Appendix

to the General Chemistry Laboratory Practice

Edited by Dr. Gábor Schuszter

Contents

1 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). Use the correct (even grammatically correct!) terminology. Indicate your name and date as well.
- The ticks on the axes should be aesthetic and allow easy reading of the (x,y) values of the data points. Attempt to minimize the empty spaces on the graph. The 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 not. The omitted points should be presented with a different mark. Indicate also the fitted curve along with the fitted parameters.
- When plotting more curves/data points, these should be clearly distinguishable from each other (even in grayscale).

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 panels of Figure 1 illustrate the linear fitting to the same data set. The one in the top panel fully meets the criteria detailed above, while the one in the bottom shows (based on our experience) the most common mistakes. These can be easily avoided by having adequate knowledge of your computer software. The next parts of this appendix aim 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 to guide the eyes to visualize trends better, mostly on graphs in the economy. In natural science, it is customary (and wise) to add a continuous curve because in case the data points do not increase successively (as in 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 the coordinates of the data point, 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 software render the minimum and maximum axis values to the respective minimum and maximum data points. For example, on the bottom panel, the range of the 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 minimum and maximal axis values and the tick labels.

Automatic choice of axis range: Because of this choice, false axis ticks and titles may occur. In the bottom panel, the tick labels of the x-axis are missing, and the meaningless automatic axis titles are composed

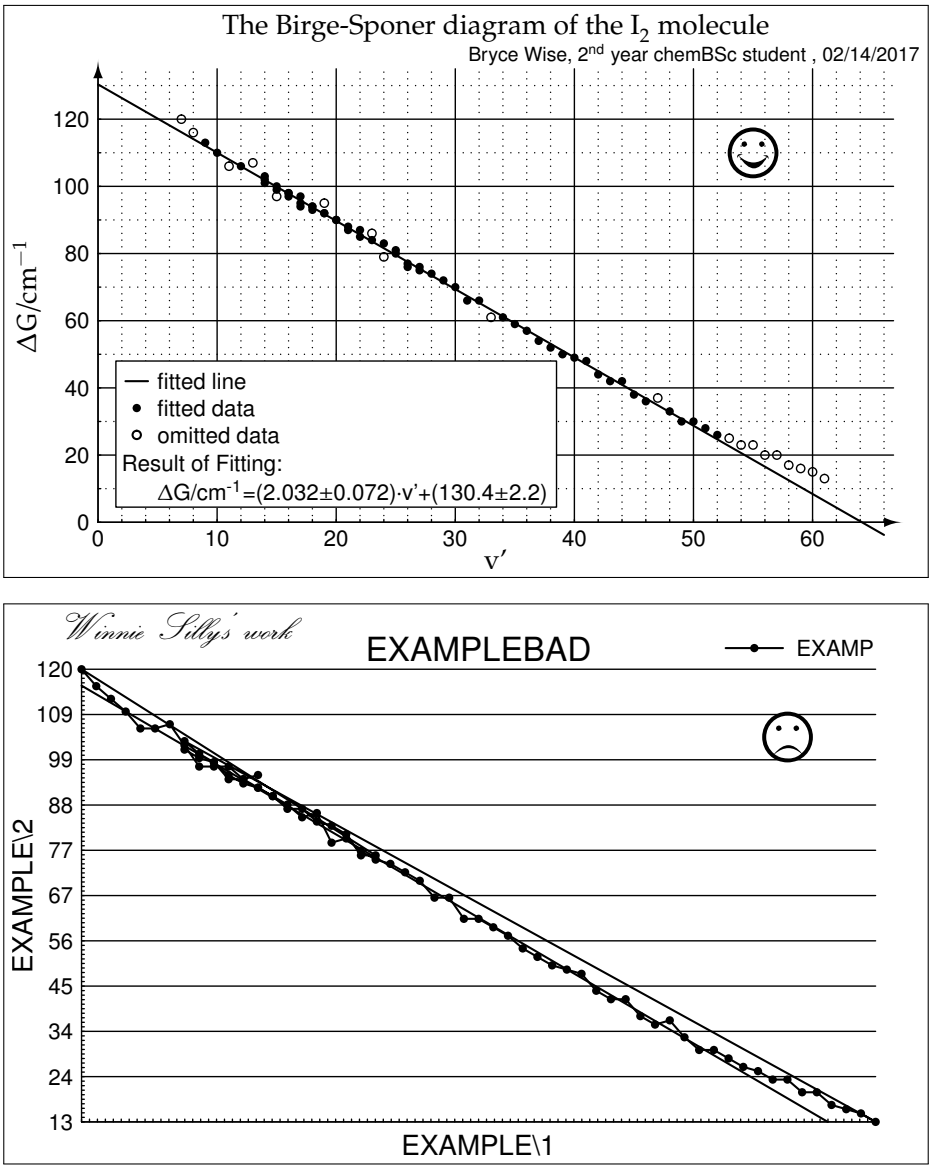

Figure 1: A quality (top panel) graph and one featuring the typical errors (bottom panel).

of the name of the data file and the enumerated columns, while the data at the side almost "fall" down from the figure.

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 information loss. 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 also lead to a loss of information. In the demonstrated case, half of the explanation box (legend) is unseen.

Automatic legend: The automatic legend is generally not informative. Better if we do not use it or fill it with true content. It makes use in the cases where 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 carefully adjusted. 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 are too loose, the data points are more difficult to 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 in 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 lose 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 is seen on the wrong figure), the reproduction of the calculated data will be problematic.

2 Linear regression and its statistical characteristics in MS Excel

2.1 Basics

2.1.1 Principle of linear regression

- By *"fitting a line"* , we mean linear regression, or in other words, linear parameter fitting (Fig. 2, left panel).
- There are two options for linear regression: 1) $y = a \cdot x$ (*i.e.*, y-intercept is the origin) and 2) $y = a \cdot x + b$ (*i.e.*, y-intercept to be fitted). The principle of the measurement determines which of those should be used. For example, consider the Beer–Lambert law. Since $A = \varepsilon c \ell$, for an $A - c$ function $y = 0$ if $x = 0$ (if a compound is not present, it should not absorb light).
- Most software uses the method of least squares to find the best fitting line. This is done through the minimization of the sum of the squares of the residuals, where a residual is the difference between a measured value and the fitted value provided by a model (see Fig. 2, right panel). The method is not exclusive for linear regression (*e.g.*, polynomial, exponential etc. can be fitted as well).
  Recall: During General Chemistry Laboratory Practice, if you had to make a plot by hand on mm-paper, the instruction was to "draw the line in between the measured points such that the line passes equally close to them". This was a somewhat preliminary version of the least squares method.

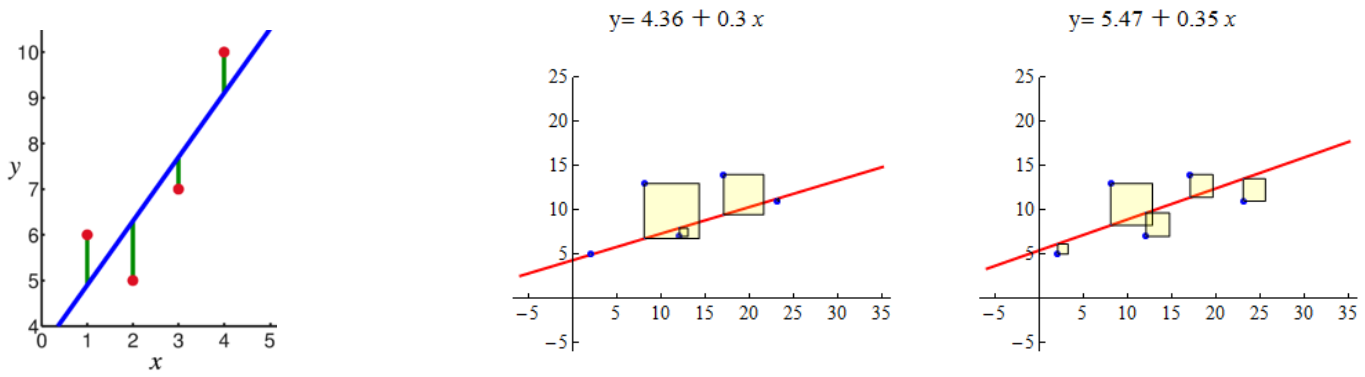

Figure 2: **Left panel**: Linear regression: measured points (red bullets), assumed relation, *i.e.*, the fitted function (blue line), and their difference with random distribution (green verticals). **Right panel**: Linear regression with least square method; bad (left) and good (right) fitting.

2.1.2 Excel trendline

- *Exclusively for illustration*! By applying the function to be fitted, plot the appropriately transformed data (*all data should be plotted, but not each of them should be fitted!*) and then apply a *linear* trendline to them.
  - Purpose of plotting before fitting
    - Check whether the dependent and independent variables are correctly assigned (data pairs follow a trend that was formerly expected)
    - Visual inspection of whether there is a linear relationship between measured data or not
    - Identifying data pairs which do not follow the trend and therefore should not be fitted
  - Identifying dependent & independent variables: Consider the first order decomposition of compound A in the reaction $A \longrightarrow B + C$. The integrated rate equation then reads $[A]_t = [A]_0 \cdot \mathrm{e}^{-kt}$. You measured data pairs $[A]_t - t$, from which $[A]_0$ and k parameters must be obtained by fitting.

There are several ways to do this. If linear regression is to be carried out, by taking the logarithm of both sides of the equation it transforms to

$$\ln[A]_t = \ln[A]_0 - kt \tag{1}$$

The next task (in each case where linear regression is to be performed) is to identify the parameters and variables according to the equation of a linear function as

$$y = a \cdot x + b \tag{2}$$

* In our example, time t passes (we might say *we let time pass*, thus we can vary it independently), and as a consequence $[A]_t$ changes. Therefore, in eq. (1), t is independent and $\ln[A]_t$ is a dependent variable.
* In eq. (2), independent variable (x) possesses a coefficient (a), which is $-k$ according to eq. (1) (be aware of the negative sign!).
* Finally, according to eq. (2), there is an additive term b which is just $\ln[A]_0$ in eq. (1).
* Following the logic mentioned above, one can identify the variables and parameters for any equation (which can be linearized), even if the mathematical formalism is much more complex.

- By applying a trendline to the plot, a best-fit linear will be displayed together with their parameters (slope, y-intercept, coefficient of determination). There is only one problem: no statistics is provided for the fitted parameters. Therefore, we cannot derive the standard deviation for the values obtained.
- Further option when a trendline is applied: one can perform linear regression with only one parameter, *i.e.*, the y-intercept can be fixed (to 0 or any arbitrary value).
- Interpretation of coefficient of correlation (r) & determination (R^2)
  - Coefficient of correlation (r): It quantifies that the linear – as a model – how accurately describes (x_i, y_i) data pairs. In other words, r tells us how strong the linear relationship is between dependent (y) and independent (x) variables. If $|r|$ is close to one, the linear relationship is more probable. *This parameter is not investigated during laboratory practice; we simply assume that the linear relationship is valid.*
  - Coefficient of determination (R^2): Calculation of R^2 and its interpretation depend on the fitted function (linear, exponential, etc.) and on the number of independent variables. *During laboratory practice, one variable (x), and one (slope) or two (slope and y-intercept) parameters are used, In such cases* $R^2 = r^2$. The coefficient of determination does *not* say anything about how appropriate a linear function is to describe the data set. It rather quantifies how the data are distributed around the fitted linear. Provided that the data set is validly fitted by a linear, an R^2 close to one represents data with narrow distribution around the linear (*i.e.*, less noisy data).
  - Fig. 3 shows the difference between selecting an appropriate function (relationship) to fit the data and testing the quality of a linear fit. Obviously, for a given R^2 we can render two totally different data sets. On the left hand side, the points truely follow a linear trend, thus R^2 close to one represents a small deviation. On the right hand side, as a contrary, we show data sets which should not be fitted as a linear. Although R^2 values are the same as on the left side, the x clearly shows a systematic bending from the linear. *As a conclusion,* R^2 *is only meaningful for the quality of the fitting, or for the deviation of the data set, if they truly follow a linear trend.* An R^2 close to one by itself does not guarantee that the trend is linear.

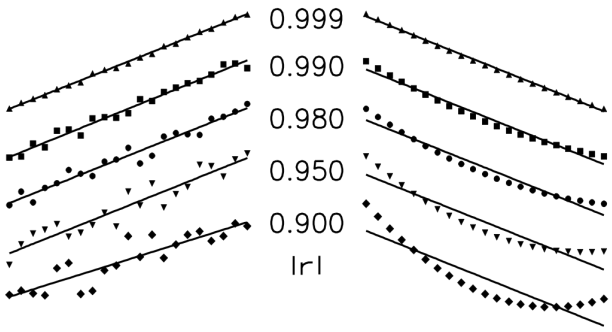

Figure 3: Representation of the absolute value of coefficient of correlation ($|r|$). **Left**: data set generated with normal error distribution; **Right**: data set with systematic bending. For these fits $R^2 = r^2$. For the interpretation treat $|r|$ as it was R^2.

2.1.3 Linear regression / parameter fitting in MS Excel

- If applying a trendline is sufficient, because we must obtain statistics for the parameters, *the LINEST* Excel function should be used.
- This is a block function, that is, we must use multiple cells for input and the results will be displayed in multiple cells as well. Also, to perform the calculations, we must run the function with the binding of the CTRL+SHIFT+ENTER key
  During laboratory practice, to perform a linear regression with one variable (x), and with one (slope) or two (slope and y-intercept) parameters, 2 (columns)$\times$ 5 (rows) blocks of cells must be highlighted before typing the following command:

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

  then hit CTRL+SHIFT+ENTER keybinding.

  - known_y's; known_x's $\longrightarrow$ Selecting the cells containing the dependent (y) and independent (x) variables. Cell selection must be continuous without any break. Therefore, if there are *bad* data in the set (based on previous visual inspection thanks to the preliminary plot), copy the data pairs to be fitted into a new columns.
  - const $\longrightarrow$ If 1 / TRUE: y-intercept (b) will be calculated during the fit; if 0 / FALSE: y-intercept (b) is taken as 0 during the calculation, and fitting is performed according to $y = a \cdot x$ equation (equation must be selected on the basis of the principle of measurement, *e.g.*, the absorbance of a solution is zero if $c = 0$ for the solute).
  - stat $\longrightarrow$ practical value is 1 (otherwise no statistical parameter is displayed).

- Statistics displayed according to the format of *LINEST* (that is, the following data will be displayed in the originally selected 2 (columns)$\times$5 (rows) cells):

	1	2
1	**slope**	**y-intercept**
2	**standard error of slope**	**standard error of y-intercept**
3	**R^2**	*not relevant for the practice*
4	*not relevant for the practice*	**degree of freedom**
5	*not relevant for the practice*	*not relevant for the practice*

  - Degree of freedom (DF): provided that the measured data are independent of each other, $DF = N - P$, where N is the number of data and P is the number of parameters.
  - Standard error ($\neq$ standard deviation!),
    their relation: standard deviation (σ) $= \sqrt{\text{degree of freedom}} \cdot$ standard error

- *Note:* Starting from the MS Office 365 Excel version, *LINEST* (and other block functions as well) can be given as a linear equation that is not required to select the block of cells in advance. Excel will automatically reserve the required cells once the command is typed in. Also, CTRL+SHIFT+ENTER keybinding is not required, simply hit ENTER. Due to compatibility issues, the *old* method described earlier also works.

2.2 Performing linear regression for experimental data: step-by-step guide

1. Identify the dependent (y) and independent (x) variables with the help of the linear equation and the description of the tusk.
2. Identify the fitted parameter(s) (slope, y-intercept) with the aid of the linear equation and the description of the tusk.
3. If the measured data are not directly the dependent (y) and/or independent (x) variables, then these must be calculated from the measured data.
4. Once the cells containing the dependent (y) and independent (x) variables are ready at hand, plot them according to the linear equation. The plot should contain all data pairs. Based on the plot, identify the *bad* data, *i.e.*, which do not follow the trend, thus they should not be fitted.
5. It is practical to copy the *good* data set (*i.e.*, which does not contain the outliers anymore) into new columns, then add it to the previous plot as a new data set besides the one showing all measured values. Label the data sets to clearly show which will be used for fitting. Apply a trendline to the data set that only contains the data to be fitted.
6. Perform linear regression by applying *LINEST* on the columns containing the data set to be fitted. By comparing the parameters obtained from the trend line and those displayed by *LINEST* one can double check whether everything was done properly. If so, values must match.
7. Make sure that the columns containing the data sets and used for calculations and the plot are well-explained and of scientific quality, including units as well.
8. Derive the standard deviation from the standard error displayed by *LINEST* for slope and y-intercept (if applicable).
9. According to the linearized equation and the corresponding description, additional calculations might be required to derive the required quantities from the fitted parameters. Pay attention to the units, both when calculations or fitting is used to obtain a quantity and derive some result.
10. If required, take error spreading into account to find the standard deviation of the final result(s).