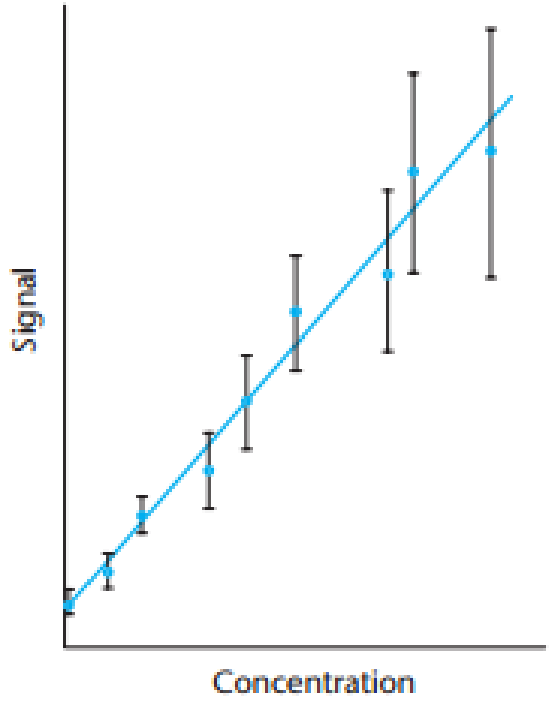
MA4605 2016 - Weighted Regression Models

5.10 Weighted regression lines

In OLS based simple linear models, we shall expect the y-direction errors in the regression curve to be approximately equal for all the points (***homoscedasticity – i.e. constant variance***), and an (unweighted) regression calculation is legitimate.

However, this assumption of homoscedascity may not be valid, and instead the variance of the residuals may be found to vary across the range of measurements. This is known as ***heteroscedasticity*** (or colloquilally the "funnel effect").



The weighting of errors in a regression calculation

In some cases the errors will be approximately proportional to analyte concentration (i.e. the relative error will be roughly constant), and in still others (perhaps the commonest situation in practice) the y-direction error will increase as x increases, but less rapidly than the concentration.

Both these types of heteroscedastic data should be treated by ***weighted regression methods***. Usually an analyst can only learn from experience whether weighted or unweighted methods are appropriate.

Predictions are difficult: examples abound where two apparently similar methods show very different error behaviour. Weighted regression calculations are rather more complex than unweighted ones, and they require more information (or the use of more assumptions).

Nonetheless they should be used whenever heteroscedasticity is suspected, and they are now more and more widely applied, partly as a result of pressure from regulatory authorities in the pharmaceutical industry and elsewhere.

The figure on the last page shows the simple situation that arises when the error in a regression calculation is approximately proportional to the concentration of the analyte, i.e. the ``error bars" used to express the random errors at different points on the calibration get larger as the concentration increases.

The regression line must be calculated to give additional weight to those points where the error bars are smallest: it is more important for the calculated line to pass close to such points than to pass close to the points representing higher concentrations with the largest errors.

This result is achieved by giving each point a weighting inversely proportional to the corresponding variance, s_i^2.

(This logical procedure applies to all weighted regression calculations, not just those where the y-direction error is proportional to x.)

Thus, if the individual points are denoted by (x_1, y_1), (x_2, y_2), etc. as usual, and the corresponding standard deviations are s_1, s_2, etc., then the individual weights, w_1, w_2, etc., are given by:

$$\text{Weights:} \qquad w_i = \frac{s_i^{-2}}{\sum_i s_i^{-2}/n} \qquad (5.14)$$

It will be seen that the weights have been scaled so that their sum is equal to the number of points on the graph: this simplifies the subsequent calculations.

The slope and the intercept of the regression line are then given by the following equations:

$$\text{Weighted slope:} \qquad b_w = \frac{\sum_i w_i x_i y_i - n\bar{x}_w \bar{y}_w}{\sum_i w_i x_i^2 - n\bar{x}_w^2} \qquad (5.15)$$

$$\text{Weighted intercept:} \qquad a_w = \bar{y}_w - b\bar{x}_w \qquad (5.16)$$

Weighted Centroid

In equation (5.16) $\bar{y}_w$ and $\bar{x}_w$ represent the coordinates of the *weighted centroid*, through which the weighted regression line must pass. These coordinates are given as expected by $\bar{x}_w = \sum_i w_i x_i / n$ and $\bar{y}_w = \sum_i w_i y_i / n$.

The unweighted centroids are simply the means of the observations for both the independent and dependent variables: .

The Simple linear regression fitted line must past through these centroids.

Example 5.10.1

Calculate the unweighted and weighted regression lines for the following calibration data. For each line calculate also the concentrations of test samples with absorbances of 0.100 and 0.600.

Concentration, μg ml^{-1}	0	2	4	6	8	10
Standard deviation Absorbance	0.001 0.009	0.004 0.158	0.010 0.301	0.013 0.472	0.017 0.577	0.022 0.739

```
Conc=(0,2,4,6,8,10)
Abso=(0.009,0.158,0.301,0.472,0.577,0.739)
```

Simple Linear Regression Model

```
> Fit1 = lm(Abso~Conc)
> summary(Fit1)

Call:
lm(formula = Abso ~ Conc)

Residuals:
         1          2          3          4          5          6
-0.0042857 -0.0003714 -0.0024571  0.0234571 -0.0166286  0.0002857

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.013286   0.010559   1.258    0.277
Conc        0.072543   0.001744  41.602    2e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01459 on 4 degrees of freedom
Multiple R-squared: 0.9977,     Adjusted R-squared: 0.9971
F-statistic:  1731 on 1 and 4 DF,  p-value: 1.995e-06
```

Regression Equation : Abso = 0.0132 + 0.0725Conc

Weighted Regression Model

```
> Abso.sd=c(0.001,0.004,0.010,0.013,0.017,0.022)
> weights=Abso.sd^(-2)/mean(Abso.sd^(-2))
> summary(Fit2)

Call:
lm(formula = Abso ~ Conc, weights = weights)

Weighted Residuals:
         1          2          3          4          5          6
-0.0001974  0.0008212 -0.0007349  0.0036841 -0.0030674 -0.0008217

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.009084   0.001048   8.671 0.000974 ***
Conc        0.073760   0.001064  69.330 2.59e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.002495 on 4 degrees of freedom
Multiple R-squared: 0.9992,     Adjusted R-squared: 0.999
F-statistic:  4807 on 1 and 4 DF,  p-value: 2.593e-07
```

Regression Equation : Abso = 0.0090 + 0.0737Conc

From the book: (Not discussed in lectures, apart from predicted values)

Results of predicted values for ***X = 0.100*** and ***0.600*** are below.

Application of equations (5.4) and (5.5) shows that the slope and intercept of the *unweighted* regression line are respectively 0.0725 and 0.0133. The concentrations corresponding to absorbances of 0.100 and 0.600 are then found to be 1.20 and 8.09 μg ml^{-1} respectively.

The *weighted* regression line is a little harder to calculate: in the absence of a suitable computer program it is usual to set up a table as follows.

x_i	y_i	s_i	$1/s_i^2$	w_i	$w_i x_i$	$w_i y_i$	$w_i x_i y_i$	$w_i x_i^2$
0	0.009	0.001	10^6	5.535	0	0.0498	0	0
2	0.158	0.004	62500	0.346	0.692	0.0547	0.1093	1.384
4	0.301	0.010	10000	0.055	0.220	0.0166	0.0662	0.880
6	0.472	0.013	5917	0.033	0.198	0.0156	0.0935	1.188
8	0.577	0.017	3460	0.019	0.152	0.0110	0.0877	1.216
10	0.739	0.022	2066	0.011	0.110	0.0081	0.0813	1.100
Sums			1083943	5.999	1.372	0.1558	0.4380	5.768

These figures give $\bar{y}_w = 0.1558/6 = 0.0260$, and $\bar{x}_w = 1.372/6 = 0.229$. By equation (5.15), b_w is calculated from

$$b_w = \frac{0.438 - (6 \times 0.229 \times 0.026)}{5.768 - [6 \times (0.229)^2]} = 0.0738$$

so a_w is given by $0.0260 - (0.0738 \times 0.229) = 0.0091$.

These values for a_w and b_w can be used to show that absorbance values of 0.100 and 0.600 correspond to concentrations of 1.23 and 8.01 μg ml^{-1} respectively.

(These calculations will not be examinable)

Comparison of the results of the unweighted and weighted regression calculations is very instructive.

Weighted Centroids:

$$\bar{y}_w = 0.1558/6 = 0.0260, \quad \text{and} \quad \bar{x}_w = 1.372/6 = 0.229.$$

Unweighted Centroids

```
> mean(Conc)
[1] 5
> mean(Abso)
[1] 0.376
```

The effects of the weighting process are clear. The ***weighted centroid*** is much closer to the origin of the graph than the ***unweighted centroid*** and the weighting given to the points nearer the origin (particularly to the first point (0, 0.009) which has the smallest error) ensures that the weighted regression line has an intercept very close to this point.

The slope and intercept of the weighted line are remarkably similar to those of the unweighted line, however, with the result that the two methods give very similar values for the concentrations of samples having absorbances of 0.100 and 0.600. It must not be supposed that these similar values arise simply because in this example the experimental points fit a straight line very well.

In practice the weighted and unweighted regression lines derived from a set of experimental data have similar slopes and intercepts even if the scatter of the points about the line is substantial.

As a result it might seem that weighted regression calculations have little to recommend them. They require more information (in the form of estimates of the standard deviation at various points on the graph), and are far more complex to execute, but they seem to provide data that are remarkably similar to those obtained from the much simpler unweighted regression method.

Such considerations may indeed account for some of the neglect of weighted regression calculations in practice. But an analytical chemist using instrumental

methods does not employ regression calculations simply to determine the slope and intercept of the calibration (i.e. regression) plot and the concentrations of test samples.

There is also a need to obtain estimates of the errors or confidence limits of those concentrations, and it is in this context that the weighted regression method provides much more realistic results.

Confidence intervals for fitted values

Previously we estimated the standard deviation and hence the confidence limits of a concentration calculated using a single y-value and an unweighted regression line. (In book - equation (5.9))

Application of this method for the example above shows that the unweighted confidence limits for the solutions having absorbances of 0.100 and 0.600 are 1.20 ± 10.65 and 8.09 ± 10.63 $\mu g\ ml^{-1}$ respectively.

These confidence intervals are very similar. In the present example, however, such a result is entirely unrealistic. The experimental data show that the errors of the observed y-values increase as y itself increases, the situation expected for a method having a roughly constant relative standard deviation.

We would expect that this increase in standard deviation, with increasing y would also be reflected in the confidence limits of the determined concentrations: the confidence limits for the solution with an absorbance of 0.600 should be much greater (i.e. worse) than those for the solution with an absorbance of 0.100.

In weighted recession calculations, the standard deviation of a predicted concentration is given by:

In weighted recession calculations, the standard deviation of a predicted concentration is given by:

$$s_{x_{0w}} = \frac{s_{(y/x)w}}{b}\left\{\frac{1}{w_0} + \frac{1}{n} + \frac{(y_0 - \bar{y}_w)^2}{b^2\left(\sum_i w_i x_i^2 - n\bar{x}_w^2\right)}\right\}^{1/2} \qquad (5.17)$$

In this equation, $s_{(y/x)w}$ is given by:

$$s_{(y/x)w} = \left\{\frac{\sum_i w_i(y_i - \hat{y}_i)^2}{n-2}\right\}^{1/2} \qquad (5.18)$$

and w_o is a weighting appropriate to the value of y_o.

[Remark : Equations (5.17) and (5.18) are clearly similar in form to equations (5.9) and (5.6).]

Equation (5.17) confirms that points close to the origin, where the weights are highest, and points near the centroid, where the difference between the observed value for y and the centroid is small, will have the narrowest confidence limits (see image on next page).

The major difference between equations (5.9) and (5.17) is the term $1/w_o$ in the latter. Since w_o falls sharply as y increases, this term ensures that the confidence limits increase with increasing y_o, as we expect.

Application of equation (5.17) to the data in the example above shows that the test samples with absorbance of 0.100 and 0.600 have confidence limits for the calculated concentrations of 1.23 ± 0.12 and 8.01 ± 0.72 µg ml^{-1} respectively.

The widths of these confidence intervals are proportional to the observed absorbances of the two solutions.

In addition the confidence interval for the less concentrated of the two samples is smaller than in the unweighted regression calculation, while for the more concentrated sample the opposite is true.

All these results accord much more closely with the reality of a ***calibration experiment*** than do the results of the unweighted regression calculation.

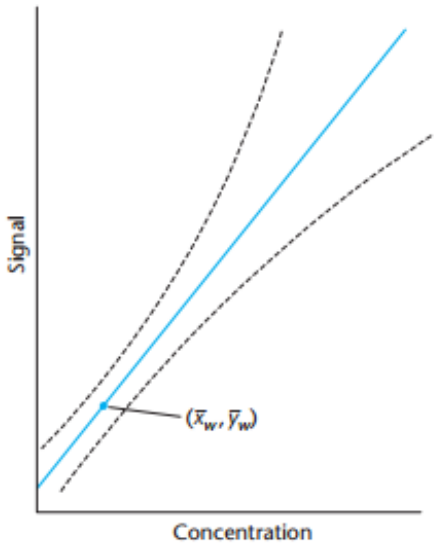


Figure 5.13 General form of the confidence limits for a concentration determined using a weighted regression line.

In addition, weighted regression methods may be essential when a straight line graph is obtained by algebraic transformations of an intrinsically curved plot.

Key points:

- Homoscedascity and Heteroscedascity
- When is it appropriate to use Weighted Linear Regression
- Comparison of width of confidence Intervals for fitted values under unweighted and weighted linear regression