LS method

Math model

Interpolation methods require that approximation curve rigorously includes all the empirical data points. Least-square fitting method requires that approximation curve *is as close to all the data points as possible*. In other words, LS method helps in fitting known data points by a known standard curve, which is proven to be the best possible fit by this standard curve.

Task: In the table, sets of $\{x_i\}$ and $\{y_i\}$ are given:

x	x_1	x_2	…	x_n
$f(x)$	y_1	y_2	…	y_n

Task: find a function $Y=F(X)$, which describes the observed dependence and having values $Y_i=F(x_i)$, as close as possible to measured $y_1 \dots y_n$ at $x_1,\dots x_n$.

LSM uses the least squares in the way where sum of squares of the differences between the measured and calculated values is taken a measure to validate the approximation quality. Introducing $F(x_i)=\hat{y}_i$, then:

x	x_1	x_2	…	x_n
$f(x)$	y_1	y_2	…	y_n
$F(x)$	$\hat{y}_1$	$\hat{y}_2$	…	$\hat{y}_n$

Condition of being as close as possible mathematically is written as follows:

$$(y_i - \hat{y}_i)^2 \rightarrow min \Leftrightarrow \Sigma (y_i - \hat{y}_i)^2 \rightarrow min$$

Standard regressions used for approximation of data points:

LS method guarantees that the found fit would be the best possible with the given type of dependence.
Standard dependences:

A. linear:	$y=ax+b$	E. fractionally-linear:	$y=1/(ax+b)$
B. polynomial (example – square-law):	$y=ax^2+bx+c$	F. logarithmic:	$y=a\ ln(x)+b$
C. power-law:	$y=ax^m$	G. inversely-proportional:	$y=a/x+b$
D. exponential:	$y=a\ exp(kx)$	H. rational:	$y=x/(ax+b)$

To choose a proper dependence type:

perform visual analysis of the data points;
consider physical processes the dependence should be described with
form a table as below and use analytical algorithm (Demidovitch):

№	x_s'	y_s'	type of dependence
1	$(x_1+x_n)/2$	$(y_1+y_n)/2$	y = ax + b
2	$(x_1*x_n)^{1/2}$	$(y_1*y_n)^{1/2}$	$y=ax^m$
3	$(x_1+x_n)/2$	$(y_1*y_n)^{1/2}$	y=a exp(b x)
3-a	$(1/x_1+1/x_n)/2$	$(y_1*y_n)^{1/2}$	y=a exp(b/x)
4	$(2*x_1*x_n)/(x_1+x_n)$	$(y_1+y_n)/2$	y=a/x+b
5	$(x_1+x_n)/2$	$(2*y_1*y_n)/(y_1+y_n)$	y=1/(ax+b)
6	$(2*x_1*x_n)/(x_1+x_n)$	$(2*y_1*y_n)/(y_1+y_n)$	y=x/(ax+b)
7	$(x_1*x_n)^{1/2}$	$(y_1+y_n)/2$	y=a ln(x)+b

- Find the row where $x_s' \approx y_s'$ (better than in other rows)
- Or, if impossible, find the row where: $|\hat{y}_s - y_s'| \approx 0$, $\hat{y}_s = y_i + ((y_{i+1} - y_i)/(x_{i+1} - x_i)) * (x_s' - x_i)$,
- here $x_i < x_s' < x_{i+1}$, x_i and x_{i+1} – are the closest to x_s'.
- Find required parameters. *a, b, c, m, k.*

Validation of the method:

If it is known that two quantities are related by a certain functional dependence, but the parameters of this dependence remain undefined, and assuming that the number of initial measurements is not less than the number of unknown parameters, then the definition of the latter is a purely algebraic problem. If the number of measurements is greater than the number of undefined parameters, then the system can be called redefined, since there are more equations than the variables. But, since each measurement is inherently inaccurate, then the extra equations cannot simply be dropped out, for each new observation adds its own portion of information. The equations obtained in the strict mathematical sense are inconsistent and in need of further definition. Minimizing the squares of the differences between the empirical and theoretical values is the extension of the system whose number of equations is no longer equal to the number of experiments *n*, but *m+1*, where *m* is the degree of the approximating polynomial.

Thus, when applied to a specific problem, the least-squares solution reduces to solving a system of linear equations, where the variables are the unknown coefficients of the desired functional dependence.

Linear approximation:

$\{x_i, y_i\}$, $i = 1,2,..n$

$\hat{y}_i(x_i) = a + b \cdot x_i$

$\sigma^2 = \frac{1}{n}\Sigma[(a + b \cdot x_i) - y_i]^2$, σ^2 is the variance, $\sigma^2 \to \min$.

Differentiating $\Sigma[(a + b \cdot x_i) - y_i]^2$ by each parameter, and setting differential equal to 0, we obtain a system of $m+1$ equations:

$$\begin{cases} \frac{\partial}{\partial a}[\Sigma((a + bx_i) - y_i)^2] = 0 \\ \frac{\partial}{\partial b}[\Sigma((a + bx_i) - y_i)^2] = 0 \end{cases} \Rightarrow \begin{cases} na + b\Sigma x_i = \Sigma y_i \\ a\Sigma x_i + b\Sigma x_i^2 = \Sigma x_i y_i \end{cases}$$

Hence, to find the best possible linear fit for $\{x_i;y_i\}$, one needs to compute Σx_i, Σx_i^2, Σy_i, $\Sigma x_i y_i$.

Define deviation

Deviation is taken in form of root-mean-square deviation: $\sigma = ((\Sigma\varepsilon_i^2)/n)^{1/2}$, with ε_i being the difference between $y_i(x_i)$ and $\hat{y}_i(x_i)$. The goal is to find best parameters leading to $\sigma \to 0$.

Polynomial approximation

$$\boldsymbol{y = a_0 x^m + a_1 x^{m-1} + a_2 x^{m-2} + \cdots + a_{m-1} x^1 + a_m}$$

Differentiating by free coefficients, we have a system with unknown $a_0 \ldots a_m$.

$$\begin{cases} a_0\Sigma x_i^m + a_1\Sigma x_i^{m-1} + a_2\Sigma x_i^{m-2} + \cdots + na_m = \Sigma y_i \\ a_0\Sigma x_i^{m+1} + a_1\Sigma x_i^m + a_2\Sigma x_i^{m-1} + \cdots + a_m\Sigma x_i^1 = \Sigma x_i y_i \\ a_0\Sigma x_i^{m+2} + a_1\Sigma x_i^{m+1} + a_2\Sigma x_i^m + \cdots + a_m\Sigma x_i^2 = \Sigma x_i^2 y_i \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_0\Sigma x_i^{2m} + a_1\Sigma x_i^{2m-1} + a_2\Sigma x_i^{2m-2} + \cdots + a_m\Sigma x_i^m = \Sigma x_i^m y_i \end{cases}$$

Make a table for calculation:

i	x_i	x_i^2	.	x_i^{2m}	y_i	$x_i y_i$	$x_i^2 y_i$	.	$x_i^m y_i$
1	x_1	x_1^2	.	x_1^{2m}	y_1	$x_1 y_1$	$x_1^2 y_1$	.	$x_1^m y_1$
2	x_2	x_2^2	.	x_2^{2m}	y_2	$x_2 y_2$	$x_2^2 y_2$	.	$x_2^m y_2$
...	...	...	....	...	...	...	...	....	...
n	x_n	$x_n^2 y_n$	.	$x_n^2 y_n$	y_n	$x_n y_n$	$x_n^2 y_n$	.	$x_n^m y_n$
Σ	Σx_i	Σx_i^2	.	$\Sigma x_i^2 y_i$	Σy_i	$\Sigma x_i y_i$	$\Sigma x_i^2 y_i$	.	$\Sigma x_i^m y_i$

Example: *finding of coefficients a_0, a_1 for fit $y = a_0 x + a_1$:*

Data table:

x_i	1	2	3	4	5	6
y_i	2.0	4.9	7.9	11.1	14.1	17.0

Calculation table:

i	x_i	x_i^2	y_i	$x_i y_i$
1	1	1	2.0	2.0
2	2	4	4.9	9.8
3	3	9	7.9	23.7
4	4	16	11.1	44.4
5	5	25	14.1	70.5
6	6	36	17.0	102
Σ	21	91	57.0	252.4

Hence, system of equations:

$$\begin{cases} a_0 \cdot 91 + a_1 \cdot 21 = 252.4 \\ a_0 \cdot 21 + a_1 \cdot 6 = 57 \end{cases}$$

Solution:

$$\begin{cases} a_0 = 3.023 \\ a_1 = -1.081 \end{cases}$$

Other standard regressions. Linearization by variable(s) change

Dependence type	Regression	Rearranging	Variable(s) change	Linear regression
C. power-law:	$y=ax^m$	$lg(y) = m \cdot lg(x) + lg(a)$	$Y = lg(y)$, $X = lg(x)$	$Y = m \cdot X + lg(a)$
D. exponential:	$y=a\ exp(kx)$	$ln(y) = ln(a) + kx$	$Y = ln(y)$	$Y = kx + ln(a)$
E. fractionally-linear:	$y=1/(ax+b)$	$1/y = ax + b$	$Y = 1/y$	$Y = ax + b$
F. logarithmic:	$y=a\ ln(x)+b$		$X = ln(x)$	$y = a \cdot X + b$
G. inversely-proportional:	$y=a/x+b$		$X = 1/x$	$y = a \cdot X + b$
H. rational:	$y=x/(ax+b)$	$1/y = a + b/x$	$Y = 1/y$, $X = 1/x$	$Y = a + b \cdot X$

LS method. Exercises.

Exercise 1. In the "Fundamentals of Chemistry" by D.I. Mendeleyev, data are presented on the solubility of sodium nitrate ($NaNO_3$) as a function of water temperature. In a hundred parts of water the following number of parts of the substance dissolves at the appropriate temperatures:

t, °C	0	4	10	15	21	29	36	51	68
n	66.7	71.0	76.3	80.6	85.7	92.9	99.4	113.6	125.1

Plot data points and find linear approximation: $n = a + bt$.

Exercise 2. When studying the street traffic, observations were made on the distance traveled by the vehicle by inertia (after braking), depending on the speed. The results of observations are:

υ, km/h	S_1, m	S_2, m	S_3, m	S_4, m	S_5, m	<S>, m
6.44	0.61	3.05				
11.26	1.22	6.71				
12.87	4.88					
14.48	3.05					
16.09	7.93	5.49	10.37			
17.70	8.54	5.18				
19.31	6.10	4.27	7.32	8.54		
20.92	10.37	7.93	10.37	14.03		
22.53	10.98	7.93	18.30	24.40		
24.14	16.47	7.93	6.10			
25.74	9.76	12.20				
27.35	15.25	12.20	9.76			
28.96	17.08	25.62	23.18	12.81		
30.57	20.74	14.03	10.98			
32.18	14.64	17.08	19.52	15.86	9.76	
35.40	20.13					
37.01	16.47					
38.62	28.36	21.35	36.60	28.06		
40.23	25.92					

Plot data points and find linear/polynomial approximation.

Exercise 3. Data of laboratory experiments on the determination of gravity with the help of a device with a falling load, in which the load positions at the ends of consecutive intervals in 1/30 second were noted by the spark method, are given in the table. The dependence *s(t)* has the form: $s = s_0 + \upsilon_0 t + {}^1/_2\, gt^2$. Find g.

Peg's data for determination of the free-fall acceleration:

time, in 1/30 sec	1	2	3	4	5	6	7	8	9	10	11	12	13	14
S, cm	11.86	15.67	20.60	26.69	33.71	41.93	51.13	61.49	72.90	85.44	99.08	113.77	129.54	146.48

Exercise 4. In a laboratory work on refractometry, it is required to calculate the unknown concentration of glycerin solution by its refractive index using the coefficients of solutions with known concentration. The refractive index in the work is determined with use of a refractometer:

Sample		1	2	3	4	5
Concentration, %	X	25	50	75	100	x
Refractive index, n	У	1.3734	1.3943	1.4244	1.4538	1.3746

Find linear dependence $y=a_0x + a_1$ and determine the unknown concentration.

Exercise 5:

Choose a proper regression.
Linearize the problem by change of variable(s).
Find approximating regression with use of LS-method.
Transform your solution back to the original variables.
Determine the activation energy (see hint below).
Plot data points and fitting curve both as $R(T)$ and $lnR(1/T)$.

Resistance of a semiconductor was measured at different temperatures:

t, °C	100	95	90	85	80	75	70	65	60	55	50	45	40	35	30
R, Ohm	380	436	479	530	590	644	718	797	880	989	1114	1251	1406	1604	1810

Hint:

From the theory, resistance of a semiconductor is:

$$R = A*exp(W_a/(kT)),$$

A – constant, it depends on the semiconductor size and its intrinsic properties (namely, concentration of the valence electrons);
k – $0.87 \cdot 10^{-4}$ eV/K – Boltzman's constant;
T – temperature (Kelvin);
W_a – activation energy (electron-volts).