APPENDIX 3G: Regression Plots

The regression plots in Chapter 3 and Appendices 3A to 3F all incorporate the following common characteristics.

When variables of different sampling durations are compared, the measurements with the shorter duration are averaged to produce a concentration for the longer duration.

The plots exclude all points when either variable is below the minimum detectable limit (mdl). When two or more measurements are averaged for a longer duration, the point is excluded if any measurement is below the mdl.

Most of the plots include the 1:1 line and the regression line. The 1:1 line passes through the origin and the upper right corner.

Most of the plots include statistics associated with the data of the plot. The values shown are for the points in the plot and not for all data for that variable. Because concentrations below the mdl are excluded, the mean values shown are generally larger than the actual mean values. The parameters at the top of the plot are

N	=	number of points
R	=	correlation coefficient
(y variable)	=	slope * (x variable) + intercept
σ_a	=	standard error in the intercept of the regression line
σ_b	=	standard error in the slope of the regression line
z	=	$y - x$
$\bar{x}, \bar{y}, \bar{z}$	=	mean value of x, y, z
$\sigma_{\bar{x}}, \sigma_{\bar{y}}, \sigma_{\bar{z}}$	=	standard error in x, y, z
$\sigma_x, \sigma_y, \sigma_z$	=	standard deviation in x, y, z

All regression lines are based on a least-squares fit that minimizes the perpendicular distances between the points and the line. This perpendicular method is preferred over the ordinary least-squares method when both variables have uncertainty. It has the clear advantage that the regression line will be simply reflected about the 1:1 line when the variables are interchanged. Thus, it makes no difference which variable is x or y.

For intercomparison of data derived from two measurement methods, the usual least-squares regression is often not appropriate. This is due to the fact that data obtained from both measurement methods are contaminated with measurement errors. The more appropriate approach is to use the so called 'errors in variables regression model.' There is a vast amount of literature on this subject, but an adequate background for applications can be obtained from a paper by Mandel.[1]

Let (X_i, Y_i) $i = 1, 2, \ldots, n$ denote the data obtained from two measurement methods where X_i denotes readings from instrument A and Y_i from instrument B. Suppose that the true values

being measured are denoted by w_i. If both instruments have a zero shift and a bias, then:

$$X_i = \alpha_1 + \beta_1 w_i + \nu_i \tag{3G.1}$$

and

$$Y_i = \alpha_2 + \beta_2 w_i + \varepsilon_i \tag{3G.2}$$

where ν_i, ε_i denote random errors. Eliminating w_i between the above two equations, the relationship between Y_i and X_i can be written in the form

$$y_i = \alpha + \beta x_i \tag{3G.3}$$

where

$$Y_i = y_i + \varepsilon_i \tag{3G.4}$$

and

$$X_i = x_i + \nu_i \, . \tag{3G.5}$$

Here y_i and x_i denote the readings from instrument A and instrument B respectively in the absence of random errors. The coefficients α and β stand for the relative zero shift and the relative bias, respectively. If the two instruments agree on the average, the α should be zero and β should be 1.

Suppose that the random error ε_i associated with the Y_i has mean zero and variance σ_ε^2 and the random error ν_i associated with X_i has mean zero and variance σ_ν^2. Let the ratio $\sigma_\varepsilon^2/\sigma_\nu^2$ be represented by λ. We assume that this ratio can be experimentally determined and hence, for all practical purposes, known. The formulas for calculating estimates a and b of α and β, respectively, along with their standard errors are shown below. The formulas for the standard errors are approximate and exact formulas are not available.

$$b = \frac{(s_{YY} - \lambda s_{XX}) + \{(s_{YY} - \lambda s_{XX})^2 + 4\lambda s_{XY}^2\}^{1/2}}{2 s_{XY}} \tag{3G.6}$$

$$a = \bar{Y} - b\bar{X} \tag{3G.7}$$

where

$$s_{YY} = \sum (Y_i - \bar{Y})^2 \tag{3G.8}$$

$$s_{XX} = \sum (X_i - \bar{X})^2 \tag{3G.9}$$

$$s_{XY} = \sum (X_i - \bar{X})(Y_i - \bar{Y}) \tag{3G.10}$$

$$\bar{X} = \text{the mean of } X_1, \ldots, X_n$$

and

$$\bar{Y} = \text{the mean of } Y_1, \ldots, Y_n$$

To compute the standard errors of these coefficients, it is convenient to first calculate the following intermediate quantities.

$$s_{UU} = s_{XX} + 2b\, s_{XY} + b^2 s_{YY} \tag{3G.11}$$

$$s_{VV} = s_{YY} - 2b\, s_{XY} + b^2 s_{XX} \tag{3G.12}$$

and

$$s_e = [s_{VV}/(n-2)]^{1/2} \tag{3G.13}$$

We then have

$$s.e.(a) = s_e \left[\frac{1}{n} + \frac{\bar{x}^2(1+b^2)^2}{s_{UU}} \right]^{1/2} \tag{3G.14}$$

and

$$s.e.(b) = s_e \left[(1+b^2)^2 / s_{UU} \right]^{1/2}. \tag{3G.15}$$

Approximate 95% confidence intervals on α and β can be obtained in the usual manner by computing the limits as estimate plus two standard errors and estimate minus two standard errors.

In all of the regressions, we assumed that the variances of the two variables, σ_ε and σ_ν, are equal, so that $\lambda = 1$. The calculation of a and b then reduces to computing the estimates of α and β by minimizing the sum of the squares of the perpendicular distances from the data points to the fitted line. When a variable has been multiplied by a constant, such as sulfur times 3 when comparing to sulfate, the concentrations are multiplied before calculating the regression; the assumption is that the variance in $3 \cdot S$ is equal to the variance in SO_4.

The primary weakness of the perpendicular method is the assumption that the uncertainties for each variable are constant. However, the uncertainties of aerosol concentrations generally vary significantly with the concentration. The general form of the uncertainty in concentration as a function of concentration c is

$$\sigma(c) = \sqrt{\sigma^2 + kc + (fc)^2}, \tag{3G.16}$$

where σ is the quadratic sum of all constant absolute uncertainties (such as artifact precision), f is the quadratic sum of all constant fractional uncertainties (such as flow rate and analytical calibration precisions), and k is a small constant whose form depends on the analytical method. The middle term, kc, is zero for gravimetric analysis and never large for any method. The uncertainty is nearly constant for small concentrations but increases linearly with concentration for $fc >> \sigma$. In most cases, the uncertainties for points near the origin are much less than those for points near the maxima. For example, at Page the precision of SO_2 varied by a factor of 80, S and SO_4 by 30, and Ca by 8. A few variables, such as mass, had nearly constant precision. Ideally, the fit would require smaller deviations near the origin that near the maxima. Because the perpendicular method minimizes all deviations equally, more weight is given to deviations at larger concentrations than is warranted by the uncertainties. This effect was checked using an alternative regression procedure that minimized the geometric means of the deviations,and allowed the individual propagated precisions to be included. In most cases, including the uncertainties slightly decreased the intercept, but did not significantly change the slope.

References

[1]Mandel, J., "Fitting straight lines when both variables are subject to error," *J. Quality Technology*, **16**, 1 (1984).