Weighted Calibration: some ideas.

Most of HPLC detector give a "linear" response Y where linear means proportional to the concentration X:

$Y = b.X$

We all know that there are random errors associated to measurements, thus is not possible to know the true response Y^* , all we can do is to estimate it. That is, given X, in theory one could take an infinite number of measurements of Y and (ignoring instrumental limitations to give a limited set of significant digits, and of course our time) we would observe a distribution of values around the true Y^*. As the population of possible values of Y is infinite, the probability of measuring Y^* is exactly zero (Oops!!).

We could carry out a number of measurements and take the average. Probability theory shows that the average will be closer to the true response Y^* as the number of measurements increases. The model that is usually assumed is

Assumption 1: Y^* depends on X, so that we'll have a $Y^*|X$ given a value of X. Therefore Y is a random variable having a mean Y^* which depends on X and variance $Var(Y^*|X)$ which could also depends on X

Assumption 2: Independence. A given Y is independent of one another

Assumption 3: The straight-line assumption. $Y^* = m^*.X + b^*$ where m^* is the true slope and b^* is the true intercept. Due to the impossibility to know Y^* is not possible to know m^* and b^*, we must estimate them with the amount

$$\hat{Y} = mX + b ,$$

The difference between the Y measured and the estimate is the residual $E = Y - \hat{Y}$ therefore the model is

$$Y = mX + b + E$$

X is not considered a random variable, the only random component is the residual E, which has a distribution with mean 0 and the same variance than Y.

Assumption 4 (Homoscedasticity) The variance of Y is the same for any X, meaning that variance does not depend on X. (This will be the key point in discussion regarding weighted regression)

Assumption 5. Errors are normally distributed. In fact we could require errors being symmetrically distributed and finite variance.

When this assumption are accomplished, Ordinary Least Square (OLS) method provides unbiased estimators for both the intercept and the slope, with the classical formulas that all we know. OLS minimizes the sum of square residuals (SSR), that is

$$SSR = \sum E^2 = \sum \left(Y - mX - b\right)^2$$

Now when variance for a given X_1 is different from that of another X_2 then Assumption 4 is violated and variance is a function of X. This situation is called hetroscedasticity, and part of the problem was discussed in the previous "Calibration curve" message .

Therefore we can write $Var(Y|X) = Var(E) = f(X)$. We saw that SSR is formed by terms with higher values as concentration increases. For a given X, $Var(E)$ is a constant and of course is different from zero. If we divided

$$e = \frac{E}{\sqrt{Var\ (E)}}$$

we will have a random variable e, a sort of "normalized" residual, with $Var(e) = 1$
The idea is to find m and b by minimizing the SSR calculated with the normalized residuals e instead of E. This procedure is called weighted least squares (WLS). The idea is that all e^2 terms has the same "weight" for all X.

$$WSSR = \sum e^2 = \sum \frac{E^2}{Var(E)} = \sum w_X \left(Y - mX - b\right)^2$$

where w_X is the weighing factor given by

$$w_X = \frac{1}{Var(E)} = \frac{1}{Var(Y|X)}$$

In order to keep calculation as simple as possible is better to define

$$w_X = \frac{n.(1/s^2_X)}{\sum_X (1/s^2_X)}$$

where n is the number of conc. levels and I wrote

$$s^2_X = Var(Y|X)$$

but we don't want to worry about such calculations here.
The problem is that we cannot know $s^2_X = Var(Y/X)$, we would need an infinite number of measurements for every X. But we have 2 alternatives

1) to estimate it for every concentration in the calibration curve. For instance to perform six replicated injections of each std. solution of conc. X and calculate the sample variance for every X. (6 replicated x 6 conc. = 36 inject. Too much work !!, we can't do it routinely, only for validation)
2) To model the variance as a function of X. I'll try this by my own responsibility !!

Let's write the Taylor's expansion of $s^2_X = f(x) = k_o + k_1.x_i + k_2.x_i^2 + \cdots$

Let's assume the response Y proportional to X. This hypothesis establishes that the true intercept b^* is zero.
RSD will be given by

$$RSD\% = 100 \times \frac{s_X}{Y} = 100 \times \frac{\sqrt{k_0 + k_1.x + k_2.x^2 + \cdots}}{m^*.x}$$

but we neither know s_X nor *Y*. Now, if we are near but just above LOQ we can take the two first terms of the Taylor's expansion and

$$s_X^2 \approx k_0 + k_1.x_i$$

But we expect that errors in $s_X^2 = Var(Y|X)$ trend to zero as X and therefore Y becomes smaller. That is

$$s_X^2 \approx k_1.x_i$$

The RSD will be given by

$$RSD\% = 100 \times \frac{s_X}{Y} \approx 100 \times \frac{\sqrt{k_1.x}}{m^*.x} \approx constant \times \frac{1}{\sqrt{x}}$$

and the weighing factor will be $w_X = \frac{n.(1/s_X^2)}{\sum_X (1/s_X^2)} \approx constant_2 \times \frac{1}{X}$

proportional to 1/X as appears in some integration software

But if we move far from LOQ, we can't take only two terms of the Taylor's expansion of s_X^2. We should take $s_X^2 \approx k_0 + k_1.x + k_2 x^2$. As X becomes larger, RSD would be

$$RSD\% = 100 \times \frac{s_X}{Y} \approx 100 \times \frac{\sqrt{k_0 + k_1.x + k_2.x^2}}{m^*.x} \xrightarrow[X \to \infty]{} constant$$

Therefore, if we move far from LOQ we should see RSD approximately constant.