Is there still a place for linearization in the chemistry curriculum?

Andrew R. McCluskey*,†,‡

†*School of Chemistry, University of Bristol, Cantock's Close, Bristol, BS8 1TS, United Kingdom*

‡*European Spallation Source ERIC, Ole Maaløes vej 3, 2200 København N, Denmark*

E-mail: andrew.mccluskey@bristol.ac.uk

Abstract

The use of mathematical transformations to reduce non-linear functions to linear problems, which can be tackled with analytical linear regression, is commonplace in the chemistry curriculum. The linearization procedure, however, assumes an incorrect statistical model for real experimental data; leading to biased estimates of regression parameters and should therefore not be used in formal data analysis. This fact is overlooked in many chemistry degrees, students do not yet have the mathematical knowledge to appreciate why linearization leads to bias when it is introduced. I hope that this commentary will start a discussion around the place of linearization in the chemistry curriculum, and more broadly around how mathematical and statistical training is currently provided to chemistry students.

Keywords: linearization, maths for chemists, data skills, statistics, mathematics.

In chemistry, non-linear relationships are commonly found between dependent and independent variables. These relationships can be simplified by the process of "linearization", where some mathematical transformation is used to reduce the non-linear problem to a linear

one. By linearizing a function, analytical linear regression can be used to quantify parameters of interest, rather than relying on numerical optimisation. We see this process in chemistry textbooks[1,2] and undergraduate degree programs: for example where it is applied to first- and second-order rate equations, and the Clausius-Clapeyron and Arrhenius equations.[1,3,4]

While mathematically sound for noise-free measurements, linearization can introduce errors in the analysis process for real experimental data. Specifically, it can lead to biased estimates of regression parameters; the gradient and intercept of the straight line – as has been noted in this Journal and others.[3,5–12] Therefore in formal analysis, where accurate and precise estimates of the parameters of interest are desired, the use of linearization should be avoided. Despite this, linearization is still included in a general chemistry education, without discussion of the problems caused by it, resulting in a vicious cycle; linearization is taught to students as it appears in the research literature and is then used in the research literature as the practitioners are not aware of the problems. The problems of linearization are rarely discussed when introduced, as at this stage students are not familiar with the mathematical or statistical concepts required to appreciate what causes the problems and unlike other "convenient truths" (e.g. the Rutherford model for the atom), linearization is typically not revisited at a later stage in the chemical education. This author believes that instead of introducing linearization as a data analysis tool, students should receive a more complete training in the relevant data analysis skills, and linearization should be kept for basic visualisation of data.

Although it has been covered by others, it is valuable to restate the problem that results from linearization. For this we can consider the decomposition of hydrogen peroxide, H_2O_2, in the presence of excess cerium(III) ion, which follows first-order rate kinetics with the form[1]

$$[H_2O_2]_t = [H_2O_2]_0 \exp(-kt), \tag{1}$$

where, $[H_2O_2]_t$ is the concentration of hydrogen peroxide at time t, $[H_2O_2]_0$ is the initial concentration and k is the rate constant (representative data is shown in Figure 1a). Lin-

earization of Equation 1 involves taking the natural logarithm of both sides to produce

$$\ln [H_2O_2]_t = -kt + \ln [H_2O_2]_0. \tag{2}$$

The gradient and intercept from linear regression, of $\ln [H_2O_2]_t$ on t, are therefore equal to $-k$ and $\ln [H_2O_2]_0$, respectively (Figure 1b).

If we were to perform repeated measurements of the concentration of H_2O_2 as a function of reaction time and analyse each repeat, we can build up a distribution of estimates of k (Figure 1c & 1d). The simplest way to analyse a linearized function is by ordinary least squares (OLS) linear regression, which we can compare with unweighted non-linear optimization. Non-linear fitting gives a normal distribution of estimated values of k, with a mean centred on the true value, i.e. the estimation is unbiased. The linearized form, however, gives a biased, broad, asymmetrical distribution, where the normalised mean is 1.05. The linearized approach will, on average, overestimate the value of k and any single estimate of k has a higher probability of being further from the true value.

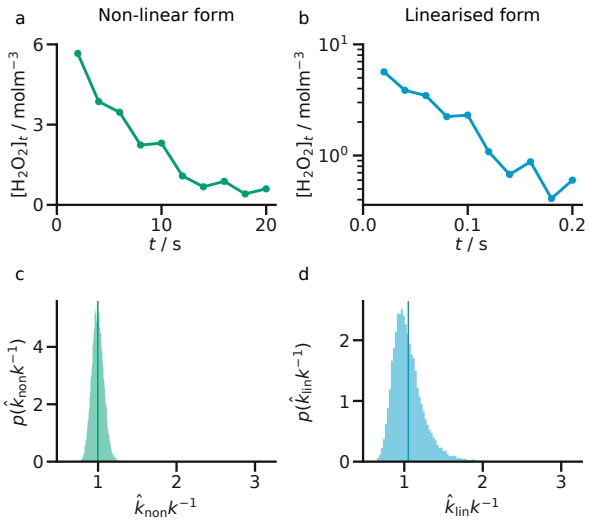

Figure 1: Representative data for first-order integrated rate equation, with a true value of $k = 0.15\,s^{-1}$ and $[A]_0 = 7.5\,moldm^{-3}$, showing (a) the non-linear and (b) the linearized forms. Estimates of k, normalised to the true value of k, from 2^{15} analyses of unique representative datasets, using (c) unweighted non-linear fitting and (d) linearization followed by ordinary least squares, with the vertical lines indicating the distribution means.

By using OLS or unweighted non-linear optimisation, we are assuming that the uncertainties in our data are all the same, i.e., they are homoscedastic. It was noted by Perrin,[3] however, these homoscedastic uncertainties may become heteroscedastic as a result of the linearization process (note the error bars in Figure 2b). Therefore, the use of OLS for linearized data is insufficient, instead weighted least squares (WLS), where the weights are determined by Gaussian error propagation, should be used. For the example in Equation 2, the correct error propagation is to divide the measured error by the nominal value (Figure 2b). WLS leads to a normal distribution of estimated k, but still, the distribution is biased (Figure 2d), with a normalised mean of 0.95. Non-linear optimisation, meanwhile, still produces an unbiased estimate.

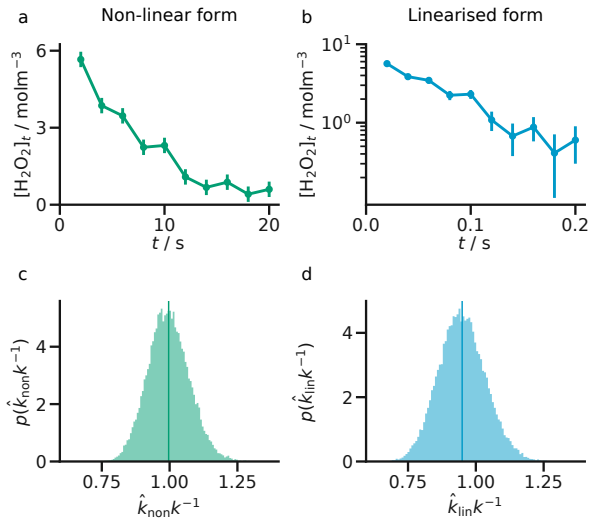

Figure 2: The same representative data as Figure 1 with error bars of 0.3 molm^{-3} (a & b). The same number of analyses were performed, however, this time using (c) weighted non-linear optimisation and (d) weighted least squares with propagated uncertainties, the vertical lines indicate the mean of the distribution.

The observed bias can be understood by recognising that the measurement of any variable, y, is only ever an estimate of the true value, $\hat{y}$, which is a random draw from a distribution of values, $p(y)$. The shape of this distribution depends on the noise or uncertainty in the measurement. It is commonly assumed that random uncertainty sources will lead to a normal distribution, $p(y) \sim \mathcal{N}(\mu, \sigma^2)$, which is defined by the mean, μ, and standard devia-

tion, σ (Figure 3).[1] When linearization is used, a mathematical transformation is performed on the dependent variable and if that transformation scales in a non-linear fashion, i.e., the reciprocal or logarithm is taken, it will cause the normally distributed variable to become non-normal (Figures 3b & 3c). Similarly, the use of Gaussian error propagation also breaks down with linearization, as Gaussian error propagation involves the application of a truncated Taylor expansion, in this case, to a non-linear function leading to a large truncation error.

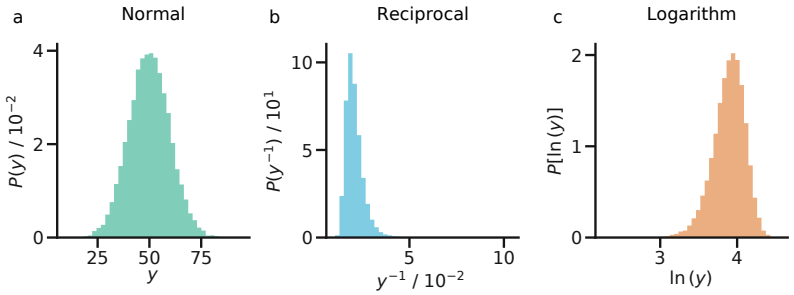

Figure 3: Histograms showing the effect on (a) a normal distribution, of mathematical transformations that scale non-linearly: (b) the reciprocal or (c) the logarithm. Produced from 2^{15} random samples from the normal distribution $\mathcal{N}(50, 10^2)$.

For normally distributed variables, both OLS and WLS produce unbiased estimates of the regression parameters. However, this is not the case when non-normally distributed variables are used. By applying OLS or WLS to non-normally distributed variables, we use the wrong statistical model for our analysis. However, the correct statistical model is being applied in the non-linear optimisation case, where the variables have not been transformed and are therefore still normally distributed.

While there is a role for linearization as a data visualization tool, e.g., the doubling of a reaction rate may be more obvious on a linearized plot, or as a qualitative classroom exercise, it should not be taught as a tool for the formal analysis of data. In addition to improving the mathematical accuracy of analysis performed by students, this will help to break down the vicious cycle that results in potentially erroneous results appearing in the research literature. Furthermore, there are additional learning outcomes in showing students

that the way that data is represented graphically may skew their interpretation. Specifically, this can be achieved by comparing the linearized and non-linear optimized solutions on the linearized and non-linear plots.

This author hopes that this commentary will both remind readers of the problems associated with linearization and inspire discussion in the community regarding how and when mathematical and statistical skills are taught to students. In my experience (which admittedly has been focused on the United Kingdom), students are rarely introduced to much statistical methodology, beyond the basics of summary statistics and Gaussian error propagation, before they are tasked with data analysis problems, e.g., students are asked to produce "lines of best fit" without being introduced to ordinary least squares. Therefore, to ensure that students appreciate why linearization is problematic, among many other benefits, they should be given a more complete training in the mathematical and statistical underpinnings of data analysis before linearization is introduced.

The importance of "data skills" in a chemical education is underestimated and should be considered similar to that of traditional mathematical concepts, such as calculus, which underpin many theoretical aspects of the chemical sciences. Data skills, including data handling and analysis, and basic programming skills, make chemists more capable in future research projects, more employable both inside and outside of the chemical industry, and can aid in the understanding of complex subjects.[13–16] Without facilitating this component of a chemist's education, we are failing to equip them to approach modern problems that they will find in the chemical sciences.

Data availability

Electronic Supplementary Information (ESI) available: A complete set of analysis/plotting scripts allowing for a fully reproducible and automated analysis workflow, using showyourwork,[17] for this work and a Jupyter Notebook showing the use of weighted non-linear op-

timisation for representative first-order rate kinetics data is available at `https://github.com/arm61/linearization-issues` (DOI: 10.5281/zenodo.7949905) under an MIT license, while the text is shared under a CC BY-SA 4.0 license.[18]

Acknowledgements

The author thanks Benjamin J. Morgan, Samuel W. Coles, Thomas Holm Rod, Gabriel Krenzer, and Kasper Tolborg for the insightful discussion that led to this work. Additionally, the author would like to thank those that engaged in discussion on Twitter, in particular Carl Poree and Fiona Dickinson, when the problem of linearization in Arrhenius modelling was initially raised.

References

(1) Monk, P.; Munro, L. J. *Maths for Chemistry: A chemist's toolkit of calculations*, 2nd ed.; Oxford University Press: London, UK, 2010.

(2) Atkins, P.; de Paula, J.; Keeler, J. *Atkins' Physical Chemistry*, 11th ed.; Oxford University Press: London, UK, 2018.

(3) Perrin, C. L. Linear or Nonlinear Least-Squares Analysis of Kinetic Data? *Journal of Chemical Education* **2017**, *94*, 669–672.

(4) Harper, J. K.; Heider, E. C. Data Linearization Activity for Undergraduate Analytical Chemistry Lectures. *Journal of Chemical Education* **2017**, *94*, 610–614.

(5) de Levie, R. When, why, and how to use weighted least squares. *Journal of Chemical Education* **1986**, *63*, 10.

(6) Rusling, J. F. Minimizing errors in numerical analysis of chemical data. *Journal of Chemical Education* **1988**, *65*, 863.

(7) Zielinski, T. J.; Allendoerfer, R. D. Least Squares Fitting of Non-Linear Data in the Undergraduate Laboratory. *Journal of Chemical Education* **1997**, *74*, 1001.

(8) Denton, P. Analysis of First-Order Kinetics Using Microsoft Excel Solver. *Journal of Chemical Education* **2000**, *77*, 1524.

(9) Vent, S. L. Don't Be Tricked by Your Integrated Rate Plot: Reaction Order Ambiguity. *Journal of Chemical Education* **2004**, *81*, 32.

(10) Rittenhouse, J.; Scarlete, M. *Annual Reports in Computational Chemistry*; Elsevier, 2005; pp 221–235.

(11) Möglich, A. An Open-Source, Cross-Platform Resource for Nonlinear Least-Squares Curve Fitting. *Journal of Chemical Education* **2018**, *95*, 2273–2278.

(12) Alamillo-Ferrer, C.; Hutchinson, G.; Burés, J. Mechanistic interpretation of orders in catalyst greater than one. *Nature Reviews Chemistry* **2022**, *7*, 26–34.

(13) Srnec, M. N.; Upadhyay, S.; Madura, J. D. A Python Program for Solving Schrödinger's Equation in Undergraduate Physical Chemistry. *Journal of Chemical Education* **2017**, *94*, 813–815.

(14) Chng, J. J. K.; Patuwo, M. Y. Building a Raspberry Pi Spectrophotometer for Undergraduate Chemistry Classes. *Journal of Chemical Education* **2020**, *98*, 682–688.

(15) Dickson-Karn, N. M.; Orosz, S. Implementation of a Python Program to Simulate Sampling. *Journal of Chemical Education* **2021**, *98*, 3251–3257.

(16) Cumby, J.; Degiacomi, M.; Erastova, V.; Güven, J.; Hobday, C.; Mey, A.; Pollak, H.; Szabla, R. Course Materials for an Introduction to Data-Driven Chemistry. *Journal of Open Source Education* **2023**, *6*, 192.

(17) Luger, R. showyourwork. `https://github.com/rodluger/showyourwork`, 2021.

(18) McCluskey, A. R. linearization-issues-0.0.1. https://github.com/arm61/linearization-issues, 2023.