Analysis of calibration curves

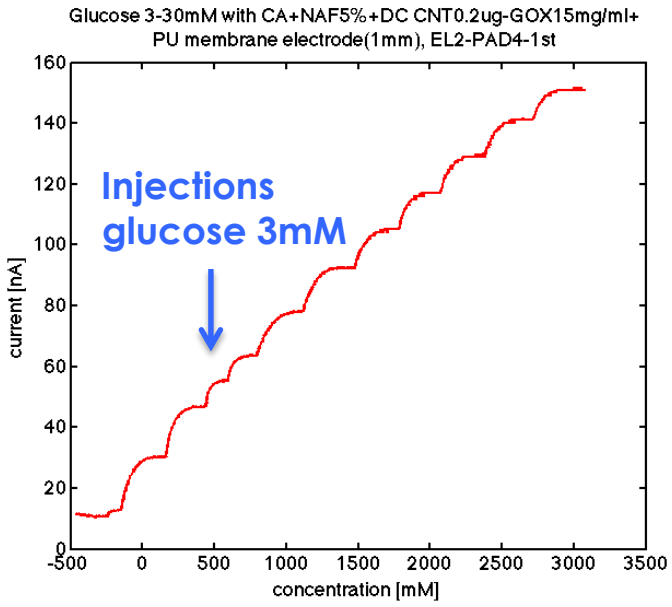
Camilla Baj-Rossi, 14.10.2014

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

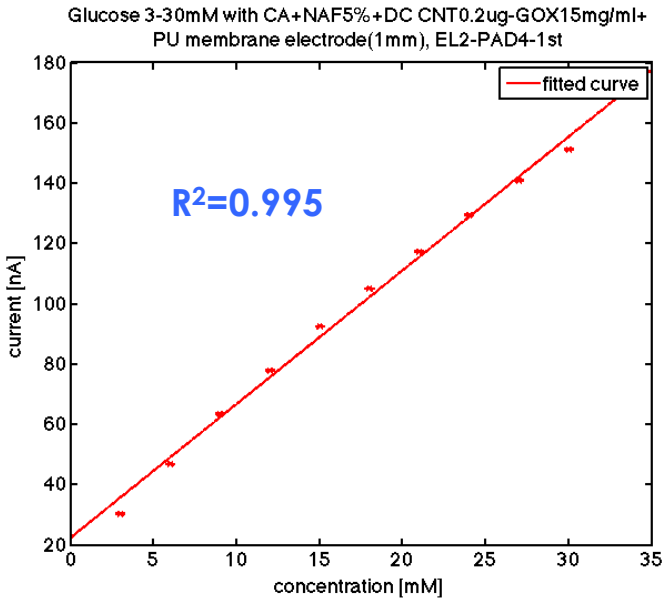
LSI
Integrated Systems Laboratory

Chronoamperometry

Raw data from chronoamperometry

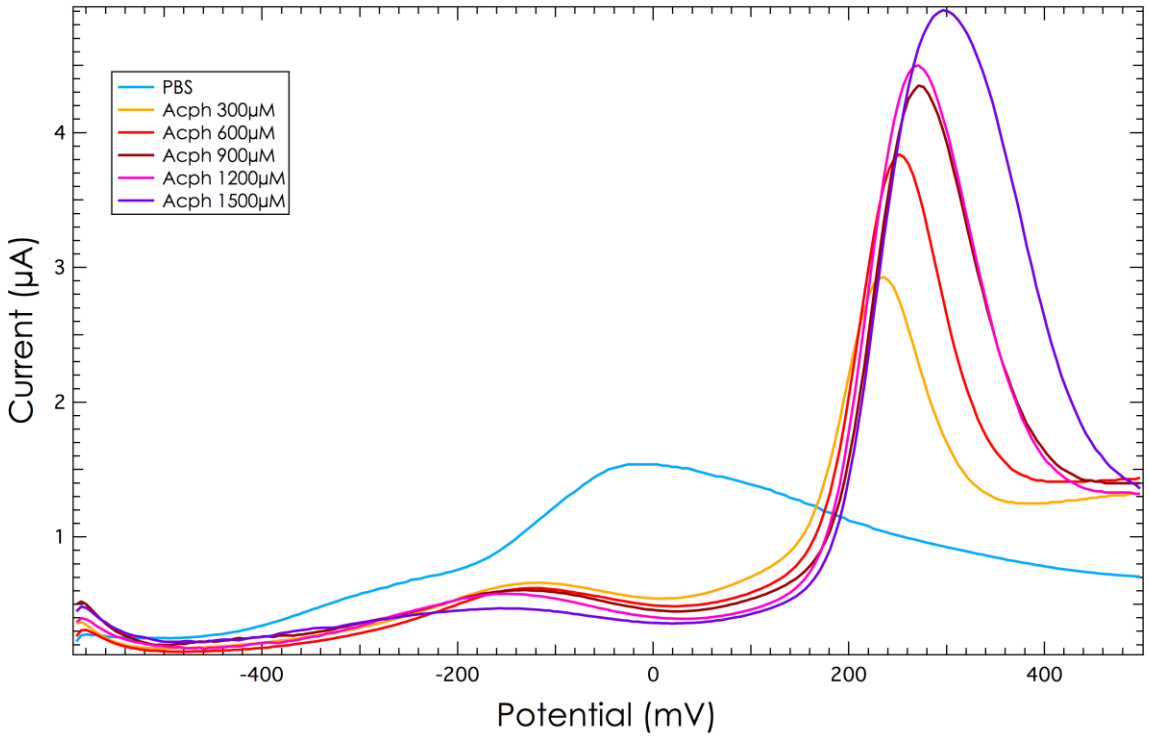


Calibration curve



Square wave voltammetry

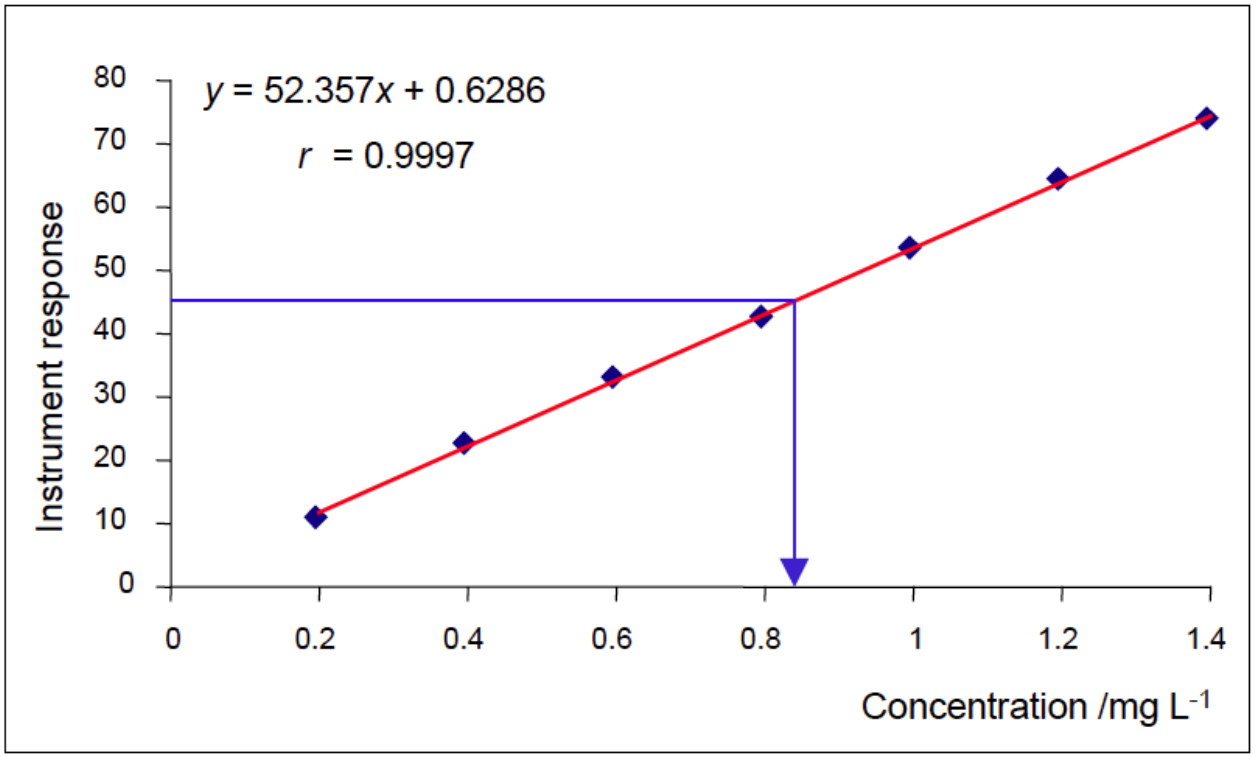
In presence of increasing concentration of the drug Acetaminophen (Acph)



Calibration curves

- Is my calibration good enough?
- How can I use my calibration to estimate the concentration of an unknown sample?

Sensor calibration



Typical calibration curve

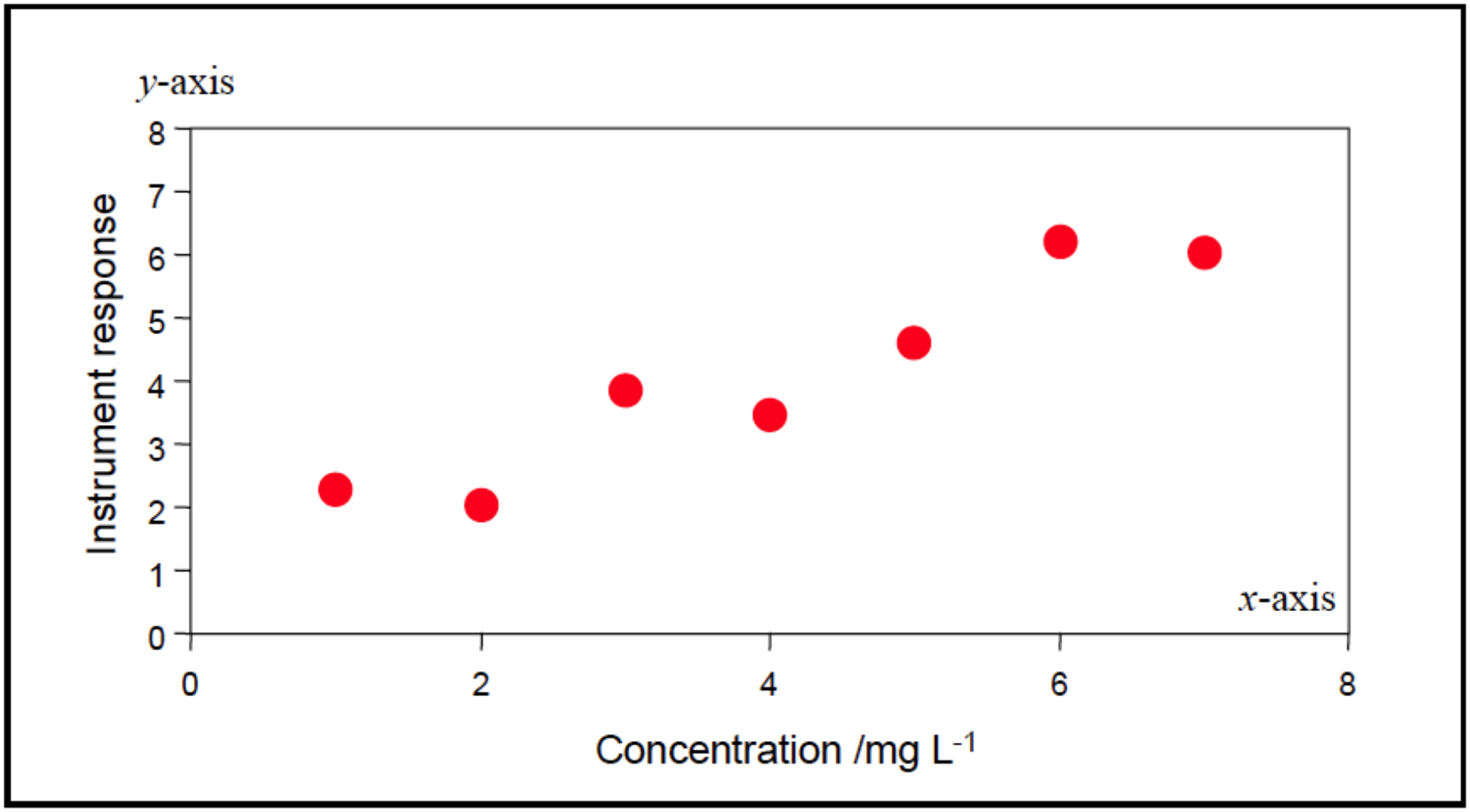
The Calibration Process

- Plan the experiments;
- Make measurements;
- Plot the results;
- Carry out statistical (regression) analysis on the data to obtain the calibration function;
- Evaluate the results of the regression analysis;
- Use the calibration function to estimate values for test samples;
- Estimate the uncertainty associated with the values obtained for test samples.

Planning the experiments

- The number of calibration standards;
- The concentration of each of the calibration standards;
- The number of replicates at each concentration;
- Preparation of the calibration standards;

Plotting the results

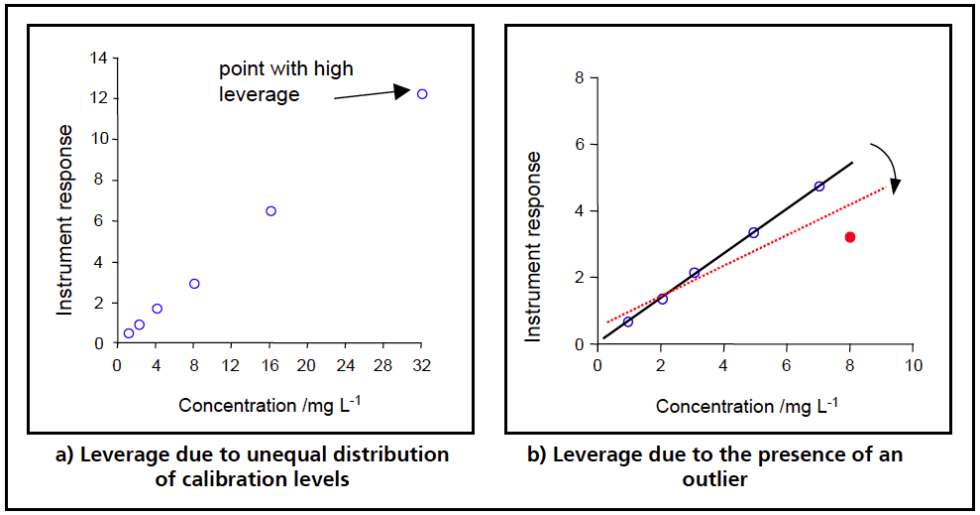


Scatter plot of instrument response data versus concentration

Evaluating the scatter plot

Points of influence:

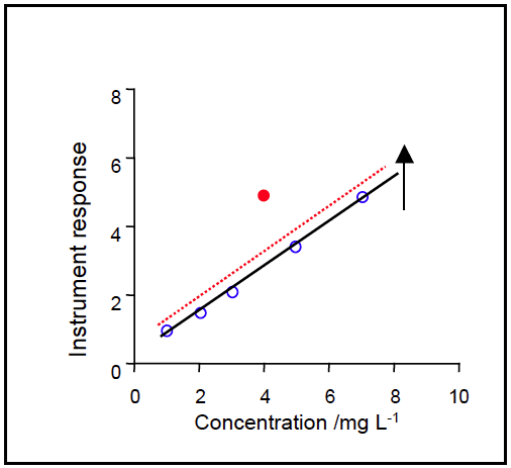
1) Leverage



a) Leverage due to unequal distribution of calibration levels

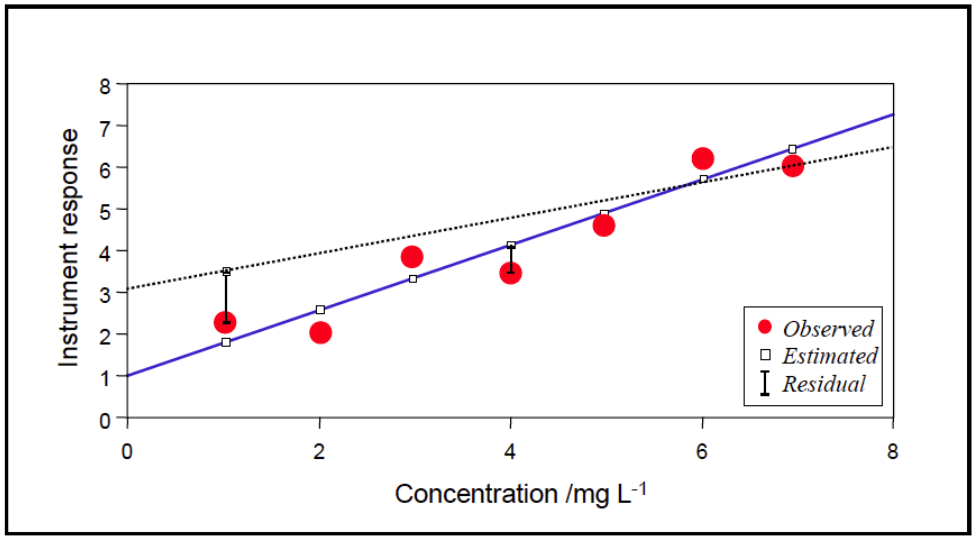
b) Leverage due to the presence of an outlier

2) Bias



Regression analysis

- The aim of linear regression is to establish the equation that best describes the linear relationship between instrument response (y) and analyte level (x). The relationship is described by the equation of the line, *i.e.*, **y = mx + c**, where **m** is the gradient of the line and **c** is its intercept with the y-axis.
- "least squares regression"→ the line that gives the smallest sum of the squared residuals best represents the linear relationship between the x and y variables



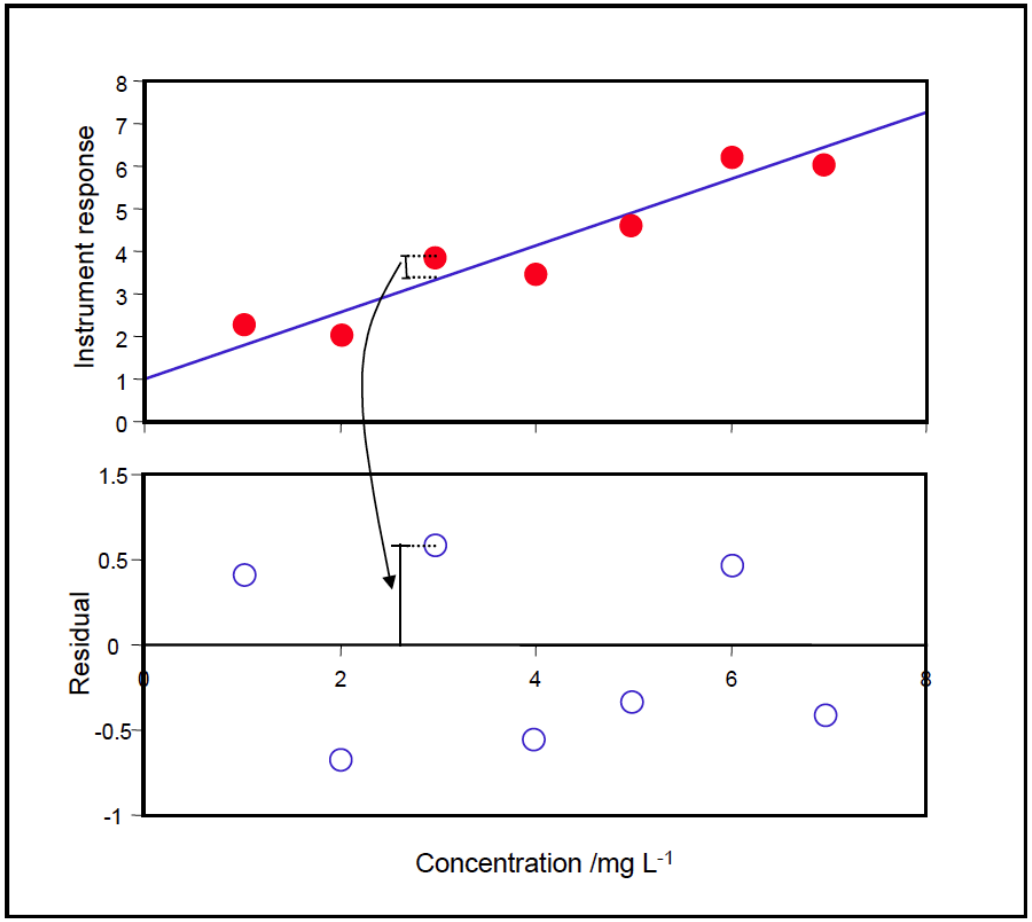
Least squares linear regression – calculating the best straight line

Assumptions

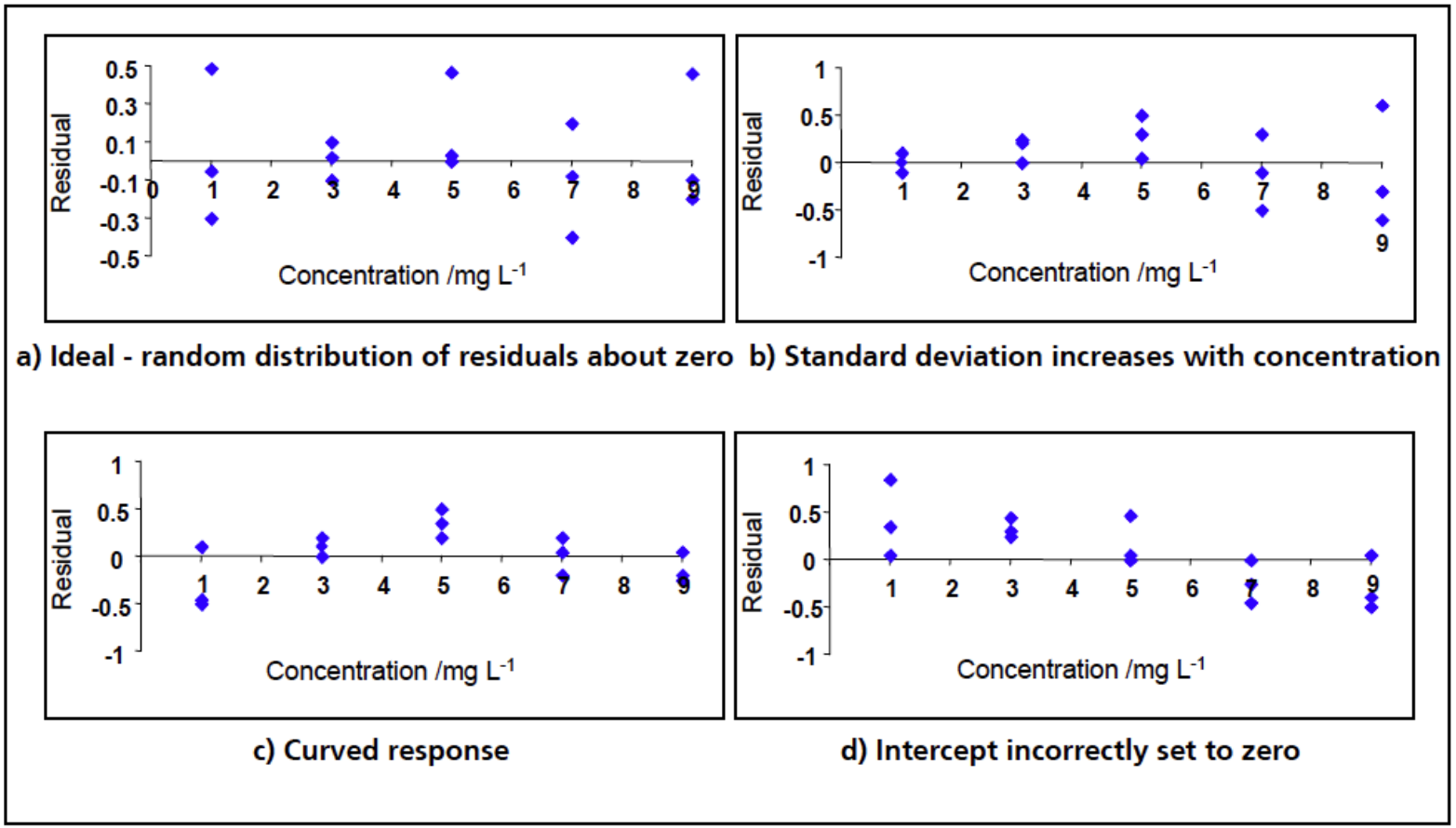
- The error in the x values should be insignificant compared with that of the y values.
- The error associated with the y values must be normally distributed.
- The magnitude of the error in the y values should also be constant across the range of interest, *i.e.* the standard deviation should be constant.
- The general solution to this problem is to use weighted regression, which takes account of the variability in the y values.

Evaluate the Regression analysis

- Plot of the residuals



Evaluate the Regression analysis

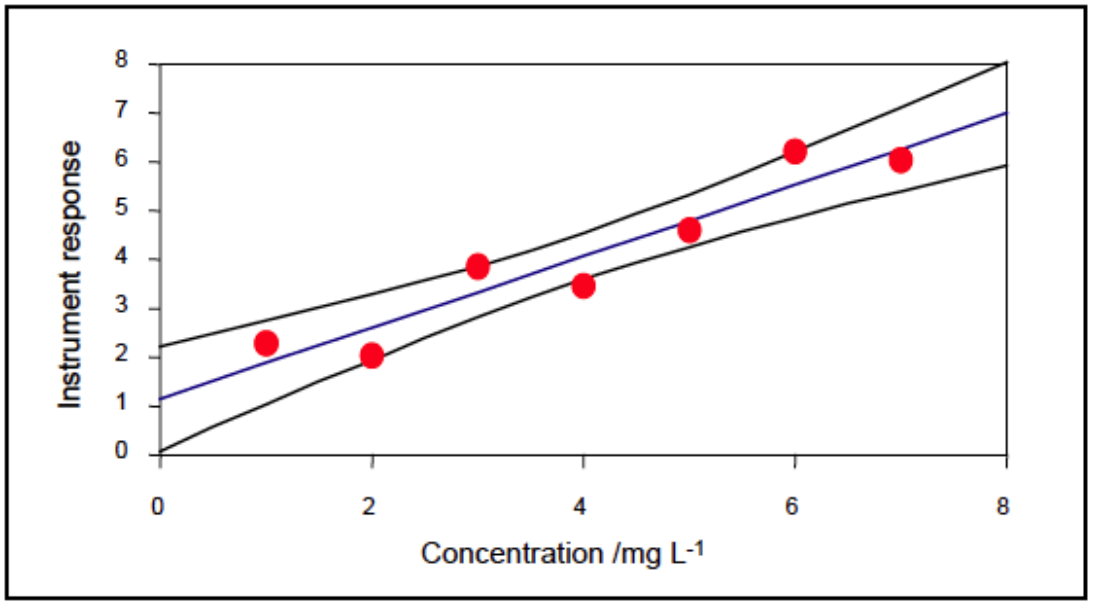


Examples of residuals plots

Evaluate the Regression analysis

- Confidence interval

It gives an indication of the range within which the 'true' line might lie.



95% confidence interval for the line

Evaluate the Regression analysis

Regression Statistics	
Multiple R	0.999955883
R Square	0.999911768
Adjusted R Square	0.999889709
Standard Error	0.005164622
Observations	6

1

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	1.2091	1.2091	45330.79	2.93×10^{-9}
Residual	4	0.00010669	2.67×10^{-5}		
Total	5	1.2092			

2

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0.0021129	0.0037548	0.56270	0.60368	-0.008312	0.012538
X Variable 1	0.10441	0.00049038	212.91	2.92×10^{-9}	0.10304	0.10577

3

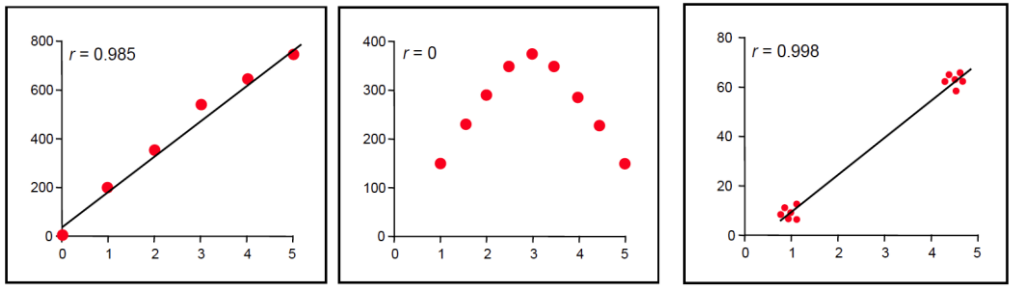
Typical output from a regression analysis using Excel

1. Regression statistics

- **The correlation coefficient, r**

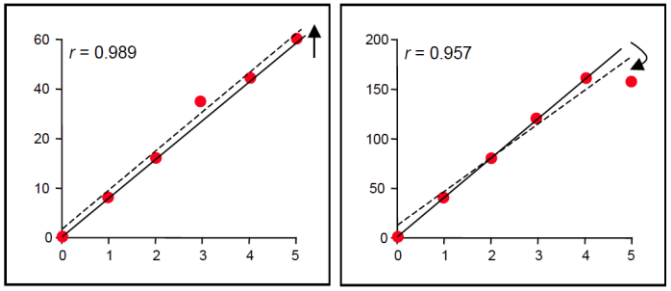
The correlation coefficient, r (and the related parameters r^2 and adjusted r^2) is a measure of the strength of the degree of correlation between the y and x values. In Excel output it is described as 'Multiple R'. r can take any value between +1 and –1; the closer it is to 1, the stronger the correlation.

Regression Statistics	
Multiple R	0.999955883
R Square	0.999911768
Adjusted R Square	0.999889709
Standard Error	0.005164622
Observations	6



a) Non-linearity

b) Relationship other than linear



c) An outlier causing bias

d) An outlier causing leverage

e) Poor experimental design

1. Regression statistics

- **r^2 and adjusted r^2.**

r^2 is often used to describe the fraction of the total variance in the data which is contributed by the line that has been fitted. Ideally, if there is a good linear relation, the majority of variability can be accounted for by the fitted line. r^2 should therefore be close to 1.

The adjusted r^2 value is interpreted in the same way as r^2 but is always lower. It is useful for assessing the effect of adding additional terms to the equation of the fitted line (*e.g.*, if a quadratic fit is used instead of a linear fit).

Regression Statistics	
Multiple R	0.999955883
R Square	0.999911768
Adjusted R Square	0.999889709
Standard Error	0.005164622
Observations	6

1. Regression statistics

- **Residual standard deviation (standard error)**

The residual standard deviation (also known as the residual standard error) is a statistical measure of the deviation of the data from the fitted regression line. It is calculated using the equation:

$$s(r) = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n-2}}$$

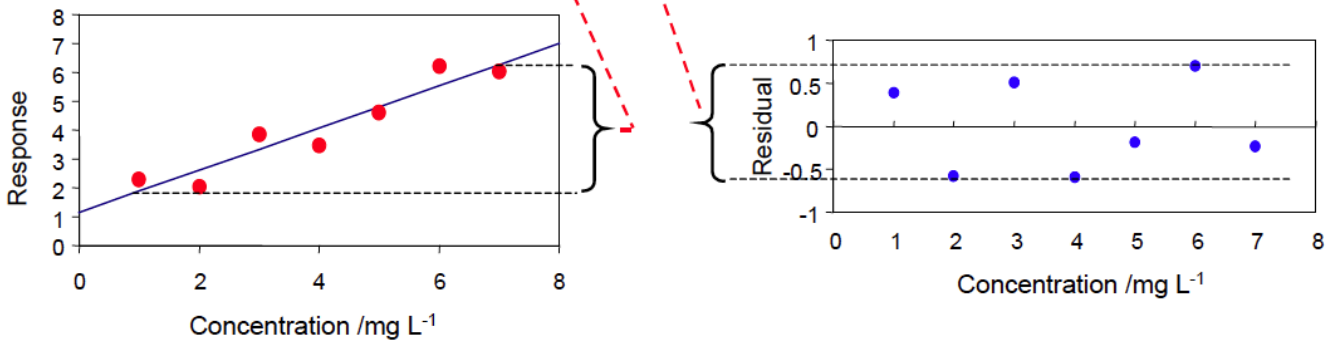
where y_i is the observed value of y for a given value of x_i, $\hat{y}$ is the value of y predicted by the equation of the calibration line for a given value of x_i, and n is the number of calibration points.

Note that there are (n-2) degrees of freedom in calculating s(r) One way of understanding the degrees of freedom is to note that we are estimating two parameters from the regression – the slope and the intercept. Therefore, v = n − 2 and we need at least three points to perform the regression analysis.

2. ANOVA table

2. Anova table: analysis of variance

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	14.90	14.90	**43.98**	**0.001**
Residual	5	1.69	0.34		
Total	6	16.60			

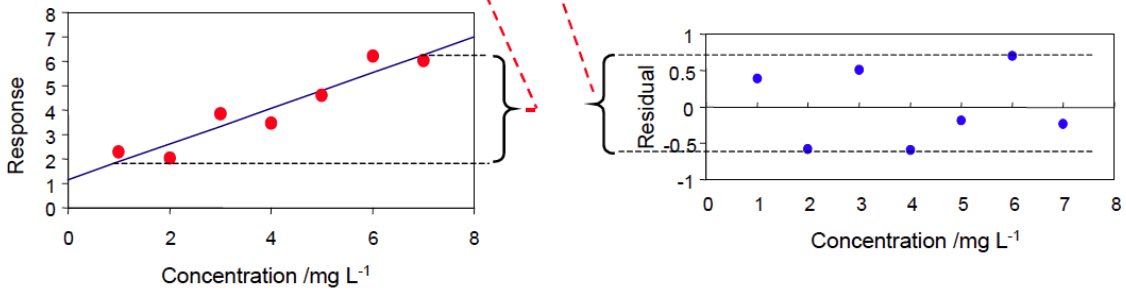


Origin of sum of squares terms in regression analysis

2. ANOVA table

- **The sum of squares terms (SS)** represent different sources of variability in the calibration data. The *regression term* represents the variability in the data that can be accounted for by the fitted regression line. Ideally this should be **large**; if there is a good linear relationship, the fitted line will describe the majority of the variability in response with concentration. The *residual term* is the sum of the squared residuals. This value should be small compared to the regression sum of squares terms because if the regression line fits the data well, the residuals will be **small**.

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	14.90	14.90	**43.98**	**0.001**
Residual	5	1.69	0.34		
Total	6	16.60			

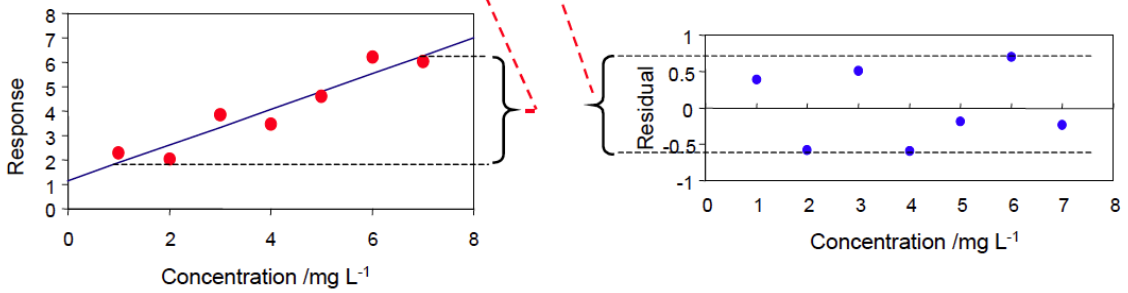


2. ANOVA table

- **The sum of squares terms (SS)** represent different sources of variability in the calibration data. The *regression term* represents the variability in the data that can be accounted for by the fitted regression line. Ideally this should be **large**; if there is a good linear relationship, the fitted line will describe the majority of the variability in response with concentration. The *residual term* is the sum of the squared residuals. This value should be small compared to the regression sum of squares terms because if the regression line fits the data well, the residuals will be **small**.

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	14.90	14.90	**43.98**	**0.001**
Residual	5	1.69	0.34		
Total	6	16.60			

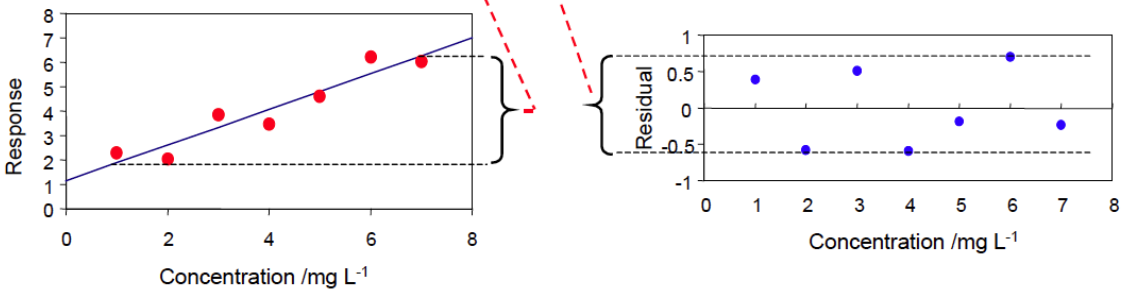
The **mean square (MS)** term is the sum of squares term divided by its degrees of freedom.



2. ANOVA table

- The **F value is the ratio of the regression MS term to the residual MS term**. Ideally this ratio should be very **large**; if there is a good linear relationship the regression MS term will be much greater than the residual MS term.

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	14.90	14.90	**43.98**	**0.001**
Residual	5	1.69	0.34		
Total	6	16.60			



- **The significance F value** represents the probability that there is no correlation between y and x values, *i.e.*, obtaining the results by chance. For a calibration curve to be of any use the significance F value should be **extremely small**. This value is also known as the **p-value**.

3. Regression coefficients

The table gives information about the slope and intercept

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0.0021129	0.0037548	0.56270	0.60368	-0.008312	0.012538
X Variable 1	0.10441	0.00049038	212.91	2.92×10^{-9}	0.10304	0.10577

- The **standard errors** (also know as the standard deviations) for each coefficient. These values give an indication of the ranges within which the values for the gradient and intercept could lie:

$$s_m = \frac{s(r)}{\left\{\sum_{i=1}^{n}(x_i - \bar{x})^2\right\}^{\frac{1}{2}}}$$

$$s_c = s(r)\left\{\frac{\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i - \bar{x})^2}\right\}^{\frac{1}{2}}$$

Where s(r) is the residual standard deviation (residual error in slide 15).

3. Regression coefficients

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0.0021129	0.0037548	0.56270	0.60368	-0.008312	0.012538
X Variable 1	0.10441	0.00049038	212.91	2.92×10^{-9}	0.10304	0.10577

- **The t-stat and p-value** relate to the significance of the coefficients, *i.e.* whether or not they are statistically significantly different from zero.

For the slope: In a calibration experiment we would expect the gradient of the line to be very significantly different from zero. The *t-value* should therefore be a **large number** (for a calibration with 7 data points the t-value should be much greater than 2.6, the 2-tailed Student t value for 5 degrees of freedom at the 95% confidence level) and the *p-value* should be **small** (much less than 0.05 if the regression analysis has been carried out at the 95% confidence level).

For the intercept: Ideally, we would like the calibration line *to pass through the origin*. If this is the case then the intercept should not be significantly different from zero. In the regression output we would expect to see a **small value** *for t* (less than 2.6 for a calibration with 7 data points) and a *p-value* **greater than 0.05** (for regression at the 95% confidence level). Whether the calibration line can reasonably be assumed to pass through zero can also be judged by inspecting the confidence interval for the intercept. If this spans zero, then the intercept is not statistically different from zero, as in the example.

3. Regression coefficients

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0.0021129	0.0037548	0.56270	0.60368	-0.008312	0.012538
X Variable 1	0.10441	0.00049038	212.91	2.92×10^{-9}	0.10304	0.10577

▲ The **lower and upper confidence limits** for the gradient and intercept.
These represent the extremes of the values that the gradient and intercept could take, at the chosen level of confidence (usually 95%). The confidence limits can then be calculated from the *t-statistic* for n–2 degrees of freedom, as:

$$c_m = ts_m \qquad c_c = ts_c$$

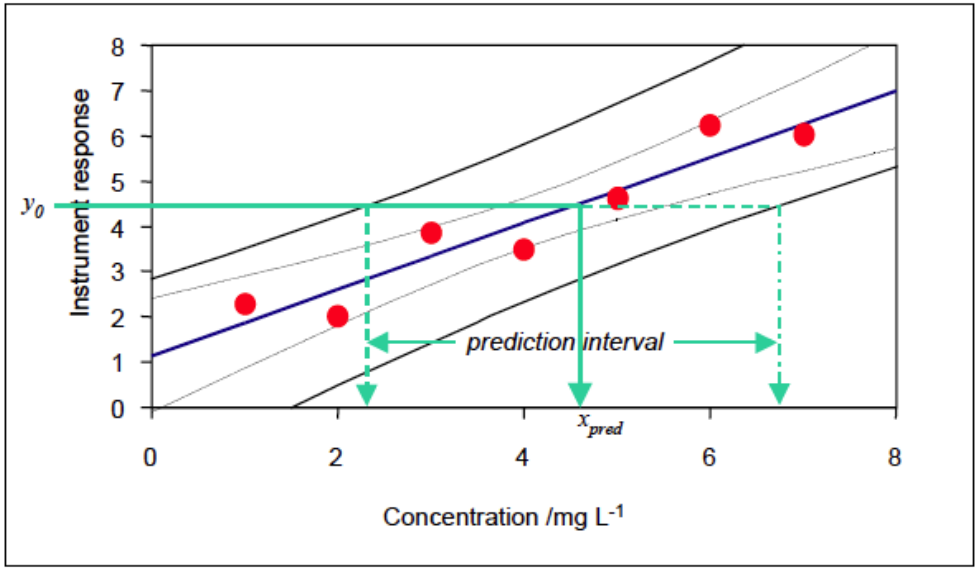
Where *t* is the 2-tailed Student's t value for n-2 degrees of freedom, and s_m is the standard error for the slope (s_c for the intercept).

How to represent the values for the slope and intercept:

$$slope = m \pm t \cdot s_m$$

$$intercept = c \pm t \cdot s_c$$

Using the calibration function to estimate values for test samples



Prediction interval

In addition, it is possible to calculate a confidence interval for values predicted using the calibration function. This is sometimes referred to as the 'standard error of prediction' and is illustrated in the Figure. The prediction interval gives an estimate of the uncertainty associated with predicted values of x.

Using the calibration function to estimate values for test samples

We can calculate the prediction interval for predicted values, as:

$$s_{x_0} = \frac{s(r)}{m}\sqrt{\frac{1}{N}+\frac{1}{n}+\frac{(\bar{y}_o-\bar{y})^2}{m^2\sum_{i=1}^{n}(x_i-\bar{x})^2}}$$

Is the difference between the mean of N repeat measurements for the sample and the mean of the y values for the calibration standards

where

s(r) is the residual standard deviation
n is the number of paired calibration points (x_i,y_i)
m is the calculated best-fit slope of the calibration curve
N is the number of repeat measurements made on the sample (this can vary from sample to sample and can equal 1)

A **confidence interval** is obtained by multiplying x_0 s by the 2-tailed Student t value for the appropriate level of confidence and n-2 degrees of freedom.

$$c_{x_0} = ts_{x_0}$$

Limit of detection (LOD)

There is always some error associated with any instrumental measurement. This also applies to **the baseline (or background or blank) measurement**, *i.e.* the signal obtained when no analyte is present. One very important determination that must therefore be made is how large a signal needs to be before it can be distinguished from the background noise associated with the instrumental measurement.

Formally, then, the **Limit-of-Detection (LOD)** is defined as the concentration of analyte required to give a signal equal to the background (blank) plus three times the standard deviation of the blank. That is, we first calculate the instrument response obtained with no analyte:

$$y_L = y_{BLANK} + 3 \cdot STDEV_{BLANK}$$

We convert it to the concentration LOD from linear regression analysis of the calibration data!

Limit of detection (LOD)

- Obtaining the LOD from the Regression Line:

In defining the LOD, IUPAC states that

$$y_L = \bar{y}_B + ks_B$$

Where y_L is the smallest detectable signal, $\bar{y}_B$ is the mean value of the blank responses, *k* is a numerical factor chosen in accordance with the confidence level desired, and S_B is the standard deviation of the blank signal. The LOD is a function of y_L and therefore

$$LOD = \frac{(y_L - \bar{y}_B)}{m}$$

where *m* is the analytical sensitivity (the slope of the regression line). Because the mean blank reading, $\bar{y}_B$, is not always 0, the signal must be background corrected. By combining the equations we obtain:

$$LOD = k\frac{s_B}{m}$$

Where S_B is the standard deviation of the blank signal and k = 3 allows a confidence level of 99.86% that $y_L \geq (y_B + 3S_B)$

Limit of detection (LOD)

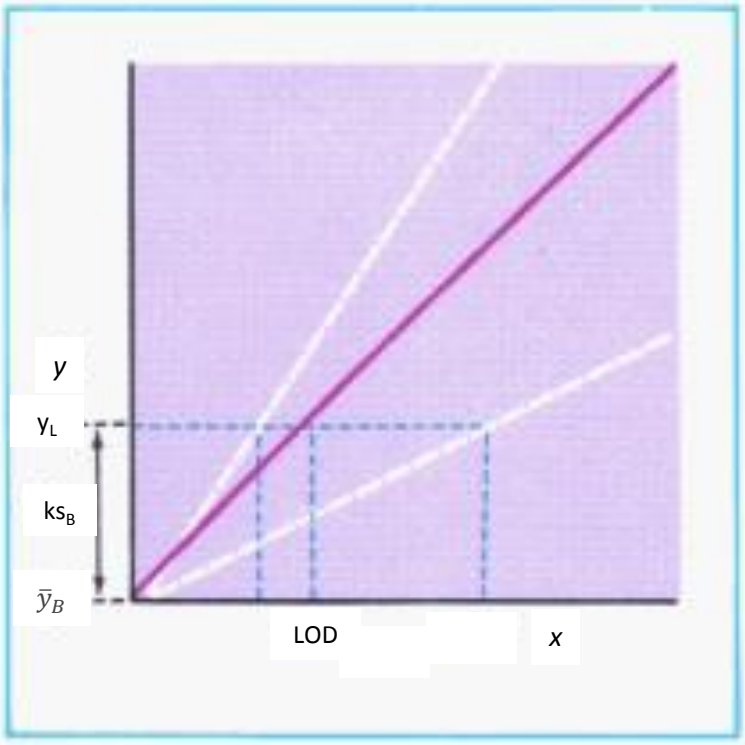


Figure 3. Analytical calibration curve of signal, x, vs. concentration, c, showing the relationship of ks_B to the limit of detection, c_L