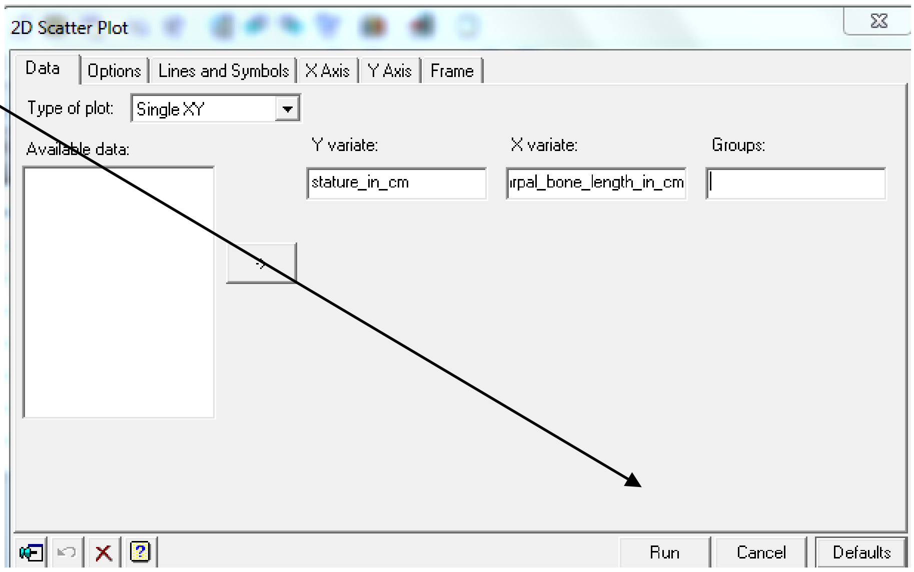
Bivariate Data Analysis using Linear Regression and Genstat

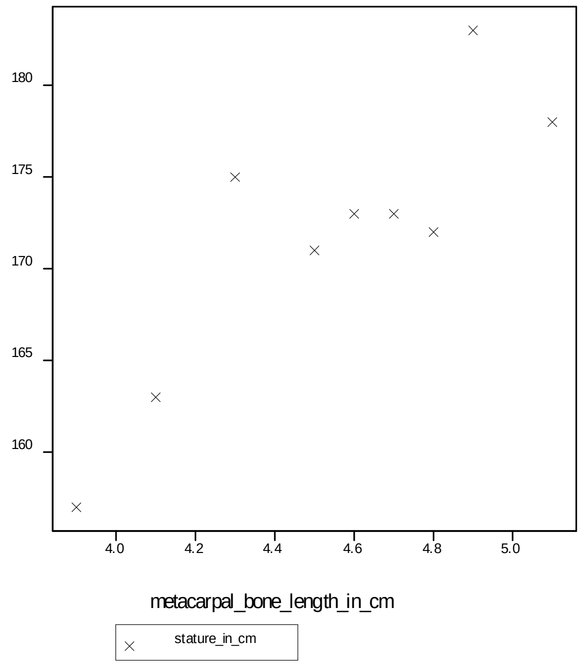
1. Open Genstat 
2. *Open the file metacarpal*

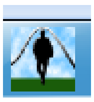
Row	metacarpal_bone_length_in_cm ch: This data was presented in the American J	stature_in_cm
1	4.5	171
2	5.1	178
3	3.9	157
4	4.1	163
5	4.8	172
6	4.9	183
7	4.6	173
8	4.3	175
9	4.7	173

3. To draw a scatterplot of the data, use the pull-down **Graphics** menu and select **2D Scatter Plot**
4. Fill in as shown by double clicking on the variables and then clicking **Run.**

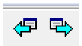


5. You should now get the graph! A **right click** will give the option to copy or click on  and the graph can be pasted into a Word document.

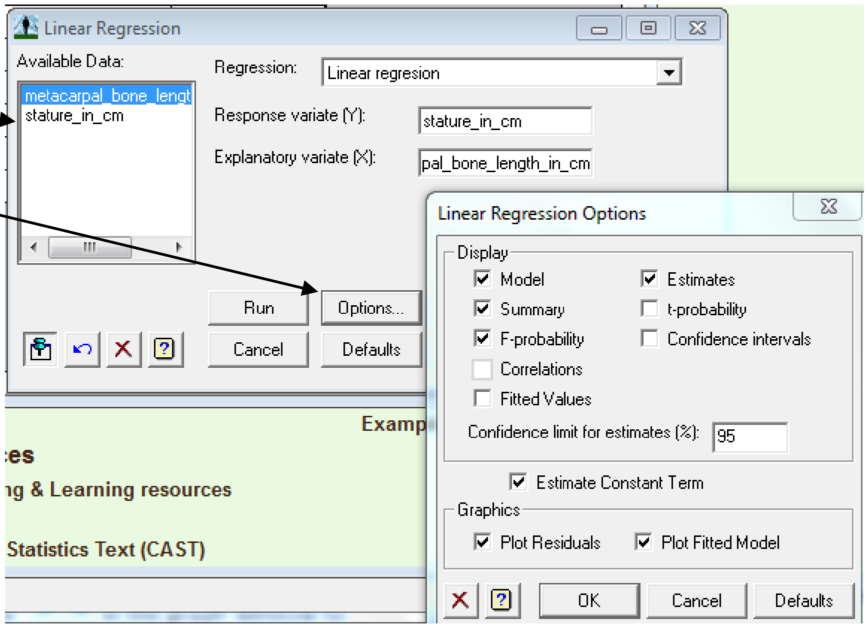


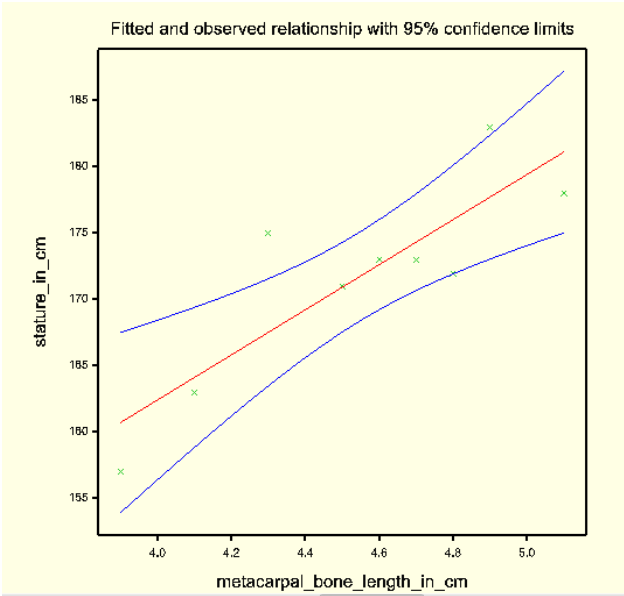
6. To return to the Spreadsheet, click on the  icon along the task bar at the bottom of the screen.

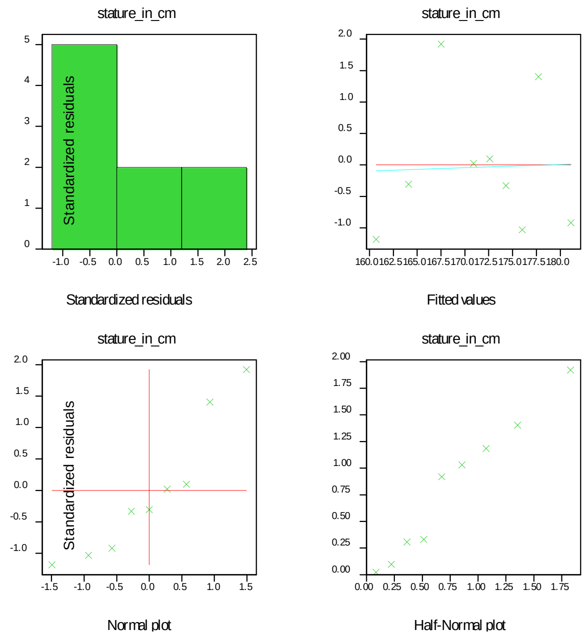
7. To perform the linear regression, use the **Stats** menu and select **Linear Regression.**

8. Fill in the dialogue box as shown, double clicking on the variables to select them. Click on **Options** to select further options and select by clicking. Fill in as shown.
9. Click **OK** and then **Run.**
10. You will now get a graph of the fitted model, the residual graphs as well as the linear regression. Use the  in the graph window to move between graphs.

    To find the output, click on the  and under the **Window** menu, select **Output.** This can be copied into Word, though you will need to select the regression output you require first.







To return to the graphs at any time just click on the 

Regression analysis

Response variate: stature_in_cm
Fitted terms: Constant, metacarpal_bone_l_length_in_cm

Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	1	347.3	347.29	19.19	0.003
Residual	7	126.7	18.10		
Total	8	474.0	59.25		

Percentage variance accounted for 69.4
Standard error of observations is estimated to be 4.25.

Adjusted R^2 > 0.5 so strong correlation

Message: the following units have high leverage.

Unit	Response	Leverage
3	157.00	0.46

this means that this point has a big effect on the trend line and hence the regression equation.

Estimates of parameters

Parameter	estimate	s.e.	t(7)	t pr.
Constant	94.4	17.7	5.34	0.001
metacarpal_bone_length_in_cm	17.00	3.88	4.38	0.003

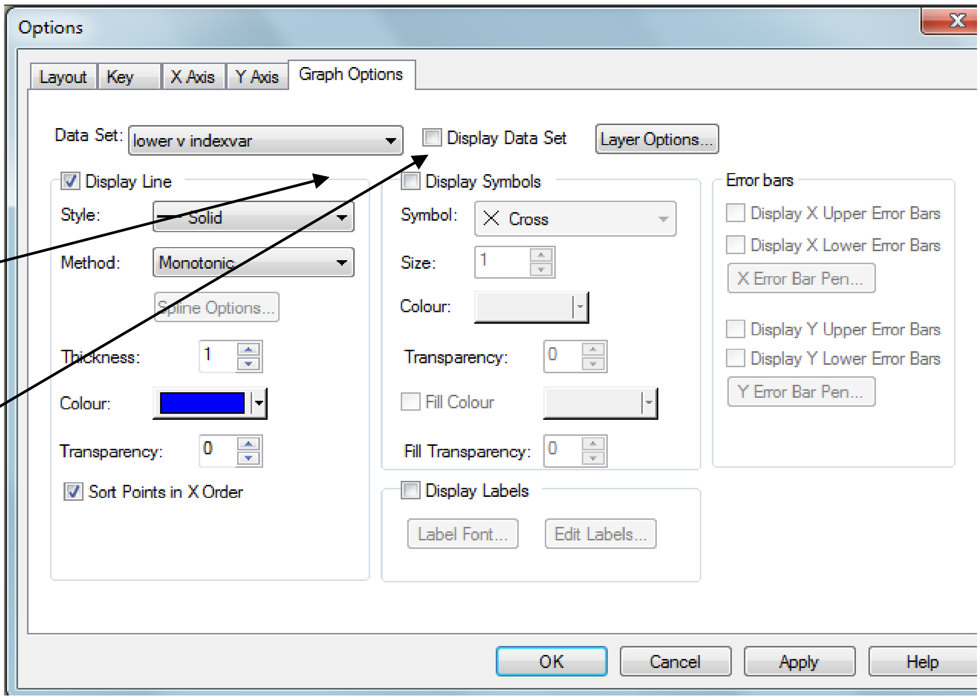
if the t probability is <0.05 then this variable is significant

The model is stature = 17 x metacarpal bone length + 94.4 cm

The graphs can be edited to remove the confidence levels if desired. In the **Graph** window, chose **Edit** and then **Edit Graph.** You now choose **Edit** and then **Graph Options**
By choosing the two Data set

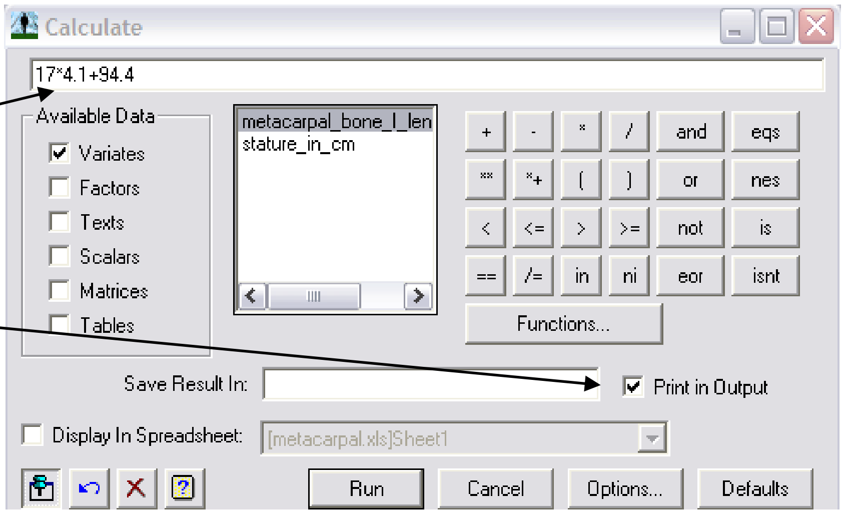
- Lower v indexvar
- Upper v indexvar

and clicking off **Display data set** you remove the lines.



Predictions

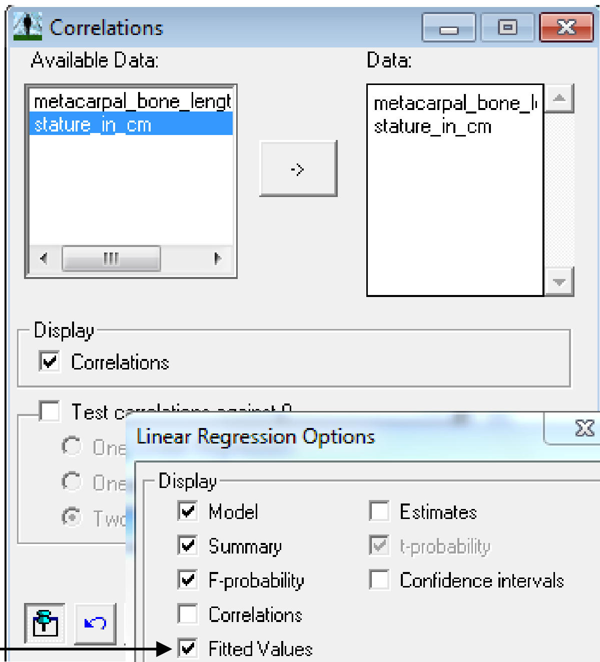
You can use your model to predict the height when given the length of the metacarpal bone for other skeletons. Using the Genstat Calculator by typing in as shown
you should get (17*4.1)+94.4 and selecting print in Output
you will get 164.1 in the Output



Correlations

To find **r** under the **Stats** menu choose **correlations** and then **correlation coefficient**

1. Click on [->] to put your variables in the Data column, tick on **Correlations** to ensure that you get the correlations
2. Click Run

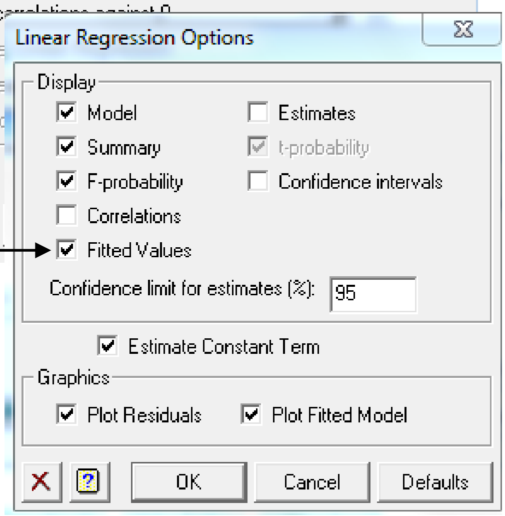


Note: Genstat gave you the ***adjusted*** $\boldsymbol{R^2}$ earlier, if you want the normal R^2, square the **r** value, or take the regression ss and divide by the total ss
(347.3 ÷474 for the metacarpal example)

Correlations between parameter estimates

Parameter	ref	correlations	
Constant	1	1.000	
metacarpal_bone_length_in_cm	2	-0.997	1.000
		1	2

Genstat will printed out all the predicted values if you ticked **Fitted Values** when you did the Linear regression



Genstat would have also printed out the standardized residuals if you ticked **Fitted Values**

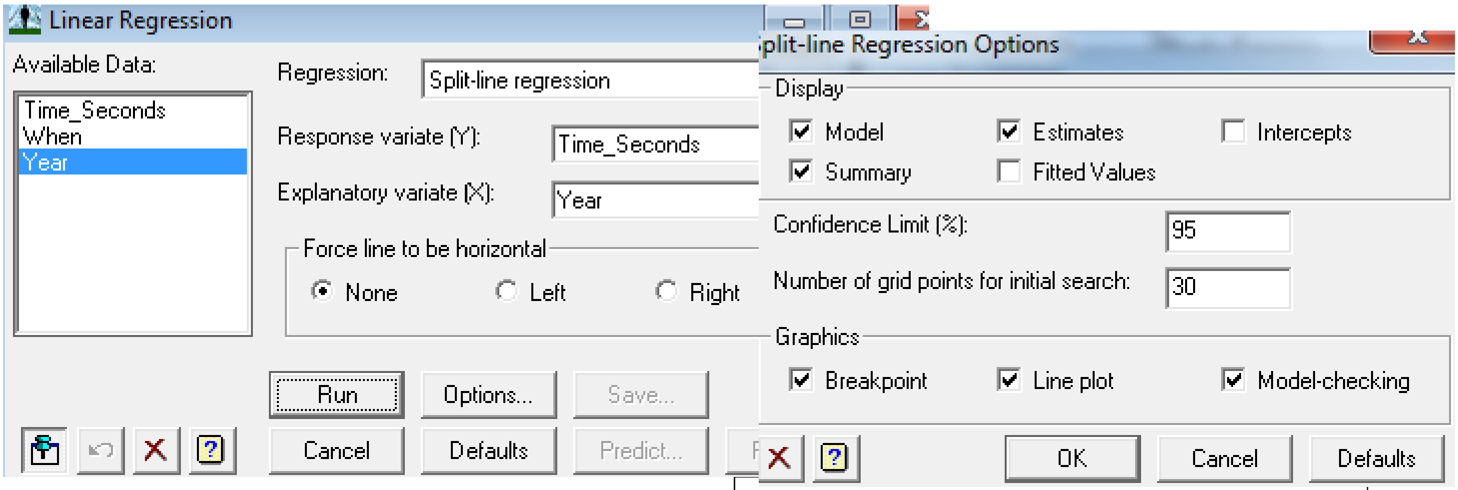
Fitted values and residuals

Unit	Response	Fitted value	Standardized residual	Leverage
1	171.00	170.91	0.02	0.11
2	178.00	181.11	-0.92	0.37
3	157.00	160.71	-1.18	0.46
4	163.00	164.11	-0.31	0.28
5	172.00	176.01	-1.03	0.17
6	183.00	177.71	1.40	0.22
7	173.00	172.61	0.10	0.11
8	175.00	167.51	1.92	0.16
9	173.00	174.31	-0.33	0.13
Mean	171.67	171.67	-0.04	0.22

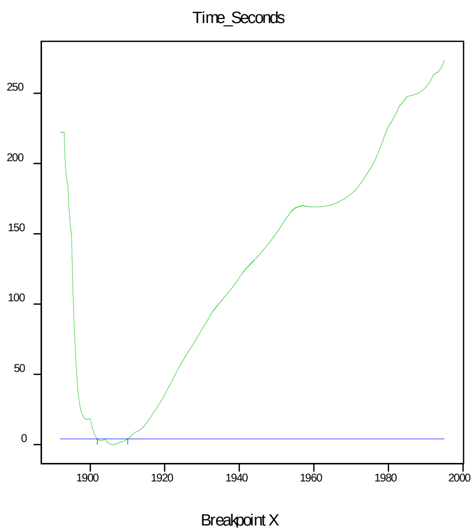
Piecewise Functions

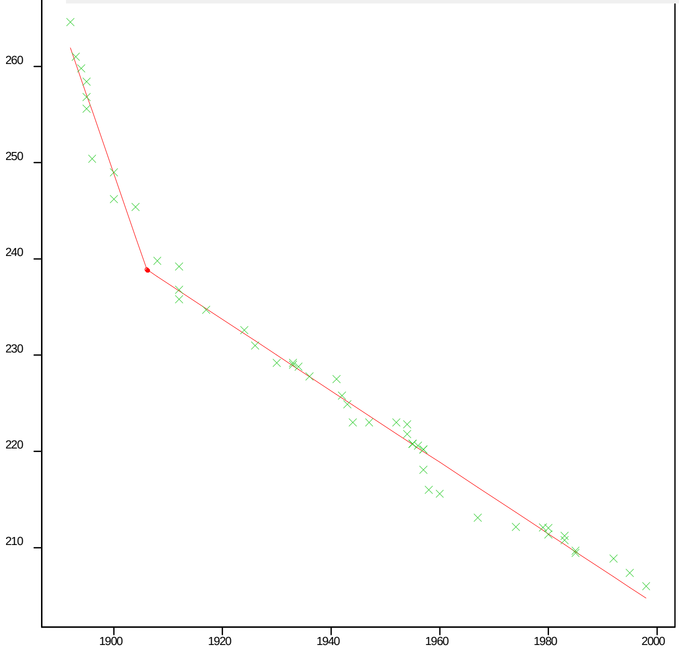
If you think your model would be better as two straight lines rather than one (or even three lines!) you can fit a piecewise model. Genstat will fit the model and even find the best breakpoint (where to split the model) for you.

1. Open the file mens 1500m
2. Choose **Stats** menu then **Linear Regression** then change the regression type to **Splitline regression**
3. Choose the options shown



You can see that there is a split in the data around 1910. Looking at the output you can see that it is at 1906

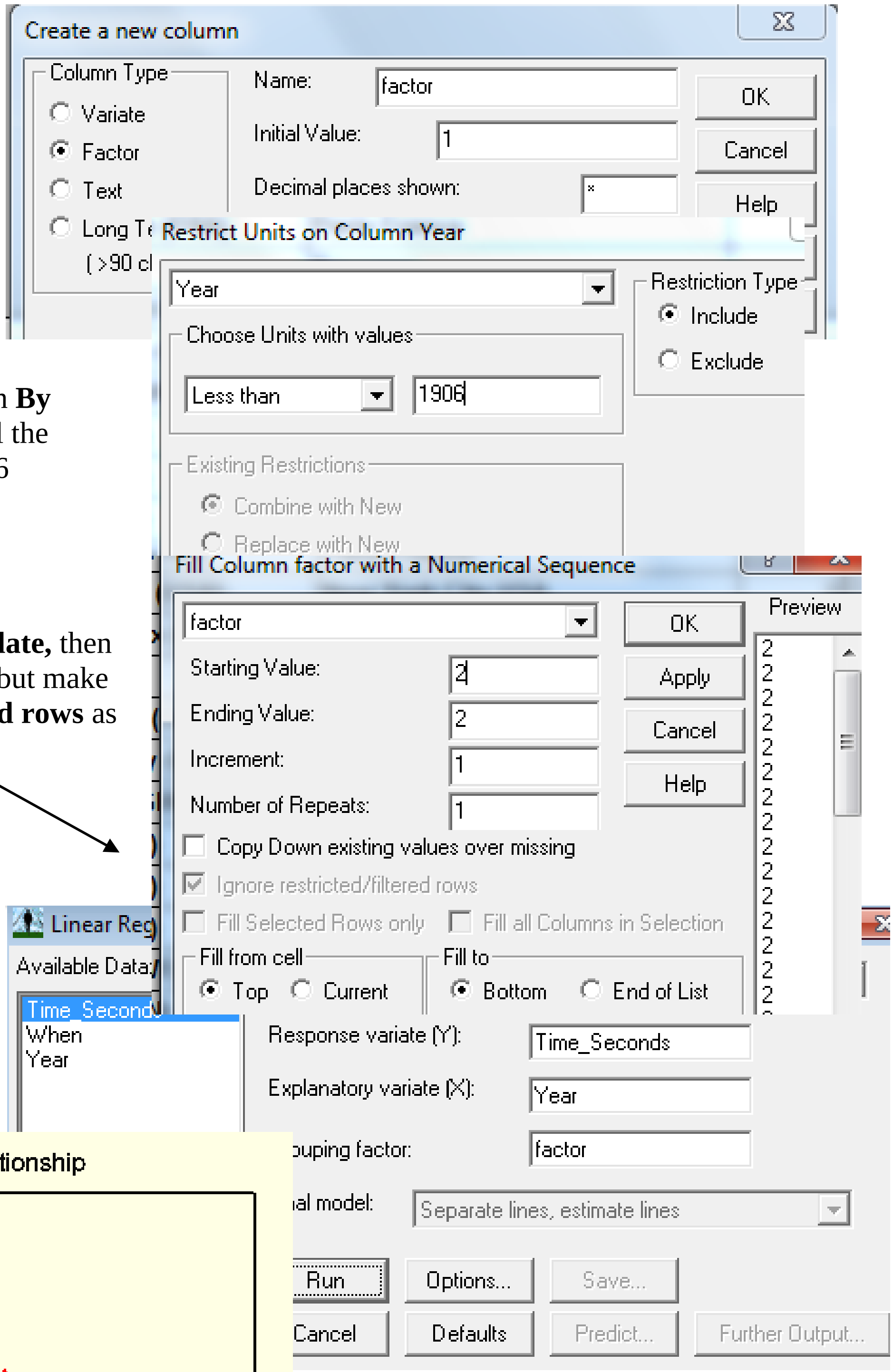




Estimates of parameters

Parameter	estimate	s.e.
Breakpoint_X	1906.07	1.26

To split the data and graph both models and get the equation for both you will need to divide the data into two groups.

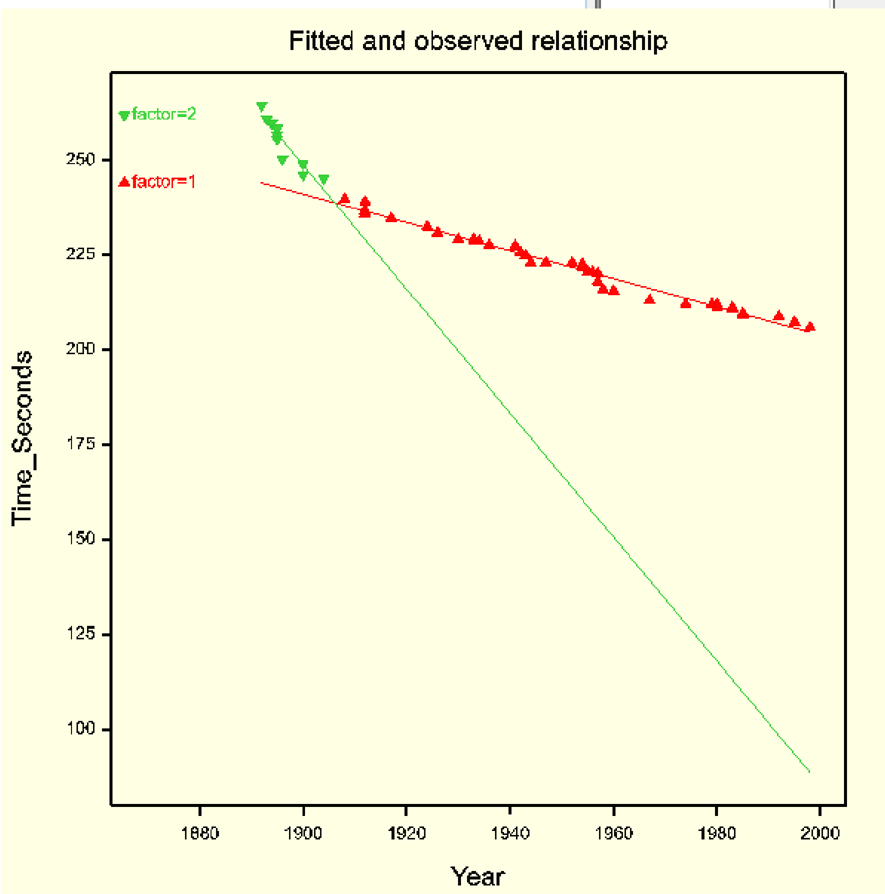


4. Create a factor column  and call it **factor**

5. Go to **Spread** then **Restrict/Filter** then **By value**: - here the data is restricted to all the values where the year is less than 1906

6. From the **Spread** menu, choose **Calculate,** then **Fill** and fill with the value 1 as shown but make sure you tick **Ignore restricted/filtered rows** as shown

7. Remove the filter with 
8. Now you can use **Linear Regression** but use **Linear Regression with groups**



Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.
Regression	3	371.3	123.77	6.03
Residual	5	102.7	20.54	
Total	8	474.0	59.25	

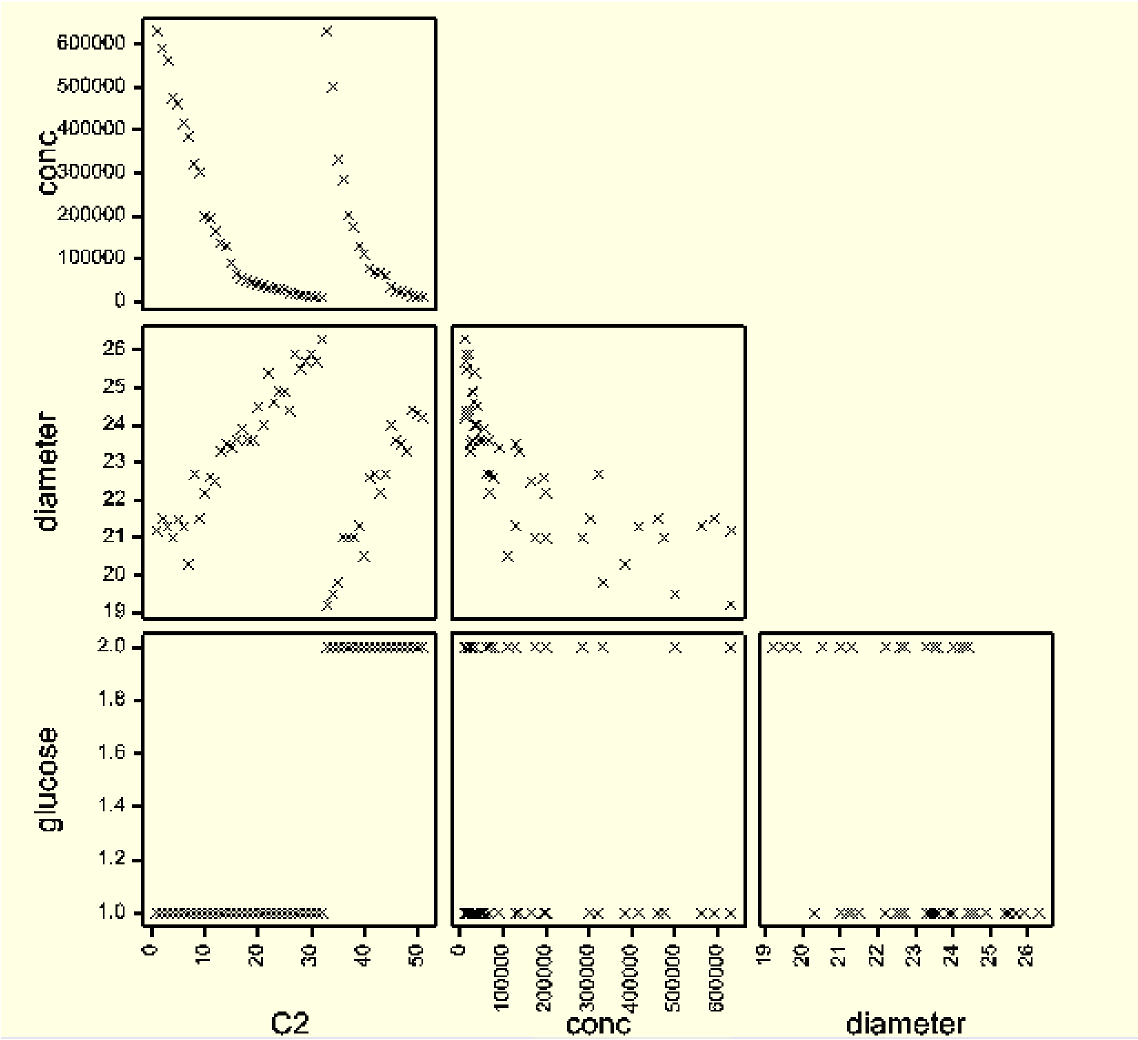
Percentage variance accounted for 65.3
Standard error of observations is estimated to be 4.53.

Estimates of parameters

Parameter	estimate	s.e.	t(5)	t pr.
factor 1	53.1	42.6	1.25	0.268
factor 2	103.5	56.8	1.82	0.128
metacarpal_bone_length_in_cm.factor 1	27.0	10.1	2.66	0.045
metacarpal_bone_length_in_cm.factor 2	15.0	11.8	1.27	0.259

While you don't have an **r** value, you do have the t probabilities and as you can see they are higher than 0.05 and before they were only 0.03 so as mentioned earlier, this data set would be better not as a piecewise model!

More than one pair of variables in your data set



When you have more than one pair of data variables, you can plot all the possible data pair combinations by using **Graphics** and then **Scatterplot Matrix** and choosing all the data variables – this now gives you a plot like the one below. Here there are 3 pairs of variables, and the six combinations are plotted – the first row gives X on the x axis and concentration (middle graph) on the y axis and diameter (right graph) on the y axis. The second row has concentration on the x axis with X(left) and diameter (right) on the x axis. The third row has diameter on the x axis and X (left) and concentration (right) on the y axis. This data is from the file cell (but the second column has been deleted – it had 2 values 1 or 2 for glucose)

Non- Linear Models

You can fit polynomial, exponential, power, square root or piecewise models using Genstat. Once the regression has been fitted, you can compare the scatterplot, the residual analysis, the R/R^2 value as well and the p-value of the F statistic and the significance of the t-test for β_1 value in your model to decide which of the models appears to be the best.

If a more complicated model is only slightly better than another, it is usual to use the more simple model as its interpretation is easier.

Remember to consider also, the number of data points you have - at least 30 is considered enough for a reliable model.

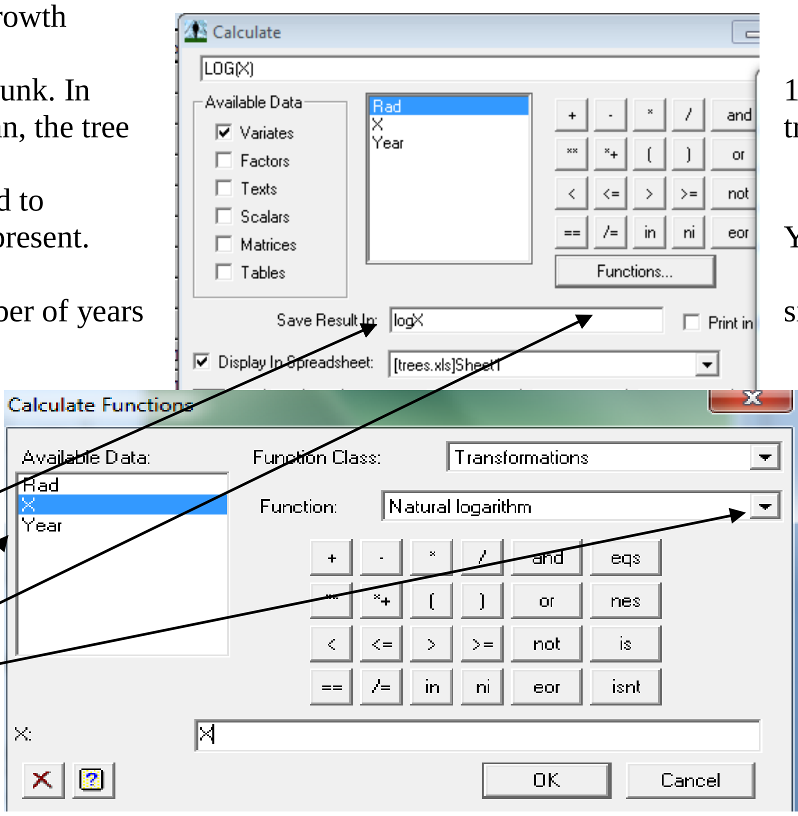
Exponential Function

$y=Ae^{kx}$ (also can be written $y=Ak^x$) e.g. $y=2e^{3x}$ or $y=3^x$

Where A is the original amount, r = rate or growth factor, x is time

The file trees has the cross section of a tree trunk. In 1990, when the recording of the cross sections began, the tree trunk which had a cross section of 2cm.

Before you can use linear regression you need to transform the data so a linear relationship is present. You can use Natural logarithms to do this.



8. Open the file *trees*. Note: X is the number of years since recording began i.e. 1990.
9. Use the calculator as before
10. This time we are going to save the results in the spreadsheet.
    a. Enter in a name for the column of the spreadsheet
    b. Click on **Functions**
    c. Use the arrow to select Natural logarithm
    d. Double Click on X
    e. Click Ok Twice

You will have got a warning message and you can see the new column is highlighted and an * put in Row 1.

Checking the output, there is a warning message

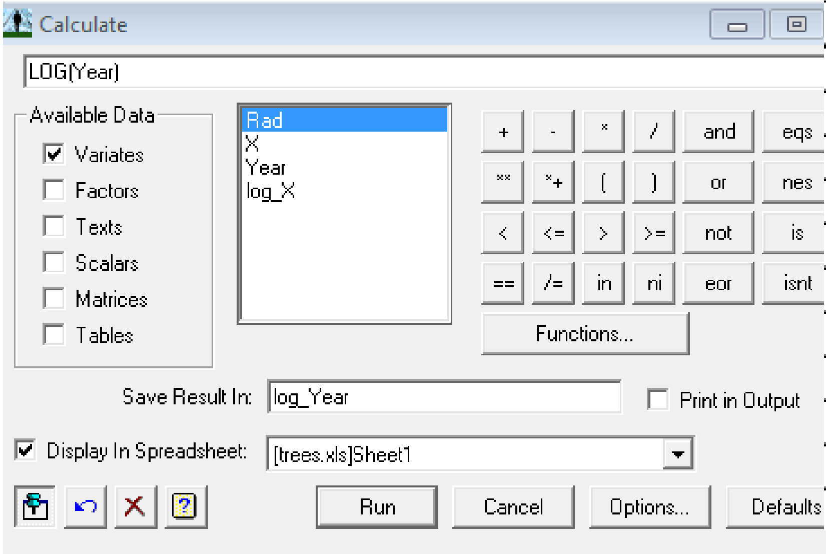
Warning 2, code CA 7, statement 1 on line 66

Command: CALCULATE log_X=LOG(X)
Invalid value for argument of function.
The first argument of the LOG function in unit 1 has the value 0.0000

As you would expect!

Row	Year	X	Rad	log_X
1	1990	0	2	*
2	1991	1	2.4	0
3	1992	2	2.88	0.693147
4	1993	3	3.46	1.09861
5	1994	4	4.15	1.38629
6	1995	5	4.98	1.60944
7	1996	6	5.97	1.79176
8	1997	7	7.17	1.94591
9	1998	8	8.6	2.07944
10	1999	9	10.32	2.19722
11	2000	10	12.38	2.30259

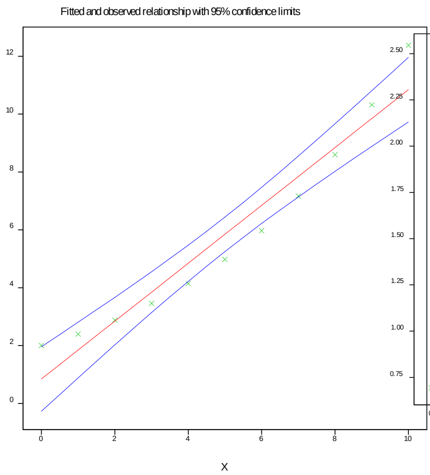
Repeat the transformation for the radius, ensuring you have a new name for the column where the results are to be displayed.

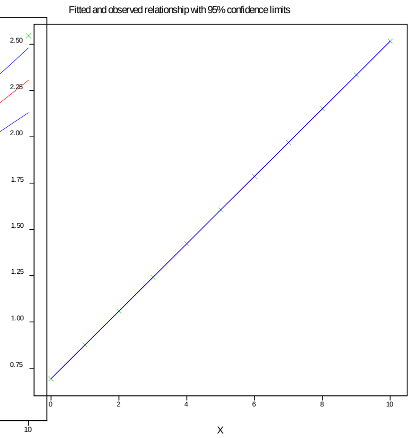


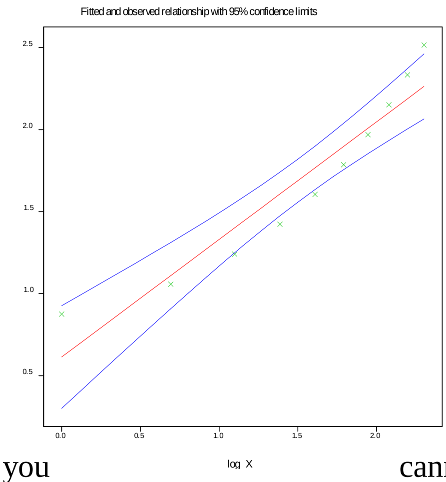
Row	Year	X	Rad	log_X	log_Radiu
1	1990	0	2	*	0.693147
2	1991	1	2.4	0	0.875469
3	1992	2	2.88	0.693147	1.05779
4	1993	3	3.46	1.09861	1.24127
5	1994	4	4.15	1.38629	1.42311
6	1995	5	4.98	1.60944	1.60543
7	1996	6	5.97	1.79176	1.78675
8	1997	7	7.17	1.94591	1.96991
9	1998	8	8.6	2.07944	2.15176
10	1999	9	10.32	2.19722	2.33408
11	2000	10	12.38	2.30259	2.51608

Now perform **Linear Regression** as you have done previously but try different combinations

a. X explanatory, Radius Response
b. X explanatory, log (Radius) Response
c. log (X) explanatory, log (Radius) Response





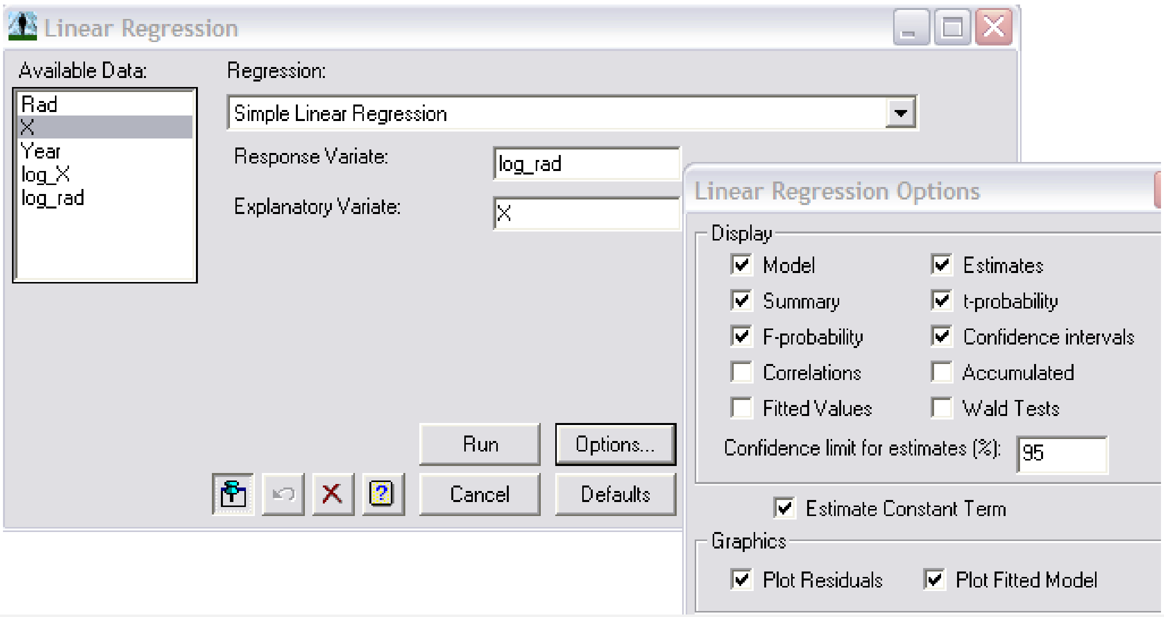


The second graph is obviously the best – it's the straightest, also notice the * in the log X column, that's because you cannot log 0, so you cannot use log X to create a model

Year	X	Rad	log_X	log_rad
1990	0	2	*	0.693147
1991	1	2.4	0	0.875469
1992	2	2.88	0.693147	1.05779

This means that an **exponential model** is possibly a very suitable model.

Now you can perform Linear Regression using X as the explanatory variable and log Radius as Response variable as you can see there is a linear relation between the two.



Regression analysis

Response variate: log_rad
Fitted terms: Constant, X

Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	1	3.655284175	3.655E+00	16215141.71	<.001
Residual	9	0.000002029	2.254E-07		
Total	10	3.655286204	3.655E-01		

Percentage variance accounted for 100.0
Standard error of observations is estimated to be 0.000475.

Estimates of parameters

Parameter	estimate	s.e.	t(9)	t pr.	lower 95%	upper 95%
Constant	0.693528	0.000268	2589.56	<.001	0.6929	0.6941
X	0.1822906	0.0000453	4026.80	<.001	0.1822	0.1824

Therefore the linear relationship is : **Ln(radius) = 0.1823 x X+ 0.6935**

Transforming this

$$e^{\text{Ln(radius)}} = e^{0.1823 \times X + 0.6935}$$
$$= e^{0.1823 \times X} \times e^{0.6935}$$
$$\text{radius} = e^{0.6935}\ e^{0.1823 \times X}$$
$$= 2.007\ e^{0.1823X}$$

We can predict that after seven years, the radius of the tree will be

$$\text{Radius} = 2.007\ e^{0.1823X}$$
$$= 2.007\ e^{0.1823 \times 7}$$
$$= 7.168 \text{ (4sf)}$$

This compares well with the observed value of 7.17.

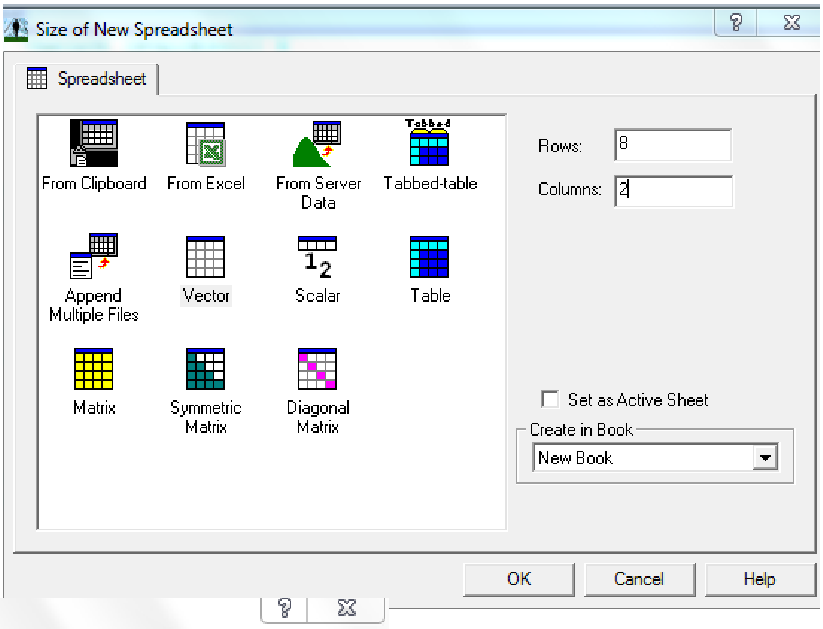
Power function

$y=kx^a$ (e.g. $y=3x^2$)

A certain type of glue needs a hardener added to set. The amount of hardener added affects the time taken for the glue to set, as shown in the table above

Hardener g	5	10	15	20	25	30	35	40
Time taken min	8.8	3.1	1.7	1.1	0.8	0.6	0.5	0.4

While this file is available as *glue*, this time we will enter the data in manually. You may wish to clear the data from the last file first (**Data, Clear All data**)

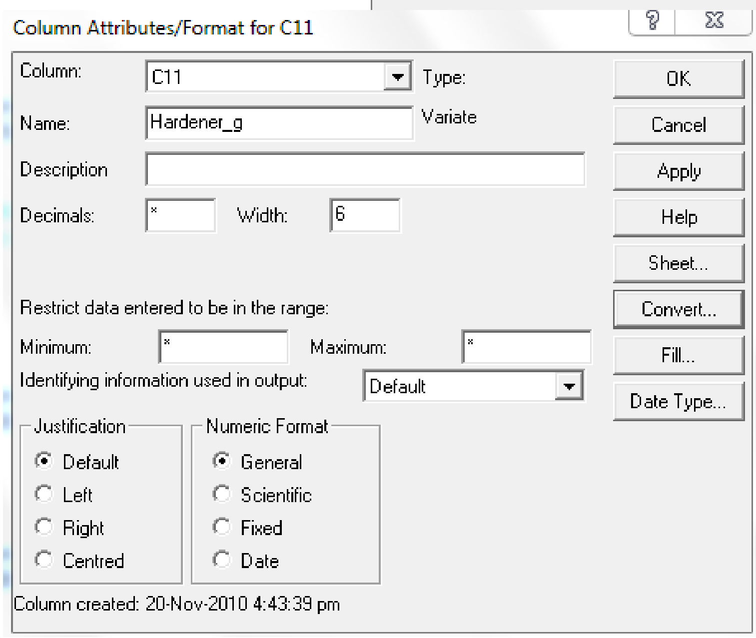


e. Click on , you will need 8 rows and 2 columns

f. Type in the hardener values in the first column and the time taken values in the second column

g. Right click in the first column and choose **Column Attribute**.

h. Fill in the dialogue box as shown below. This is where you can also change the type of data by using



Convert if it is the wrong type (variate when is should be date etc.) and where you can change the **Date Type.** You can alter the width here or by manually dragging in the spreadsheet window.

- Repeat for the other column, naming it **Time_taken - min**

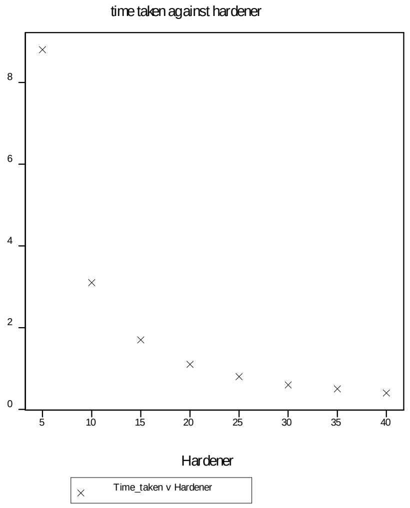
Now you can transform the data as before. (Remember to use **Natural Logarithms**) and graph the three possible models

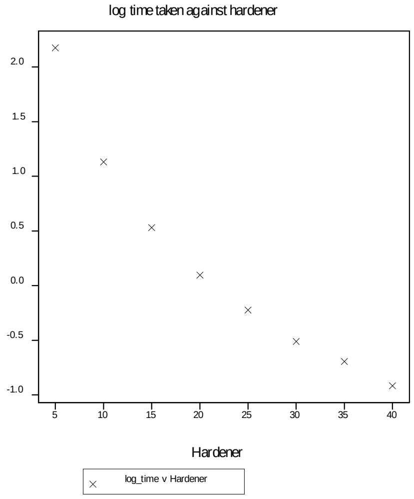
- Explanatory : Hardener, Response: Time taken
- Explanatory : Hardener, Response: log (Time taken)
- Explanatory : log(Hardener), Response: log(Time taken)

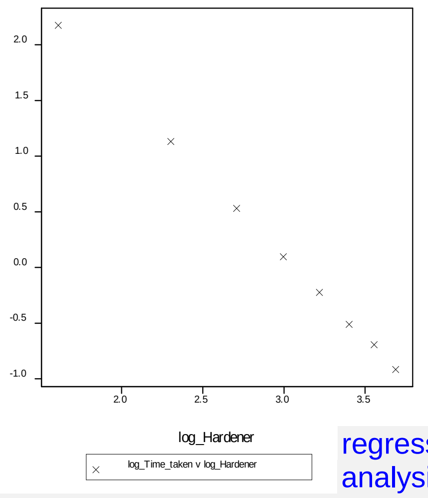
Row	Hardener_g	Time_take	log_Hardener	log_Time_taken
1	5	8.8	1.60944	2.17475
2	10	3.1	2.30259	1.1314
3	15	1.7	2.70805	0.530628
4	20	1.1	2.99573	0.0953102
5	25	0.8	3.21888	-0.223144
6	30	0.6	3.4012	-0.510826

Now graph the three possible models.

The last graph looks the most linear, so perform Linear Regression on Explanatory : log(Hardener), Response: log(Time taken) to find the equation for the power model







regression analysis

Response variate: log_time
Fitted terms: Constant, log_hardener

Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	1	7.614093	7.6140927	31269.41	<.001
Residual	6	0.001461	0.0002435		
Total	7	7.615554	1.0879362		

Percentage variance accounted for 100.0
Standard error of observations is estimated to be 0.0156.

Estimates of parameters

Parameter	estimate	s.e.	t(6)	t pr.
Constant	4.5504	0.0252	180.42	<.001
log_hardener	-1.48273	0.00838	-176.83	<.001

Parameter	lower95%	upper95%
Constant	4.489	4.612
log_hardener	-1.503	-1.462

In this glue example, the y intercept is 4.55 and the gradient -1.48

$\ln(\text{hardener}) = -1.48\ln(\text{time}) + 4.55$
$e^{\ln(\text{hardener})} = e^{-1.48\ln(\text{time}) + 4.55}$
$\text{hardener} = e^{-1.48\ln(\text{time})} \text{ x } e^{4.55}$
$\text{hardener} = e^{4.55} \text{ x } e^{-1.48\ln(\text{time})}$
$= 94.6 \ \text{time}^{-1.48}$ ($-1.48\ln(\text{time}) = \ln(\text{time})^{-1.48}$)

We can test this model to by substituting in a hardener value e.g. 35 and checking the time taken.
$\text{Time} = 94.6\ (35)^{-1.48}$
$= 0.49$ very close to the observed 0.5

Now we can use this to predict the time taken for 50g
$\text{Time} = 99.48\ (50)^{-1.5}$
$=0.28$ minutes

Polynomial

You may fit any **polynomial** in Genstat

- Choose Linear Regression but this time change the Regression to Poynomial Regression, then choose whether you want a quadratic, cubuc etc, you will get a similar output to before
- Regression analysis
-
- Response variate: stature_in_cm

- Fitted terms: Constant + metacarpal_bone_I_length_in_cm
- Submodels: POL(metacarpal_bone_I_length_in_cm; 2)

Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	2	370.4	185.20	10.73	0.010
Residual	6	103.6	17.27		
Total	8	474.0	59.25		

- Percentage variance accounted for 70.9
- Standard error of observations is estimated to be 4.16.

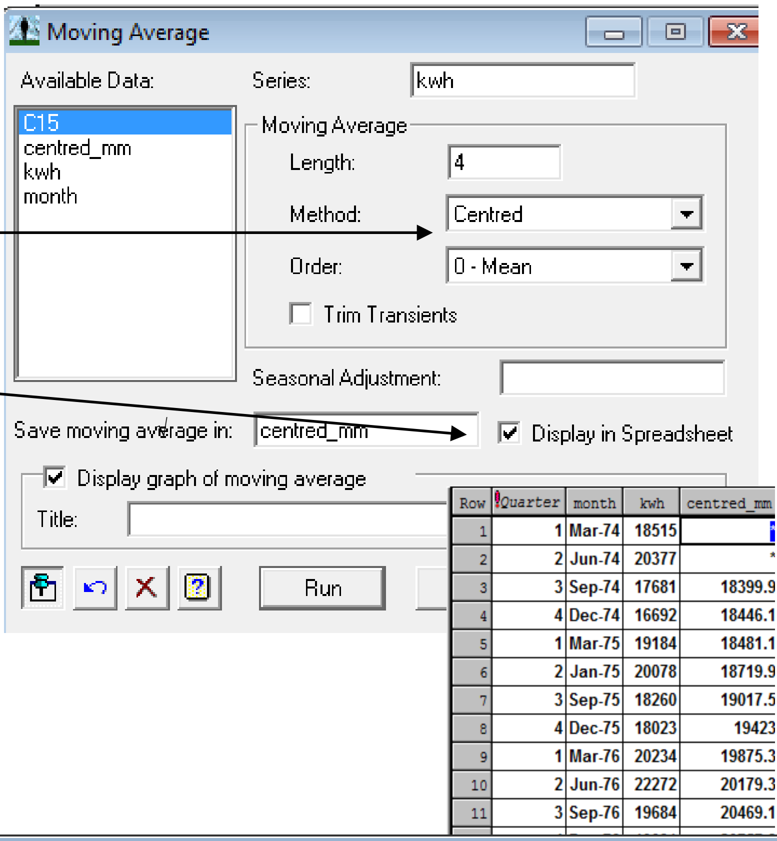
Message: the following units have high leverage.

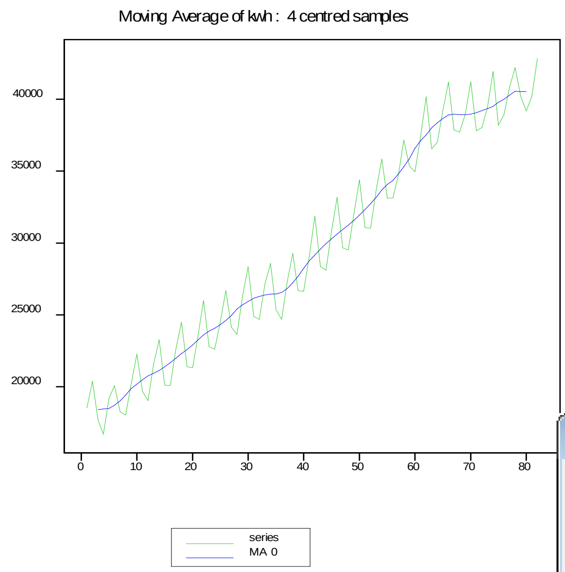
Unit	Response	Leverage
2	178.00	0.73
3	157.00	0.74

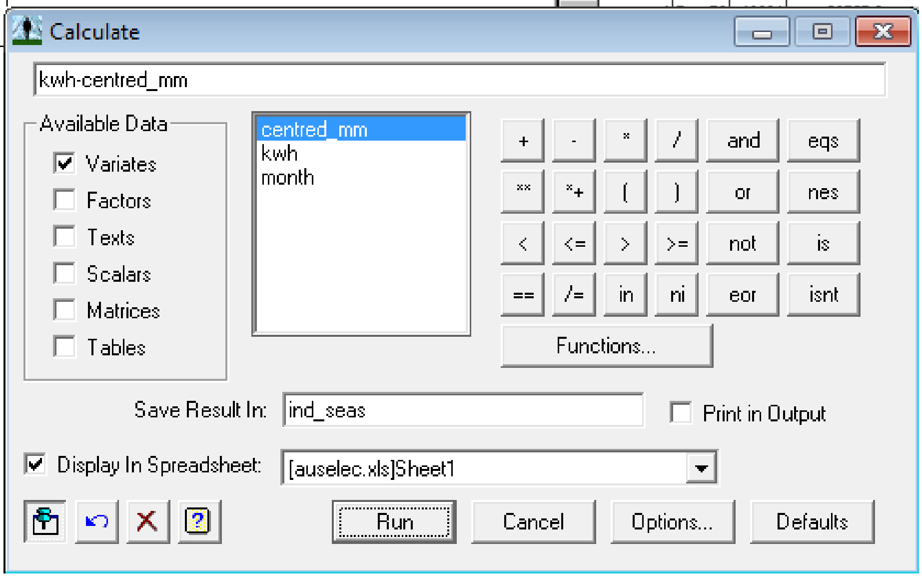
Time Series using Genstat

Open the file *Auselec*

- From the **Stats** menu choose **Time Series** and then **Moving Average**
- the series will be *kwh*
- Length will be 4 as quarterly data
- Method will be centred
- Type in a name for the column
- Tick trim transients
- Click Display in Spreadsheet





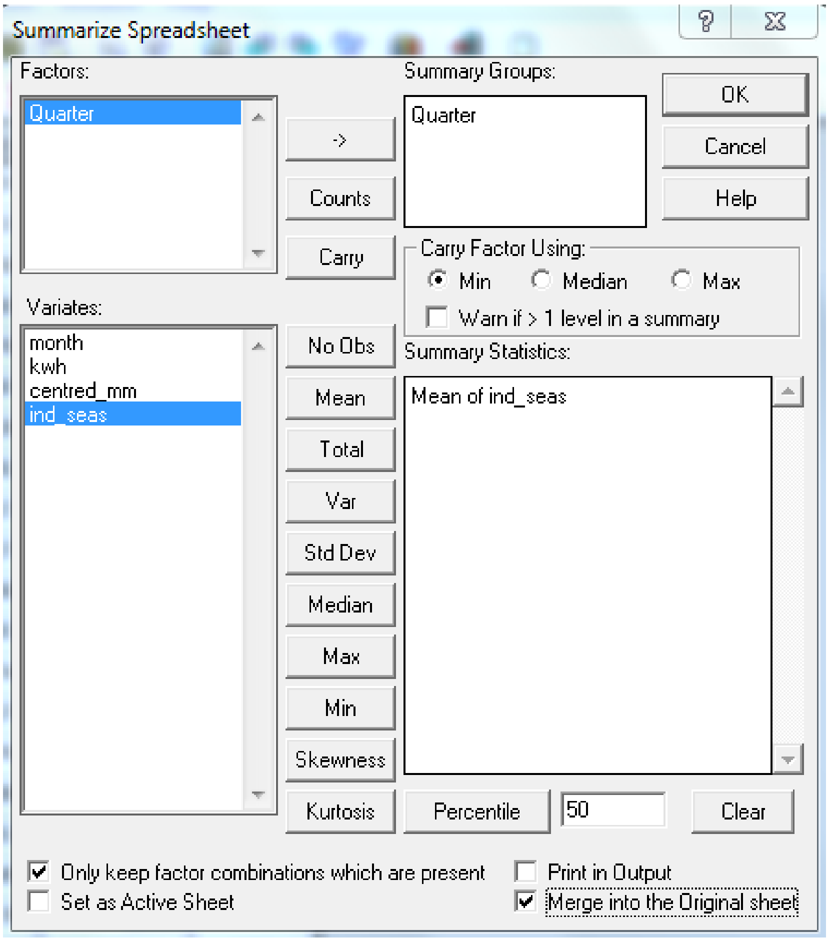


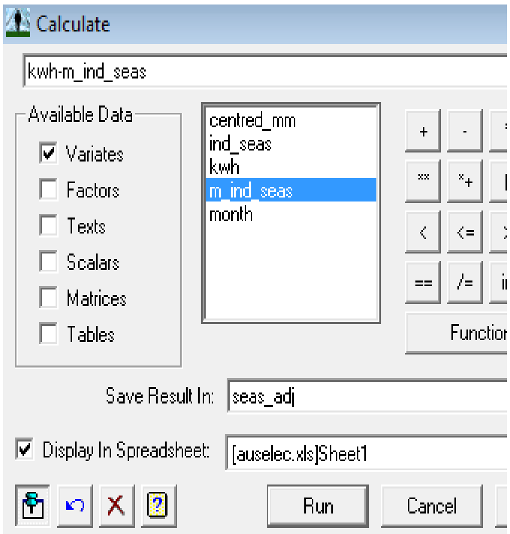
To find the Individual seasonal value, use the calculator 

To find the average seasonal value, the *Quarter* column needs to be a factor. This is indicated by the ! in front. If it is not a factor, right click and select **Convert to Factor**.

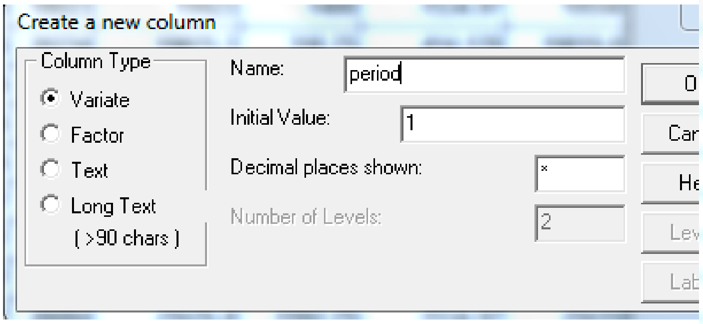
Now to get the average seasonal effect, choose **Calculate** from the **Spread** menu and then **Summary Statistics.** Remember to click **Merge**!

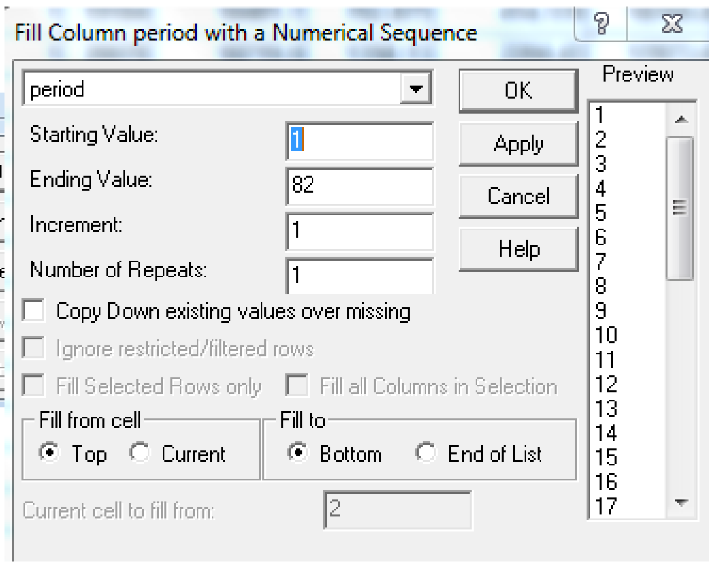
If the quarter column isn't there, just insert a column (from the ***Spread*** menu) with the required number of factors and use Fill from the ***Spread*** menu



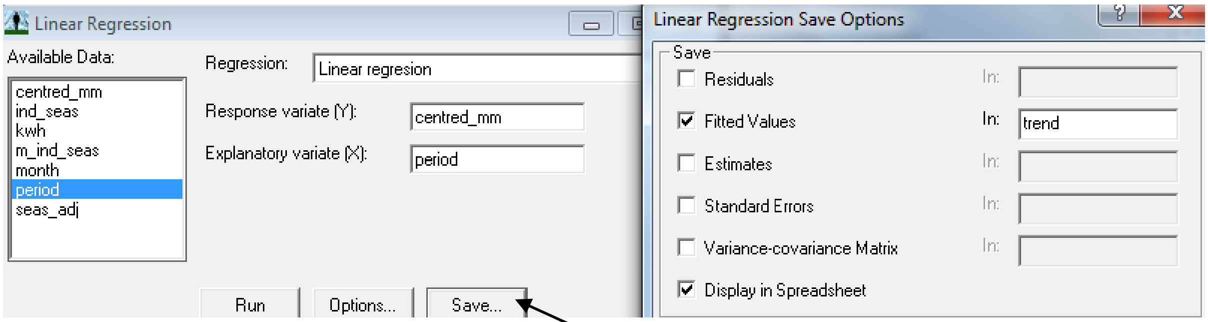


You can now also find the seasonally adjusted data using the **Calculator** 
To get the trend line and its equation you
need to perform **Linear Regression.** You need to know how many time periods have passed. You can insert a new column (Choose **Insert, Column** from the **Spread** menu) 
To fill it easily choose **Calculate** then **Fill** from the **Spread** menu





Now run the **Linear Regression**



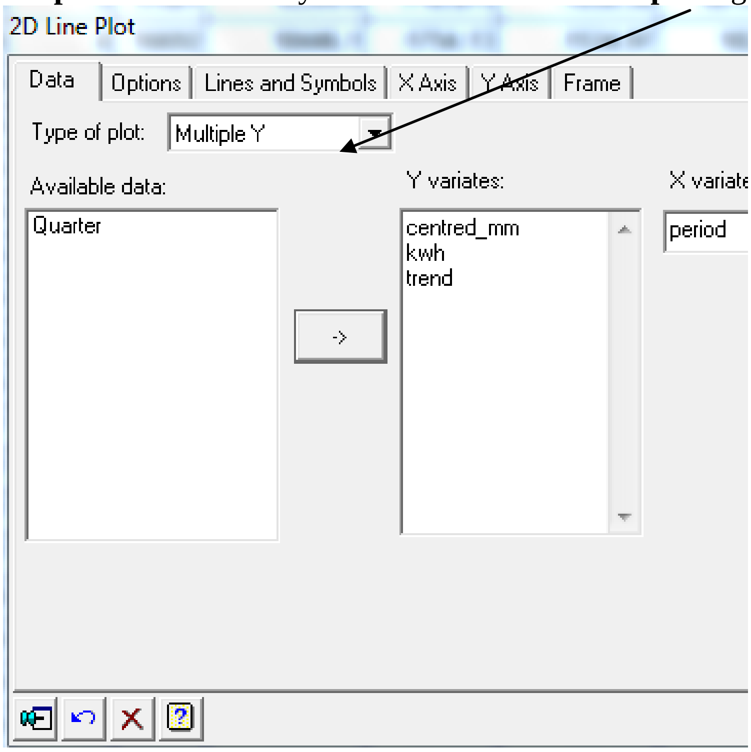
To save the fitted values, you click on the **Save** option when you run the **Linear Regression**

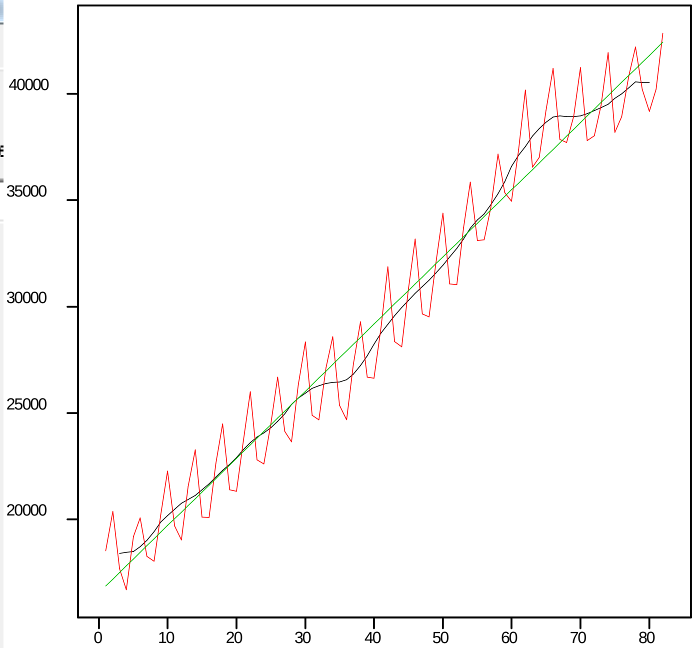
Estimates of parameters

Parameter	estimate	s.e.	t(76)
Constant	16557.	168.	98.40
period	315.57	3.56	88.55

So the model is kwh = 315.57 * quarter period + 16557

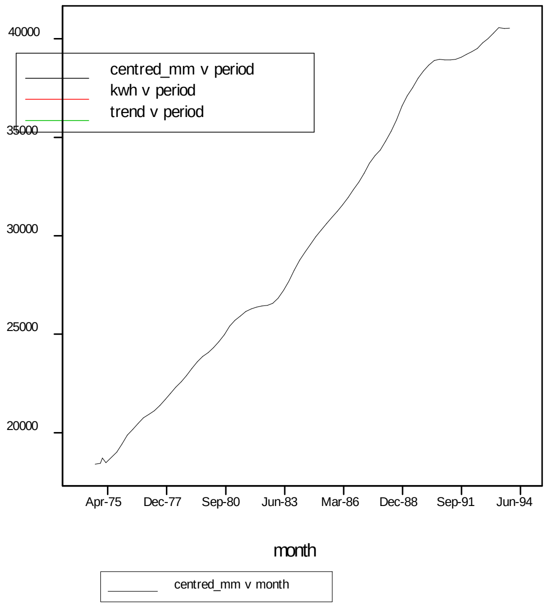
To graph the raw data, the trend and the smoothed data on the same graph, you choose **Line 2D** from the **Graphics** Menu. Then you need to choose a **Multiple Y** graph

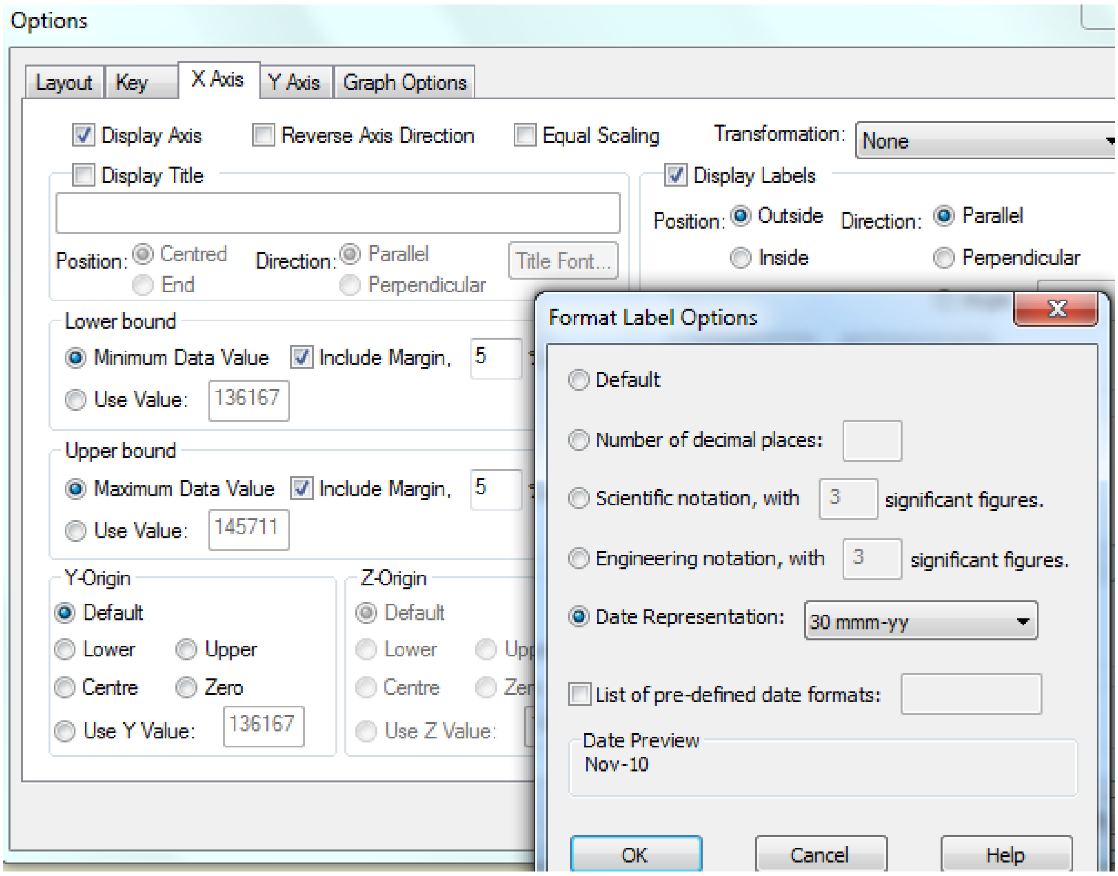




If you prefer a graph with the dates along the bottom graph just select month rather than period for the X variate, however you will need to edit the graph to change the axis to read in dates...

Choose **Edit** then **Edit graph** as you did earlier and change the **x-axis** as shown.





To make predictions, you can just use the formula for the trend line and then add on the average seasonal effect.

You can use the computer to work out the moving means (or medians) and produce a graph with a trendline and find the equation of the trend line.