$ is a straight-line regression model with constants a_0 and a_1
$f(x) = a_0e^{a_1x}$ is an exponential model with constants a_0 and a_1
$f(x) = a_0 + a_1x + a_2 x^2$ is a quadratic model with constants a_0 , a_1 and a_2

Least squares method II

A measure of goodness of fit, that is how the regression model f (x) predicts the response variable y is the magnitude of the residual, E_i at each of the n data points.

$$E_i = y_i - f(x_i),\ i = 1,2,....n$$

Ideally, if all the residuals E_i are zero, one may have found an equation in which all the points lie on a model.

In the least squares method, estimates of the constants of the models are chosen such that minimization of the sum of the squared residuals is achieved, that is minimize

$$\sum_{i=1}^{n} E_i^{\,2}$$

from where the name least squares.

Least squares method III

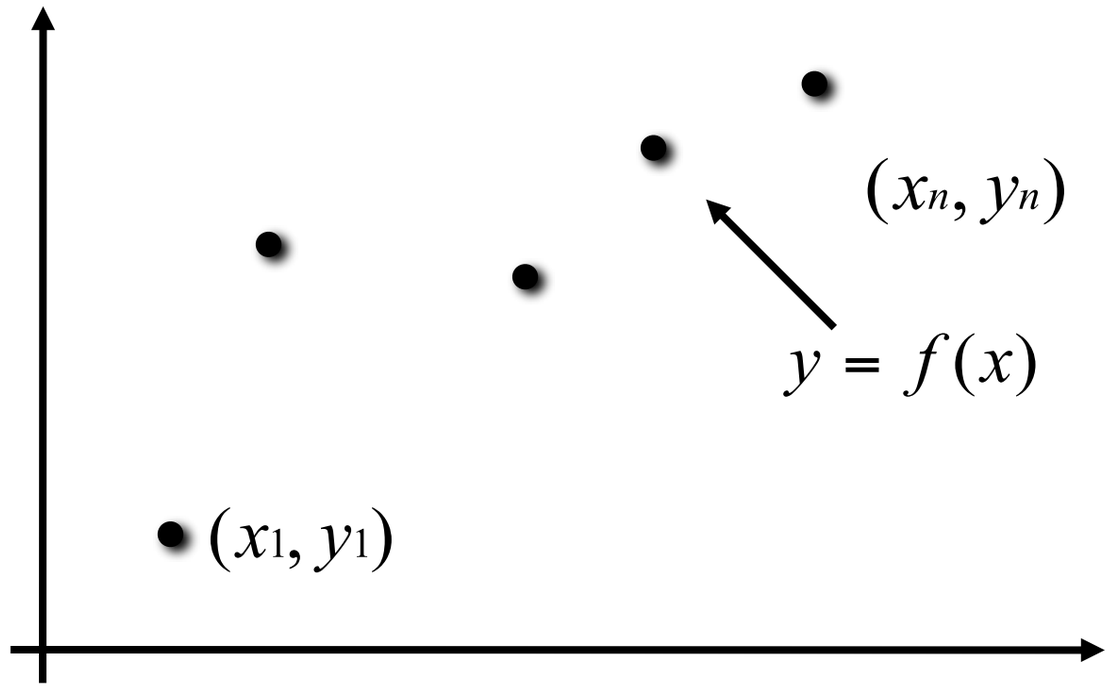
In other words, given n data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ best fit $y = f(x)$ to the data. The best fit is based on minimizing the sum of the square of the residuals, S_r .

Residual at a point is

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

Sum of the square of the residuals

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



Basic model for regression

3 Linear Regression

Linear Regression

Given n data points $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ best fit

$$y = a_0 + a_1 x$$

to the data.

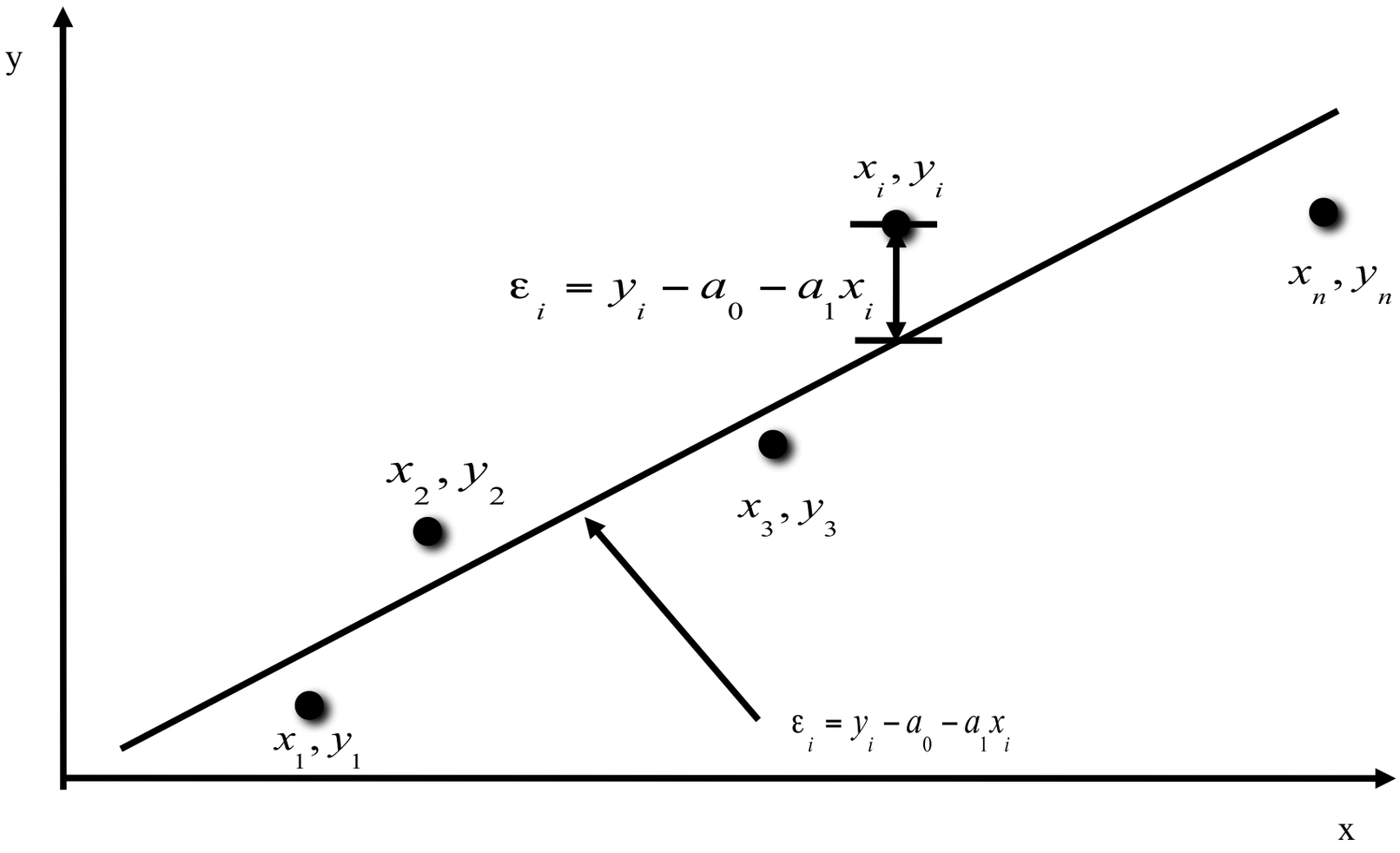


Illustration with Mathematica

Parameter determination I

The least squares criterion minimizes the sum of the square of the residuals in the model, and produces a unique line.

$$S_r = \sum_{i=1}^{n} \varepsilon_i^{\,2} = \sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i\right)^2$$

Our task is to minimize the sum of the square of the residuals by determining the best regression (or fit) parameters a_0 and a_1.

At a minimum, we must have (using the chain rule):

$$\frac{\partial S_r}{\partial a_0} = -2\sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i\right)(-1) = 0$$

$$\frac{\partial S_r}{\partial a_1} = -2\sum_{i=1}^{n} \left(y_i - a_0 - a_1 x_i\right)(-x_i) = 0$$

Parameter determination II

This gives

$$\sum_{i=1}^{n} a_0 + \sum_{i=1}^{n} a_1 x_i = \sum_{i=1}^{n} y_i$$

$$\sum_{i=1}^{n} a_0 x_i + \sum_{i=1}^{n} a_1 x_i^{\,2} = \sum_{i=1}^{n} y_i x_i$$

Noting that $\sum_{i=1}^{n} a_0 = a_0 + a_0 + \ldots + a_0 = n a_0$

$$n a_0 + a_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

Parameter determination III

$$na_0 + a_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$a_0 \sum_{i=1}^{n} x_i + a_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

This later can be understood as a system of two linear equations, which we can write in a 2 x 2 matrix form as we have seen in the last lecture, with the unknowns a_0 and a_1 (all x_i and y_i are known). For such a small matrix, one quickly find the result

$$a_0 = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

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

Parameter determination IV

Defining

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

$$S_{xy} = \sum_{i=1}^{n} x_i\, y_i - n\,\bar{x}\,\bar{y} \qquad S_{xx} = \sum_{i=1}^{n} x_i^2 - n\,\bar{x}^2$$

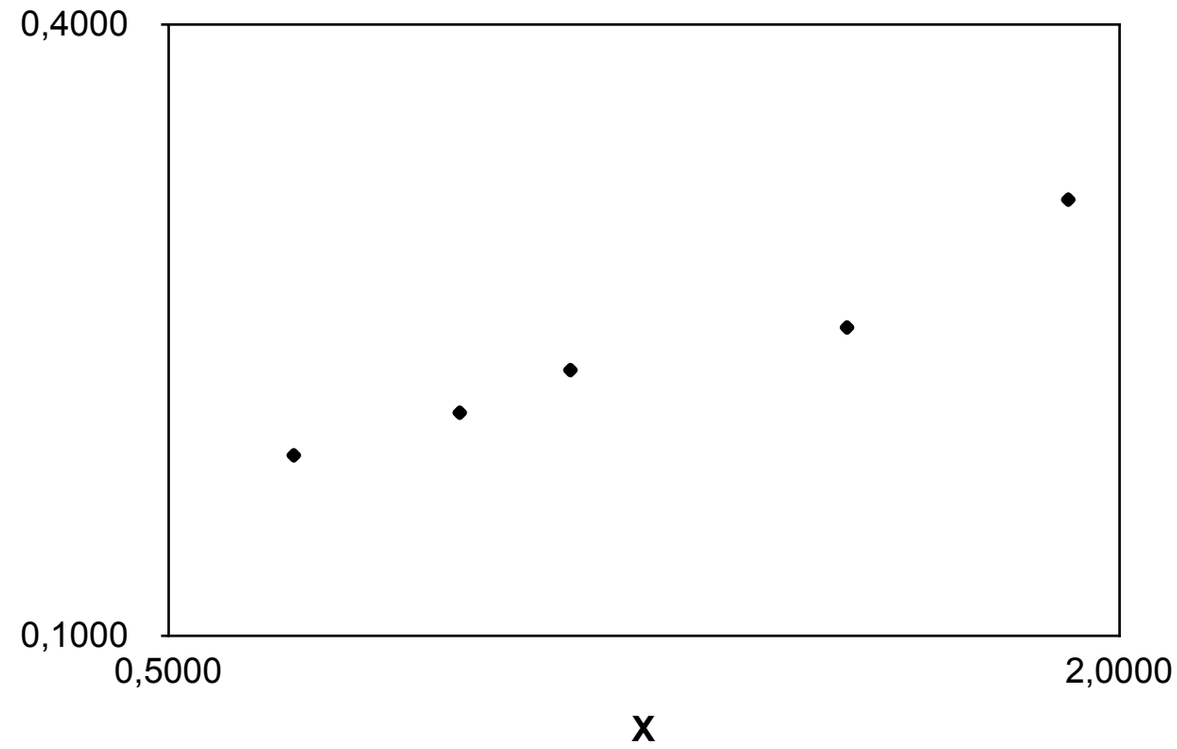
we can write the least squares linear fit parameters as

$$a_1 = \frac{S_{xy}}{S_{xx}} \qquad a_0 = \bar{y} - a_1\,\bar{x}$$

Example I

Given the following arbitrary data set, find the best fitting last square model.

x	y
0.698132	0.188224
0.959931	0.209138
1.134464	0.230052
1.570796	0.250965
1.919862	0.313707



Find the constants a_0 and a_1 for the linear model given by

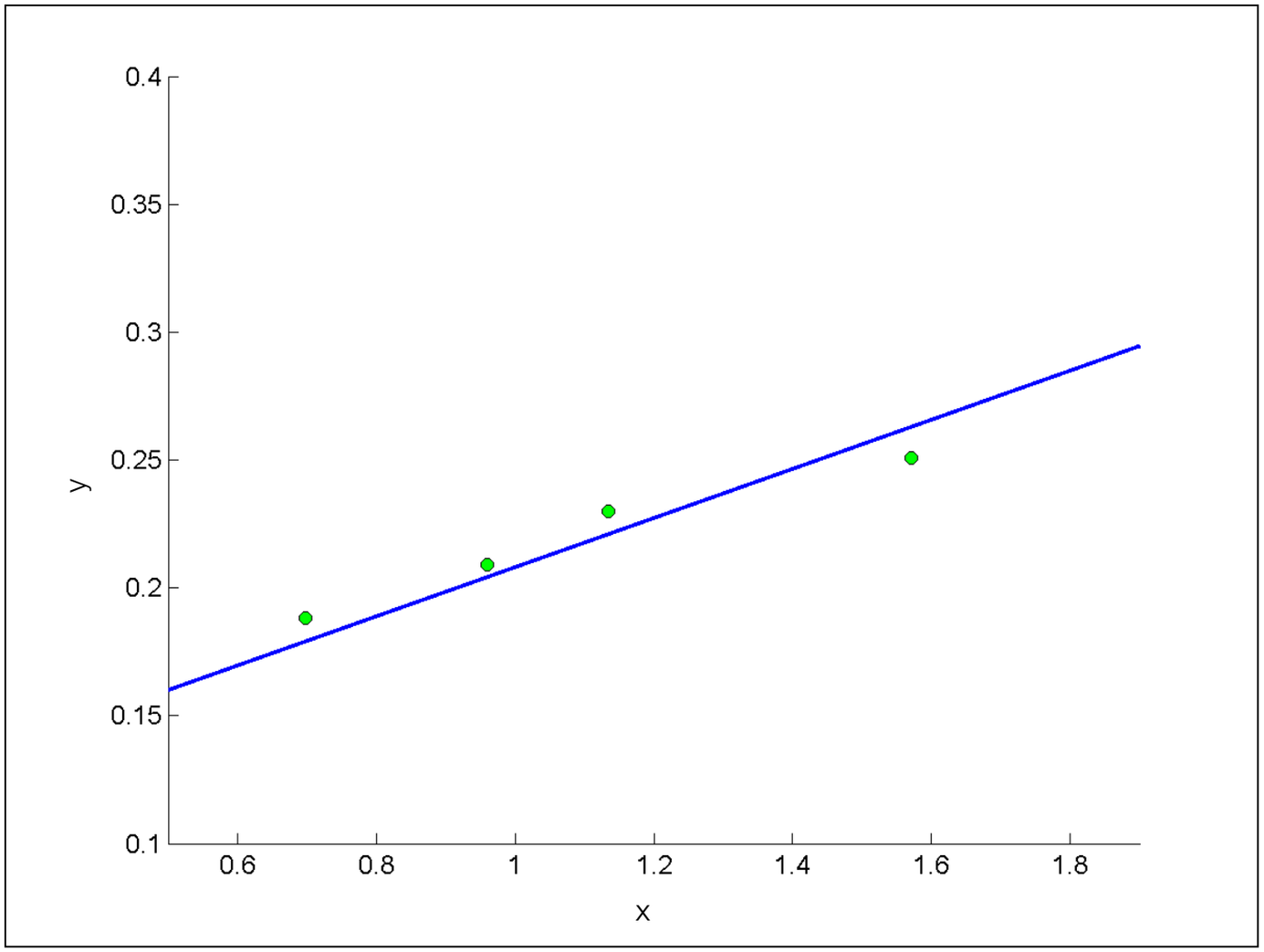
$$y = a_0 + a_1 x$$

Example II

Using the equations described for a_0 and a_1 we rapidly find

$$a_0 = 1.1767 \times 10^{-1}$$

$$a_1 = 9.6091 \times 10^{-2}$$



4 Non-Linear Regression

Nonlinear Regression

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: $(y = a_0 + a_1 x + \ldots + a_m x^{m})$

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.

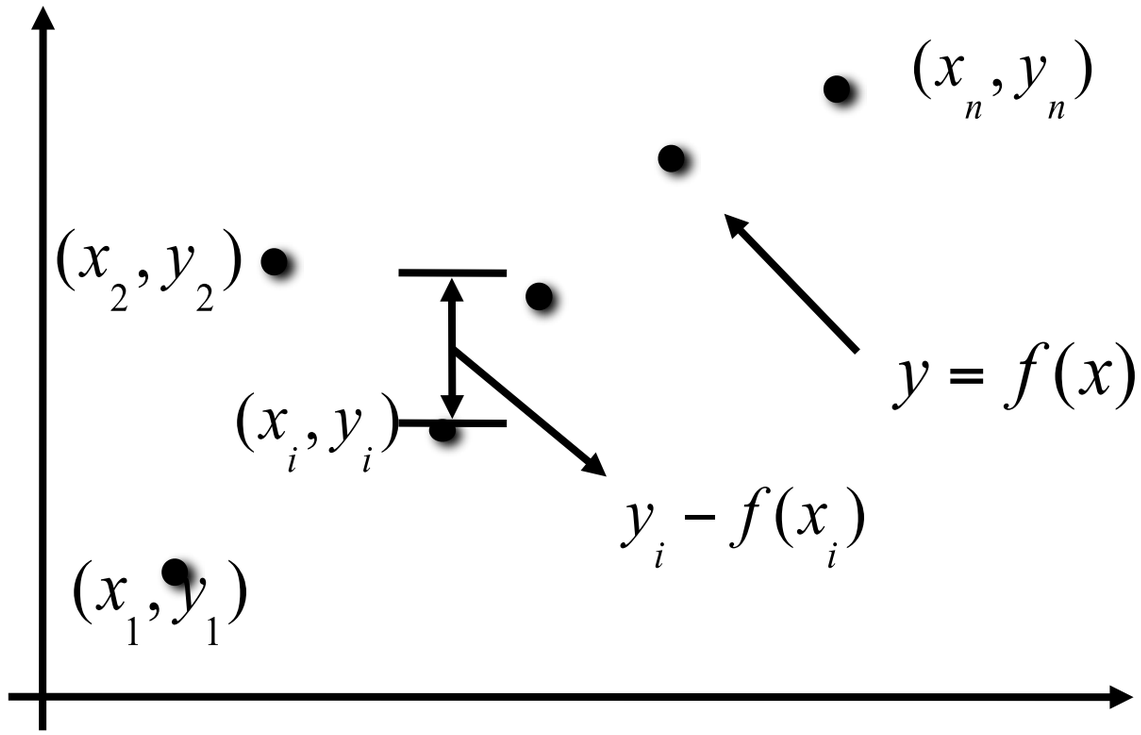


Figure. Nonlinear regression model for discrete y vs. x data

Exponential Model

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

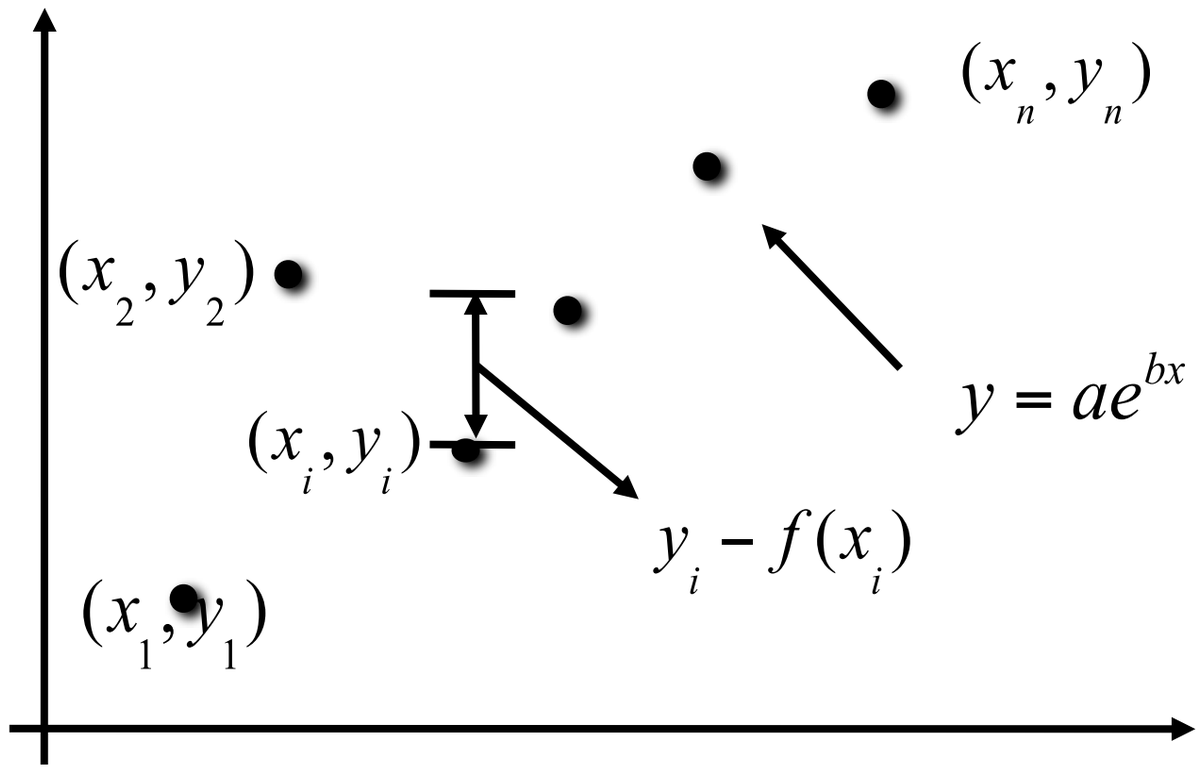


Figure. Exponential model of nonlinear regression for y vs. x data

Finding Constants of Exponential Model I

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 II

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 III

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. Once it is found, we can also calculate a.

Example - Exponential Model I

Many patients get concerned when a test involves injection of a radioactive material. For example for scanning a gallbladder, a few drops of Technetium-99m isotope is used. Half of the Technetium-99m would be gone in about *6* hours. It however takes about *24* hours for the radiation levels to reach what we are exposed to in day-to-day activities. Below is given the relative intensity of radiation as a function of time.

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 II

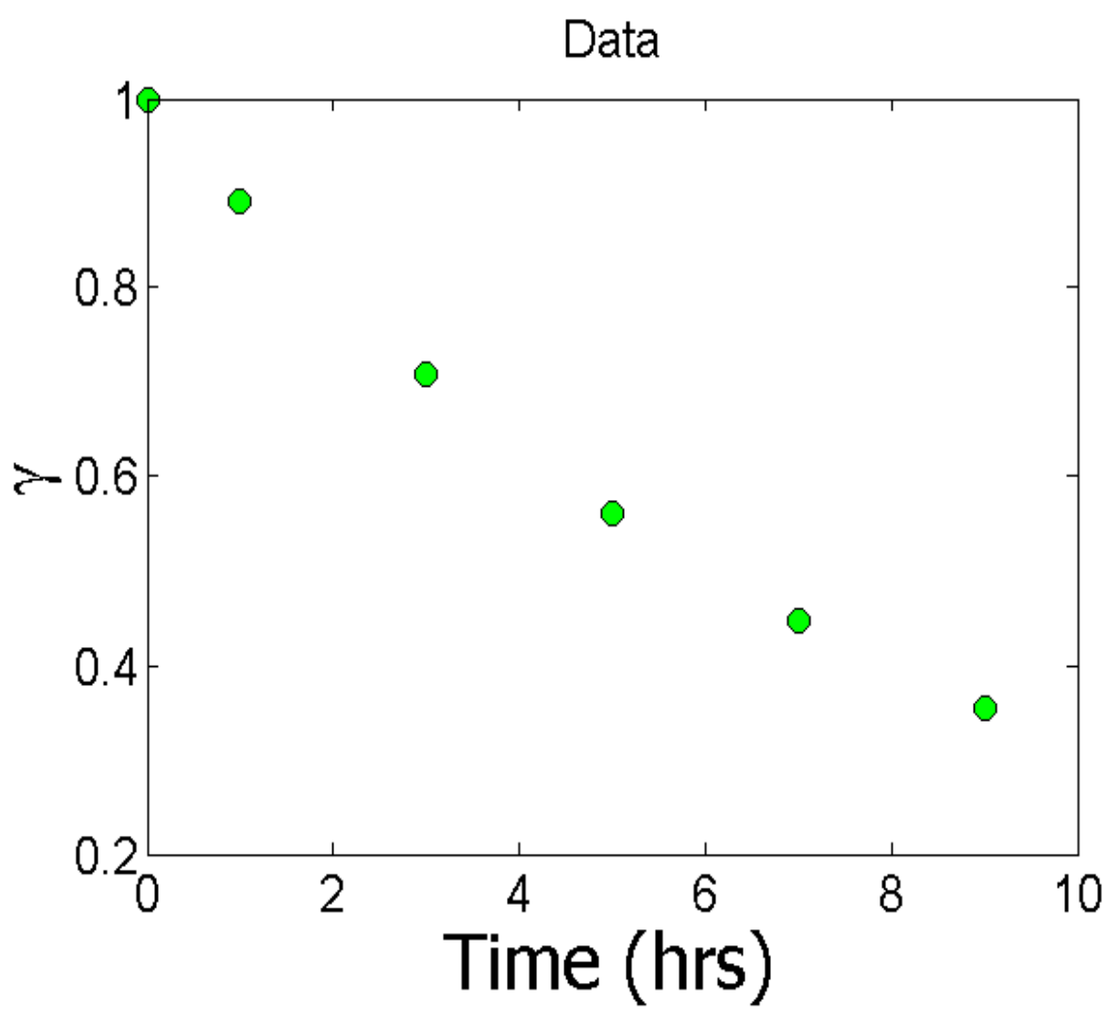
The relative intensity is related to time by the equation $\gamma = Ae^{\lambda t}$

Find:

a) The value of the regression constants A and λ

b) The half-life of Technetium-99m

c) Radiation intensity after 24 hours



Constants of the Model

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

The value of λ is found by solving the nonlinear equation

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

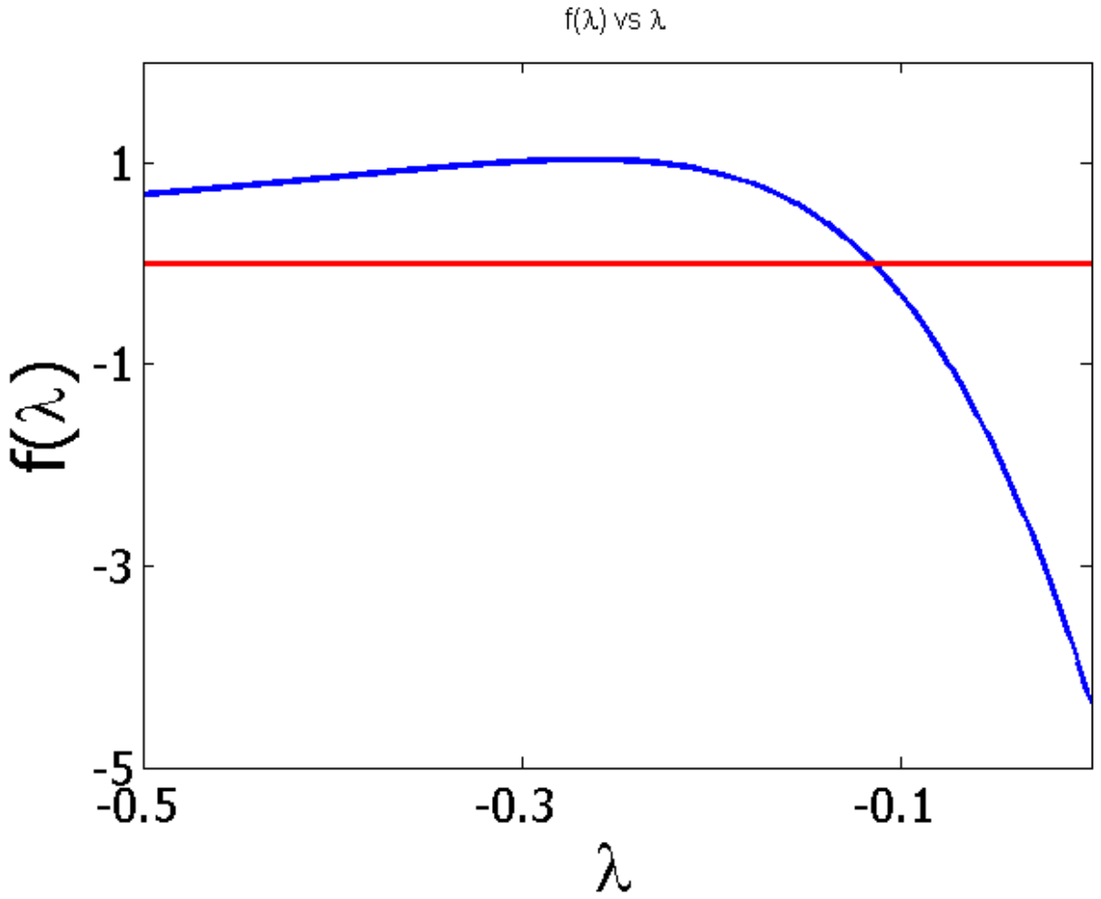
Once it is found,

A is given as

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

Solving the non-linear equation

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



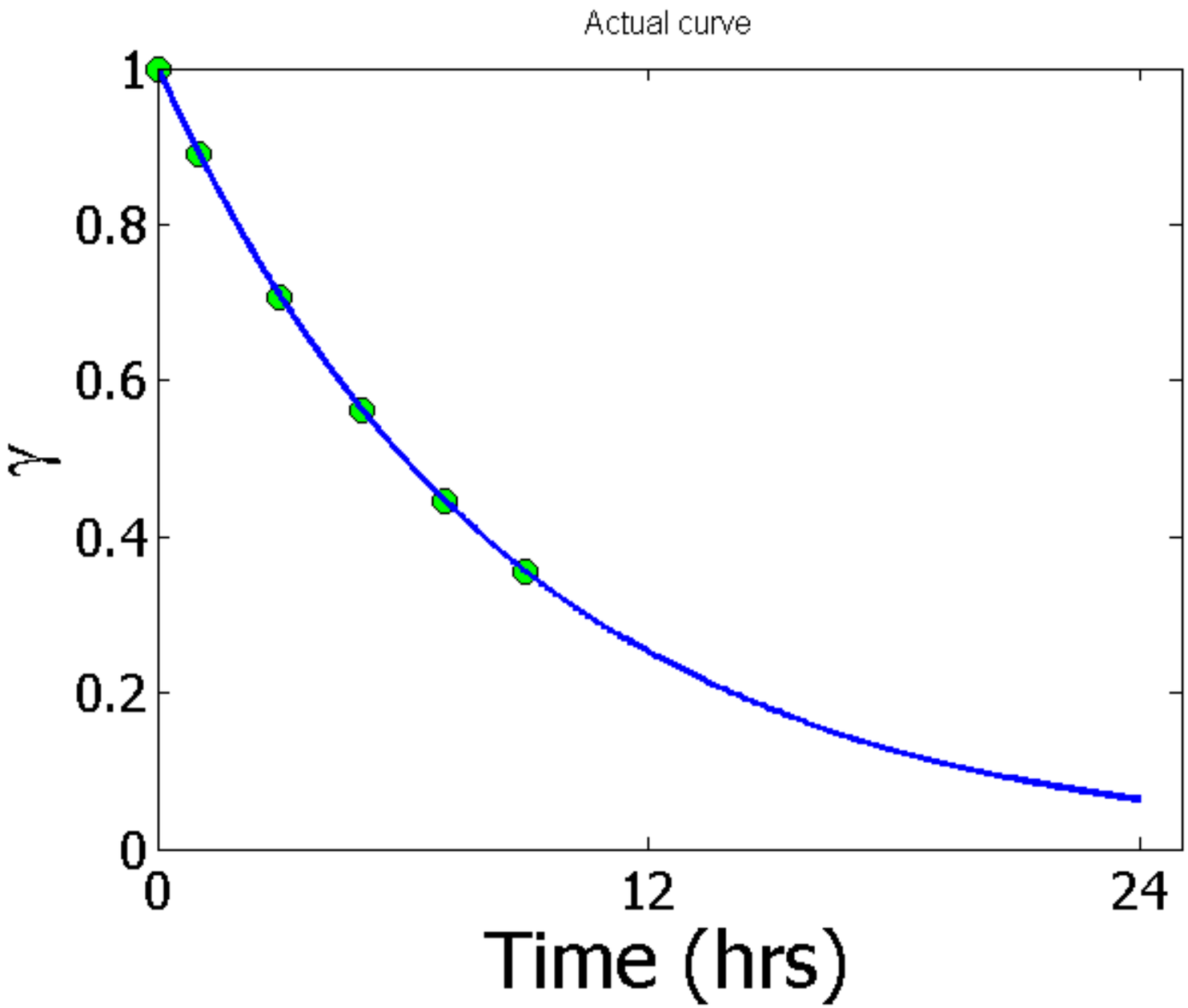
$$\lambda = -0.1151$$

The value of *A* can now be calculated also:

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

Plot of data and regression curve

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



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.

Linearization of data I

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 linearized. For example, the data for an exponential model can be linearized.

As shown in the previous example, many chemical and physical processes are governed by the 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 II

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

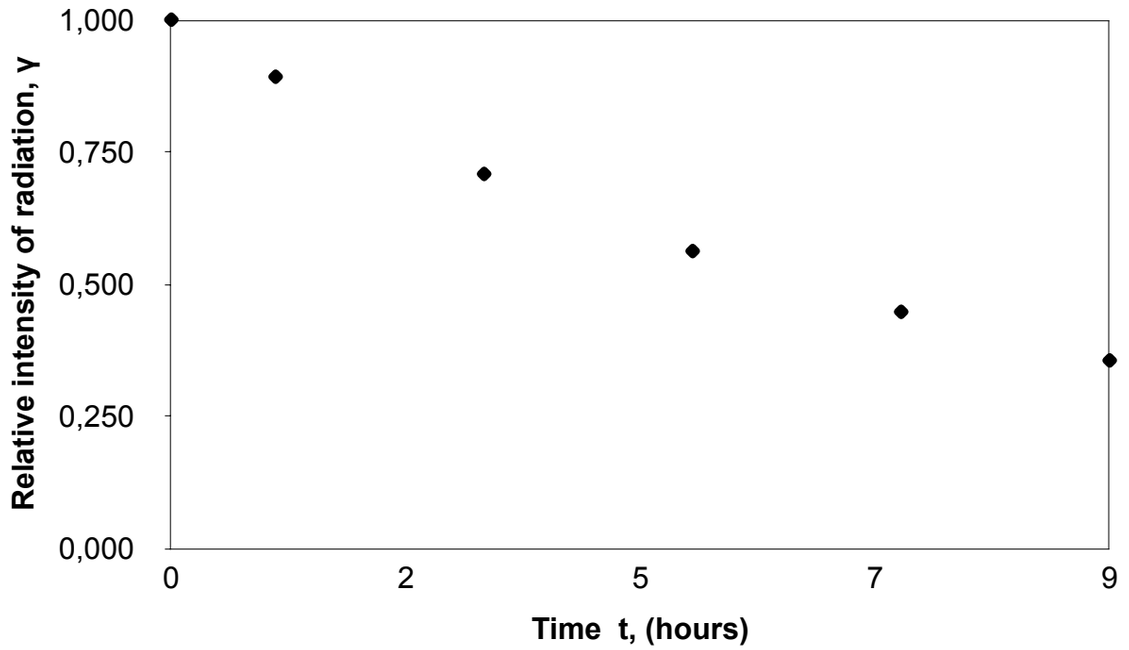
$$b = a_1$$

$$a = e^{a_0}$$

Example - Linearization of data I

Same example of radioactive decay as before

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



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 II

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

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

$$\lambda = a_1$$

$$A = e^{a_0}$$

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 III

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

$$a_0 = \ln(A)$$

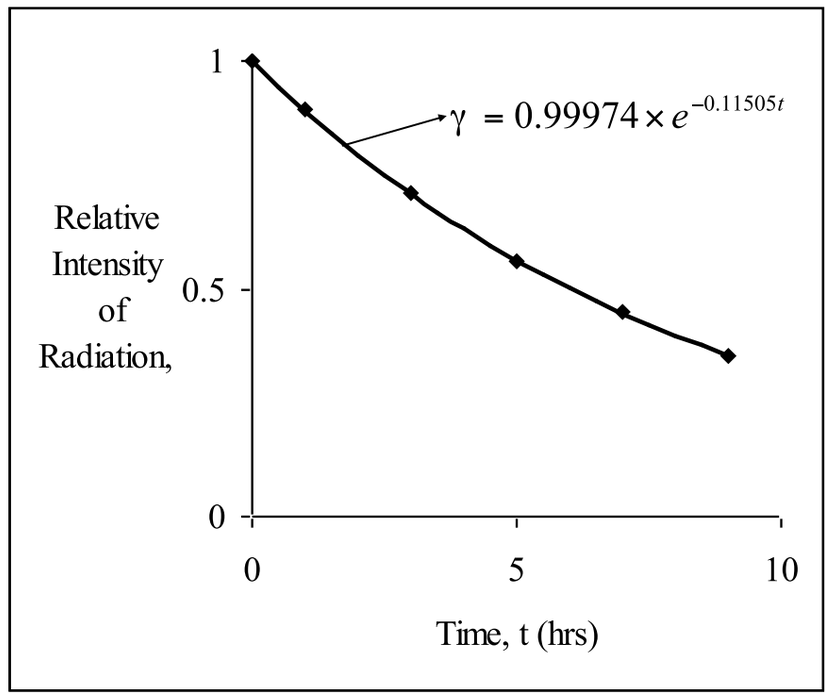
$$A = e^{a_0}$$

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

also

$$\lambda = a_1 = -0.11505$$

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



Example - Linearization of Data IV

The regression formula is $\gamma = 0.99974 \times e^{-0.11505t}$

The Half life of Technetium 99 is when $\gamma = \frac{1}{2}\gamma\Big|_{t=0}$

$$0.99974 \times e^{-0.11505t} = \frac{1}{2}(0.99974)e^{-0.11505(0)}$$

$$e^{-0.11508t} = 0.5$$

$$-0.11505t = \ln(0.5)$$

$$t = 6.0248 \ hours$$

Literature value: 6.01 hours

Comparison

Comparison of exponential model with and without data linearization:

	With data linearization	Without data linearization
A	0.99974	0.99983
λ	−0.11505	−0.11508
Half-Life (hrs)	6.0248	6.0232
Relative intensity after 24 hrs.	6.3200×10^{-2}	6.3160×10^{-2}

The values are very similar so data linearization was suitable to find the constants of the nonlinear exponential model in this case.

References

- This script is based on http://numericalmethods.eng.usf.edu by Autar Kaw, Jai Paul and Numerical Recipes (2nd/3rd Edition) by Press et al., Cambridge University Press
  http://www.nr.com/oldverswitcher.html

Statistics and Fitting

Return by 9:15 a.m. tomorrow

Assignment for the Afternoon / Homework

All data files required for this exercise sheet can be downloaded from the UKNum homepage https://www2.mpia-hd.mpg.de/~klahr/UKNUM2022.html

- **Exercise 1, 4 points:** Statistics (I).
  The data in the file `dice.dat` (available on the UKNum homepage) represent the results of experiments rolling two dice and adding the pips ("Augenzahl"). The first and second column in the file give the number of the experiment and the result, respectively. **(a)** Write a program code which reads in the data file and calculates the mean, the median and the standard deviation of the data. What is the relative difference between the mean and the median?

- **Exercise 2, 6 points:** Statistics (II).
  The data in the file `grades.dat` represents the grades of an exam (with results between 0 (worst) and 100 (best). The number in the first column stands for the student's number, with his grade given in the second column. **(a)** Find the mean, the median and the standard deviation of the data. **(b)** What is the relative difference between the mean and the median? How does it compare to the result from 1? What does it tell you about the symmetry of the underlying parent distribution of the data?

- **Exercise 3, 10 points:** Least-Squares Fit to a Line
  A student has measured the potential difference (in volts) along a conducting nickel-silver wire using an analog voltmeter. The voltage is measured between the negative end of the wire and various positions along the wire (in cm). Uncertainties in the positions are less than 1 mm and can be neglected. Assume the the uncertainty in the voltage measurement to be the same for each measurement (i.e., set σ_i= 1.0). **(a)** Write a program code to read in the data in the file `data1.dat` and fit a line $f(x) = a + bx$ through the data using a Least-Squares Fit as presented in the Lecture. **(b)** Estimate the standard deviation of an individual measurement using

$$\sigma^2 \sim s^2 = \frac{1}{N-2} \sum (y_i - a - bx_i)^2.$$

Continued on next page/backside

- **Optional Task:** Least-Squares Fit to an Exponential
  The file `data2.dat` contains the measurements of an exponential decay, e.g., of a radioactive element (in some arbitrary units). The last column gives the (constant) uncertainty in each measurement. Fit an exponential function $f(x) = a\exp(bx)$ to the data using **(a)** the *direct* exponential fit as presented in the Lecture. Use a bracketing root-finding algorithm, e.g., bisection. In the current case, initial guesses of $b_1 = -1$ and $b_2 = 0$ do bracket the root $b_\star$. **(b)** Fit a straight line to the linearized equation using a least-squares fit. Compare the result to the exponential fit in (a). Plot the results of (a) and (b) on linear and on logarithmic scale ($\log y$ vs. x) including the errorbars.