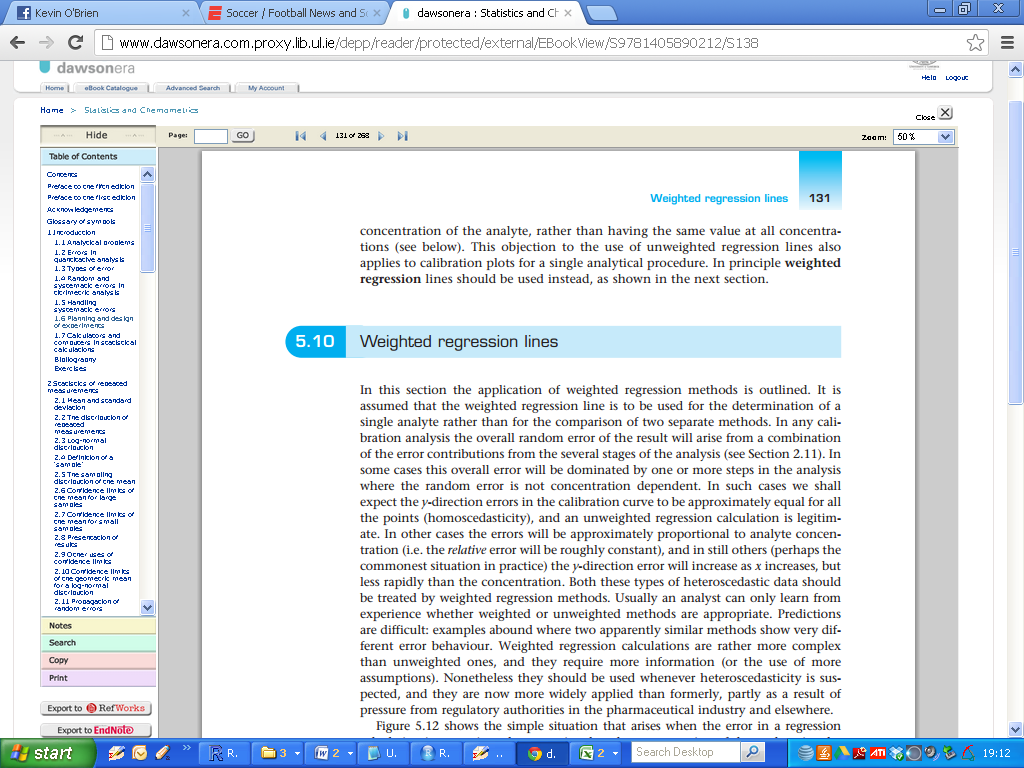
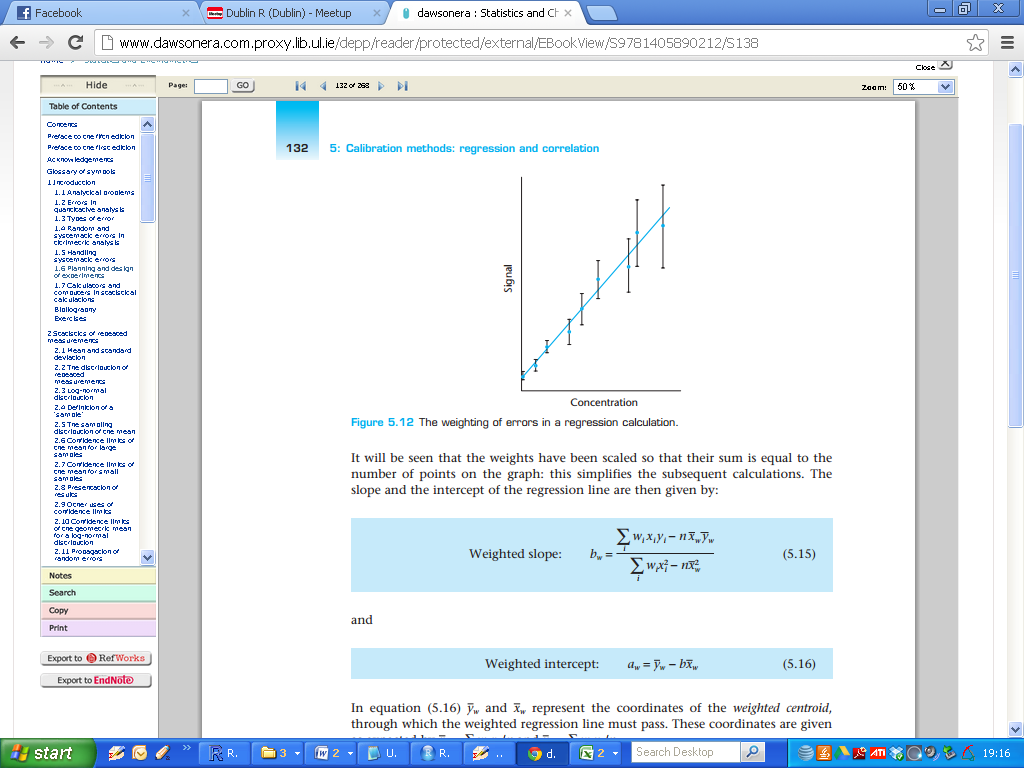
**MA4605 2016 - Weighted Regression Models**



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.

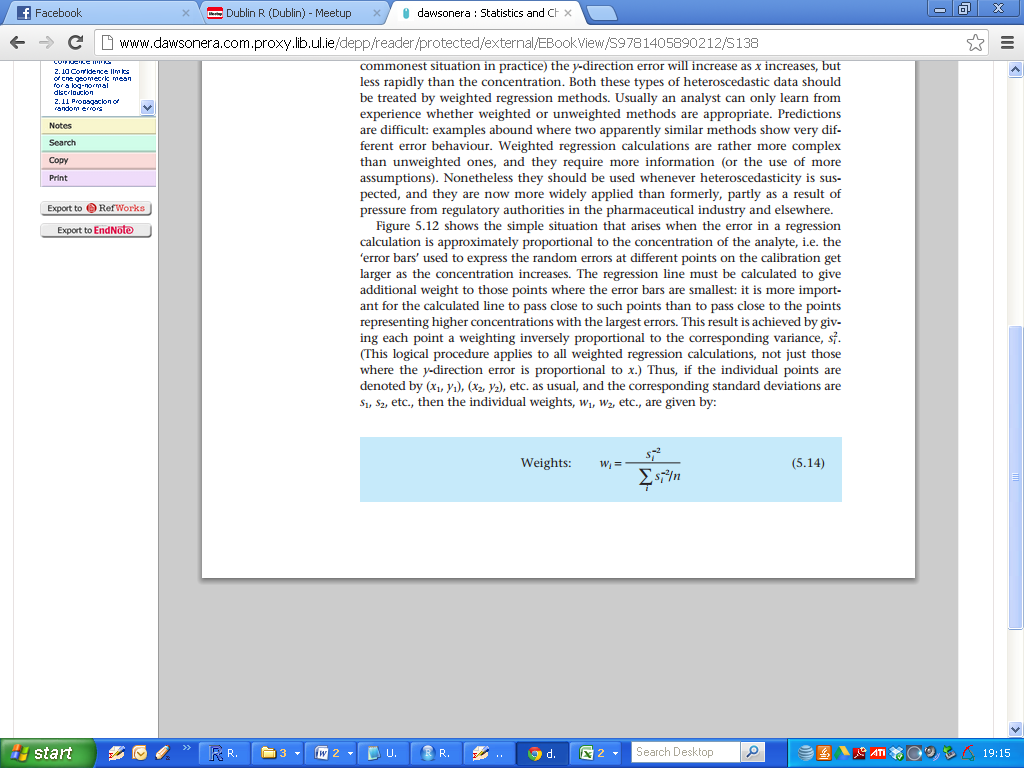
The figure on the last page shows the simple situation that arises when the error in a regression calculation is approximately proportional to the concentration of the analyte, i.e. the ``error bars” used to express the random errors at different points on the calibration get larger as the concentration 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, ***s2i***.

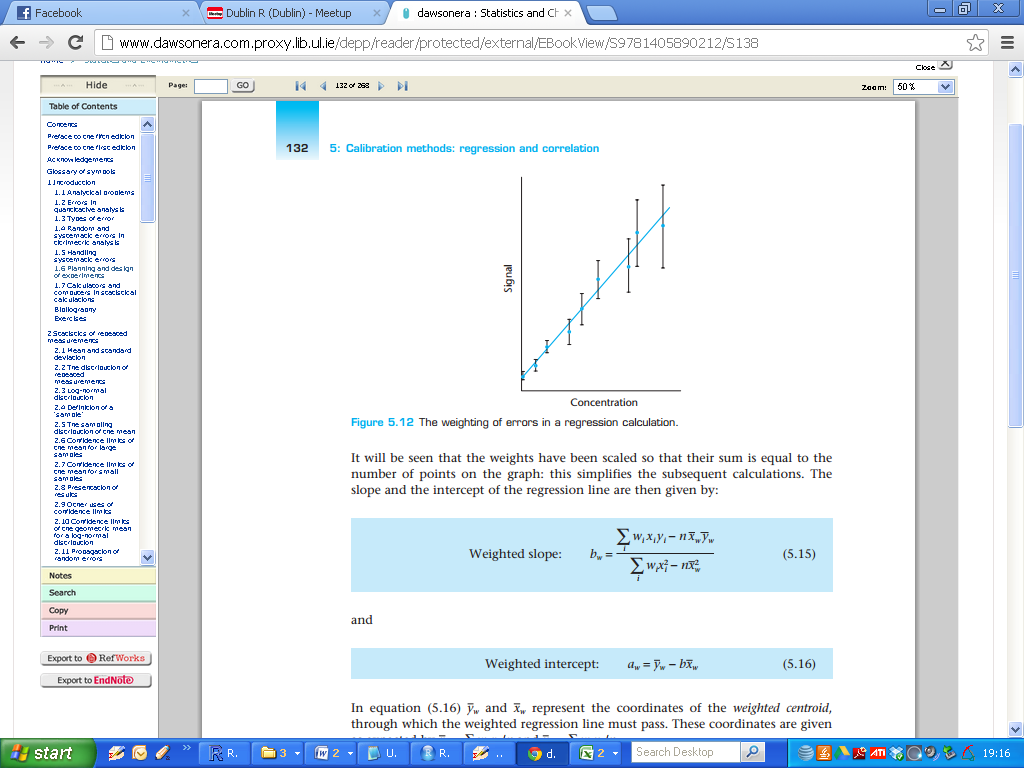
(This logical procedure applies to all weighted regression calculations, not just those where the y-direction error is proportional to x.)

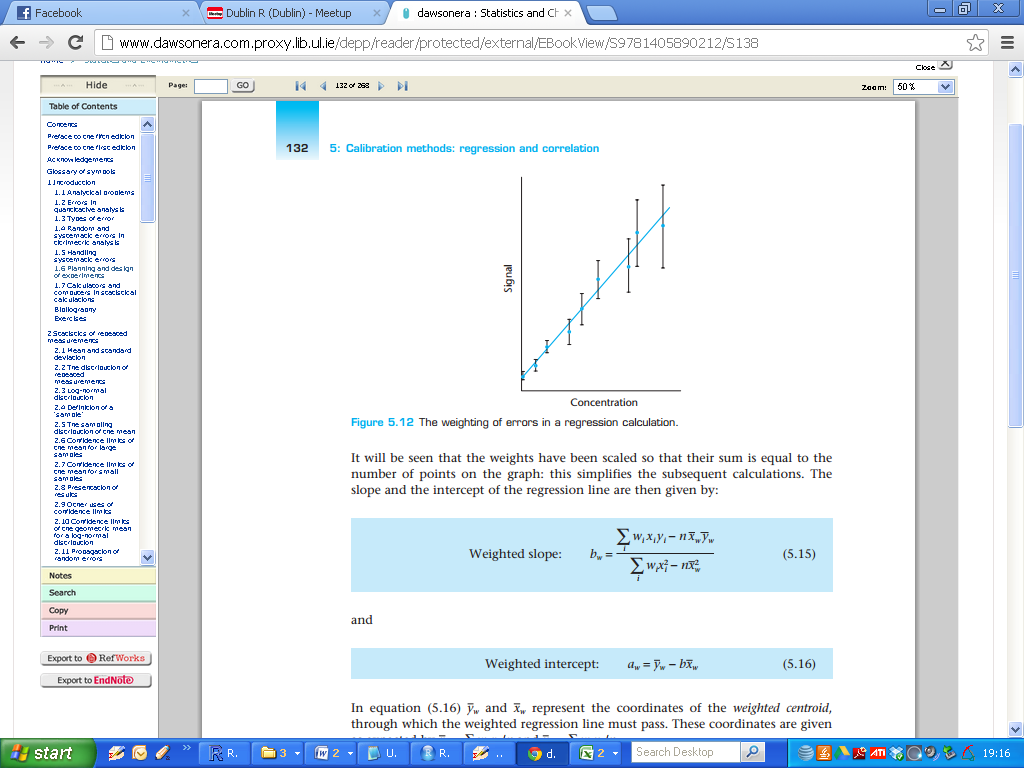
Thus, if the individual points are denoted by (x1, y1), (x2, y2), etc. as usual, and the corresponding standard deviations are s1, s2, etc., then the individual weights, w1, w2, etc., are given by:



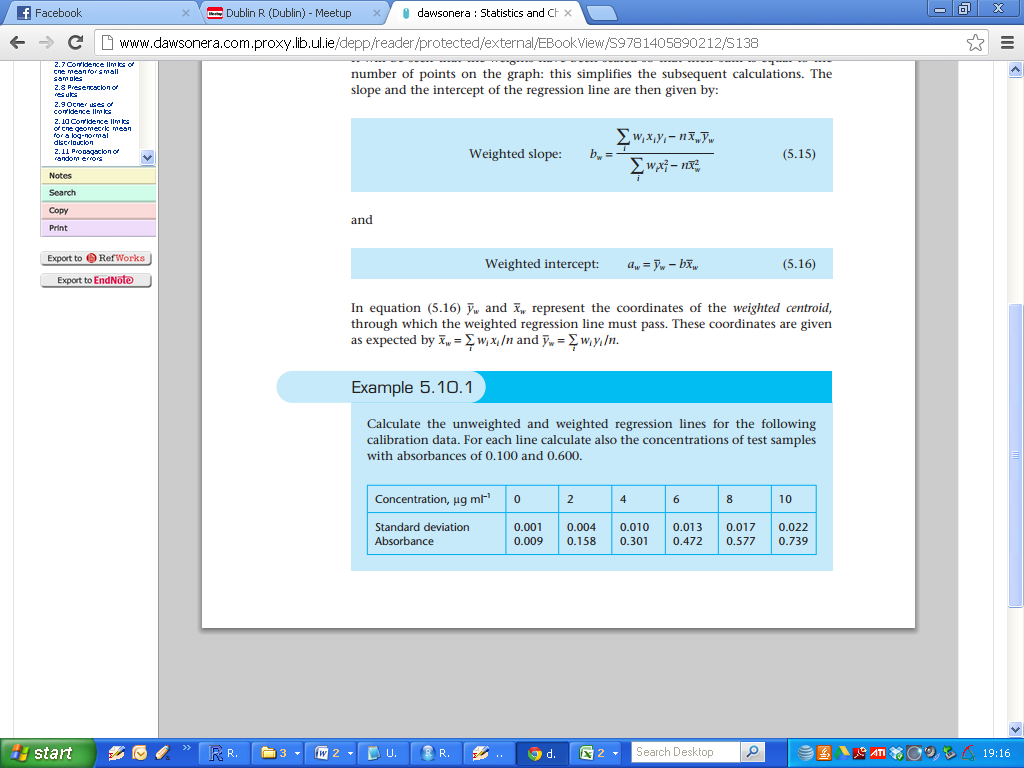
It will be seen that the weights have been scaled so that their sum is equal to the number of points on the graph: this simplifies the subsequent calculations.

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



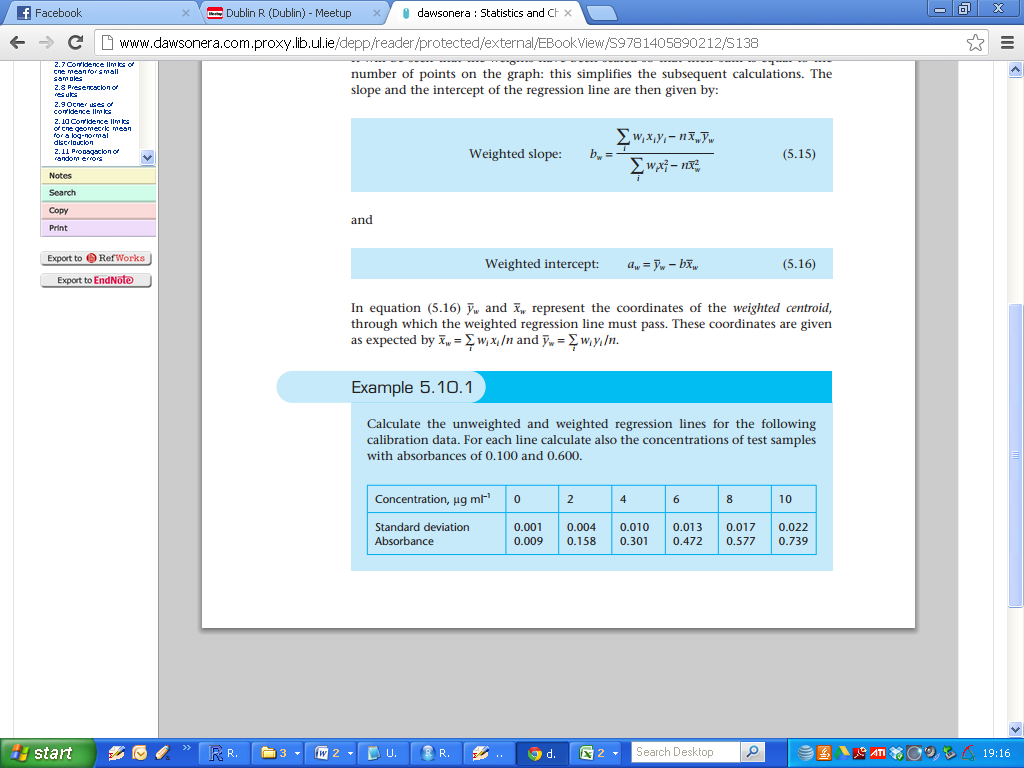


**Weighted Centroid**



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.



|  |
| --- |
| 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***

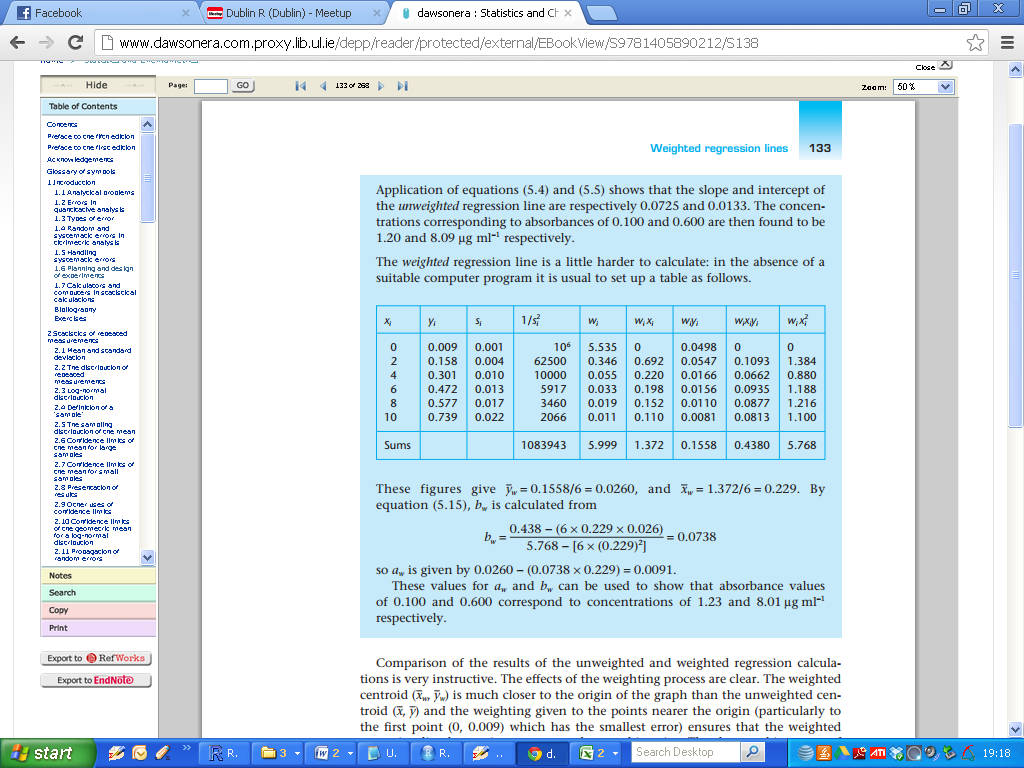
**Weighted Regression Model**

|  |
| --- |
| > 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.



(These calculations will not be examinable)

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

|  |  |
| --- | --- |
| ***Weighted Centroids:***    ***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**

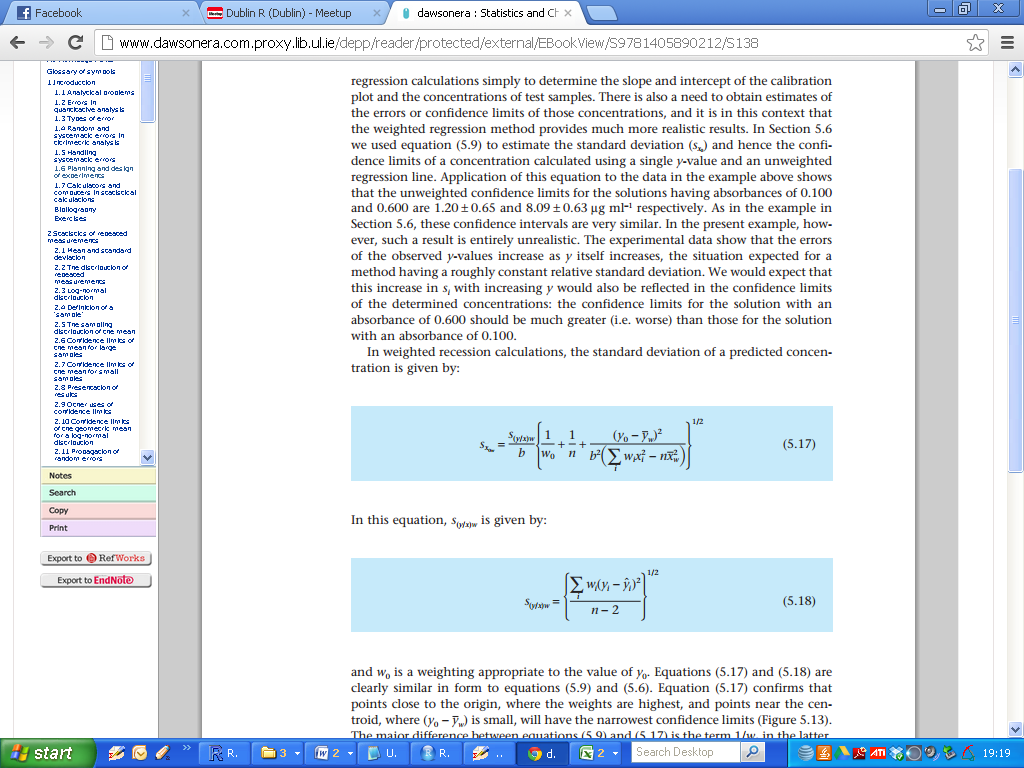
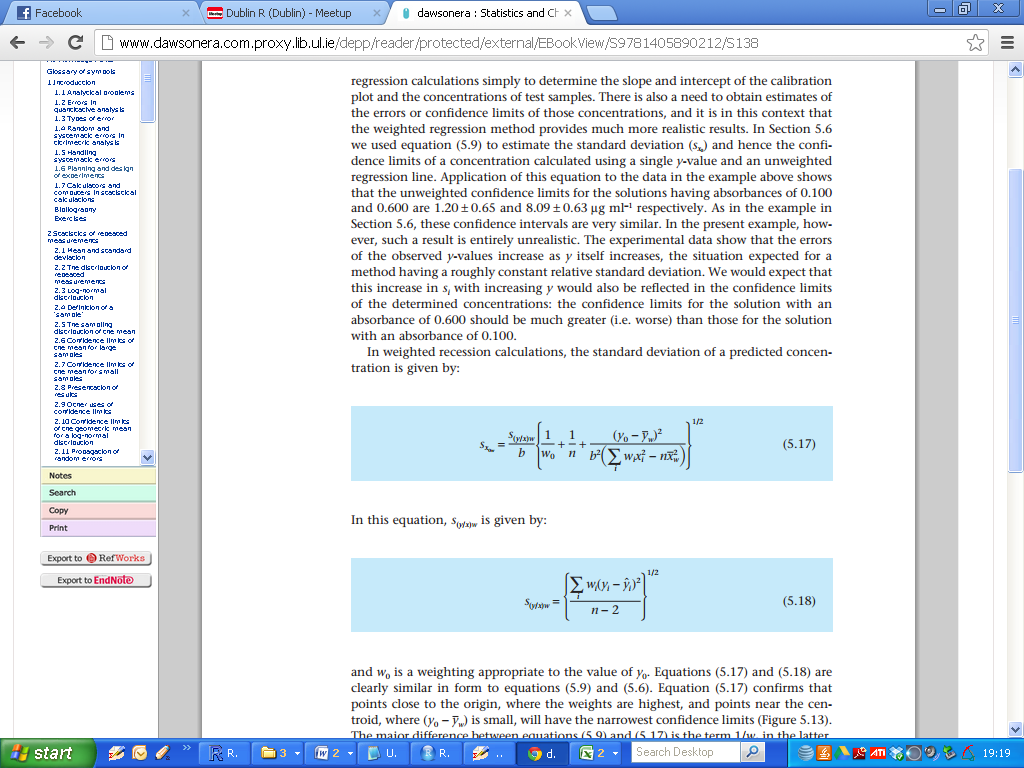
Previously we estimated the standard deviation and hence the confidence limits of a concentration calculated using a single y-value and an unweighted regression line. (In book - equation (5.9) )

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 μg 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:



and ***wo*** is a weighting appropriate to the value of ***yo***.

[ 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).

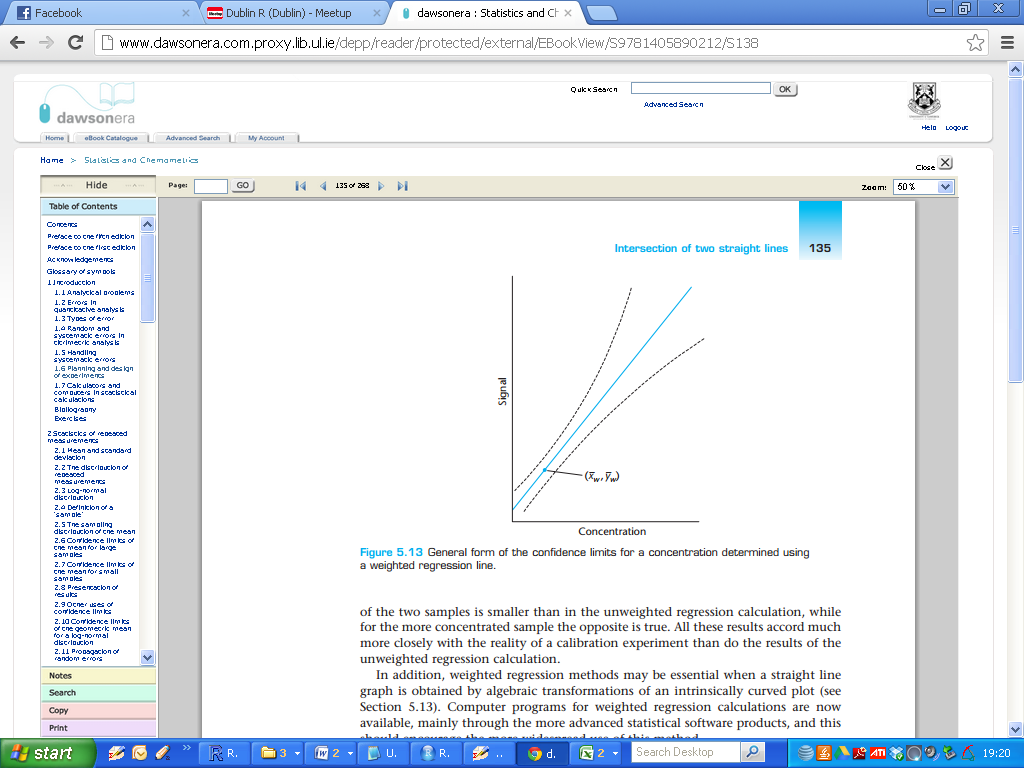
The major difference between equations (5.9) and (5.17) is the term ***1/ wo*** in the latter. Since ***wo*** falls sharply as y increases, this term ensures that the confidence limits increase with increasing ***yo***, as we expect.

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.



ln 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