CALIBRATION

Introduction.
When applying analytical chemistry one is often interested in examining how one variable or condition affetcs another. Typically, a response (UV-absorbtion, an area in a chromatogram etc.) changes as a function of the change of the concentration of one or more components (analytes) in the sample.
The simplest – useful – assumption is that the response varies dirctly proportional with the change of a single component in the sample:

(1) response = constant▪concentration

The constant is often called the sensitivity. Often equation (1) actually applies athough for practical purposes it is usually sufficient that the response varies sufficiently linearly as a function of the change. This can be the case, if the prerequisites of the analytical method is adhered to, e.g. the validated concentration range, measuring method, measuring equipment etc.
A practical example where equation (1) applies would be Beer-Lamberts law:

(2) $A = \epsilon \cdot l \cdot c$

In real life it can be necessary to extend equation (1) a little:

(3) response = a_1▪concentration + a_0

Where a_1 is the sensitivity and a_0 is the bias. The bias could be due to e.g. the phenomenon that sometimes a response is observed even though no analyte is present. On the other hand, sometimes it is not possible to observe a response until a certain (hopefully small) amount of the analyte is present, se figure 1. This could be due to e.g. adsorption of the analyte to glassware, filters etc.

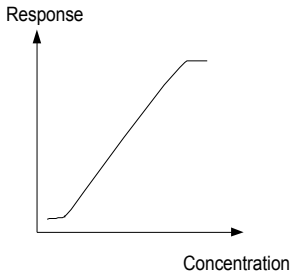

Figure 1. Expected response function for an analytical method

It is known from working with Beer-Lamberts law in spectrophotometry that there is a limit below which no changes in absorbtion is observed when the concentration of the analyte is changed and likewise a higher limit where the measuring equipment – even with an infinitely strong light source – can not deliver a linear change in the measured absorbtion when the concentration of the analyte is increased.

It should be noted that a linear relation between response and concentration is not required (see e.g. ICH guideline: Validation of Analytical Procedures: Methodology, section 2, Linearity), allthough it is usually preferred. If the response is not (sufficiently) linear, it is necessary to know the theoretical relation between response and concentration, or perform a meticulous validation of a nonlinear model.

References
The content of this chapter is mainly based on Draper og Smiths work: “Applied Regression Analysis” (reference 1), which is recommended if a deeper understanding of regression analysis is sought.
It is also recommended to read: "On the misinterpretation of the correlation coefficient in pharmaceutical sciences". J. M. Sonnergaard. International Journal of Pharmaceutics 321, 2006, 12-17.

Linear regression
Given that the measuring conditions has been determined, the next step is to determine the (empirical/estimated) values of a_1 and a_0. For a spectrophotometrical method this can be done by measuring the absorbance for a series of solutions suitably distributed over the selected concentration range (the range of the analysis). The actual calculation of a_1 and a_0 is done by performing a linear regression analysis, e.g. by means of a pocket calculator, a spreadsheet on a computer or a statistics program on a computer

Real data often suffers from random errors -noise – so that the analytical response should be desribed by the following equation[1]:

(4)
$$Y = A_1 \cdot X + a_0 + error$$

Where a_1 and a_0 applies to all data and *error* is a random error depending on other factors than concentration.

When the linear regression is calculated, a_1 and a_0 are determined so that for any given X (in the concentration range) a Y-value can be estimated, i.e. the response that would have been measured if a solution of a concentration corresponding to the X-value was examined. This is desribed by the following equation:

(5) $\hat{Y} = a_1 \cdot X + a_0$ (so that the measured $Y = \hat{Y} + error$)

where $\hat{Y}$ is the estimated value. Note that in analytical chemistry one is really more interested in the "reverse" estimate, that is, we measure the Y-values for an unknown sample and wish to estimate X (e.g. a concentration) with a suitable accuracy and prescision[2], corresponding to equation 6:

(6) $$\hat{X} = \frac{(Y - a_0)}{a_1}$$

When the notation corresponding to equation 5 is used it is based on the assumption, that the measurement error is on the Y-values while the X-values are considered to be "error free".

If n concentrations has been measured, a data set consisting of n concentrations and n response values is available:

(7) $\{(X_1, Y_1), (X_2, Y_2),, ..., (X_n, Y_n)\}$

For a given datum consisting of an X- and a Y-value, e.g. The i^{th}, equation 4 is written:

(8) $$Y_i = a_1 \cdot X_i + a_0 + error_i$$

or:

(9) $$error_i = Y_i - a_1 \cdot X_i - a_0$$

i can take the values 1,2,...,n usually written thus:
$i \rightsquigarrow \{1,2,...,n\}$.

---

1 The equation: $Y = a_1 \cdot X + a_0 + error$ is linear in the parameters (a_1, a_0) in the context of regression analysis.

2 Rule of thumb: It is possible to miss with unfailing precision/Man kan ramme ved siden af med usvigelig præcision

One can then write the sum of the squares of the deviation from the "true" line as:

(10) $$S=\sum_{i=1}^{n} error_i^2=\sum_{i=1}^{n}(Y_i-a_0-a_1\cdot X_i)^2$$

Your calculator or computer adjusts a_1 and a_0 so that S becomes as small as possible. This method is therefore called the method of least sqaures. The derivation of the eqautions does not interest us here, the equations for a_1 and a_0 becomes[3]:

(11) $$a_1=\frac{\sum X_i\cdot Y_i+\frac{(\sum X_i)\cdot(\sum Y_i)}{n}}{\sum X_i^2-(\sum X_i)_2}$$

or

$$a_1=\frac{\sum (X_i-\bar{X})\cdot(Y_i-\bar{Y})}{\sum (X_i-\bar{X})^2}$$

and

(12) $$a_0=\bar{Y}-a_1\cdot\bar{X}$$

Mean values are defined as usual:

(13)

$$\bar{X}=\frac{(X_1+X_2+...+X_n)}{n}=\frac{1}{n}\sum X_i$$

$$\bar{Y}=\frac{(Y_1+X_2+...+Y_n)}{n}=\frac{1}{n}\sum Y_i$$

To make later calcuations easier the following variables are defined (they can be written in many ways, see the appendix to this chapter):

(14) $$S_{XY}=\sum X_iY_i-n\,\bar{X}\,\bar{Y}$$

(15) $$S_{XX}=\sum X_i^2-n\,\bar{X}^2=\sum (X_i-\bar{X})^2$$

---

3 Note that for convenience,

$\sum_{i=1}^{n} something_i$ is written as $\sum something_i$ in the following text

(16) $$S_{YY}=\sum Y_i^2-n\bar{Y}^2=\sum (Y_i-\bar{Y})^2$$

Which means that the equation for a_1 can be written in a more compact form:

(17) $$a_1=\frac{S_{XY}}{S_{XX}}$$

If we insert equation (12) into equation (5) we get:

(18) $$\hat{Y}=\bar{Y}+a_1\cdot(X-\bar{X})$$

By inserting $X=\bar{X}$ into equation (18) it can be shown, that $(\bar{X},\bar{Y})$ is on the regression line.
The *Residuals* are defined as:

(19) $$R_i=Y_i-\hat{Y}_i$$

The following is valid for the sum of the residuals:

(20) $$\sum_{i=1}^{n} R_i=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)=0$$

In principle, this is always true, but due to round off errors, the sum of the residuals will not always be exactly 0.

Precision of the estimated regression line.

In real life, one can always fit a straight line to any data set even though it may not make much sense. It is therefore necessary to be able to estimate the precision of the fitted line, espcially in highly regulated contexts such as the pharmaceutical.

Equation (19) can be written as:

(21) $$R_i=Y_i-\hat{Y}_i=Y_i-\bar{Y}-(\hat{Y}_i-\bar{Y})$$

which rearranged gives:

(22) $$(Y_i-\bar{Y})=(\hat{Y}_i-\bar{Y})+(Y_i-\hat{Y}_i)$$

And squared:

(23) $$\sum (Y_i-\bar{Y})^2=\sum (\hat{Y}_i-\bar{Y})^2+\sum (Y_i-\hat{Y}_i)^2$$

$(Y_i-\bar{Y})$ is the deviation of the ith measured response from the mean value of all measured responses, so the left side of equation (23) is the sum of the squares of the deviations from the mean value and is abbreviated SS about the mean (corrected sum of squares of the Y's). As $\hat{Y}_i-\bar{Y}$ is the deviation of the ith estimated value from the mean value and $Y_i-\hat{Y}_i$ is the deviation of the ith measured value form the estimated or fitted value (the ith residual, R_i) Equation (23) can be expressed in words as:

Sum of squares about the mean	=	Sum of squares due to the regression	+	Sum of squares about the regression line

In other words, part of the variation of the Y's about their mean can be ascribed to the regression line and part - $\sum (Y_i - \hat{Y}_i)^2$ - can be ascribed to the fact that not all measured values lies exactly on the regression line. If they did, the sum of squares about the regression line would be 0.
A way to estimate how useful the regression line is to estimate the measured values would be to calculate how much of the SS about the mean originates from the SS due to the regression line and how much is due to the SS about the regression line. For a "good quality" regression line the SS due to the regression line must be much larger than the SS about the regression line, or, put in another way:

(24) $$R^2 = \frac{SS_{regression line}}{SS_{about the mean}}$$

Should be close to 1. The equation for R^2 then becomes:

(25) $$R^2 = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2}$$

R^2 can also be interpreted as the fraction of the total variance of the Y-values, that can be explained by the regression line.
If all X-values are different, R^2 can become 1, but if two or more X-values are identical (replicates), R^2 can never become 1.

The correlation coefficient.
The correlation coefficient can for at straight line be calculated as:

(25b) $$K = (sign of\ the slope) \cdot \sqrt{R^2}$$

Degrees of freedom
Any sum of squares has got a number of degrees of freedom that describes how many independent "pieces of information" that are present in the data used to calculate the sum. The sum of squares about the mean has n-1 degrees of freedom $(Y_1 - \bar{Y}, Y_2 - \bar{Y}, \ldots, Y_n - \bar{Y})$, as one is used to calculate the mean. Put a nother way, the sum of the numbers in the parenthesis is 0 (due to the definition of the mean value), so there is one equation that limits the data set.
The sum of squares due to the regression line is calculated by means of a single function of $Y_1, Y_2, \ldots, Y_n$ (a_1). This sum therefore has one degree of freedom.
The sum of squares about the regression line (hereafter called the residual sum of squares) has n-2 degrees of freedom (because 2 parameter has been used - 2 restrictions on the data set has been imposed - as two parameters are necessary to determine a straight line, slope and intercept).

Analysis of variance
It is possible to setup an analysis of variance table for the regression analysis, see table 1. The "Mean square" column is calculated by division of the relevant sum of squares with the corresponding number of degrees of freedom.
s^2 (see table 1), with n-2 degrees of reedom, is an estimate of s_{YX}, the variance about the regression line. This entity is a measure of the error with which a Y-value is predicted from a given X-value by means of the calculated regression line.

Source of variance	Degrees of freedom	Sum of squares (SS)	Mean squares (MS)
Due to the regression	1	$\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2$	MS_{reg}
About the regression (residual)	n-2	$SS=\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$	$s^2=\frac{SS}{n-2}$
Total, corrected for the mean, $\bar{Y}$	n-1	$\sum_{i=1}^{n}(Y_i-\bar{Y})^2$	

Table 1. Analysis of variance table for linear regression.

Standard deviation of the slope
The estimated standard deviation of the slope can be expressed as follows:

(26) $$est.std(a_1)=\frac{s}{\sqrt{\sum(X_i-\bar{X})^2}}=\frac{s}{\sqrt{S_{XX}}}$$

If one assumes that the variation of the observations about the regression line can be described by a normal distribution, it is possible to construct the following equation for the confidence interval of the slope:

(27) $$a_1\pm\frac{t\left(n-2,1-\frac{1}{2}\alpha\right)\cdot s}{\sqrt{\sum(X_i-\bar{X})^2}}$$

where $t\left(n-2,1-\frac{1}{2}\alpha\right)$ is $100\cdot\left(1-\frac{1}{2}\alpha\right)$ percent point of a t-distribution with (n-2) degrees of freedom (follows from s^2).

From equation (26) and (27) it follows that when planning an experiment with the object of getting the best estimate of a_1, the denominator in equation (26) resp. (27) should be as large as possible in order to assure the smallest standard deviation of the slope/the smallest confidence interval. **In other words, one should include concentration values corresponding to the smallest and largest concentration values to be determined.**

Standard deviation of the intercept
The estimated standard deviation of the intercept with the Y-axis can be calculated using the following equation:

(28) $$est.std(a_0)=s\cdot\sqrt{\frac{\sum X_i^2}{n\cdot\sum(X_i-\bar{X})^2}}$$

And for the confidence interval:

(29) $$a_0\pm t\left(n-2\,,1-\frac{1}{2}\alpha\right)\cdot s\cdot\sqrt{\frac{\sum X_i^2}{n\cdot\sum(X_i-\bar{X})^2}}$$

Standard deviation of the estimated Y-value
The esimated standard deviation for the estimated Y-value, $\hat{Y}$ can be calculated using the following equation:

(30) $$est.std\left(\hat{Y}_k\right)=s\cdot\sqrt{\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum(X_i-\bar{X})^2}}$$

Where $\hat{Y}_k$ is the estimated Y-value, corresponding to the X-value X_0. The corresponding confidence interval becomes:

(31) $$\hat{Y}_k \pm t\left(n-2, 1-\frac{1}{2}\alpha\right)\cdot s\cdot\sqrt{\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum(X_i-\bar{X})^2}}$$

There equations are valid for the mean value of the estimated Y-value. For a single observation the following equation applies:

(32) $$\hat{Y}_k \pm t\left(n-2, 1-\frac{1}{2}\alpha\right)\cdot s\cdot\sqrt{1+\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum(X_i-\bar{X})^2}}$$

The standard deviation of an estimated Y-value becomes small:
- if $X_0 = \bar{X}$
- if n is large

In other words, the regression line is "best" at estimating Y, if one is in the middle of the X-interval/concentration range.

If one ore more replicates are measured for a sample equation (32) becomes:

(33) $$\hat{Y}_k \pm t\left(n-2, 1-\frac{1}{2}\alpha\right)\cdot s\cdot\sqrt{\frac{1}{q}+\frac{1}{n}+\frac{(X_0-\bar{X})^2}{\sum(X_i-\bar{X})^2}}$$

q is the number of replicates.

F-test for significance of the regression
As the Y_i-values are random variables, any function of them will also be a random variable. Two relevant parameters in this context are MS_{reg} (mean square due to the regression line) and s^2 (mean square due to residual variation), see table 1. It is possible to show that the fraction

(34) $$F = \frac{MS_{reg}}{s^2}$$

follows an F-distribution with respectively 1 and (n-2) degrees of freedom, on the condition that $a_1 = 0$. It is therefore possible to test if one can consider a_1 as being different from 0 based on the given data.

Lack of Fit
If the variance of the pure, random errors of the measured Y-values is not known, this can be estimated by repeating the measurement because the only factor that can then be influencing the result is the "pure" error. These repeated measurements, replicates, corresponding to the individual X-values must be real replicates. It is not enough to read a value several times, the analysis must be repeated the prescribed number of times.

Our notation is therefore expanded a little:

$Y_{11}, Y_{12},..., Y_{1n1}$ are n_1 replicated observations at X_1.
$Y_{21}, Y_{22},..., Y_{2n2}$ are n_2 replicated observations at X_2.
Y_{ju} is the u^{th} observation at X_j.
$Y_{k1}, Y_{k2},..., Y_{knm}$ arer n_m replicated observations at X_k.

All in all, there will be:

(35) $$n=\sum_{j=1}^{m}\sum_{u=1}^{n_j}1=\sum_{j=1}^{m}n_j$$

observations. It is now possible to calculate the "pure error" sum of squares for each X_i, e.g. for X_1:

(36) $$\sum_{u=1}^{n_1}\left(Y_{1u}-\bar{Y}_1\right)^2=\sum_{u=1}^{n_1}Y_{1u}^2-\left(\sum_{u=1}^{n_1}Y_{1u}\right)^2$$

If this is done for all X_i's the total "pure error SS" is obtained

(37) $$\sum_{j=1}^{m}\sum_{u=1}^{n_j}\left(Y_{ju}-\bar{Y}_j\right)$$

with the following number of degrees of freedom(each X_i costs one degreee of fredom):

(38) $$n_e=\sum_{j=1}^{m}(N_j-1)=\left(\sum_{j=1}^{m}n_j\right)-m$$

The "pure error" mean square then becomes:

(39) $$s_e^2=\frac{\sum_{j=1}^{m}\sum_{u=1}^{n_j}\left(Y_{ju}-\bar{Y}_j\right)^2}{\left(\sum_{j=1}^{m}n_j\right)-m}$$

This is an estimate of s^2 (the variance of the "pure error") whether or not the fitted model is valid (the model is not part of the calculation).

s_e^2 is in other words a measure of the pure error, that can not be fitted. If one wishes to examine how well the linear regression describes the data, s_e^2 is subtracted form the mean square of the residuals, MS_{res} and it is tested if they are significantly different. This diffference is named the mean square due to "lack of fit", MS_L.

(40) $$MS_L=MS_{RES}-s_e^2$$

MS_L has (n_r-n_e) degreees of freedom.
Test for significance:

(41) $$F=\frac{MS_L}{s_e^2}$$

with 100·(1-a)% point of an F-distribution with respectively (n_r-n_e) and n_e degrees of freedom.

If the F-test is significant the linear model appears to be insufficient. A way to examine the origin of this insufficiency can be to create a plot of residuals (see later).
If the F-test is not significant, there is no immediate reason to discard the linear model. Both s_e^2 and MS_L can be used as estimates of s^2. A pooled estimate of s^2 can be calculated using MS_{RES} (s^2).

When validating a method of analysis it is customary (even required) to perform a certain number of replicate measurements for each concentration.

Standard deviation of estimated X-values
As mentioned before, as an analytical chemist one is usually more inerested in the "reversed" relation called inverse regression), where one initially determines a calibration curve by means of a suitably chosen series of standards. Subsequently, one measures a series of Y-values corresponding to the response from a series

of solutions to be tested. Based on these one wishes to determine the X-values, concentrations, with suitable confidence intervals.
If n solutions are measured q times each, the following equation applies. S_{XX} is defined earlier (equation 15)

(42a)
$$X_U = \hat{X}_0 + \frac{\left(\hat{X}_0 - \bar{X}\right)\cdot g + \left|\left(\frac{t\cdot s}{a_1}\right)\cdot\sqrt{\frac{\left(\hat{X}_0-\bar{X}\right)^2}{S_{XX}} + \frac{(n+q)\cdot(1-g)}{n\cdot q}}\right|}{1-g}$$

(42b)
$$X_L = \hat{X}_0 + \frac{\left(\hat{X}_0 - \bar{X}\right)\cdot g - \left|\left(\frac{t\cdot s}{a_1}\right)\cdot\sqrt{\frac{\left(\hat{X}_0-\bar{X}\right)^2}{S_{XX}} + \frac{(n+q)\cdot(1-g)}{n\cdot q}}\right|}{1-g}$$

(43)
$$g = \frac{t\left(\nu, 1-\frac{1}{2}\alpha\right)^2\cdot s^2\cdot S_{XX}}{a_1^2}$$

X_U is the upper limit of the confidence interval, X_L the lower. n is the number of degrees of freedom for s^2 (n-2). If the equation is expected to deliver meaningfull results, the calibration curve should be welldefined, which means that g should be smaller than about 0.20 (t should be approx. 2.236).
From equation (43) it can be seen that if a_1 (the slope) is large, g becomes small, which again means that the confidence interval is narrower for the estimated value of X. In other words, if S_{XX} is large compared to S_{XY} - corresponding to a small value of the slope, a_1, - the confidence interval will be broad. Intuitively this makes a lot of sense as a small slope – low sensitivity – tells us something about the ability of the analytical method to distinguish X-values, and poor ability to distinguish X-values is the same as the uncertainty of the X-values will be large. However, one should not be tempted to make S_{XX} smaller as this defines the analytical range. Also, a large S_{XX} is only "bad" if the sensityvity is low. From equation (43) one can also see, that if the Y-values – the responses – are not well-determined g becomes larger, affording a broader confidence interval. This is not all that surprising as a badly determined calibration curve is expected to give a less precise detemination of the estimated concentration.
From equation (42a) and (42b) one can see, that the longer one is from the mean of the X-values the broader the confidence interval. This is in complete analogy with the relations for the Y-values and it is therefore worth to take note of the following generalised rules:

> When one is planning to determine a calibration curve 3 parts of the are to be determined with high precision i.e. many replicates:
> The two ends and the mean value of the analytical range:
>
> $$\left(\frac{C_{min}+C_{max}}{2}\right)$$

This gives the best utilization of the replicates in terms of a narrower confidence interval for the extimated X-value. You may wish to compare with chapter on validation.

Residual plots
The residuals are defined in equation (19). The residual contains information about the deviations from the fitted model, in this case linear. These deviation can be caused by drift of the apparatus, erroneous preparation of standard solutions etc.
We operate with two slightly different versions of residual plots:

Type a.

The residuals are plotted against the concentration values. A plot of this type will show i the linear model is not well suited to the data or if e.g. the standard deviation of the analytical response is concentration dependent. If the analytical response in reality is "curved" (single curve), the residuals will predominatly deviate to "one side" at high and low concentrations while the results for the intermediate concentrations will deviate in the opposite "direction".
However, this type of residual plot is not well suited to disclose a constant drift of the apparatus. This is especially true if the sequence of measurement is not randomized, and a drift can manifest itself as a change of the slope.

Type b
Another type of residual plot can be made by plotting the residuals against the measurement number in the measurement sequence and is best formatted as a histogram, where the first column shows the residual for the first measurement, the next for the second etc. Combined with randomization of the measurement sequence, this type of residual plot is better suited to disclose drift, but not whether the calibration curve is linear or not.

One should therefore make both types of residual plots.
It can also be recommended to create another version of the residual plots, where each residual is divided by the nominal concentrration.

Appendix 1. Measuring the standard curve/the calibration curve.
Please compare with the chapter on validation.
When a calibration curve is to be determined, a series of standard solutions are prepared in accordance with the following strategy:

Lowest concentration: 2·LOQ
Highest concentration: (maximal concentration)·1.25

Standard solution	1	2	3	4	5
Concentration	2·LOQ	(C1+C3)/2	(C1+C5)/2	(C3+C5)/2	C_{max}·1.25
Replicates	6	2	6	2	6

Appendix 2. "Things" that should be calculated for a calibration curve obtained using linear regression.
- Calculate slope, intercept e.g. using equations (11) og (12).
- Calculate residuals after equation (19) and create type a and type b residual plots.
- A plot (drawing, graph) of the regression line with all data points included.
- R^2 after equation (25).
- The correlation coefficient calculated as the square root of R^2, with sign, using equation (25b).
- Optionally an analysis of variance table like table 1.
- The estimated standard deviation of the slope, equation (26), and the confidence interval, equation (27).
- The estimated standard deviation of the intercept, equation (28), and the confidence interval, equation (29).
- the standard deviation of the estimated Y-values using equation (30) and the confidence interval, equation (31), (32) or (33), optionally displayed as confidence bands on the plot of the regression line.
- F-test for significance of the regression, equation (34).
- Test for significance of the regression/lack of fit, equation (41).
- Estimation of the quality of the regressions equation using equation (43).
- Check whether the calibration curve goes through 0,0.
- LOD and LOQ determined from the calibration curve, the parameter s (table 1).

If the calibration curve is used for estimating X-values (concentrationc) from the Y-values (measured response), the confidence intervals for the results are calculated using equations (42a) and (42b). The results are displayed as both:

x ± (standarddeviation) and as x ± (confidence interval)