Residual Weighting Within PhazeComp

Aaron A. Zick, Ph.D.
Zick Technologies

January 22, 2007

PhazeComp allows regression on equation-of-state (EOS) parameters to minimize a sum of squares of weighted residuals between observed data and their corresponding EOS predictions. This document describes how the residuals are calculated and weighted.

The objective function that PhazeComp attempts to minimize is defined as

$$\chi^2 = \sum_{i=1}^{N} (w_i r_i)^2 , \tag{1}$$

where r_i is the residual between observed measurement i and its corresponding EOS prediction, while w_i is the user-assigned weighting factor for that residual. Ideally, the weighting factors should be inversely proportional to the standard deviations of the residuals. Then, statistically speaking, if the residuals are independent and normally distributed, the minimization of χ^2 will give the maximum likelihood estimation of the model parameters. PhazeComp does not enforce any requirement on the weighting factors, however, and assigns default weighting factors equal to 1. Because the weighting factors are arbitrary, it's a little difficult to assign a physical significance to χ^2. Therefore, instead, PhazeComp reports a root-mean-square (RMS) residual, $\bar{r}$, defined as

$$\bar{r} = \sqrt{\frac{\chi^2}{\sum_{i=1}^{N} w_i^2}}, \tag{2}$$

which is obviously related to χ^2 but is more easily interpreted.

The residuals themselves are calculated as relative percentages by the formula

$$r_i = 100(y_i^c - y_i^e)/y_i^r, \tag{3}$$

where y_i^e is the experimental value, y_i^c is the calculated value, and y_i^r is the reference value for observation i. For an isolated measurement, such as the saturation pressure of an experiment's incoming feed stream, the reference value will be the same as the experimental value, i.e., $y_i^r = y_i^e$. For a collection of related measurements, however, such as those within the same column of an experiment's data table, the reference value will be taken as the experimental value with the largest magnitude from among that particular collection of related measurements (counting only those with non-zero weighting factors). Whenever physical units are involved, the same absolute units are used consistently. If y_i^r turns out to have a value of zero, it will be replaced by 1 (in whatever units, if any, apply).

The weight factor for an isolated measurement is normally specified with the measurement. For a measurement within an experiment's data table, however, the weight factor is taken as the product of the weight factor assigned to the measurement's column and the weight factor assigned to its row. Again, all weight factors default to 1 unless the user specifies otherwise.

If the standard deviation, σ_i, is known for each measurement error, $y_i^c - y_i^e$, then it might be desirable to use the standard statistical formula for χ^2, namely

$$\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i^c - y_i^e}{\sigma_i} \right)^2 . \tag{4}$$

If so, then one can simply use

$$w_i = y_i^r / \left(100\sigma_i\right) \tag{5}$$

for the weighting factors. It is usually more practical, however, to assign the weighting factors heuristically, especially if the standard deviations are not easily estimated.