

Weighted Least-Squares Regression in Practice: Selection of the Weighting Exponent

Hans-Joachim Kuss, Neurochemical Dept, Ludwig-Maximilians University, Munich, Germany.

Calibration is based on least squares regression analysis. It can be used under the assumption of homogeneous variances. Using chromatographic methods variances are often heteroskedastic. Weighting of variances is a valuable and simple tool to get realistic prediction intervals (y deviations) and uncertainties of the result (x deviations).

Calibration

Many analytical processes, including chromatography, relate detection signal y linearly to concentration x . For the purpose of calibration a straight line is determined with

$$y = bx + a$$

where b = the slope and a = the intercept.

A straight line is selected, for which the squares deviation between measured signal and the y value on the straight line (known as the residual standard deviation or RSD) is a minimum:

$$RSD^2 = \frac{\sum(y - y(x))^2}{n - 2} = \text{minimum}$$

where $y(x)$ = y value, which can be estimated from the straight line at concentration x , and n is the number of calibration points. To calculate the slope:¹

$$b = \frac{\sum(x - Mx)(y - My)}{\sum(x - Mx)^2} = \frac{SAW_{xy}}{SAW_{xx}}$$

where Mx = mean x , My = mean y and SAW = sum of squares deviations.

The straight line always leads through the centroid Mx, My . Therefore, with known slope b , the intercept a can be estimated:

$$a = My - b \cdot Mx$$

Residual Standard Deviation

From the overall variance of the y values (SDy^2), the percentage part of the variance can be determined, which is explained by the correlation between the x and y values:²

$$\text{Percentage fit: } r^2 = \frac{SDy^2 - RSD^2}{SDy^2} * 100$$

If there is no correlation between x and y , then $RSD^2 = SDy^2$, and $r^2 = 0$. With increasing correlation, RSD becomes smaller than SDy . For $RSD = 0$ a total correlation is given ($r^2 = 1 = 100\%$, with the correlation coefficient $r = 1$ or -1).

RSD is required to estimate the standard deviation of the slope and the intercept, but also for the limits of detection and quantification. Furthermore, RSD is essential for estimating the

prediction interval PI (y deviation) and the uncertainty of the result UOR (x deviation).

Computer Programs

Statistics programs are clearly able to find the slope and the intercept of the least squares straight line. In Excel it is possible to directly estimate b and a by using <paste><function> <slope> and <function> <intercept>. In the same way <function> <correl> and <function> <sterror> can be used to find the correlation coefficient r and the RSD. It is even easier, if analysis functions are installed, to enter <extras> <analysis functions> <regression> for a summary of the characteristics of a straight line.

Coefficient of Variation of the Straight Line

With known a and b, the unknown concentration of an analysis sample can be estimated from the measured signal value:

$$x = \frac{y - a}{b}$$

RSD is the standard deviation in the y direction. Analogous to the conversion from y into x values by division through slope b, the standard deviation in the x direction can be found. This is termed the process standard deviation (VSD), which is required by validation guidelines. Dividing VSD by Mx and multiplying by 100 leads to the process coefficient of variation, VVK:³

$$VSD = \frac{RSD}{b} \quad VVK = \frac{VSD * 100}{Mx}$$

VVK is an illustrative value that is not included in most statistic programs; however, in Excel it is relatively simple to obtain both the VSD and VVK values.

F-Test

Another requirement of validation guidelines is to determine the standard deviation of a signal at its lowest concentration x_u (S_{du}) and the highest concentration x_o (S_{do}) of the working range. Using the F-test, if significant differences can be found then the variances are heterogeneous:⁴

$$ABx = \frac{x_o}{x_u} \quad ABy = \frac{y_o}{y_u} \quad F = \frac{SD^2_o}{SD^2_u}$$

The least-squares regression is based on the assumption that the variances are homogeneous. With a significant F-test this precondition is proved incorrect, and the validation guidelines requesting a non-significant F-test must be neglected.

Realistically, in chromatographic processes using the working range $AB = x_o/x_u$ of more than 5, this will often happen. The critical F-value for six measured signals with a confidence of 95% is nearly 5.

Weighting Variances

To reach homogeneity of variances, they can be weighted with a weighting factor w:⁵

$$RSSw^2 = \frac{\sum w(y - y(x))^2}{n - 2} = \text{minimum}$$

The centroid is now:

$$M_{xw} = \frac{\sum xw}{n} \quad M_{yw} = \frac{\sum yw}{n}$$

$$b = \frac{\sum w(x - M_{xw})(y - M_{yw})}{\sum w(x - M_{xw})^2} \quad a = M_{yw} - b * M_{xw}$$

The linear regression can be seen as a special instance of weighted linear regression using a weighting factor of 1.

The usual integration programs for chromatography contain the possibility of weighting with $1/x$, $1/x^2$ or $1/y$, $1/y^2$. Because of the high correlation between y and x, there is only a

Sample preparation is normally a combination of volumetric steps and often the most variable part of the chromatographic method.

slight difference between y and x weighting, especially in instances when there is no significant difference between the intercept and zero. The same is true for pharmacokinetic programs, which may include a weighting with complex models. If the first concentration x is zero, then weighting by $1/x$ or $1/x^2$ is impossible. Weighting using y seems more practical, because one can assume that the standard deviation of y is more highly correlated to y than to x. Additionally, the effect of outliers decreases with weighting;⁶ weighting the signal of the analysis samples for the estimation of UOR is necessary.

Let us assume the deviations of a chromatographic process only come from the uncertainty of the injection. Then we have only volumetric deviations. Sample preparation is normally a combination of volumetric steps and often the most variable part of the chromatographic method. With dominating volumetric deviations we expect the same cv at all concentrations, which means increasing absolute standard deviations with higher concentrations. With $AB = 10$ we have nearly 10-fold standard deviations and, therefore, 100-fold variances at the upper concentration. This will be balanced by weighting with $1/x^2$ or $1/y^2$. The weighted quotient of the variances will be in the region of 1. The weighting, for example, with $1/y^2$ is performed in two steps:⁶

$$g = \frac{1}{y^{WE}} \quad w = \frac{g * n}{\sum g}$$

Excel

At this point it is no longer possible to use the Excel functions <slope> and <intercept>. One alternative is to create a table as described below:

Naming: Introduce the description of the procedure into cell A1: "Regression with weighting exponent=", into A2: "y=" and into C2: "x+". To expand the table with more input data, reserve the upper rows for results. Begin, for example, in row 5 and name the columns: x, y, xw, yw, g, w, Dxyw, Dxxw, yx, R, RRw. Excel does not accept y(x), therefore yx is used here. Only introduce text into cells A5 to K5. If no more than 10 concentrations will be used, mark A6 to A15 and assign this

region the name in cell A5. Using the standard symbols, find the default name of the upper left-hand cell (e.g., A6 to K6). Click on the cell and introduce the name to rename the cells with something more recognizable. Then use the names in the equations using indirect (A1) or direct (\$A\$1) addresses. Not using names will quickly render the equations difficult to understand. To see the number of introduced concentrations, input in F3 $|\text{=number}(x)|$ and call this cell "n". Insert in cell E3: "number=".

Weighting: To choose the weighting, enter into E1 the weighting exponent and call it WE. "0" means "no weighting"; "1" means "weighting with $1/y$ " and "2" means "weighting with $1/y^2$ ". Use "0" as the first choice. Column A and B are the input fields for x and y, which should contain example data. Now enter into E6: $|\text{=If}(x>0;y^{\wedge}\text{-WE};\text{""})|$ and pull the calculation rule from E6 to E15. The English version of Excel may use ";" (comma) as a delimiter for the functions instead of "," (semicolon). If in doubt use <paste><function> etc. Cell E4 should contain: $|\text{=sum}(g)|$ and is termed Sg. In F6 enter $|\text{=If}(x>0;n*g/Sg;\text{""})|$, which is expanded to F15. Now, the final weighting (w) will be in column F. Cell F4 contains: $|\text{=sum}(w)|$ called Sw, which must be identical with n (H1). Try WE=1 and WE=2 and look at the results for plausibility.

Calculation: Into cells A4 to D4, introduce $|\text{=mean}(x)|$ etc., and name them accordingly (i.e., Mx etc.). Then introduce $|\text{=sum}(Dxyw)|$, etc. into G4 to K4 and name them SAWxyw, SAWxxw, SAWyx, SR, SRRw. Into C6 should be entered $|\text{=If}(x>0;x*w;\text{""})|$ and into D6 $|\text{=If}(x>0;y*w;\text{""})|$ with the calculation specification pulled down. Into G6 must be entered $|\text{=If}(x>0;w*(x-Mxw)*(y-Myw);\text{""})|$ and into H6 $|\text{=If}(x>0;w*(x-Mxw)^2;\text{""})|$ and expanded to row 15. Calculation of the quotient in cell B2 ($|\text{=SAWxyw}/\text{SAWxxw}|$) gives the slope, which we call b. Then B4 will be filled out with $|\text{=Myw}-b*Mxw|$, which we call a. With known values of a and b, yx in column I: $|\text{=If}(x>0;b*x+a;\text{""})|$ can be calculated. Table 1 (below) shows the result of this.

Next, calculate the difference between calculated and measured y values: $|\text{=If}(x>0;yx;\text{""})|$ in column J. Into column K introduce the weighted squares of these differences: $|\text{=If}(x>0,R*R*w;\text{""})|$. K4 will contain the sum of weighted squares deviations SRRw. Into K3 insert $|\text{=Root}(SSRw/(n-2))|$ and call it RSD. Insert $|\text{=RSD}/b|$ into K2 to get the standard deviation of x and call it VSD. Insert into K1 $|\text{=VSD}*100/Mxw|$ to get the coefficient of VVK.

Graph: Mark the columns A, B and I between rows 6 to 15 (use Ctrl button) and create a new table with a point diagram using Excel's diagram assistant. Click on the straight line and choose the colour of the line (without highlighting the points), then click on the measured points and choose "without line" for a graph (Figure 1).

Table 1: Ordinary least squares regression using weighting exponent WE = 0.

Residuals: Marking column J will lead to a column diagram, which shows the residuals $R=y-y(x)$. The distribution of the residuals below and above the zero line gives the information, if a quadratic fit could give better results. Looking at the distribution of the residuals in the used concentration range illustrates whether variance homogeneity exists (Figure 2).

The data used for the example are test data for DIN 32645.⁷ This is standard input for the Excel program DINTTEST, available on the internet (www.rzuser.uni-heidelberg.de/~df6/tox/dintest.htm). Included in this program is a data file "Testdaten.xls", which compares the results of the test data for different programs. The above shown results are in accordance with these results. Unfortunately, no information is given concerning weighted regression.

The weighting exponent: Input of 1 or 2 for the weighting exponent (WE) in cell E1 only leads to small changes for the results of a and b, meaning that the coefficients of the straight line is very similar with and without weighting. The confidence interval is the result of the weighting only increased at high concentrations and decreased at low concentrations compared with unweighted regression. The straight line is not influenced as much. The point is that calculation of the uncertainty of the measurement in this way is more realistic. The uncertainty will be lower at low concentrations using the weighting. Therefore, the limits of detection and quantification are also (realistically) lower.

In the integration systems, the used praxis choosing weighting exponent (WE) (0, 1 or 2) is arbitrary and achieved by trial and error, which could be one reason for the low acceptance of weighted regression today. However, as shown above, the calculation of the regression line is only slightly more difficult and, therefore, this cannot be the reason.

Figure 1: Graph signal against concentration according to Table 1.

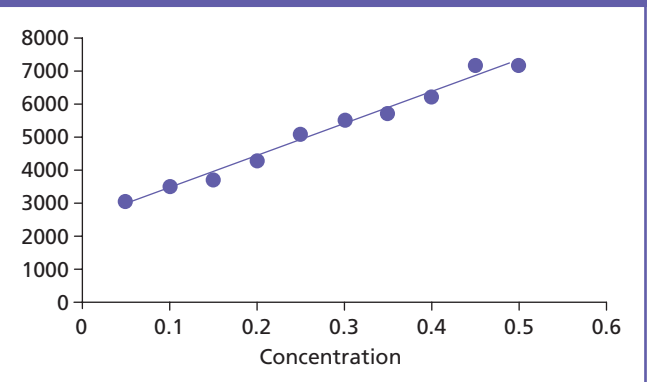
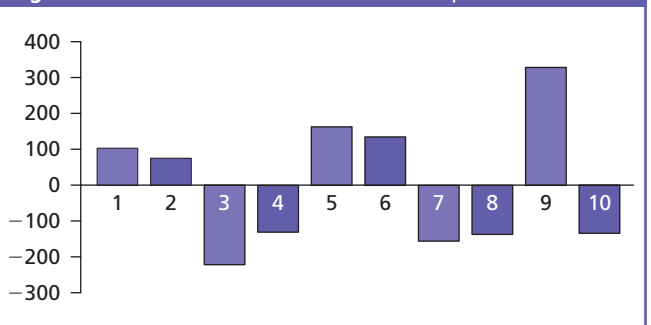


Figure 2: Residuals of the 10 concentration points.



Using the validation guidelines one has to measure the highest (xo) and the lowest (xu) concentration more than once (mostly 10-fold), for the F-test. The F-value can be used to determine an individual weighting exponent:⁸

$$WE = \frac{\log F}{\log ABy} = \frac{\log SDo^2 - \log SDu^2}{\log yo - \log yu}$$

The weighting exponent gives a standardization of the standard deviations SDo and SDu (only at these concentrations the standard deviations are known) in such a way that $F = 1$. Therefore, this method uses the available information optimally and the conditions for linear regression are fulfilled.

Uncertainty of the result: To calculate the uncertainty of the result UOR (see last two parts of the equation), the main term is the standard deviation in the x direction ($VSD = RSD/b$) multiplied by the student's t value. Then follows a correction term with three expressions under a square root. The first two expressions are $1/w + 1/n$. For 10 concentrations and a weighting factor of 1 these two terms are 1.1 (root 1) and the square root is 1.05. The complete equation⁹ is:

$$X_a = \frac{y_a - a}{b} \pm \frac{t \cdot RSD}{b} * \sqrt{\frac{1}{w} + \frac{1}{n} + \frac{(y_a - My)^2}{b^2 \sum (x - Mx)^2}}$$

root 1 root 2

This demonstrates that we must be able to calculate the required weighting factor for the measured signal y.

The third term (root 2) of the equation is zero at the centroid point and only the above mentioned factor is valid. At the other y values there is an additional slight widening towards the ends of the calibration curve. The weighting has the additional effect of widening the uncertainty of the result UOR at high concentrations and narrowing UOR at low concentrations, as it has been found experimentally.

UOR with Excel

For clarity, open a second table in the same working sheet (perhaps name it 'Uncert'). Insert into cell A1 |UOR interval with weighting exponent=| and in F1: |=WE|. Row 5 should

Table 2: Input data to calculate the uncertainty of the result.

Table 3: Results found for weighted regression using $WE = 2$ and uncertainty of the result UOR.

again be reserved for the names of columns, and calculation should take place in rows 6 to 15. The inputs are:

The signal of 3500 was chosen to compare the results with Testdaten.xls, in which an error probability of 1% was used. A result of 0.105 ± 0.074 is not acceptable and below the limit of quantification, for which a maximum interval of 33% is allowed. With a signal of 6000 one finds 0.364 ± 0.071 ($\pm 20\%$); a VVK of 7.24% should be multiplied by 3 to reach a significance of 99%.

Table 3 shows the results of the example using a weighting exponent of 2. Slope and intercept show only slight differences. The UOR (in Table 3 transferred from Excel Table 2 to the end of Excel Table 1 to show more details.) is smaller at low signals and larger at high signals in comparison to ordinary least squares regression.

One should try the effect of weighting with one's own examples to decide if this method will achieve better results. In all instances, a significant F-test offers an elegant option for overcoming this problem and calculating results and UOR more realistically. The calculation of the prediction interval PI and more practical examples are shown elsewhere.^{10,11}

An intensive reflection of weighted regression can be found by Miller,¹² who concluded: "All [these] results accord much more closely with the reality of a calibration experiment than do the results of the unweighted regression calculation."

Especially with pharmacokinetic studies for the registration of drugs, the weighting deduced from the measurement of concentrations could possibly be carried forward to the weighting for pharmacokinetic calculations.

References

1. R. Caulcutt and R. Boddy, *Statistics for Analytical Chemists*, (Chapman and Hall 1983), 80.
2. H. Motulsky, *Intuitive Biostatistics*, (Oxford University Press 1995), 173.
3. W. Funk, V. Dammann and G. Donnevert, *Qualitätssicherung in der Analytischen Chemie*, (VCH-Verlag 1992), 14.
4. S. Burke, *LC*GC Europe Statistics and Data Analysis Online Supplement*, p.6.
5. R. Caulcutt and R. Boddy, *Statistics for Analytical Chemists*, (Chapman and Hall 1983), 103-104.
6. S. Burke, *LC*GC Europe Statistics and Data Analysis Online Supplement*, p.17.
7. DIN 32645, Beuth-Verlag 1994.
8. H.J. Kuss, in *Handbuch Validierung in der Analytik*, S. Kromidas, Ed. (Wiley-VCH, 2000), 182.
9. R. Caulcutt and R. Boddy, *Statistics for Analytical Chemists*, (Chapman and Hall 1983), 106.
10. H.J. Kuss, *Quantitative Auswertung*, in *HPLC Tips II*, S. Kromidas, Ed., (Hoppenstedt 2003), in press.
11. H.J. Kuss, *Quantitative Evaluation*, in *More Practical Problem Solving in HPLC*, S. Kromidas, Ed., (Wiley-VCH, 2004), in press.
12. J.N. Miller and J.C. Miller, *Statistics and Chemometrics for Analytical Chemistry*, (Prentice Hall 2000), 135.

Hans-Joachim Kuss trained as a physicochemist. He has been engaged in research activities since 1976 in the Neurochemical Department of the psychiatric clinic at Ludwig-Maximilians University, Munich, Germany. He has developed chromatographic methods for the determination of catecholamines, indolamines, their metabolites and of psychoactive drugs in body fluids. He conducts lectures in GC, GC-MS, HPLC, pharmacokinetics and the interaction of psychopharmaca. In 1986 and 1987, he was regional manager for Waters in Bavaria.
E-mail: Kuss@med.uni-muenchen.de