Non-linear fitting of TGA data

Non-linear fitting

As we have seen the matrix formula $\mathbf{(X^TX)^{-1}X^TY}$ allows us to calculate the least squares estimates in a variety of models, provided these models are *linear in the parameters* $\boldsymbol{\beta}$. In many cases we cannot linearize our fitting problem. Fortunately you can still minimize the residuals (actually their sum of squares SS) with a very similar formula $\mathbf{(J^TJ)^{-1}J^TY}$.

The difference between the two is that **X** only contains information on where we take our data (our *independent* variables). **J** however also depends on the parameters themselves. In fact **J** contains the derivatives of the fit function f(x; $\boldsymbol{\beta}$) versus each parameter in each chosen data point.

This means that we need to have an idea of what $\boldsymbol{\beta}$ is before we can compute **J**. It also means that $\mathbf{(J^TJ)^{-1}J^TY}$ will only give us a *better* estimate of $\boldsymbol{\beta}$, not the *best*. That's no problem: we can keep applying the process until no more improvement is observed. This iteration process is called ***refinement***.

1. make guess of parameters
2. calculate the **J** matrix based on that guess
3. calculate $\mathbf{(J^TJ)^{-1}J^TY}$ to get better parameters $\boldsymbol{\beta}$,
4. if this is an improvement go to step 1; if not stop process

What refinement does is look for the minimum in the SS function. However this function is now like a landscape with hills and valleys, not a single well. Therefore the procedure will only work ***if*** your initial guess for $\boldsymbol{\beta}$ is close enough to the final minimum. Otherwise the procedure gets lost in the hills.

The final $\mathbf{(J^TJ)^{-1}}$ matrix and Sum of Squares tells you the uncertainties in the final parameters, just like its cousin $\mathbf{(X^TX)^{-1}}$ did. Because you have to terminate the refinement process somewhere (if improvement is less that some criterion) these values are known as *asymptotic standard errors* and $\mathbf{(J^TJ)^{-1}}$ produces an asymptotic variance covariance matrix.

Excel contains an add-in that will do all this for you and we will fit some data with it. The add-in is called the ***solver***. The instructor/TA will help you to make sure it is loaded. Unfortunately the Excel solver does not produce values for the uncertainties, but the book Excel for Chemist by J. Billo has another module on its CD that remedies that.

The data

The data we will be fitting comes from the TGA, in fact it is a decomposition run on Calcium oxalate monohydrate ($CaC_2O_4.H_2O$). It represents a series of decomposition reactions the oxalate as the temperature is ramped to 1000°C.:

$CaOx.H_2O \rightarrow CaOx \rightarrow CaCO_3 \rightarrow CaO$

The first step produces water vapor, the second CO and the third one CO_2.

Open the Non-lin excel spreadsheet. Select the data block and make a graph (line only) of the data in the C column (weight) versus the A column (time). We will work in time, but as the oven was ramped at constant heating rate we could easily translate time into temperature (in the B column).

The Model Function

The biggest problem with fitting data is always to formulate a reasonable *model*. More often than not you do not know '***the'*** model and this is a good example: a good physical model is not known for this kind of data. The graph certainly makes it obvious a straight line will not do! Deviations from straight can often be modeled by adding higher order terms (a polynomial) but this is not recommended for this type of data.

Sigmoidal (S-shaped) step functions are notorious for requiring an infinite number of terms to fit well. This violates the ***parsimony*** principle: always try to retain as many degrees of freedom as you can, or stated differently: is you can fit something with three variables, do not fit it with 300. If you throw in enough parameters you can fit even the kitchen sink! This is why we fit this with a function that already has a sigmoidal shape, the ***logistic function lgt(x):***

$$lgt(x) = \frac{e^x}{1+e^x}$$

We can fit every decomposition event as:

$$W_o lgt(a - bt)$$

As you see, there are three parameters (not 300!) per event: W_o, a and b. W_o stands for the amplitude of the change (the entire weight loss of the event). The time around which the event happens is t_{event}= -a/b and the value of b represents the slope in the weight curve at this time (how sudden the event takes place). As we are losing weight its value is always negative in our case. As we have three events, but do not lose all weight the total fit function becomes

$$W(t) = W_{o,1} lgt(a_1 - b_1 t) + W_{o,2} lgt(a_2 - b_2 t) + W_{o,3} lgt(a_3 - b_3 t) + W_{baseline}$$

As you see we have a total of ten parameters: 3+3+3+1.
It is advisable to build up such a fit job systematically in your sheet and not write out a function like this in one cell as you are bound to make mistakes.

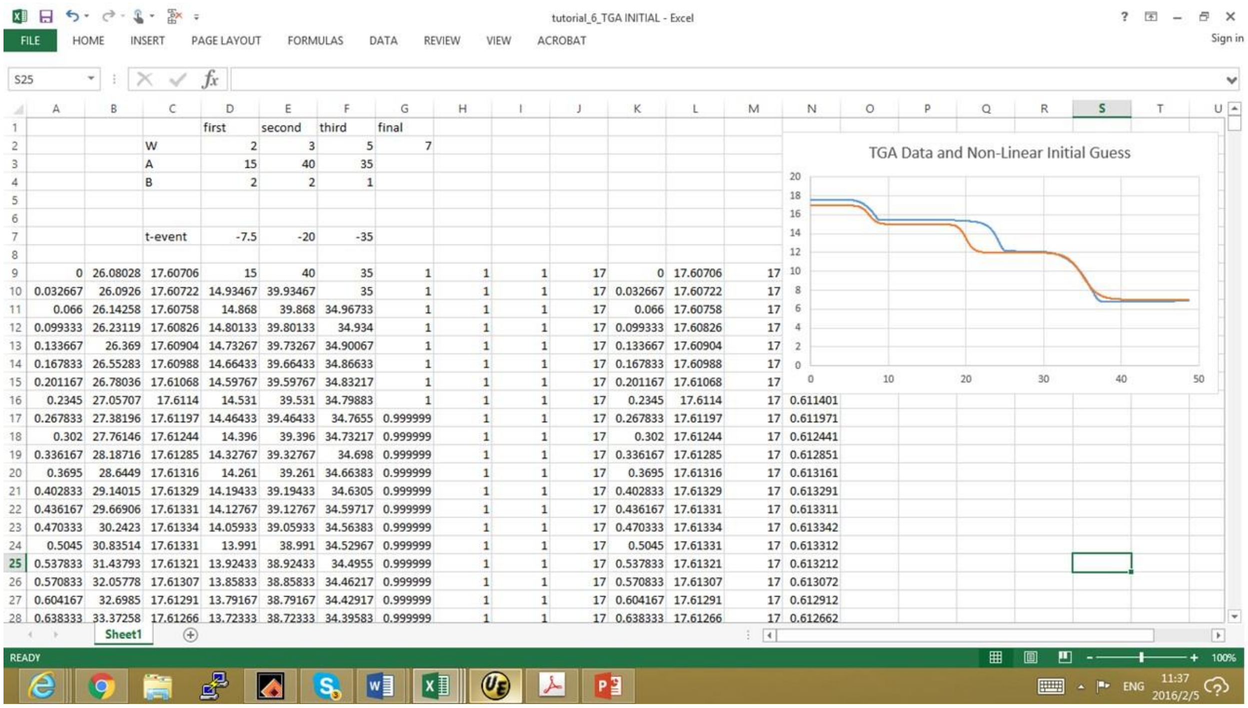
- First copy the entire data block (three columns) and open a new sheet and paste it in cell A8, B8 and C8. It consists of 1452 points so that the three columns should be filled down to line 1460.
- Now put the following initial guesses for the parameters in the range C1:G4 including the labels. The numerical parameters should be in D2:G4.

	First	second	Third	Final
W	2	3	5	7
A	15	40	35	
B	2	2	1	

1. Type in C6: t-event
2. Type in D6: -D3/D4
3. Copy D6 over D6:F6
4. Type in D8: =D$3-D$4*$A8 (this calculates *a-bt* for the first event)
5. Copy D8 over D8:F8
6. Type in G8: =EXP(D8)/(1+EXP(D8)) (The lgt function)
7. Copy G8 over G8:I8
8. Type in J8: =$D$2*G8+$E$2*H8+$F$2*I8+$G$2 (This computes the W(t) fit function)

9. For plotting purposes we will copy the relevant columns:
10. Type in K8: =A8; (time)
11. Type in L8: =C8 (the measured weight)
12. Type in M8: =J8 (the fit function)
13. Type in N8: =L8-M8 (the residual)
14. Now select D8:N8 and use the double click on the + symbol that appears on the bottom right corner of N8 to double click and fill all your calculations over the whole data set.
15. Select the K,L and M functions and make a chart with only lines of the measured and calculated data.

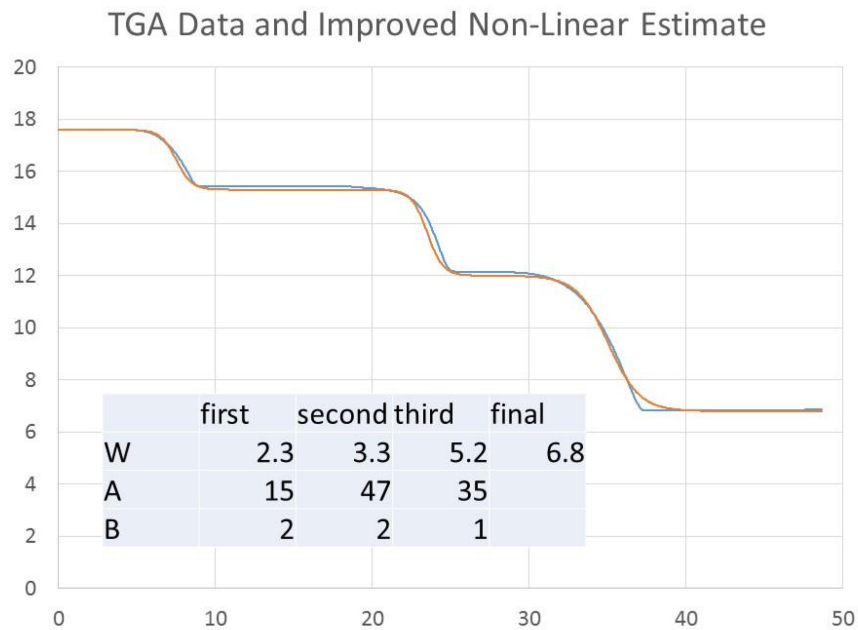
    Your worksheet should look like this.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1				first	second	third	final							
2			W	2	3	5	7							
3			A	15	40	35								
4			B	2	2	1								
5														
6														
7			t-event	-7.5	-20	-35								
8														
9	0	26.08028	17.60706	15	40	35	1	1	1	17	0	17.60706	17	
10	0.032667	26.0926	17.60722	14.93467	39.93467	35	1	1	1	17	0.032667	17.60722	17	
11	0.066	26.14258	17.60758	14.868	39.868	34.96733	1	1	1	17	0.066	17.60758	17	
12	0.099333	26.23119	17.60826	14.80133	39.80133	34.934	1	1	1	17	0.099333	17.60826	17	
13	0.133667	26.369	17.60904	14.73267	39.73267	34.90067	1	1	1	17	0.133667	17.60904	17	
14	0.167833	26.55283	17.60988	14.66433	39.66433	34.86633	1	1	1	17	0.167833	17.60988	17	
15	0.201167	26.78036	17.61068	14.59767	39.59767	34.83217	1	1	1	17	0.201167	17.61068	17	
16	0.2345	27.05707	17.6114	14.531	39.531	34.79883	1	1	1	17	0.2345	17.6114	17	0.611401
17	0.267833	27.38196	17.61197	14.46433	39.46433	34.7655	0.999999	1	1	17	0.267833	17.61197	17	0.611971
18	0.302	27.76146	17.61244	14.396	39.396	34.73217	0.999999	1	1	17	0.302	17.61244	17	0.612441
19	0.336167	28.18716	17.61285	14.32767	39.32767	34.698	0.999999	1	1	17	0.336167	17.61285	17	0.612851
20	0.3695	28.6449	17.61316	14.261	39.261	34.66383	0.999999	1	1	17	0.3695	17.61316	17	0.613161
21	0.402833	29.14015	17.61329	14.19433	39.19433	34.6305	0.999999	1	1	17	0.402833	17.61329	17	0.613291
22	0.436167	29.66906	17.61331	14.12767	39.12767	34.59717	0.999999	1	1	17	0.436167	17.61331	17	0.613311
23	0.470333	30.2423	17.61334	14.05933	39.05933	34.56383	0.999999	1	1	17	0.470333	17.61334	17	0.613342
24	0.5045	30.83514	17.61331	13.991	38.991	34.52967	0.999999	1	1	17	0.5045	17.61331	17	0.613312
25	0.537833	31.43793	17.61321	13.92433	38.92433	34.4955	0.999999	1	1	17	0.537833	17.61321	17	0.613212
26	0.570833	32.05778	17.61307	13.85833	38.85833	34.46217	0.999999	1	1	17	0.570833	17.61307	17	0.613072
27	0.604167	32.6985	17.61291	13.79167	38.79167	34.42917	0.999999	1	1	17	0.604167	17.61291	17	0.612912
28	0.638333	33.37258	17.61266	13.72333	38.72333	34.39583	0.999999	1	1	17	0.638333	17.61266	17	0.612662

Note that the initial guess is not great. Hopefully it is close enough that you can refine it. In Computational Lab 6 we will learn how to refine this using the SOLVER. However, in this lab you will simply apply some trial-and-error to improve the estimate. This will be a good way to get a feeling for how such functions depend on the input parameters.

16. You can now change the values in the parameter block to make the function fit a bit better. It is useful to change them by hand to get a feel for what they do. If the new guess is really bad, just hit Ctrl+z to correct your mistake. Here is a trial example using the input data set from the website.



The model is not perfect. The data (blue curve) has sharper features than the model (brown curve). In theory we could look for a different fitting function. However, in practice the fitting function we are using is the best one for this type of data.

17. In e.g. M4 type =SUM(N8:N1459^2)/1450/SE. To activate this formula use <Shift> <ctrl> <Enter>. This calculates the sum of squares SS, where SE is the standard error in your data. You may estimate SE using a flat section of the data to generate a series of numbers. Then you can calculate the STDEV for this section. Remember that SE is equal to $\frac{\sigma}{\sqrt{N}}$, where N is the number of data points in the short section you are using to calculate the SE.
18. Try a change in one of the parameters and look at the value of SS. It should get lower as the fit gets better.
19. For this project simply try to improve your estimate by trial-and-error, but keeping an eye on the SS function to minimize it.
20. On the pop up click the icon with the red dot of the *set target cell* option and select the cell that contains the SS (M4)
21. Then click the *Min* option of the *Set equal to* set. (We want to minimize SS!!)
22. Click the red dot icon of the *By changing cells* block an select the cell that contains the final weight. Then click solve.

You can choose which parameters you want to run first. Often it is wise not to take too many parameters at once, but once you are close enough to a good fit you can select then all at once. E.g. you can click the red dot icon and select D2:F4 then type a comma and then click G2 to get them all.

The fit is not bad but not ideal either. Make a plot of the residuals to see how bad it is! Often this is the best you can do, but it may still be useful in case you have overlapping events.