Documenting Data Analysis

Barry Linkletter

You have analyzed data and presented your conclusions. Can other researchers obtain to the same values by using your data? Could someone check your methods in detail? If you perform your analysis using *Python*, the code will reveal your methods. We will explore using an interactive *Python* environment to analyze your data and to document your methodology. Along the way we will hopefully agree to never use "cubic spline" again.

This document was produced using the LaTeX typesetting language with the Tufte-handout document class. Chemical diagrams were created in *ChemDoodle*. Plots were typeset using *MatPlotLib* and *Python* tools. Diagrams were created and edited using *Affinity Designer*.

The Seminar

In the seminar supported by this document we will explore the following ideas:

1. We all love **spreadsheets**, but **have we outgrown them**?
2. *Python* is a tool for **data analysis**. As an example, we will use the **Erying plot** data described below and explore the following concepts:
   (a) Linearized plots and their strengths and weaknesses.
   (b) Handling error propagation for strongly correlated parameters (e.g. when small changes in slope have an outsized effect on the intercept.)
   (c) Fitting the non-linear model of the Erying equation to your data.
   (d) Graphical approaches to expressing confidence intervals in your plots.
3. Using *Jupyter* notebooks as a method for using these *Python* tools for **documenting your methods and reasoning**. Allow others to read your code and read your mind.[1]

[1] Won't it be great to be able to figure out exactly what you were thinking when you did that thing that time.

The Eyring Equation

The **Eyring equation** is a classic of physical chemistry.[2] The free energy of activation for a reaction can be determined and, if a series of rate constant values are available at different temperatures, we can determine the enthalpy and the entropy of activation. These parameters are very helpful in **interpreting reaction mechanism**. But can these results be trusted? How could you estimate their reliability? Unless you are calculating the standard deviations from your plot, you may be unaware of a low quality relationship.

The Erying equation was developed almost 100 years ago[3,4] by Henry Eyring,[5] Michael Polanyi[6] and Meredith Evans.[7] It was analytically derived from statistical thermodynamics and transition state theory. Unlike the Arrhenius equation, which was just a guess based on empirical observations, the Erying equation provides useful information. The equation is expressed as…

$$k = \kappa \frac{k_B}{h} T \cdot e^{\frac{-\Delta G^{\ddagger}}{RT}}$$

…where κ is the transmission coefficient (assumed to be unity, but that's not always the case), R is the gas constant, k_B is the Boltzmann constant and h is the Planck constant. k is the observed **rate constant** and T is the **absolute temperature**.

The free energy term is an expression of enthalpy and entropy…

$$\Delta G^{\ddagger} = \Delta H^{\ddagger} - T\Delta S^{\ddagger}$$

…and so the Erying equation can be expressed as…

$$k = \kappa \frac{k_B}{h} T \cdot e^{\frac{-\Delta H^{\ddagger}}{RT}} e^{\frac{\Delta S^{\ddagger}}{R}}$$

This can be presented in a linear form…

$$\ln \frac{k}{T} = \frac{-\Delta H^{\ddagger}}{R} \frac{1}{T} + \frac{\Delta S^{\ddagger}}{R} + \kappa \frac{k_B}{h}$$

The famous **Eyring plot** is a plot of $\ln k/T$ vs. $1/T$. The **slope** will be the value of $-\Delta H^{\ddagger}/R$ and the **intercept** will the the sum $\Delta S^{\ddagger}/R + \kappa k_B/h$. Assuming that $\Delta H^{\ddagger}$ and $\Delta S^{\ddagger}$ do not change with temperature in the range being studied, the plot should produce a straight line.

[2] Read chapter 7 of "Modern Physical Organic Chemistry" by Anslyn & Dougherty, *University Science Books*, **2006** or any textbook that discussed chemical kinetics.

[3] "The activated complex in chemical reactions." H. Eyring, *J. Chem. Phys.*, **1935**, *3*, 107–115, https://doi.org/10.1063/1.1749604

[4] "Some applications of the transition state method to the calculation of reaction velocities, especially in solution." M.G. Evans, M. Polanyi, *Trans. Faraday Soc.*, **1935**, *31*, 875–894, https://doi.org/10.1039/TF9353100875

[5] Eyring was born in Mexico in 1901 and his family fled to the USA in 1912 during the Mexican revolution. Erying was a Mormon. His father practiced polygamy and was married to the two great-aunts of former US presidential candidate Mitt Romney at the same time. You could say that Erying "had two moms." He was awarded just about every honour in chemistry except the Nobel prize.

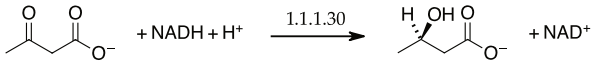
[6] Polanyi fled Germany in 1933 and was appointed a professor of physical chemistry at the University of Manchester. He was a deep thinker and was appointed to a chair in social science at U. Manchester in 1948. He passionately argued that human nature was greater than the sum of its chemical parts. Two of his students won Nobel prizes: Eugene Wigner (a nuclear physicist who helped draft the famous "Einstein letter" that led to the creation of the atomic bomb - 1963 for physics) and Melvin Calvin (a biochemist and a leader in the elucidation of carbon fixation in photosynthesis - 1961 for chemistry.) His son, John Polanyi, is a professor emeritus of chemistry at U. Toronto and won the Nobel prize in 1986 for chemistry.

[7] It wasn't a team effort; it was a race. All three proposed the equation around the same time in 1935. Evans was a student of Eyring and a colleague of Polanyi; perhaps he was the link?

An Example

To demonstrate the Erying plot, we must first obtain some **experimental data**. Here we will present some results from an enzyme kinetics experiment.[8] The authors determined the Michaelis-Menten k_{cat} value for a reaction over a series of temperatures. As long as the chemical change is the rate-determining-step, and not some other step such as substrate binding or product release,[9] then the Erying equation will apply.[10]

The authors reported rate/temperature data for the reduction of the β-ketoacids 3-oxobutanoate and 3-oxopentanoate by NADH in the presence of *hydroxybutyrate dehydrogenase* (EC 1.1.1.30) from *Acinetobacter baumannii*.



The rate of reaction was followed by measuring the change in NADH concentration using UV-vis spectrometry. This is a precise method and should give repeatable results.

The Data

Below are two sets of data from the publication (always check the **supplementary material** for experimental details and data tables not presented in the publication.)

	k_{cat} / s^{-1}			
Temp. (/ *K*)	3-oxobutanoate		3-oxopentanoate	
293	7.6	±0.2	0.46	±0.02
298	11.7	±0.3	0.60	±0.03
303	15.2	±0.1	1.16	±0.02
308	21.3	±0.9	1.47	±0.03
313	27.8	±0.9	2.07	±0.09

Table 1: Michaelis-Menten turnover number, k_{cat}, vs. temperature for reaction of *hydroxybutyrate dehydrogenase* with 3-oxobutanoate and 3-oxopentanoate.

Data taken from tables S-1 and S-3 of the supplementary material for the the paper.[8]

[8] "Linear Eyring Plots Conceal a Change in the Rate-Limiting Step in an Enzyme Reaction", Teresa F. G. Machado, Tracey M. Gloster, and Rafael G. da Silva, *Biochemistry*, **2018**, *57*, 6757-6761. https://doi.org/10.1021/acs.biochem.8b01099.

[9] It is an investigation into this very possibility that is the focus of this paper.

[10] The Erying equation requires a reaction that is a "single step." Multistep reactions with high-energy intermediates will also work as long as there is no significant concentration of the intermediate present during the reaction. If a significant portion of the reactant is converted to an intermediate that then converts to product in a subsequent r.d.s., such as in enzyme burst kinetics, the Erying equation would not be appropriate. If one could determine the separate rate constants, then the Eyring equation could be successfully applied to each step.

Your Challenge

Think about the following activities as you approach this seminar.

1. Construct an Eyring plot using the data above and perform a **linear curve fit**. Then complete the following activities.

   (a) Perform the curve fit with and without consideration of the estimated experimental errors. When using the experimental error, present error bars on the plots. How did you include the error from the rate data in your curve fit?

   (b) Report the slope and intercept, their respective standard deviation and the squared Pearson correlation constant (r^2). Why are these parameters slightly different when we do or do not include the experimental errors?

   (c) Calculate the values for $\Delta H^{\ddagger}$ and $\Delta S^{\ddagger}$ with error-propagation.[11,12] Comment on the precision of each parameter.

   (d) Using the calculated activation parameters, predict the value of k_{cat} at a selected temperature. Using error-propagation, determine the estimated standard deviation for your predicted results.

   (e) Most Erying plots in the literature use four or five data points. Select any four data points from the sets above and repeat the above exercises. What happened to your precision when using less data? More rates and temperatures were reported in the paper. Add some data points and see what happens to the precision.

2. Linearizing mathematical models for the convenience of using linear regression analysis is very 1980's.[13] Perhaps we could try fitting the data directly to k_{cat} vs. T using **non-linear curve fitting**? Repeat the exercises described above. How is the precision now?

3. **Document your methods** for creating the plots, performing the curve fits and performing calculations with correct error-propagation. Explain how your method of documentation provides an unambiguous description of your math and methods. Does your documentation provide a tool that can be easily reused by yourself or your lab-mates in analyzing other sets of rate data at different temperatures? What if you didn't have to write detailed documentation because your data analysis documented itself?

4. Where in the above process did **Microsoft Excel** become **useless** to you?

[11] Error propagation for the Erying equation is expressed using the following two relationships:

$$\sigma_{\Delta H^{\ddagger}} = \sqrt{2\frac{R^2 T_{max}^2 T_{min}^2}{(T_{max} - T_{min})^2}\left(\frac{\sigma_{k_{cat}}}{k_{cat}}\right)^2}$$

$$\sigma_{\Delta S^{\ddagger}} = \sqrt{\frac{R^2\left(T_{max}^2 + T_{min}^2\right)}{(T_{max} - T_{min})^2}\left(\frac{\sigma_{k_{cat}}}{k_{cat}}\right)^2}$$

[12] Error propagation formula adapted from "A Static Agostic α-CH $\cdots$ M Interaction Observable by NMR Spectroscopy: Synthesis of the Chromium(II) Alkyl $[Cr_2(CH_2SiMe_3)_6]^{2-}$ and Its Conversion to the Unusual 'Windowpane' Bis(metallacycle) Complex $[Cr(\varkappa^2\text{-}C, C' - CH_2SiMe_2CH_2)_2]^{2-}$." P.M. Morse et al., *Organometallics*, **1994**, *13*, 1646-1655. https://doi.org/10.1021/om00017a023

[13] Without a doubt, the '80's was the greatest decade – except for the computers. I miss my Commodore 64, but my current computer is one million times more powerful in every metric: 64 kilobytes vs. 16 gigabytes; no storage (or whatever an audio cassette tape could hold) vs. one terabyte; 4528 transistors vs. 16 million transistors; 1.5 MHz vs. 3.2 GHz; dial-up bulletin-board services using 54 kb/s modems vs. the internet over optical fibre at 300 mb/s.

We have enormous computing power at our fingertips now. Why do we stick with methods meant for graph paper?