STEPS TO TREAT DATA FOR CALIBRATION CURVE BY LINEAR REGRESSION

170524 /jp

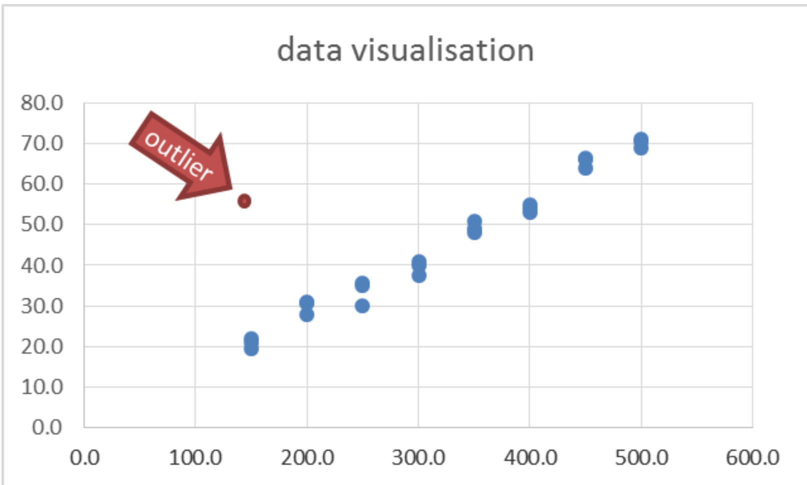
1/ DATA VISUALISATION

Input is a table of X-Y data:

c(mg/L)	signal (mV)		
150.0	22.0	21.0	19.5
200.0	30.5	31.0	28.0
250.0	35.0	35.5	30.0
300.0	40.0	41.0	37.5
350.0	51.0	49.0	48.0
400.0	55.0	53.0	54.0
450.0	64.0	66.5	66.0
500.0	70.5	71.0	69.0

The independent variable (X) are: concentrations of standard (mg/l, ug/ml or mol/l), volumes of standard, milligrams or millimols of standard etc. The number of the levels is typically 5-7. Since experiments should be replicated (at least 3x), first question is: how many points should be shown in a graph? The answer is: **all data measured** should be in the graph (N=24). **Do not average** the values at the same x-levels (rows)! This approach would hide outliers and also decrease degree of freedom! If replicates are present, rewrite the data into two columns (to make a graph in Excel easy) and plot a scattered (x-y) graph. The visualisation helps to reveal obvious outliers (gross errors) that should be removed.

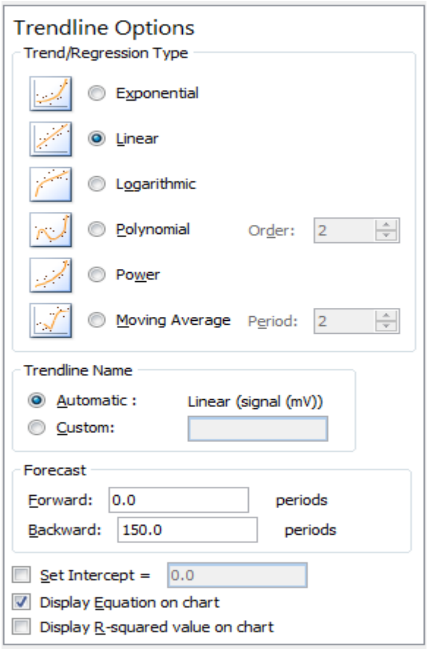
150.0	22.0
200.0	30.5
250.0	35.0
300.0	40.0
350.0	51.0
400.0	55.0
450.0	64.0
500.0	70.5
150.0	21.0
200.0	31.0
250.0	35.5
300.0	41.0
350.0	49.0
400.0	53.0
450.0	66.5
500.0	71.0
150.0	19.5
200.0	28.0
250.0	30.0
300.0	37.5
350.0	48.0
400.0	54.0
450.0	66.0
500.0	69.0

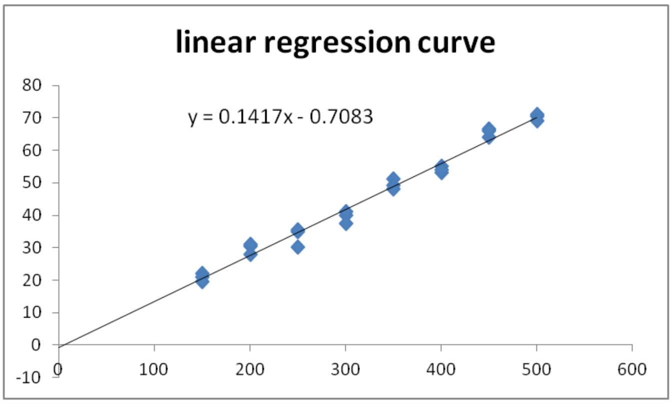


Note: in this step it become also clear that there is a strong correlation between Y and X, which is not surprising: since the beginning we intended to construct a calibration curve! Therefore, it is not necessary to prove a strong correlation according to Spearman.

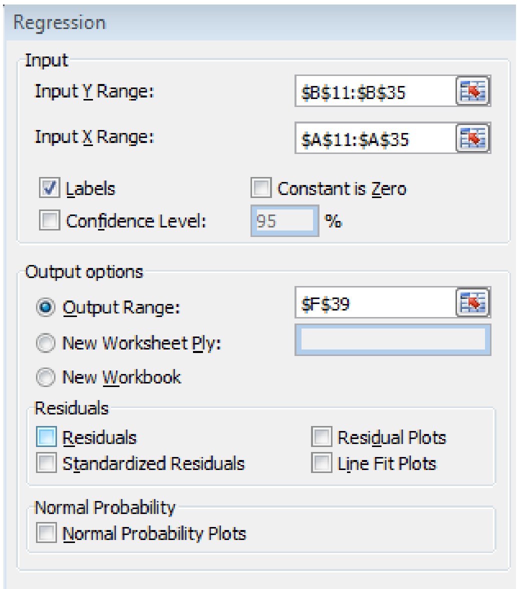
2/ THE FIRST MODEL OF LINEAR REGRESSION Y = b * X + a

Our ultimate goal is to find an equation of a line that fits best the points, which will be called “calibration line”. For the first estimation we may use either right button click on the graph points (**Add trendline + linear + show the equation in the graph** – check the red arrows below)





or using a dialog Data - Data analysis – Regression:



SUMMARY OUTPUT				
Regression Statistics			=correl()	
Multiple R	0.991338		0.99133811	
R Square	0.982751			
Adjusted R Square	0.981967			
Standard Error	2.245787			
Observations	24			
ANOVA				
	df	*SS*	*MS*	*F*
Regression	1	6321.875	6321.875	1253.455
Residual	22	110.9583	5.04356061	
Total	23	6432.833		
	Coefficients	*andard Erro*	*t Stat*	*P-value*
Intercept	-0.70833	1.378892	-0.5136976	0.612584
c(mg/L)	0.141667	0.004001	35.4041629	6.82E-21

$t = \frac{|a|}{s_a} = \frac{0.7083}{1.3789} = 0.5137$

t(crit.)=T.INV.2T(0.05;N-2)=	2.07387
t(crit.)=TINV(0.05;N-2)=	2.07387

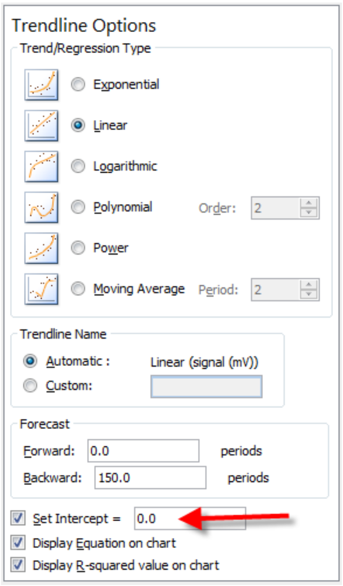
The latter way is recommended because it helps a lot **in the following**, because **THIS IS NOT THE END OF THE STORY!!!**

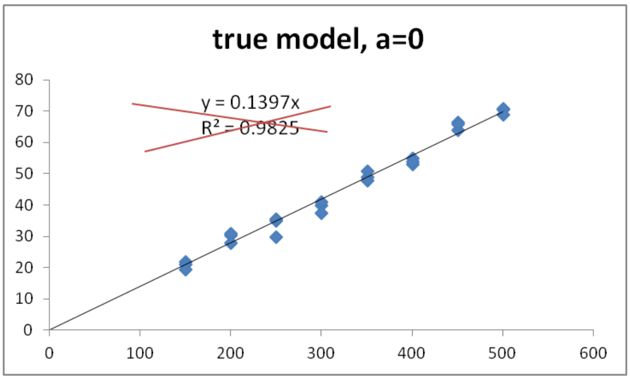
3/ IS THE INTERCEPT SIGNIFICANT?

Since our measurements are subjected to random errors, it may happen that another series of data obtained under the same conditions would give an equation of e.g. Y=0.1417*x **+ 0.001**, and another series for example Y=0.1417*x **- 0.111**. This suggests that the value (and sign) of intercept has no meaning. If this is true, very important consequence follows: IT SHOULD BE SET TO ZERO. The table generated by **Data analysis** gives information, if the intercept -0.7083 of our regression model Y = 0.1417 * X **- 0.7083** is significantly different from 0.

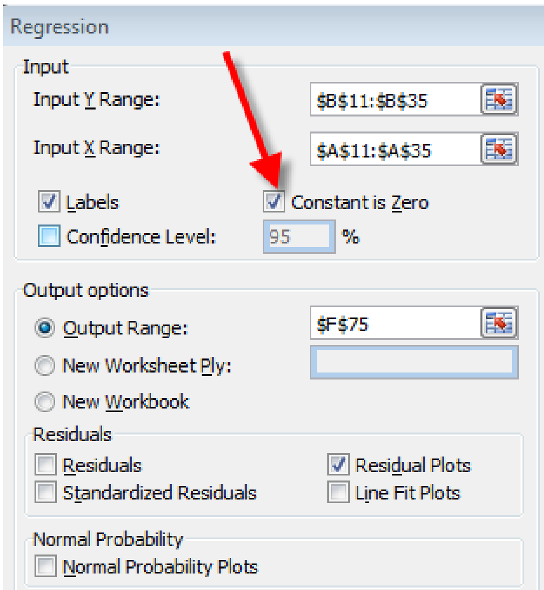
To confirm the null hypothesis a=0 (on the level α=0.05), we should compare ratio **|a|/s_a** to **t(crit.)** (see the yellow parts above **0.532<2.0739**) or, more simply, compare *P*-value to 0.05 (**0.613 > 0.05**). In our case, both the values **confirm** the null hypothesis. An easy rule to remember is: we set the intercept to zero if the *P*-value is **higher** than 0.05.

In this case we have to FORGET about the first regression model and to CHANGE it to a more simple equation Y = b*X, which means that the procedure of finding the true equation has to be repeated with Y(0)=0 (the slope must slightly change, the statistics must also change):





Note that the graphical approach **Add trendline gives a wrong value of R-squared**. Therefore, use Data - Data analysis – Regression instead, which yields both correct R=0.998993 and slope=0.139728:



SUMMARY OUTPUT			
Regression Statistics			=correl()
Multiple R	0.998993		0.99133811
R Square	0.997986		
Adjusted R Square	0.954508		
Standard Error	2.209557		
Observations	24		
ANOVA			
	df	*SS*	*MS*
Regression	1	55643.21	55643.2107
Residual	23	112.2893	4.8821415
Total	24	55755.5	
	Coefficients	*andard Err*	*t Stat*
Intercept	0	#N/A	#N/A
c(mg/L)	0.139728	0.001309	106.758118

4/ CALCULATION OF THE AMOUNT IN AN UNKNOWN

No matter if the intercept was set to zero or not, the final task with calibration curve is to calculate the amount (concentration) in an unknown. We take the correct regression equation and calculate X for a given Y:

X = Y / b (if the intercept was set to zero)

Or, if the intercept **a** was significant: **X = (Y – a) / b**

Usually, the unknown was also measured repeatedly; then we find Y as an average of the multiplicates, after testing if there is not an outlier.