Appendix

to the Physical Chemistry Laboratory Practice

Contents

1 The Atomic Weights 2

2 Fundamental Constants for the Calculations 2

3 Temperature and Concentration Dependence of the Potential of the Calomel Reference Electrode 3

4 Specific Conductivity Values of KCl Solutions at Different Temperatures and Concentrations 3

5 Temperature Dependence of the Density of Water 3

6 Temperature and Ionic Strength Dependence of the Ionic Product of Water 4

7 Preparation of Starch Solution 4

8 Standard Deviation of Data 5

9 Calculation of the Error Propagation 5

10 Slope, intercept and their statistical characteristics of a fitted straight line. Using Excel to determine fitted parameters. 7

10.1 Adding a Trendline 10

10.2 Using the LINEST function 10

10.3 Regression analysis 10

10.4 Using the Solver Add-in 12

11 Plotting mathematical functions for scientific purposes or engineering 15

1 The Atomic Weights

Periodic Table of the Elements

IUPAC ←group number→ CAS

period

atomic number
element symbol
atomic weight (IUPAC, 2007, with up to 5 significant figures)
name

legend

	1 IA	2 IIA	3 IIIB	4 IVB	5 VB	6 VIB	7 VIIB	8 VIIIB	9 VIIIB	10 VIIIB	11 IB	12 IIB	13 IIIA	14 IVA	15 VA	16 VIA	17 VIIA	18 VIIIA
1	1 H 1.0079 hydrogen																	2 He 4.0026 helium
2	3 Li 6.941 lithium	4 Be 9.0122 beryllium											5 B 10.811 boron	6 C 12.011 carbon	7 N 14.007 nitrogen	8 O 15.999 oxygen	9 F 18.998 fluor	10 Ne 20.180 neon
3	11 Na 22.990 sodium	12 Mg 24.305 magnesium											13 Al 26.982 aluminium	14 Si 28.086 silicon	15 P 30.974 phosphorus	16 S 32.065 sulfur	17 Cl 35.453 chlorine	18 Ar 39.948 argon
4	19 K 39.098 potassium	20 Ca 40.078 calcium	21 Sc 44.956 scandium	22 Ti 47.867 titanium	23 V 50.942 vanadium	24 Cr 51.996 chromium	25 Mn 54.938 manganese	26 Fe 55.845 iron	27 Co 58.933 cobalt	28 Ni 58.693 nickel	29 Cu 63.546 copper	30 Zn 65.38 zinc	31 Ga 69.723 gallium	32 Ge 72.61 germanium	33 As 74.922 arsenic	34 Se 78.96 selenium	35 Br 79.904 bromine	36 Kr 83.80 krypton
5	37 Rb 85.468 rubidium	38 Sr 87.62 strontium	39 Y 88.906 yttrium	40 Zr 91.224 zirconium	41 Nb 92.906 niobium	42 Mo 95.96 molybdenum	43 Tc (98) technetium	44 Ru 101.07 ruthenium	45 Rh 102.91 rhodium	46 Pd 106.42 palladium	47 Ag 107.87 silver	48 Cd 112.41 cadmium	49 In 114.82 indium	50 Sn 118.71 tin	51 Sb 121.76 antimony	52 Te 127.60 tellurium	53 I 126.90 iodine	54 Xe 131.29 xenon
6	55 Cs 132.91 cesium	56 Ba 137.33 barium	57 La† 138.91 lanthanum	72 Hf 178.49 hafnium	73 Ta 180.95 tantalum	74 W 183.84 tungsten	75 Re 186.21 rhenium	76 Os 190.23 osmium	77 Ir 192.22 iridium	78 Pt 195.08 platinum	79 Au 196.97 gold	80 Hg 200.59 mercury	81 Tl 204.38 thallium	82 Pb 207.2 lead	83 Bi 208.98 bismuth	84 Po (209) polonium	85 At (210) astatine	86 Rn (222) radon
7	87 Fr (223) francium	88 Ra (226) radium	89 Ac‡ (227) actinium	104 Rf (267) rutherfordium	105 Db (268) dubnium	106 Sg (271) seaborgium	107 Bh (272) bohrium	108 Hs (270) hassium	109 Mt (276) meitnerium	110 Ds (281) darmstadtium	111 Rg (280) roentgenium							

† lanthanoids	58 Ce 140.12 cerium	59 Pr 140.91 praseodymium	60 Nd 144.24 neodymium	61 Pm (145) promethium	62 Sm 150.36 samarium	63 Eu 151.96 europium	64 Gd 157.25 gadolinium	65 Tb 158.93 terbium	66 Dy 162.50 dysprosium	67 Ho 164.93 holmium	68 Er 167.26 erbium	69 Tm 168.93 thulium	70 Yb 173.04 ytterbium	71 Lu 174.97 lutetium
‡ actinoids	90 Th* 232.04 thorium	91 Pa* 231.04 protactinium	92 U* 238.03 uranium	93 Np (237) neptunium	94 Pu (244) plutonium	95 Am (243) americium	96 Cm (247) curium	97 Bk (247) berkelium	98 Cf (251) californium	99 Es (252) einsteinium	100 Fm (257) fermium	101 Md (258) mendelevium	102 No (259) nobelium	103 Lr (262) lawrencium

* The atomic weight cannot be given for the elements have no stable nuclides. For these elements, the value enclosed in parentheses indicates the mass number of the longest-lived isotope of the element. However, there are three exceptions (Th, Pa and U) because they have characteristic composition in the crust of Earth so their atomic weights can be given.

2 Fundamental Constants for the Calculations

Symbol	Value	Quantity
c^0	1 M	standard concentration
F	96485 C/mol	Faraday's constant
p_0	101325 Pa	1 atm pressure expressed in SI unit
p^0	10^5 Pa	standard pressure
R	$8.314 \frac{\mathrm{J}}{\mathrm{mol\,K}}$	gas constant
T_0	–273.15 °C	absolute zero degree

3 Temperature and Concentration Dependence of the Potential of the Calomel Reference Electrode

The potential of calomel electrode (E_{cal}, vs. SHE) can be calculated with ±0.1 mV accuracy in the 0 – 50 °C temperature range and at different KCl concentrations with the

$$\boxed{\mathrm{E_{cal}} = \mathrm{E}^{25\,^\circ\mathrm{C}} - \sum_{i=1}^{3} \mathrm{a_i} \times (\mathrm{t} - 25\,^\circ\mathrm{C})^{\mathrm{i}}}$$

expression where t is the temperature expressed in °C, furthermore empirical constants $\mathrm{E}^{25^\circ\mathrm{C}}$, a_1, a_2 and a_3 are the followings:

[KCl]/M	lg([KCl]/M)	$E^{25\,^\circ C}$/V	a_1/(V/°C)	a_2/(V/°C)	a_3/(V/°C)
0.1	-1	0.3337	8.75×10^{-5}	3.00×10^{-6}	0
1.0	0	0.2801	2.75×10^{-4}	2.50×10^{-6}	4×10^{-9}
3.5	0.5441	0.2500	4.00×10^{-4}	0	0
5.15*	0.7114	0.2412	6.61×10^{-4}	1.75×10^{-6}	9×10^{-10}

*It is the concentration of the saturated KCl solution at 25 °C.

For other concentration of KCl solution, the values of the four empirical constants must be interpolated as functions of the logarithm of the concentration. E.g., in case of [KCl] = 0.5 M, the 10-based logarithm of the concentration is −0.3010 so the

$$\frac{-0.301-(-1)}{0-(-1)} = \frac{\mathrm{E}^{25\,^\circ\mathrm{C}} - 0.3337}{0.2801 - 0.3337} = \frac{a_1 - 8.75 \times 10^{-5}}{2.75 \times 10^{-4} - 8.75 \times 10^{-5}} = \frac{a_2 - 3.00 \times 10^{-6}}{2.50 \times 10^{-6} - 3.00 \times 10^{-6}} = \frac{a_3 - 0}{4 \times 10^{-9} - 0}$$

equations are to be solved to get the appropriate values of $\mathrm{E}^{25\,^\circ\mathrm{C}}$, a_1, a_2 and a_3 in order to use the equation above.

4 Specific Conductivity Values of KCl Solutions at Different Temperatures and Concentrations

t/°C	18	19	20	21	22	23	24	25	26	27	28	29	30
0.01 M KCl	0.001225	0.001251	0.001278	0.001305	0.001332	0.001359	0.001386	0.001413	0.001441	0.001468	0.001496	0.001524	0.001552
0.1 M KCl	0.01119	0.01143	0.01167	0.01191	0.01215	0.01239	0.01264	0.01288	0.01313	0.01337	0.01362	0.01387	0.01412
1.0 M KCl	0.09822	0.10014	0.10207	0.10400	0.10554	0.10789	0.10984	0.11180	0.11377	0.11524	–	–	–

The specific conductivity values are given in $\Omega^{-1}\mathrm{cm}^{-1}$ unit in this table.

5 Temperature Dependence of the Density of Water

The density of water can be calculated with five digits accuracy after the decimal point by the help of the

$$\boxed{\rho_{\mathrm{v}}(\mathrm{t}) = 1.00026 - 5.08692 \times 10^{-6} \times \mathrm{t}^2}$$

empirical formula in the range of 15 °C $\leq t \leq$35 °C. The result is given in g/cm^3 unit at *t* temperature (expressed in °C).

If either different temperature range or higher accuracy is needed then the next (more complicated) empirical formula should be used:

$$\rho_v(t) = a_0 + \sum_{i=1}^{n} a_i \times t^i ,$$

where the next values must be substituted as empirical coefficients:

range	0 – 55 °C	0 – 31 °C	0 – 55 °C	0 – 100 °C
number of precise digits	4	6	5	5
n	3	5	5	10
a_0	0.99987	0.9998406403	0.9998419163	0.99984014
a_1	5.291×10^{-05}	6.801284×10^{-05}	6.694929×10^{-05}	6.8755×10^{-05}
a_2	-7.47×10^{-06}	-9.11644×10^{-06}	-8.91382×10^{-06}	-9.3732×10^{-06}
a_3	3.36×10^{-08}	1.02356×10^{-07}	8.77509×10^{-08}	1.38951×10^{-07}
a_4	–	-1.22323×10^{-09}	-7.80638×10^{-10}	-3.87034×10^{-09}
a_5	–	8.11007×10^{-12}	3.35582×10^{-12}	1.152421×10^{-10}
a_6	–	–	–	$-2.552887 \times 10^{-12}$
a_7	–	–	–	3.700248×10^{-14}
a_8	–	–	–	$-3.290154 \times 10^{-16}$
a_9	–	–	–	1.623754×10^{-18}
a_{10}	–	–	–	-3.3993×10^{-21}

For example, if the density of the water is required with the precision of four digits, the

$$\rho_v(t) = 0.99987 + 5.291 \times 10^{-05} \times 54 - 7.47 \times 10^{-06} \times 54^2 + 3.36 \times 10^{-08} \times 54^3 = 0.9862\,\mathrm{g/cm^3}$$

equation is suitable to calculate it.

6 Temperature and Ionic Strength Dependence of the Ionic Product of Water

The negative logarithm of the ionic product of water is given with two digits accuracy after the decimal point at a given temperature *t* (expressed in °C) and at ionic strength *I* (expressed in molar concentration) by the

$$\boxed{pK_v = 13.99 - 1.02 \times \sqrt{I} - 0.0343 \times (t - 25)}$$

empirical formula in the range of 15 °C $\leq t \leq$30 °C and at ionic strength values less than 0.05 M.

7 Preparation of Starch Solution

For the preparation of 100 cm^3 ~0.5 % starch solution, 0.1 g salicylic acid is solved in about 100 cm^3 boiling water in a ~250 cm^3 Erlenmeyer flask. ~0.5 g starch (made from potato) is shaken with about 10 cm^3 distilled water in a test tube, and this solution is infused into the boiling salicylic acid solution. This mixture is boiled until it is getting lose its translucency and the solution becomes opalescent (no more than two minutes). This solution must be cooled and filtered through cotton wad. This starch solution can be used in about two months if it is stored in fridge. If starch is made from corn then the starch solution can be used only within two weeks. If the starch solution is to be used soon (within 4 – 5 days) then the salicylic acid can be omitted from the above procedure and everything else remains the same.

8 Standard Deviation of Data

It frequently happens during the laboratory exercises that the same value is determined more times from more measurements (e.g., a pseudo-first-order rate coefficient can be calculated from any point of a kinetic curve). These values do not equal to each other completely because of experimental and other uncertainties. Assume that a value is measured m times and let denote the j^{th} data with z_j. In this case, the final (more precisely the most probable) value is regarded as the mean of the individual values, and the value ($\overline{z}$) and its standard deviation ($\sigma_{\overline{z}}$) can be given with the following formulas:

$$\boxed{\overline{z}=\frac{\sum_{j=1}^{m} z_j}{m}} \quad \text{and} \quad \boxed{\sigma_{\overline{z}}=\sqrt{\frac{\sum_{j=1}^{m}(z_j-\overline{z})^2}{m-1}}} = \sqrt{\frac{m\times\sum_{j=1}^{m} z_i^2-\left(\sum_{j=1}^{m} z_i\right)^2}{m\times(m-1)}}.$$

Those quantities above can be calculated, e.g., with MS Office Excel applying the *AVERAGE* and *STDEV* functions. Other spreadsheet softwares are also applicable; look up the appropriate functions.

9 Calculation of the Error Propagation

The calculation of the error propagation (more precisely, the standard deviation propagation) is a frequent task when measured data are evaluated. The most simple approximate rule is well known: the absolute values of the deviations are to be added in cases of addition and subtraction, and the relative values of the deviation are to be added in cases of multiplication and division. This procedure, however, always overestimates the deviation of the result, moreover, it cannot be applied even to the most common function transformations (e.g., square root, logarithm). This section gives those formulas by the help of which the calculation of the deviations can be done correctly.

We assume that there are two data and their deviations are known: $X\pm\sigma_X$ and $Y\pm\sigma_Y$. A result (Z) must be calculated by using one or both of them, and the deviation of Z (σ_Z) is also to be known. Table 1 summarizes the formulas applicable for the basic arithmetic operations and also for the most common function transformations to get the deviation of the result. If the wanted result requires the use of more operation and/or transformations then these formulas can be used one after another to get the final result. For example:

$$\begin{aligned}
\ln(2.0\pm0.1)+(0.4\pm0.02)^{0.5} &= \left(\ln 2\pm\frac{0.1}{2}\right)+\left(0.4^{0.5}\pm(|0.5\times0.02\times0.4^{-0.5}|)\right)\\
&= (0.693\pm0.050)+(0.632\pm0.016)\\
&= (0.693\pm0.632)+\left(\sqrt{0.05^2\pm0.016^2}\right)\\
&= \underline{\underline{1.34\pm0.05\ (\text{or } 1.336\pm0.052)}}
\end{aligned}$$

Table 1: Calculation of the standard deviation during the basic arithmetic operations and applying the most important functions. a denotes the constant, deviationless values in the following formulas. For the trigonometric functions, the values of the angles and their deviations should be given in radian. The other abbreviations are exlained in the text.

operation or function	result and deviation ($Z \pm \sigma_Z$)	example
multiply with a	$(a \times X) \pm (\lvert a \times \sigma_x \rvert)$	$3 \times (1.2 \pm 0.3) = (3 \times 1.2) \pm (3 \times 0.3) = \underline{\underline{3.6 \pm 0.9}}$
addition	$(X+Y) \pm \left(\sqrt{\sigma_X^2 + \sigma_Y^2}\right)$	$(2.2 \pm 0.3) + (8.4 \pm 0.5) =$
		$= (2.2 + 8.4) \pm \left(\sqrt{0.3^2 + 0.5^2}\right) = \underline{\underline{10.6 \pm 0.6}}$
subtraction	$(X-Y) \pm \left(\sqrt{\sigma_X^2 + \sigma_Y^2}\right)$	$(3.2 \pm 0.3) - (2.4 \pm 0.5) =$
		$= (3.2 - 2.4) \pm \left(\sqrt{0.3^2 + 0.5^2}\right) = \underline{\underline{0.8 \pm 0.6}}$
multiplication	$(X \times Y) \pm \left(\sqrt{Y^2 \times \sigma_X^2 + X^2 \times \sigma_Y^2}\right)$	$(2.2 \pm 0.2) \times (8.4 \pm 1.0) =$
		$= (2.2 \times 8.4) \pm \left(\sqrt{8.4^2 \times 0.2^2 + 2.2^2 \times 1.0^2}\right) = \underline{\underline{18.5 \pm 2.8}}$
division	$\left(\frac{X}{Y}\right) \pm \left(\sqrt{\frac{Y^2 \times \sigma_X^2 + X^2 \times \sigma_Y^2}{Y^4}}\right)$	$(22.0 \pm 2.0)/(8.4 \pm 1.0) =$
		$= \frac{22.0}{8.4} \pm \left(\sqrt{\frac{8.4^2 \times 2.0^2 + 22.0^2 \times 1.0^2}{8.4^4}}\right) = \underline{\underline{2.6 \pm 0.4}}$
reciprocal	$\left(\frac{1}{X}\right) \pm \left(\frac{\sigma_X}{X^2}\right)$	$\frac{1}{(0.44 \pm 0.12)} = \left(\frac{1}{0.44}\right) \pm \left(\frac{0.12}{0.44^2}\right) = \underline{\underline{2.3 \pm 0.6}}$
raising	$(X^a) \pm \left(\lvert a \times \sigma_X \times X^{a-1} \rvert\right)$	$(3.0 \pm 0.5)^{1.2} = (3.0^{1.2}) \pm \left(\lvert 1.2 \times 0.5 \times 3.0^{1.2-1} \rvert\right) = \underline{\underline{3.7 \pm 0.7}}$
exponential functions	$(\mathrm{e}^X) \pm (\sigma_X \times \mathrm{e}^X)$	$\mathrm{e}^{2.0 \pm 0.5} = (\mathrm{e}^{2.0}) \pm (0.5 \times \mathrm{e}^{2.0}) = \underline{\underline{7.4 \pm 3.7}}$
	$(10^X) \pm (\ln(10) \times \sigma_X \times 10^X)$	$10^{1.3 \pm 0.1} = (10^{1.3}) \pm (2.3 \times 0.1 \times 10^{1.3}) = \underline{\underline{20 \pm 5}}$
logarithmic functions	$(\ln X) \pm \left(\frac{\sigma_X}{X}\right)$	$\ln(2.0 \pm 0.1) = (\ln(2.0)) \pm \left(\frac{0.1}{2.0}\right) = \underline{\underline{0.69 \pm 0.05}}$
	$(\lg X) \pm \left(\frac{\sigma_X}{\ln(10) \times X}\right)$	$\lg(20 \pm 10) = (\lg(20)) \pm \left(\frac{10}{2.3 \times 20}\right) = \underline{\underline{1.3 \pm 0.2}}$
trigonometric functions	$(\sin X) \pm (\lvert \cos X \rvert \times \sigma_X)$	$\sin(60^\circ \pm 5^\circ) = \left(\sin\frac{\pi}{3}\right) \pm \left(\left\lvert \cos\frac{\pi}{3} \right\rvert \times \frac{5 \times \pi}{180}\right) = \underline{\underline{0.87 \pm 0.04}}$
	$(\cos X) \pm (\lvert \sin X \rvert \times \sigma_X)$	$\cos(60^\circ \pm 5^\circ) = \left(\cos\frac{\pi}{3}\right) \pm \left(\left\lvert \sin\frac{\pi}{3} \right\rvert \times \frac{5 \times \pi}{180}\right) = \underline{\underline{0.5 \pm 0.08}}$
	$(\mathrm{tg}\mathrm{X}) \pm \left(\frac{\sigma_X}{(\cos X)^2}\right)$	$\mathrm{tg}(45^\circ \pm 5^\circ) = \left(\mathrm{tg}\frac{\pi}{4}\right) \pm \left(\frac{5 \times \pi}{180} \Big/ \left(\cos\frac{\pi}{4}\right)^2\right) = \underline{\underline{1{,}0 \pm 0{,}2}}$
inverse trigonometric functions	$\arcsin X \pm \left(\frac{\sigma_X}{\sqrt{1-X^2}}\right)$	$\arcsin(0.87 \pm 0.08) =$
		$= \left(\arcsin(0.87) \pm \left(\frac{0.08}{\sqrt{1-0.87^2}}\right)\right) \times \frac{180}{\pi} = \underline{\underline{60^\circ \pm 9^\circ}}$
	$\arccos X \pm \left(\frac{\sigma_X}{\sqrt{1-X^2}}\right)$	$\arccos(0.5 \pm 0.08) =$
		$= \left(\arccos(0.5) \pm \left(\frac{0.08}{\sqrt{1-0.5^2}}\right)\right) \times \frac{180}{\pi} = \underline{\underline{60^\circ \pm 5^\circ}}$
	$\mathrm{arctg} X \pm \left(\frac{\sigma_X}{1+X^2}\right)$	$\mathrm{arctg}(1{,}0 \pm 0{,}2) =$
		$= \left(\mathrm{arctg}(1.0) \pm \left(\frac{0.2}{1+1.0^2}\right)\right) \times \frac{180}{\pi} = \underline{\underline{45^\circ \pm 6^\circ}}$

10 Slope, intercept and their statistical characteristics of a fitted straight line. Using Excel to determine fitted parameters.

Values to be determined at the practices are often calculated from the slope and / or the intercept obtained from fitting a straight line. In this section, we summarize, without derivation, the formulas that allow the calculation of the parameters of the fitted line, as well as their standard deviations with a simple calculator. However, before giving the formulas, two important remarks have to be made:

1. The formulas given here may seem complicated at first glance, but they are actually simple to use. This is easy to see if the reader calculates the example detailed below using a calculator. An additional ease may be that most scientific calculators already compute statistical functions. With these, the calculations can be made even faster, because the statistical method of scientific calculators automatically gives the partial results required during the calculations.
2. Many programs (e.g., spreadsheets) can fit straight lines. These are likely to be used more frequently at laboratory practices. However, these programs most often give *not* the standard deviation of the slope and the intercept, but their standard error as a statistical parameter, and they are not the same! Sometimes, (seemingly), the program calculates a standard deviation but actually results in the standard error. The relationship between standard deviation and standard error is given by

$$\text{st. deviation} = \sqrt{\text{number of data}} \times \text{st. error} \tag{1}$$

The following notations are used in the formulas below:

n the number of data pairs used,

x_i value of the *independent* variable in the *i*-th data pair ($i = 1 \dots n$),

y_i value of the *dependent* variable in the *i*-th data pair ($i = 1 \dots n$),

a slope of the fitted staight line ($y = a \times x + b$ or $y = a \times x$),

b intercept of the fitted straight line ($y = a \times x + b$),

σ_a standard deviation of the slope of the fitted line,

σ_b standard deviation of the intercept of the fitted line,

S_x is the sum of x_i,

S_y is the sum of y_i,

S_{xy} is the sum of the product of x_i and y_i,

S_{xx} is the sum of x_i square,

S_Δ is the sum of the square of the difference of y_i (measured data), and the fitted (calculated data) of $y_i = (a \times x_i + b)$.

$$\boxed{S_x = \sum_{i=1}^{n} x_i, \quad S_y = \sum_{i=1}^{n} y_i, \quad S_{xy} = \sum_{i=1}^{n} x_i \times y_i, \quad S_{xx} = \sum_{i=1}^{n} x_i^2 \quad \text{and} \quad S_\Delta = \sum_{i=1}^{n} (y_i - a \times x_i - b)^2 .}$$

Based on these, the slope of the fitted line (a) and the standard deviation of the slope (σ_a) if the line passes through the origin ($b = 0$):

$$\boxed{a = \frac{S_{xy}}{S_{xx}}} \quad = \quad \left(\sum_{i=1}^{n} x_i \times y_i\right) \Big/ \left(\sum_{i=1}^{n} x_i^2\right) \tag{2}$$

$$\boxed{\sigma_a = \sqrt{\frac{n}{n-1} \times \frac{S_\Delta}{S_{xx}}}} \quad = \quad \sqrt{\frac{n}{n-1} \times \frac{\sum_{i=1}^{n} (y_i - a \times x_i)^2}{\sum_{i=1}^{n} x_i^2}}\,. \tag{3}$$

If the intercept of the fitted line is not $b = 0$, the slope and standard deviation are:

$$\boxed{a = \frac{n \times S_{xy} - S_x \times S_y}{n \times S_{xx} - S_x^2}} \quad = \quad \frac{n \times \sum_{i=1}^{n} x_i \times y_i - \left(\sum_{i=1}^{n} x_i\right) \times \left(\sum_{i=1}^{n} y_i\right)}{n \times \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \tag{4}$$

$$\boxed{\sigma_a = \sqrt{\frac{n^2}{n-2} \times \frac{S_\Delta}{n \times S_{xx} - S_x^2}}} \quad = \quad \sqrt{\frac{n^2}{n-2} \times \frac{\sum_{i=1}^{n} (y_i - a \times x_i - b)^2}{n \times \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}}\,. \tag{5}$$

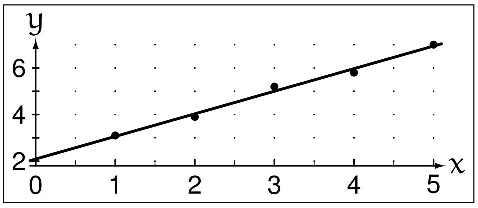
The intercept of the fitted line (b) and the standard deviation of the intercept (σ_b):

$$\boxed{b = \frac{S_{xx} \times S_y - S_x \times S_{xy}}{n \times S_{xx} - S_x^2}} \quad = \quad \frac{\left(\sum_{i=1}^{n} x_i^2\right) \times \left(\sum_{i=1}^{n} y_i\right) - \left(\sum_{i=1}^{n} x_i\right) \times \left(\sum_{i=1}^{n} x_i \times y_i\right)}{n \times \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} \tag{6}$$

$$\boxed{\sigma_b = \sqrt{\frac{n \times S_{xx}}{n-2} \times \frac{S_\Delta}{n \times S_{xx} - S_x^2}}} \quad = \quad \sqrt{\frac{n \times \sum_{i=1}^{n} x_i^2}{n-2} \times \frac{\sum_{i=1}^{n} (y_i - a \times x_i - b)^2}{n \times \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}}\,. \tag{7}$$

In the following, the calculation technique of straight-line fitting is illustrated by means of a detailed example using the notations defined above. The data pairs used for straight-line matching and their graphical representation are as follows:

i:	1	2	3	4	5
x_i:	1.0	2.0	3.0	4.0	5.0
y_i:	3.1	3.9	5.2	5.8	7.0



The partial results of the calculations:

$$\begin{aligned}
S_x &= 1.0+2.0+3.0+4.0+5.0 = 15.0\\
S_y &= 3.1+3.9+5.2+5.8+7.0 = 25.0\\
S_{xy} &= 1.0\times 3.1+2.0\times 3.9+3.0\times 5.2+4.0\times 5.8+5.0\times 7.0 = 84.7\\
S_{xx} &= 1.0^2+2.0^2+3.0^2+4.0^2+5.0^2 = 55.0\\
n\times S_{xx}-S_x^2 &= 5\times 55.0-15.0^2 = 50.0
\end{aligned}$$

Slope and intercept values from (4) and (6):

$$a = \frac{5\times 84.7-15.0\times 25.0}{50.0} = 0.97 \qquad \text{and} \qquad b = \frac{55.0\times 25.0-15.0\times 84.7}{50.0} = 2.09$$

Another partial result:

$$\begin{aligned}
S_\Delta &= (3.1-0.97\times 1.0-2.09)^2+(3.9-0.97\times 2.0-2.09)^2+(5.2-0.97\times 3.0-2.09)^2+\\
&+ (5.8-0.97\times 4.0-2.09)^2+(5.9-0.97\times 5.0-2.09)^2 = 0.091
\end{aligned}$$

Values of standard deviation of slope and intercept from (5) and (7):

$$\sigma_a = \sqrt{\frac{5^2}{5-2}\times\frac{0.091}{50.0}} = 0.12 \quad \text{and} \quad \sigma_b = \sqrt{\frac{5\times 55.0}{5-2}\times\frac{0.091}{50.0}} = 0.41$$

Instead of (or in addition to) the standard deviation of the fitted parameters, we often use the so-called correlation coefficient (R) to see the goodness of the fitting. Introducing, in a manner analogous to the previous ones, the

$$\mathrm{S_{yy}} = \sum_{\mathrm{i=1}}^{\mathrm{n}} \mathrm{y_i^2}$$

notation, the correlation coefficient

$$\mathrm{R} = \frac{\mathrm{n\times S_{xy}-S_x\times S_y}}{\sqrt{\mathrm{(n\times S_{xx}-S_x^2)\times(n\times S_{yy}-S_y^2)}}}.$$

It is trivial that the correlation coefficient can range from -1 to $+1$. The higher the absolute value, the more linear the relationship, i.e., the more correlated the data. $\mathrm{R} = 0$ corresponds to the complete absence of correlation. Based on the data in the previous example, $\mathrm{R} = 0.9952$, which is a good fit.

Instead of the correlation coefficient, it is more common to use its square, the determination (regression) coefficient (R^2):

$$\mathrm{R^2} = \frac{\mathrm{(n\times S_{xy}-S_x\times S_y)^2}}{\mathrm{(n\times S_{xx}-S_x^2)\times(n\times S_{yy}-S_y^2)}} = 1-\frac{\mathrm{n\times S_\Delta}}{\mathrm{n\times S_{yy}-S_y^2}}.$$

Based on the data in the previous example, $\mathrm{R}^2 = 0.9904$. R^2 allows quick diagnostics. If $\mathrm{R}^2 \approx 1$ (S_Δ is small, good correlation), then the relationship $\mathrm{x-y}$ is really linear. If $\mathrm{R}^2 \approx 0$, (S_Δ is very large) then there is certainly no correlation. In other words, in the case of a straight line, we minimize S_Δ, that is, we maximize R^2.

It would be time consuming to do all those calculations for all the practices manually (with a calculator) as described above. That's why a variety of spreadsheets programs was developed and everyone can use that he / she likes best or that, what is available. At the *Education Level* of the Institute of Chemistry those are QtiPlot and MS Office Excel. In the following, we present possibilities of straight-line fitting (linear regression) using MS Excel. (For other spreadsheet programs, you have to find the suitable functions.)

10.1 Adding a Trendline

Recommended for illustration mainly! If you already have a figure of the measured data, you can request different *forecasts*, *function relationships*, so-called trendlines. If the *linear* trendline is selected, the program calculates the best-fitting line for the data series, and it is also possible to print the parameters of the fitted line (slope, intercept and R^2). You can also choose a single-parameter fit (if this corresponds to the principle of measurement), where you enter the value of the intercept and the program only calculates the slope. The advantage of adding a trendline is that it is fast and eye-catching. The disadvantage, it does not provide statistics and the calculation of R^2 is incorrect in single-parameter mode (program error!). From the examples shown in Figure 1, it can be seen that the determination coefficient (the square of a real number) became negative for the second data set! In addition to the fact that this is a bug, it also draws attention to the fact that a straight line cannot be fitted to the data, and it is worth checking which points were measured – reported – typed incorrectly!

10.2 Using the LINEST function

This is the most useful / fastest / simplest version of Excel functions for line fitting, which gives statistical parameters in addition to the slope and intercept of the line. *The LINEST* function uses the least squares method to calculate the equation of the line that best fits the given data and returns the array describing the line as a result. You can also use *LINEST* with other functions to compute statistics for other types of models with linear unknown parameters (such as logarithmic, polynomial, exponential, and power series). Since this function returns an array, it must be entered as an array formula. That is, in uni-variable case (there is only one independent variable, x-data series) select an array of 2 (columns)×5 (rows) cells, enter the function:

$$= \mathrm{LINEST(known\_y's; known\_x's; const; stats)} \quad ,$$

where

known_y's is the set of y-values that you already know in the relationship y = a x+b,

known_x's is the set of x-values that you may already know in the relationship y = a x+b,

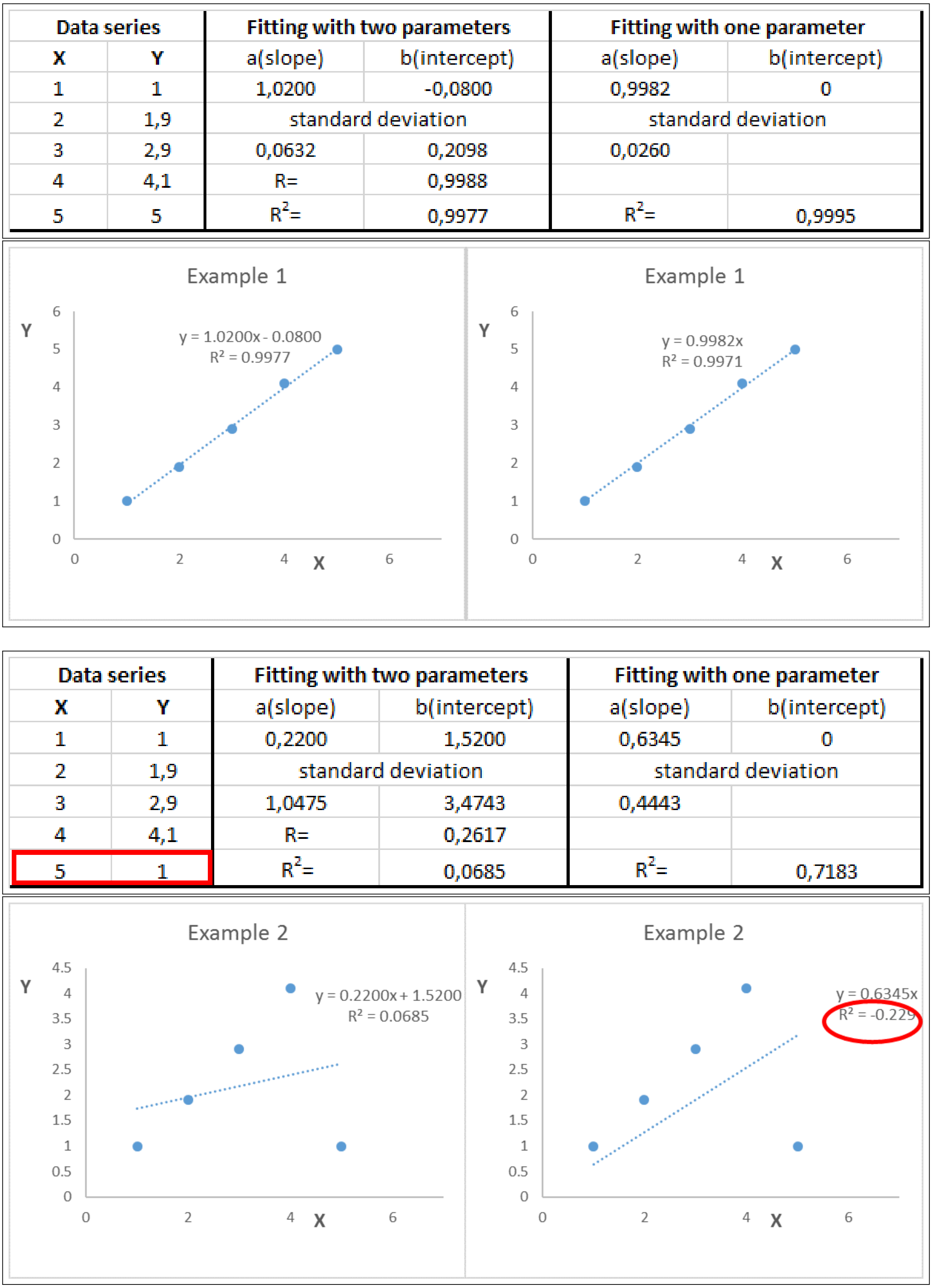
[const] is a logical value specifying whether to force the constant *b* (intercept) to equal 0. If const is TRUE (or 1), *b* is calculated normally. If const is FALSE (or 0), *b* is set equal to 0 and the slope (*a*) is adjusted to fit y = a x,

[stats] is a logical value specifying whether to return additional regression statistics. If stats is TRUE (or 1), LINEST returns the additional regression statistics. If stats is FALSE (or 0), LINEST returns only the slope (*a*) and the intercept (*b*).

Since *LINEST* is a array function, after entering the values, press CTRL+SHIFT+ENTER at the same time. The table shown in Figure 2 give help on the displayed values. The function gives the standard error of the intercept, from which the standard deviation can be calculated from the equation (1). Obviously, the function gives the same result as the *manual calculation* with statistical parameters.

10.3 Regression analysis

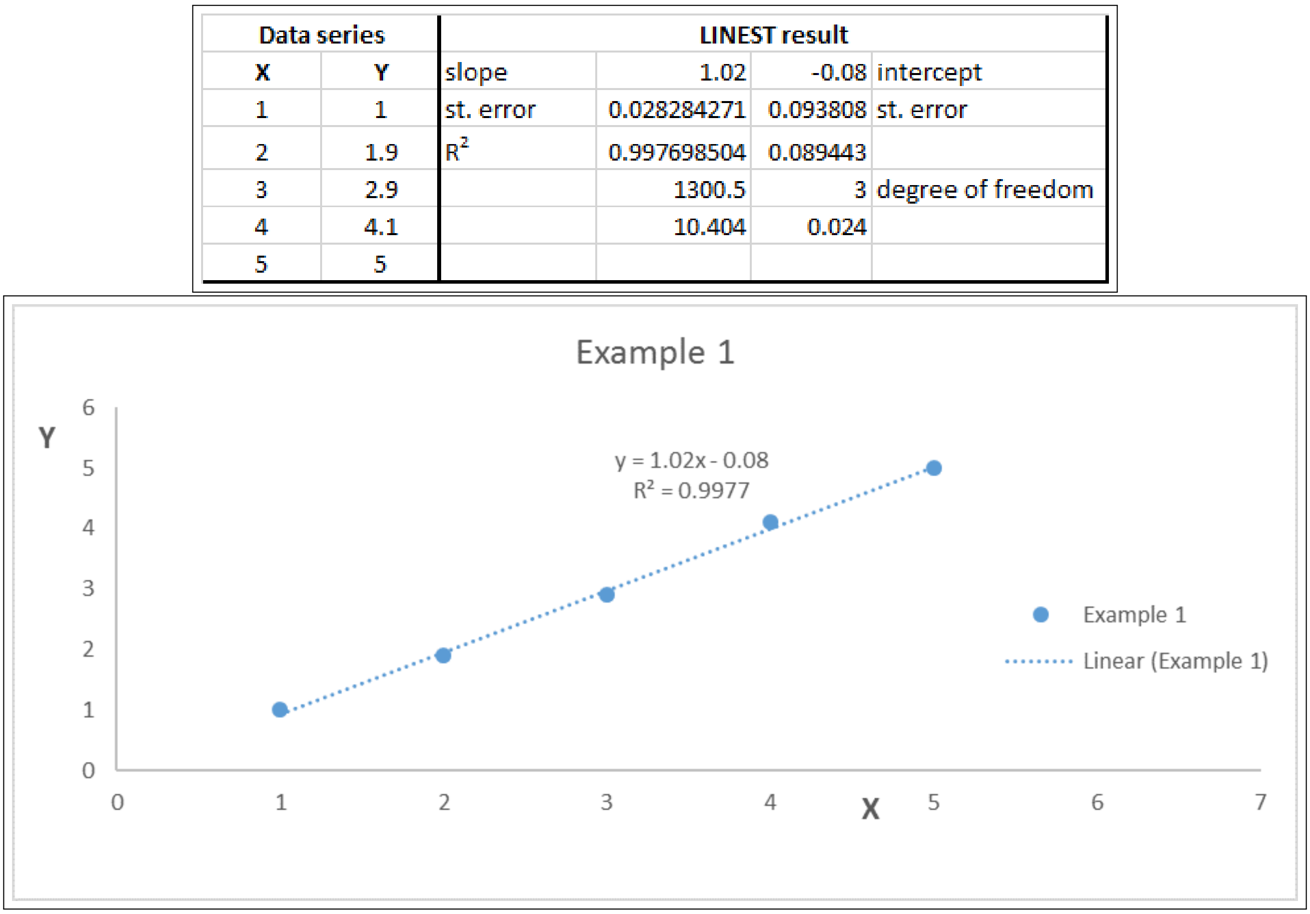
In the basic version of Excel, under the File / Options / Add-ins menu, you will find some additional applications that can be activated. One of these is the *Analysis ToolPak*, which provides data analysis tools for statistical and engineering analysis. If this is activated, opening the DATA *tab* will display a *Data Analysis* function in the last column. It contains (among many other statistical function) the *Regression*. This gives the most detailed statistics about a straight line fitting. Note that its algorithm is different from the calculation of *LINEST*, so it may give slightly different results.



Data series		Fitting with two parameters		Fitting with one parameter	
X	Y	a(slope)	b(intercept)	a(slope)	b(intercept)
1	1	1,0200	-0,0800	0,9982	0
2	1,9	standard deviation		standard deviation	
3	2,9	0,0632	0,2098	0,0260	
4	4,1	R=	0,9988		
5	5	R^2=	0,9977	R^2=	0,9995

Data series		Fitting with two parameters		Fitting with one parameter	
X	Y	a(slope)	b(intercept)	a(slope)	b(intercept)
1	1	0,2200	1,5200	0,6345	0
2	1,9	standard deviation		standard deviation	
3	2,9	1,0475	3,4743	0,4443	
4	4,1	R=	0,2617		
5	1	R^2=	0,0685	R^2=	0,7183

Figure 1: Illustration of using trend line for straight-line fitting to a good (top panel) and unsuitable (bottom panel) data series. The fits were made according to the formulas given earlier.



Data series		LINEST result			
X	Y	slope	1.02	-0.08	intercept
1	1	st. error	0.028284271	0.093808	st. error
2	1.9	R^2	0.997698504	0.089443	
3	2.9		1300.5	3	degree of freedom
4	4.1		10.404	0.024	
5	5				

Figure 2: Illustration of the use of *LINEST* in a two-parameter case for a data set corresponding to the previous straight-line fitting.

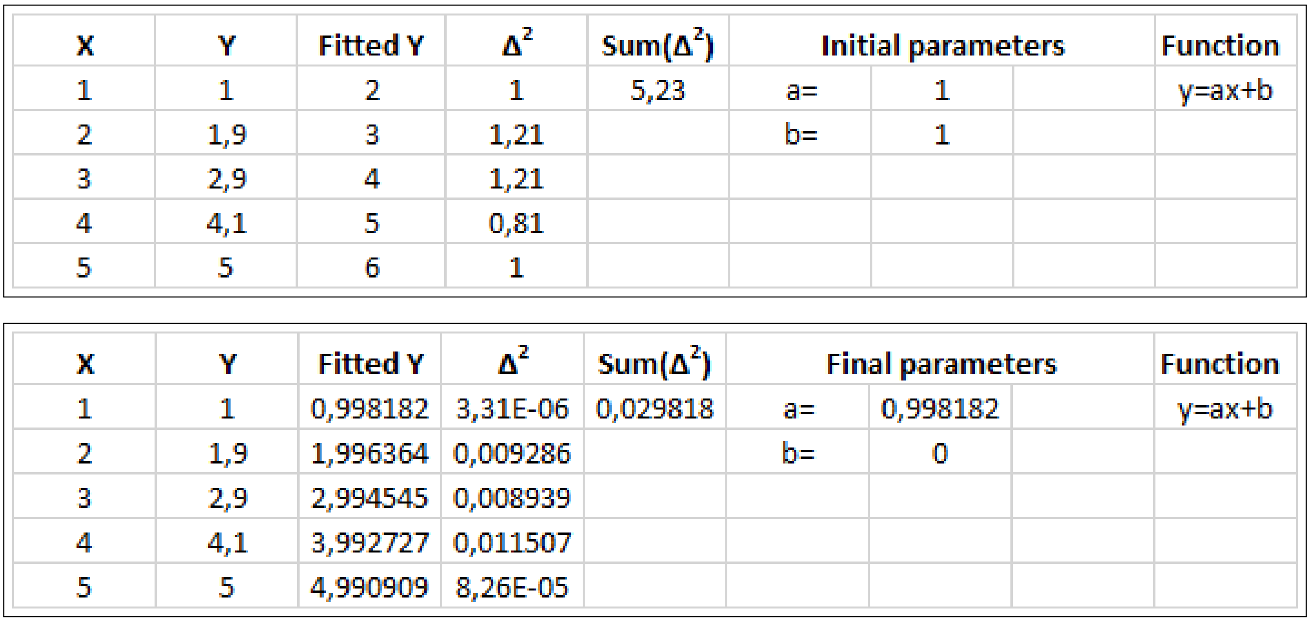
10.4 Using the Solver Add-in

The *Solver* extension can be activated in the same way as the *Analysis ToolPak* and will appear in the same place in the Excel menu. Here you have to build the analysis, e.g., using the least squares method. That is, *Solver* should be used to minimize the sum of the squares of the differences between the (measurement) data and the fitted values to obtain the two parameters (intercept and slope) that represent the best-fit line.

Figure 3 shows an example of using *Solver*. The input data includes the data set to be matched (X and Y) and the *Initial Parameters* that must be estimated (give a hint on common sense). Based on the *Function*, the program calculates the *Fitted Dependent Variable (Y)*. Δ is the difference between the fitted and the measured data. When using *Solver*, *"Set Objective"* is the *Sum*(Δ^2), which is to be minimized. *Variable Cells* are the values of *a* and *b* that are changed to reach the minimum of the *Objective*. The values obtained after running *Solver* (Figure 3, bottom panel) are about the same as previously (within the standard deviation). Now, you programmed the least squares method, and did not used the explicit formula for straight line fitting.

Despite the example presented earlier, the use of *Solver* is not about linear regression, as a line can be fitted in many ways. Note that the *Function* is defined by you (*Fitted Y*), meaning that the parameters of any nonlinear function can be calculated using the same method. There is also no limit to the number of parameters to be fitted. *Important: In the case of a linear fit, the wrong choice of the initial parameters does not cause an error, there will always be a solution. For a nonlinear fit, a bad initial parameter can mean that the calculation does not converge, there will be no real solution. The initial parameters must be chosen on a physico-chemical basis!* The nonlinear fitting is performed in a manner analogous to that described for linear regression, as illustrated in Figure 4. If the initial and fitted data differ significantly, the trend cannot be described with the specified *Function* and the result is hardly acceptable.

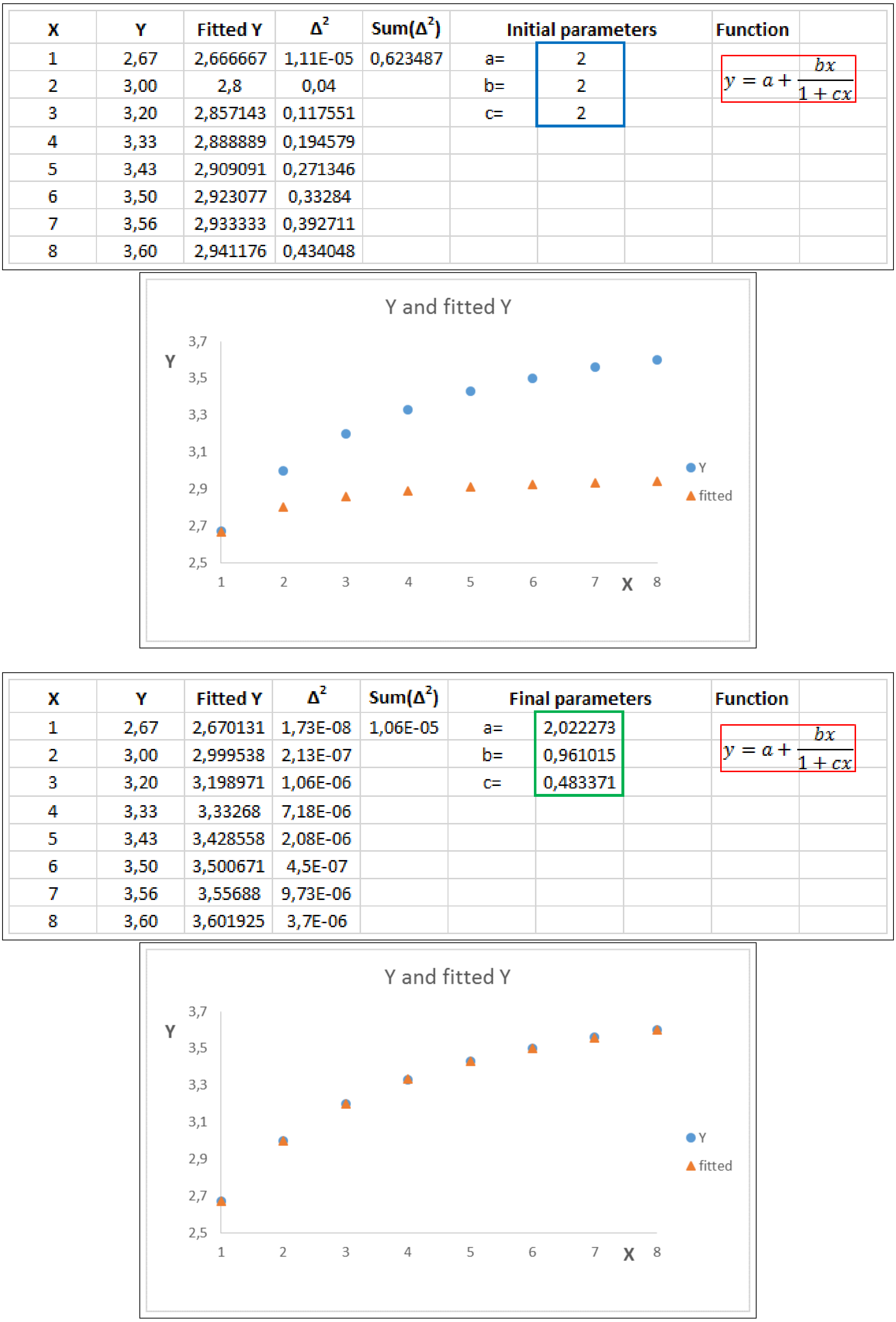
Solver can be used in the same way / similarly to solve nonlinear equations.



X	Y	Fitted Y	Δ^2	Sum(Δ^2)	Initial parameters			Function
1	1	2	1	5,23	a=	1		y=ax+b
2	1,9	3	1,21		b=	1		
3	2,9	4	1,21					
4	4,1	5	0,81					
5	5	6	1					

X	Y	Fitted Y	Δ^2	Sum(Δ^2)	Final parameters			Function
1	1	0,998182	3,31E-06	0,029818	a=	0,998182		y=ax+b
2	1,9	1,996364	0,009286		b=	0		
3	2,9	2,994545	0,008939					
4	4,1	3,992727	0,011507					
5	5	4,990909	8,26E-05					

Figure 3: Demonstration of using Excel *Solver*: before (top panel) and after running (bottom panel).



X	Y	Fitted Y	Δ^2	Sum(Δ^2)	Initial parameters			Function	
1	2,67	2,666667	1,11E-05	0,623487	a=	2		$y = a + \frac{bx}{1+cx}$	
2	3,00	2,8	0,04		b=	2			
3	3,20	2,857143	0,117551		c=	2			
4	3,33	2,888889	0,194579						
5	3,43	2,909091	0,271346						
6	3,50	2,923077	0,33284						
7	3,56	2,933333	0,392711						
8	3,60	2,941176	0,434048						

X	Y	Fitted Y	Δ^2	Sum(Δ^2)	Final parameters			Function	
1	2,67	2,670131	1,73E-08	1,06E-05	a=	2,022273		$y = a + \frac{bx}{1+cx}$	
2	3,00	2,999538	2,13E-07		b=	0,961015			
3	3,20	3,198971	1,06E-06		c=	0,483371			
4	3,33	3,33268	7,18E-06						
5	3,43	3,428558	2,08E-06						
6	3,50	3,500671	4,5E-07						
7	3,56	3,55688	9,73E-06						
8	3,60	3,601925	3,7E-06						

Figure 4: The use of *Solver* for nonlinear fitting: before (top panel) and after running (bottom panel).

11 Plotting mathematical functions for scientific purposes or engineering

The requirements related to figures / graphs are the same whether they were made by computer or on millimeter paper:

- All data or their derivatives should be presented.
- Both the figure and the axes should contain titles with the respective units (if any). Be grammarly and use correct terminology. Indicate your name and the date as well.
- The axis ticks should be aesthetic and should allow the easy reading of the (x – y) values of the data points. Aim for minimizing the empty spaces on the graph. This latter rule should be applied for the specific figure. For example, for the most appropriate demonstration of linear fitting, it is beneficial to show the intercept, even when it is outside the range of the measured values.
- When curve fitting is required, the figure should contain all data points, irrespective of whether they were included in the fitting or they were omitted. The latter points should be presented with a different mark. Indicate also the fitted curve along with the fitted parameters!
- When you plot more curves/data points, these should be clearly distinguishable from each other (even in greyscale)!

Surely, in some cases there are exceptions: when a point differs by magnitudes from the other ones, you cannot make a meaningful figure including all points. Therefore, scientific softwares, which make a first-case automatic plot on data should not be fully relied on – the consideration of the individual researcher is not supplemented by their artificial intelligence. Especially, a large part of the commercially available programs is routinely used for graphs in economy etc, and adjusts the figure for representative purposes and not to scientific precision.

In the following, we will show the general mistakes, which are most abundantly committed while preparing graphs for scientific purpose. The top and bottom panel of Figure 5 illustrates the linear fitting to the same dataset. The one on the top panel fully meets the criteria detailed above, while the one at the bottom shows (based our experience) the most common mistakes. These can be easily avoided by an adequate knowledge of your computer software. The next parts of this appendix aims to guide you in this direction:

Automatic connection of data points: The default setting for almost all programs is the connection of plotted data points by straight lines. This does not make any deep sense in most cases, they are only used for guiding the eyes to visualize trends better, mostly on graphs in economy. In natural science it is customary (and wise) to add a continuous curve, because in case the data points are not increasing successively (as for the bottom panel), a meaningless set of lines will be produced.

Wrong range of axis values: Some softwares automatically includes the origin of the coordinate system. This, depending on the span of data point coordinates, can lead to the shrinkage of the area with the data points. Now, some meaningful information cannot be retrieved from the graph.

Uneven axis range: This problem is similar to the former one, but not exactly the same. Many softwares render the minimum and maximum axis values to the respective minimum and maximum data points. For examples, on the bottom panel, the range of y-axis is not comfortable because the range of 13 – 120 cannot be divided by integer numbers, especially not to decadic values (10, 20, 30, . . . etc.). Moreover, the tick labels are incorrect because the decimals are not indicated. This results in incorrect readings of the data values, and the user needs to know how to set the minimal and maximal axis values and the tick labels.

Automatic choice of axis range: Because of this choice, false axis ticks and titles may occur. On the bottom panel, the tick labels of the x-axis are missing, the meaningless automatic axis titles are composed of the name of the data file and the enumerated columns, while the data at the side almost "fall" down from the figure.

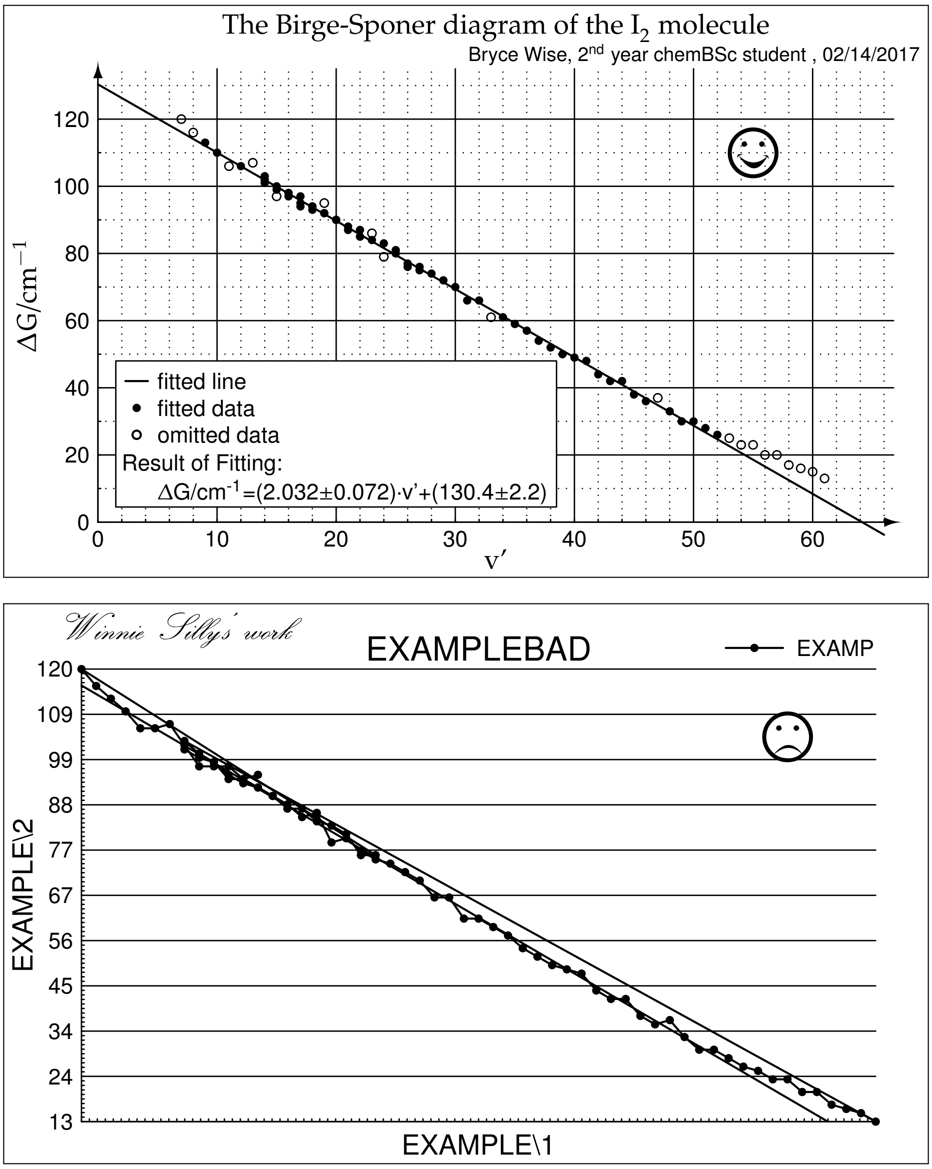


Figure 5: A quality (top panel) graph and one featuring the typical errors (bottom panel).

Meaningless main title: It may become more difficult to understand the figure with a meaningless main title, especially if there is a long time between the preparation and the interpretation of the figure. As the default for many softwares, the main title is associated with the file name containing the graphical setting.

Absence of the name, title and the date: This may cause annoying loss of information. In the given example, the date is missing.

Incorrect positioning: In a more fortunate case, this is only comical but in worst cases, it will lead also to a loss of information. In the demonstrated case, half of the explanation box (legend) is unseen.

Automatic legend: The automatic legend generally is not informative. Better if we either do not use it, or we should fill it with true content. It makes use in the cases when multiple curves are indicated in the figure and we want to help the understanding by using these short points.

Grids: When indicated, these need to be adjusted carefully. The overly dense grid system does not aid the understanding of the figure, because it practically covers the curves and data points. If the gridlines stand too loosely, the data points are more difficult to be read. In many cases, the figure is clearer in the absence of the gridlines. Not a great choice as well to indicate only the vertical or the horizontal gridlines.

Unfeasible letter style / size: In a better case, this leads only to an ugly or to a comical appearance, but in worse cases it can also cause ambiguity. The name on the lower panel is represented by such letters. It is more expedient to use simpler and thicker fonts such as Swiss, Arial, Helvetica, Tahoma, Verdana, Calibri, etc.

The omitted points should still be presented on the figure: If we do not do this, we will loose information on the precision of the measurements and the possible reasons behind the omission of the data points. If we mark the wrong points with those used for a fitting (as it is seen on the wrong figure), the reproduction of the calculated data will be problematic.