A Note on Covariance in Propagation of Uncertainty

Edwin F. Meyer
Chemistry Department, DePaul University, Chicago, IL 60614

Propagation of uncertainty from a set of measurements to a derived quantity is an important topic whose fundamental ideas have been competently addressed in this *Journal* (*1–4*). However, there is a pitfall associated with the uncritical use of the standard formulas promulgated in many texts that has not been emphasized to date. It is readily illustrated with the data collected in a common undergraduate physical chemistry experiment, and is the subject of this note.

We use spreadsheets nowadays in much of our data workup. Students learn early on how to use the data regression tool therein, and we try to instill in them the significance and importance of the estimates of uncertainty that are provided—for example, for the slope, m, and intercept, b, of the common $y = mx + b$ correlation. These should be examined for consistency with the students' own estimates of uncertainties in the measured variables, and they of course provide essential information for the propagation of uncertainty to derived quantities. A familiar example of the latter is the enthalpy of vaporization derived from the slope of a plot of the logarithm of vapor pressure versus inverse temperature.

One of the standard experiments, in fact, in our first physical chemistry lab course is the measurement of the vapor pressure of water as a function of temperature (see, e.g., ref *5*). Students adjust the pressure above the boiling liquid and wait for the temperature to stabilize. They then measure the pressure with a mercury meter-stick manometer and the temperature with a thermistor capable of 0.02 K precision. Ten to 12 data pairs are collected between 170 and 750 torr.

Pollnow (*6*) has done a beautiful job of illustrating the importance, in proper optimization of parameters, of the weighting of pressures when the dependence of vapor pressure on temperature is linearized. However, this is a subtlety unjustified by the quality of the data in question here, and our students are instructed to do a simple least squares regression of ln P vs. $1/T$:

$$\ln(P/\text{torr}) = m\,(K/T) + b$$

The spreadsheet supplies them with estimates of the "best fit" parameters m and b as well as their uncertainties. They use the information for m to estimate the enthalpy of vaporization at the midpoint of the temperature range of the data, and its uncertainty. This calculation is almost always satisfactory.

Students are told also to use this equation to predict the normal boiling point of water by substituting 760 torr for P and solving for T, and to estimate the uncertainty in T by propagation of the uncertainties in m and b. They obtain very respectable values for the boiling point, but their estimates of uncertainty are unreasonably large (1 to 2 K) if they use the familiar "propagation of error" formula

$$\sigma_T^2 = (\partial T/\partial m)^2\sigma_m^2 + (\partial T/\partial b)^2\sigma_b^2$$

where σ represents standard deviation.

How can it be that in a series of measurements characterized by roughly 0.05 K uncertainty in temperature (the uncertainty in P contributes less than 0.04 K to the uncertainty in boiling point in our experiment), the estimated uncertainty in the boiling point can be as great as 2 K? The answer is that in this calculation *both* the slope *and* the intercept are required, and these two parameters are *not* independent from one another. This violates one of the assumptions used in producing the above formula: namely, that the variables m and b must be independent. Thus the formula is invalid for this application.

Guedens et al. have recently discussed exactly this situation (*3*). We can consider the expression for the boiling point as a function of m and b:

$$T_{\text{bp}}/K = f(m,b) = m/(\ln 760 - b)$$

The variance in T_{bp} may be expressed in terms of the variances of m and b alone *only if* m *and* b *are independent from one another;* if, as in this case, they are not, we must add a covariance term to the equation above for the variance in T:

$$\sigma_T^2 = (\partial T/\partial m)^2\sigma_m^2 + (\partial T/\partial b)^2\sigma_b^2 + 2\,(\partial T/\partial m)(\partial T/\partial b)\,\sigma_{mb}$$

where σ_{mb} is the covariance of m and b. If m and b were independent, this term would be zero, and the expression would be the usual one found in most texts. It happens that the third term is opposite in sign to the first two, and of similar magnitude. Its presence lowers the estimate of uncertainty in the boiling point obtained in its absence.

Unfortunately, spreadsheets do not supply covariances as part of their data regression output, making it difficult to perform a proper analysis of uncertainty without outside help whenever more than one parameter from a fit is required to estimate a physical property. A readily available alternative to spreadsheets for rigorous parameter optimization, which supplies the necessary covariances, is Ramette's software package called FLEXFIT.[1] (The covariances do not appear as such, but in an equivalent form easily deciphered by following the clear instructions provided.)

Using a typical set of student data, we obtain the following results: $m = -5126.85$, $b = 20.3775$, $\sigma_m^2 = 161.35$, $\sigma_b^2 = .0012192$, and $\sigma_{mb} = -0.44341$. Using these numbers to evaluate the respective terms in the equation above for the uncertainty in boiling point, we get 0.85413, 0.89802, and -1.75113, which sum to 1.02×10^{-3}. It is important to note that the absolute value of the sum of the first two terms is very nearly equal to that of the third, implying that an apparently unreasonable number of digits be carried in such a calculation. This is precisely because m and b are not independent. While the value of m per se is not determined to better than about ±12 for this fit, its digits to the right of the decimal point are paired with definite digits in the value of b beyond its "significant" digits per se. Digits that are not significant per se in a parameter are significant in a different sense in this type of statistical calculation.

Were the third (covariance) term in the above equation ignored, the (erroneous) estimate of uncertainty in boiling

point would be 1.3 K for this set of data. Including the covariance term, it is 0.03 K, a perfectly reasonable estimate considering the equipment at hand.

Another example from the undergraduate physical chemistry laboratory of the requirement that both slope and intercept be used in the evaluation of a physical quantity arises in the determination of surface area using the BET model of physical adsorption (e.g., see ref *5*). The amount of adsorbate in a monolayer is given by $1/(m + b)$ after linearizing the experimental data and performing a data regression. Here also the covariance term must be included as in the above equation in order to obtain a reliable estimate of the statistical uncertainty in the amount of adsorbate in a monolayer.

Note

1. This program is available on JCE Online at *http://jchemed.chem.wisc.edu/.*

Literature Cited

1. Andraos, J. *J. Chem. Educ.* **1996,** *73*, 150–154.
2. Malinowski, E. R. *J. Chem. Educ.* **1995,** *72*, 1079–1082.
3. Guedens, W. J.; Yperman J.; Mullens, J.; Van Pouke, L. C.; Pauwels, E. J. *J. Chem. Educ.* **1993,** *70*, 776–779.
4. Rusling, J. F. *J. Chem. Educ.* **1988,** *65*, 863–866.
5. Shoemaker, D. P.; *Experiments in Physical Chemistry*, 4th ed.; McGraw-Hill: New York, 1981.
6. Pollnow, G. F. *J. Chem. Educ.* **1971,** *48*, 518–519.