Linear regression and calibration lines

1. Calibration lines

Calibration is a very important procedure because it is the standard way to remove *systematic errors* from measured data. It is also the only way to make sure that your scale of units corresponds to everybody else's. The basic idea is very simple:

1. Take a sample with ***known*** properties (a standard)
2. Measure it with your instrument
3. If your measurement gives an incorrect value, correct it

Of course you are not interested in one value but in a *whole scale* within the dynamic range of your instrument and so in general we have to use a *series* of standards and correct the whole scale. This is typically done using *simple linear regression* although there are far more elaborate schemes possible. In this lab we will explore some of the properties of calibration curves

Instructions

- Open up the workbook trumpets.xls. Make sure macros are ***enabled***. Microsoft tends to disable everything.
- Sheet 1 contains two buttons and a two column range of data points that represent a series of measured standards.
- Select the range and click the ***trumpet*** button. The button activates a macro that calculates a simple linear regression using the Linest function. The output of this procedure is summarized in the block in the J and K column. At the top of this block e.g. you will find the values for the slope and the intercept of the calibration line.
- The formulas underneath the heading: ***fit, Conf95%+,Conf95%-,Pred95%+.Pred95%-*** should be selected. Put the cursor on the bottom right corner of that range until it changes into a **+** and then double click. This should fill down to the last calibration point.
- Now select the entire data block including the empty cell above the first data point and the headers on the first row. Make a chart: a scatter plot with *only markers*
- While the chart is active, click the ***decent trumpets*** button.
- There is a green straight line. This is the ***calibration line***. It is what you use to correct you measurement with.
- Typically what is on the horizontal (X) axis here are the ***calibration values*** (the 'right' ones). Vertically you have the measurement (Y). The dimensions are *not* necessarily the same.

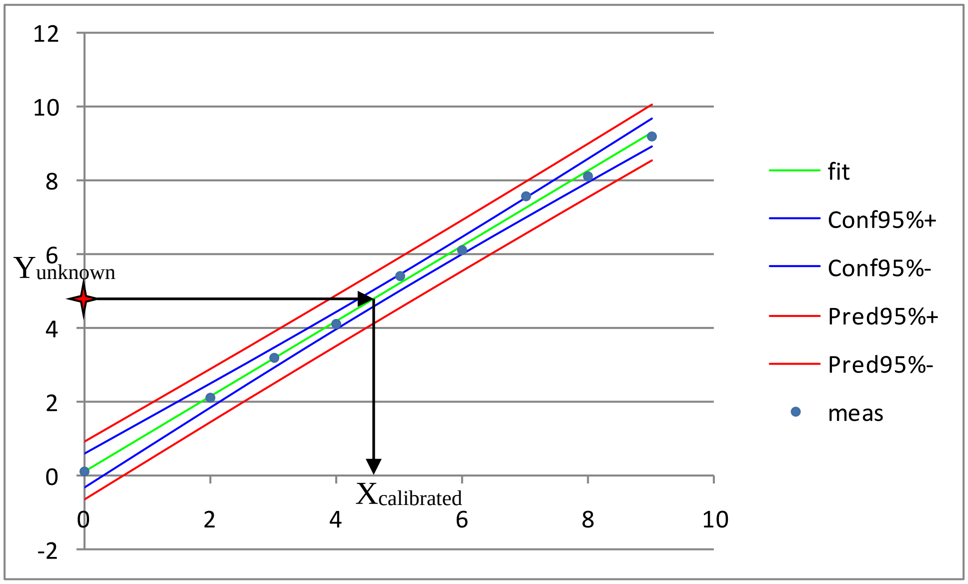
Questions:

1. Suppose we are calibrating a UV/VIS spectrophotometer and measure absorbance at a wavelength of 400 nm. We want to know if an accused person has actually

put a red poison in someone's drink. We have made up a number of solutions of that poison with known molarity and measured them. What are the units on the vertical and on the horizontal scale? What does Y represent?

2. The calibration line can be written as $Y_{cal} = b_{intercept} + m_{slope}.X_{standard}$. What are the units for the intercept and the slope in this case?
3. Inspect the formulas in C8, D8 and F8. Activate each cell and then click behind the formula that appears in the formula bar above the sheet. Write out the formulas in mathematical format in terms of the quantities given in the statistics table. Compare to the statistics handout in the CH452 manual. What is the difference between the formula for the prediction and the confidence hyperbolas?

- Notice that if I measure an ***unknown*** sample, what I do not know is the poison concentration $X_{unknown.}$ All I can do is measure its Y absorbance value. To arrive at a concentration value I have to read back, i.e. ***we need to invert***: $(Y_{unknown} - b_{intercept})/ m_{slope} = X_{calibrated}$



- Suppose $Y_{unknown}$ is measured to be 4.342. Use the slope and intercept values in the Linest block to calculate $X_{calibrated}$. (The intercept is on the top right of the range; the slope is on the left).

  Of course the above totally ignores the fact that both in the calibration measurement and in the measurement of the unknown there are *inevitable uncertainties* (read: random errors). This is why I have added the red and blue 'trumpets'. They may look like straight lines but they are really hyperbolas.

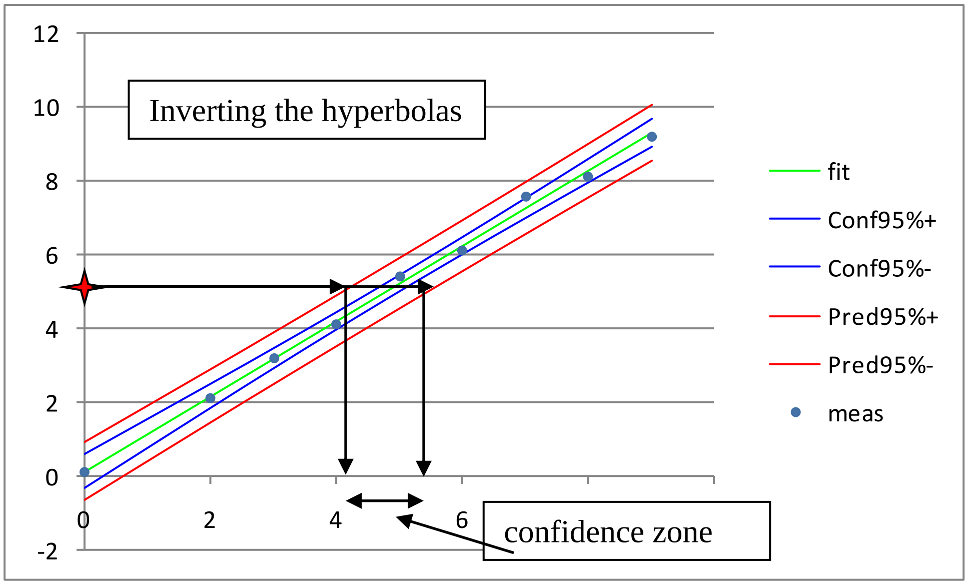
- The calibration set contains a *blank* measurement, i.e. one where X=0. Let's make a gross error there. Change its measured value to 10. As you see that really screws up things: the calibration line no longer passes through the data points but the hyperbolas become much clearer. Change the value at X=0 back (Ctrl+z).

The two trumpets

There are ***two sets*** of hyperbolas:

1. The area between the ***inner*** blue curves, known as the ***confidence limits (of the line)*** represents the zone within which you would expect ***any new calibration line*** to appear, if you measured the same standards again.
2. The area enclosed between the ***outer*** red curves, known as the ***prediction limits (around the line)*** represent the zone within which you can say ***any new data point*** will appear and be right about it 95% of the time.

So, whatever measurement we do we expect it to come within the outer trumpets, as long as the data quality and instrument settings etc. do not change. Therefore, we can use the outer curves to find the error in the calibrated X value of an unknown by a read back procedure much like what we did above:



We know that the point on the Y scale must have come from between the outer trumpet, so if we ***invert*** the hyperbolas we should get the lower and upper 95% confidence limits of the calibrated value we found above. ***Caution***: the nomenclature is very confusing: you use the *prediction limits for a point (around the line)* to find the *confidence limits (of the point)*.

You may wonder what the inner limits mean: they represent the **systematic component** or **the calibration error.** If we were to replicate our unknown measurement, we can improve our uncertainty by averaging and thus reduce the width of our confidence zone, but the calibration error would remain the same as long as we keep using the same calibration line. Thus the inner blue part does not average out no matter how many replicates we measure.

Question:

Unfortunately the exact inversion of a hyperbola leads to horrible algebra, but in a spreadsheet you can do it graphically or by preparing a look up table

- Type in M8: 0; type in M9: 0.01
- Select M8:M9 and put the cursor on the corner until **+** appears
- Drag the **+** down to say M2000. This should fill M8:M2000 with numbers increasing in steps of 0.01
- Type in N8: 0 (This is just a dummy: my buttons expect a measured value here and we are not going to put any in..)
- Go back to where the calibration data are and select the formulas in the first row beneath the heading : ***fit, Conf95%+,Conf95%-,Pred95%+.Pred95%-*** (starting under ***fit***). Hit Ctrl+c to copy
- Go to O8 and paste
- Use the **+** double click trick to copy the formulas down to the bottom of the region.
- Select M8:S2000 and make a scatter plot with only markers
- Use the ‘’Decent trumpet” button to clean it up
- (The N column just contains dummy values that create a straight line that dominates your graph if you also filled them down, click on the line in the graph and hit *delete* to remove it from your graph.)

We now have values for the calibration line and the trumpets that are not limited to where we took our calibration standards.

- Select M8:M1000 (Go to M8; hold down *Shift*; press *End*; press *Arrow-Down*) then copy (Ctrl+c)
- Goto T8 and paste

We are now ready to look up values.

First let’s say we use our calibration to measure three unknown samples

- Type in W4 = 2.33; type in W5 = 5.13, type in W6 = 6.01

Now just scroll down to find the value of 2.33 in column R and note the value of the concentration we have copied into column T. This would give you the upper confidence limit of the measurement. Obviously this is a pretty tedious way of working. Fortunately Excel has a lookup function. It is called Vlookup and you give it the value you want it to look up, then a range (table) in the first column of which it looks up your number and then the column with the values you want returned:

- type in X4: =VLOOKUP(w4, $o$8:$t$1000, 6)
- type in y4: =VLOOKUP(w4, $r$8:$t$1000, 3)
- type in z4: =VLOOKUP(w4, $s$8:$t$1000, 2)

(A handy way to convert O8:T1000 to $O$8:$T$1000 is to type the first and then hit the F4 key)

- Now select X4:Z4 and copy it and fill down to Z6 (double click +)

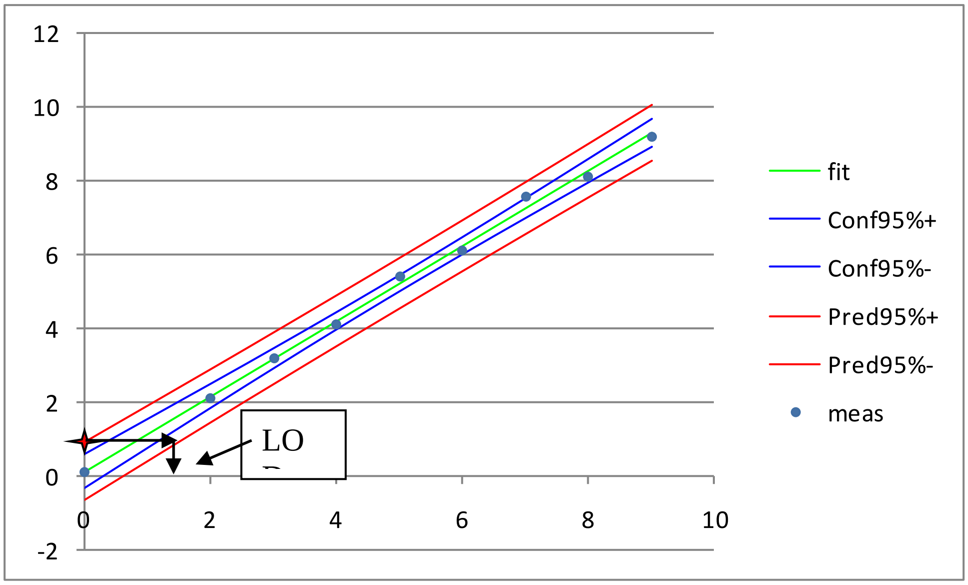
We now have the calibrated values in the X column and the lower and the higher 95% confidence limits in Y and Z. Unfortunately there is a bit of a problem. Use the AA and AB columns to calculate the distance δ_+ and δ_- from the calibrated value $X_{calibraed}$ to the upper and lower limits. (=X4-Y4 and =Z4-X4). As you see the error margins δ_+ and δ_- are *not* quite the same. This implies that the statistical distribution around $X_{calibrated}$ is no longer strictly Gaussian! This is an inconvenient truth that is conveniently *ignored in science*. Just remember: almost all data in science are obtained through calibration, so that this would mean that scientific data is generally *not* normally distributed. Fortunately the deviation from symmetrical is pretty small, particularly if the trumpets are narrow and typically people use a *symmetrical approximation formula* that can be computed from the statistics in the statistics block:
Approximate standard error =RMSE / |slope|* Sqrt(1 + (x^2 * n + sum(x^2) - 2 * x * sum(x)) / DD)
$\delta_+ = \delta_- =$ t-value*approximate standard error.
:where DD = n*sum(x2)-sum(x)2.
x =$X_{calibrated}$ and n=df+2

- Use the statistics in the regression block to compute these error margins

Notice that we use the Student t value in this approximation formula to multiply an approximate estimate for the standard error of $X_{calibrated}$ (the rest of the formula), happily assuming we can treat it as normally distributed. We then can report the result either as
: $X_{calibrated}$(approx. st. error) in 2/15 format
Or, and the ISO 9000 laws often *prescribe* that it must be given as:
: 95% confidence limits are: $X_{calibrated}$± δ.
Whether you get 95% limits or say 99% depends on what t-value you use. That is easy to change in the cell labeled as t-value: change =TINV(0.05,*xxx*) into =TINV(0.01,*xxx*) and watch what happens.

The limit of detection

Examine what the intersection point of the upper prediction hyperbola is with the Y-axis. This value has a special meaning



When you measure such an absorbance value you cannot really say much about your sample. The confidence limits now contain X=0. That means they also contain X=0.1 of 0.001 or 0.0001 or 10^{-10}. That is to say that:

1. you do not even know if your poisonous compound is actually there
2. *if* it is there, you do not even know at what scale it is there
3. all you know it is not more than the Limit of Detection

This puts you in the position of a jury trying to decide whether an accused person should be condemned or not on insufficient evidence. They can make two different kinds of errors: they can send an innocent man to jail of they can let a crook go. In either case there is a crook on the loose!

In this case CSI did not come through for you: you need to let him go, whether he has done it or not. Either your poison is not there or your data is not good enough to see it.

The limit of detection obviously depends on the confidence level I take. If I opt for 99% the bands will be broader. So what do I take? There is a *trade off* here. If I am pickier and insist upon 99% or 99.9% confidence my chances of sending an innocent man to jail will diminish (LOD will be higher), but I'll let more crooks go. The only way to diminish both types of error is to get better data.

(There is an awful lot of confusion on this point. Don't be surprised to be told by ISO 9000 people that you must take '6 sigma' to lower *both* risks at once. This is a common fallacy.)

Standard addition

Go to Sheet 2 of the workbook. It contains a number of measurements of the atomic absorption of calcium in milk. A known amount of calcium solution was added to the sample, a method called ***standard addition.*** The concentration given is the *added* concentration. In addition each sample contains a contribution from the sample itself.

Select the data range and use the two buttons to make a calibration graph with decent error trumpets.

To find the concentration of the unknown you need to back extrapolate the calibration line to where it intersects with the concentration (x-) axis, so:

intercept + slope.x=0

x= -intercept/slope

Calculate this value and generate a column with x values in small steps around it. Then copy the functions for the fit line and the two inner confidence limit trumpets and use them to calculate their values and make a graph. The idea is to find the points where the *inner* trumpets cross the horizontal axis graphically to find the 95% confidence limits of the concentration of our unknown sample.

There is also a symmetrical approximation formula you can use.

Approximate standard error =RMSE / |slope|* Sqrt((x^2 * n + sum(x^2) - 2 * x * sum(x)) / DD)

$\delta_+ = \delta_-$= t-value*approximate standard error.

:where DD = n*sum(x2)-sum(x)2.

Use it to check your results against the graphical method. All the statistics are already in your sheet.