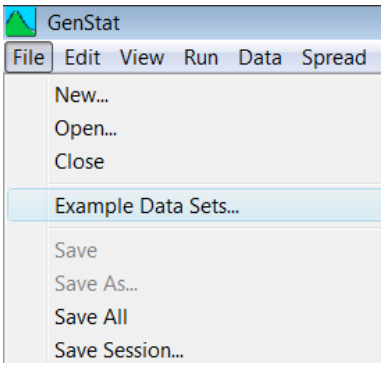
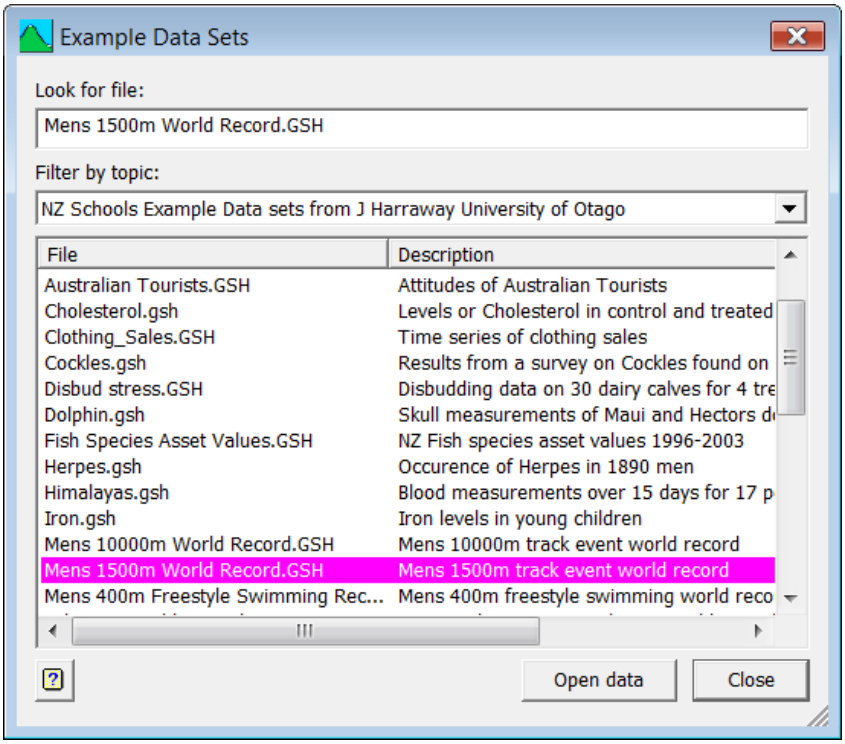
The GenStat for Schools program has an example s directory with a number of example datasets. We are adding to this as we get input from teachers. It has a number of datasets that are accompanied by videos produced by John Harraway at University of Otago. It is envisioned to distribute the videos with the DVD. The program will be available for download from the web, and students will be able to install in on their home PCs.

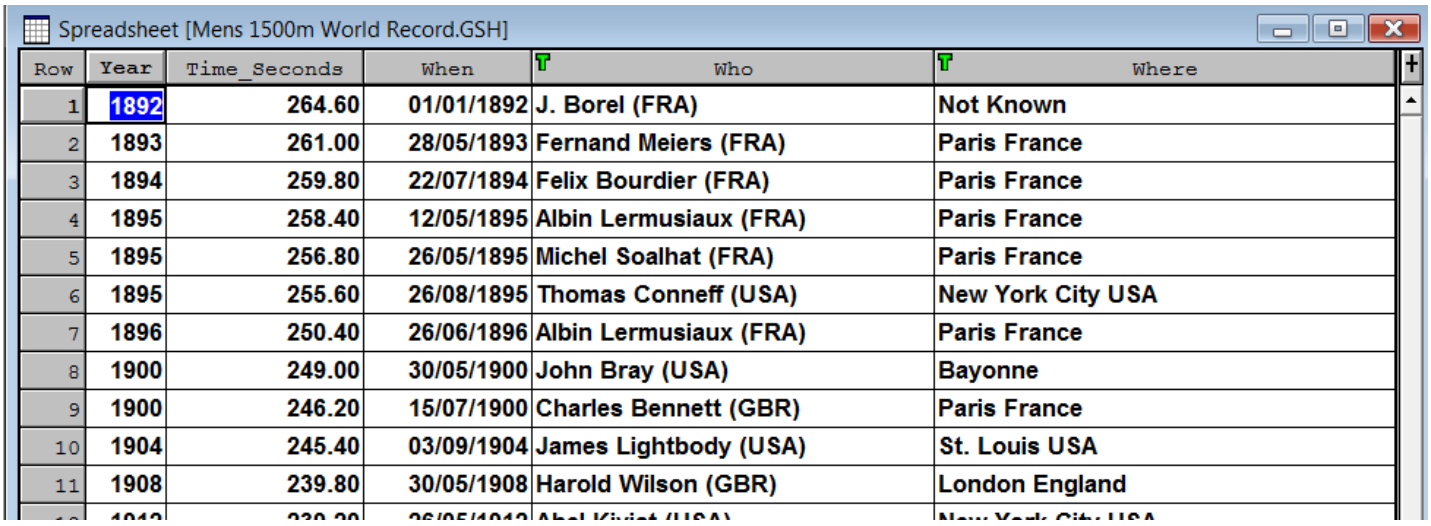
The following shows linear regression on the world record times in the men's 1500 track race over the last century.



On the Files menu, there is an examples data sets item. Subset the list of datasets by selecting NZ Schools in the Filter by topic, and select the Mens 1500m World Record data set:



This opens the spreadsheet:



Row	Year	Time_Seconds	When	Who	Where
1	1892	264.60	01/01/1892	J. Borel (FRA)	Not Known
2	1893	261.00	28/05/1893	Fernand Meiers (FRA)	Paris France
3	1894	259.80	22/07/1894	Felix Bourdier (FRA)	Paris France
4	1895	258.40	12/05/1895	Albin Lermusiaux (FRA)	Paris France
5	1895	256.80	26/05/1895	Michel Soalhat (FRA)	Paris France
6	1895	255.60	26/08/1895	Thomas Conneff (USA)	New York City USA
7	1896	250.40	26/06/1896	Albin Lermusiaux (FRA)	Paris France
8	1900	249.00	30/05/1900	John Bray (USA)	Bayonne
9	1900	246.20	15/07/1900	Charles Bennett (GBR)	Paris France
10	1904	245.40	03/09/1904	James Lightbody (USA)	St. Louis USA
11	1908	239.80	30/05/1908	Harold Wilson (GBR)	London England

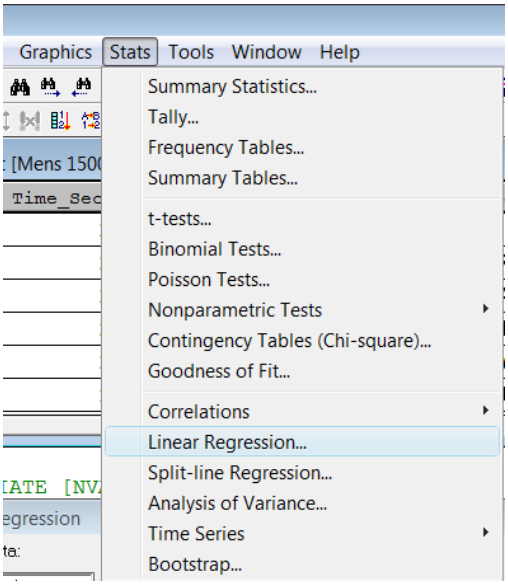
Note the Spread menu allows the spreadsheet to be manipulated in various ways.

The Graphics menu allows data to be graphed and contains all the common graphics.

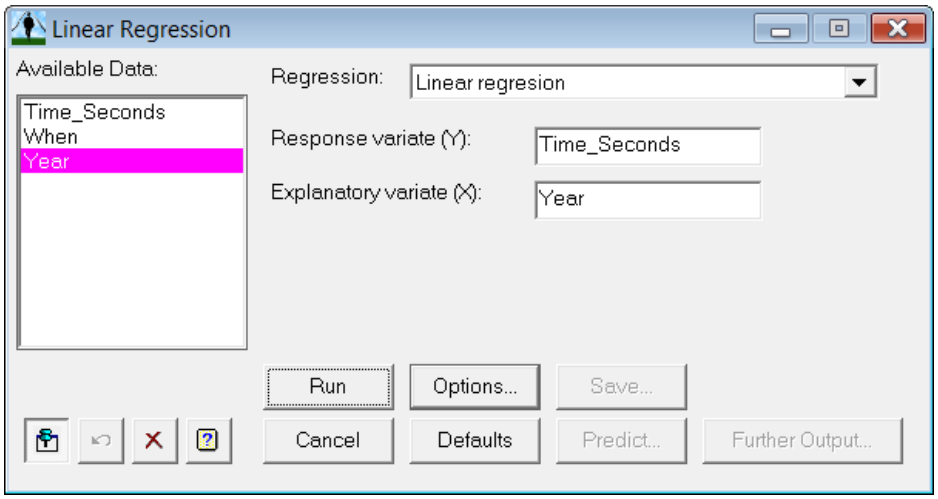
The following is a text file that describes the data:

```
"The data is the set of world records in the men's 1500 metre
track race from the years 1892 to present.

Year         - Year of record being broken
Time_Seconds - Record time in seconds
When         - The date of the record being broken
Who          - The athlete who broken the record and his country
Where        - Where the record was broken

Source:
http://en.wikipedia.org/wiki/World_record_progression_1500_metres

Notes: Before 1912 the world record was not maintained by IAAF.
Before 1912 times were recorded to nearest fifth of a second.
In 1967 times started being recorded to nearest 100th of a second.

Analysis:

Fit a linear regression model to the times.
Use the Options button and check the two graphics items
for the model checking and fitted value plot.
Examine the residual plots and fitted model to decide if a
linear model fits this data appropriately.

Use the Predict button to estimate what the record time should
be in 2010.

Does a quadratic model fit better (drop down the regression
type list to polynomial and re-enter Year as the explanatory?
How does this change the prediction in 2010?

Fit a two-part model using the split-line regression menu.
This will allow for the faster decrease in the earlier record,
by fitting separate slopes for the early part and later part
of the series with an estimated change over point.
Use the Options button to select the 3 graphics items
for break-point, line-plot and model checking before running
the menu.

Does the change point in 1912 when the IAAF started ratifying
the records?

For consideration:
This must be a strictly decreasing series so would you expect
the residuals to be from a normal distribution? Is there any
evidence for this in the residual plots."
```

Now to do a linear regression of Time_Seconds with Year, Open the Stats | Linear Regression menu:



This opens the menu below. To enter the variables to be analysed, double click Time_Seconds for Y, and Year for X:



Clicking the Run button produces the output in the Output window:

Regression analysis

Response variate: Time_Seconds
Fitted terms: Constant, Year

Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.	F pr.
Regression	1	11806.7	11806.69	655.91	<.001
Residual	49	882.0	18.00		
Total	50	12688.7	253.77		

Percentage variance accounted for 92.9
Standard error of observations is estimated to be 4.24.

Message: the following units have large standardized residuals.

Unit	Response	Residual
1	264.60	3.04

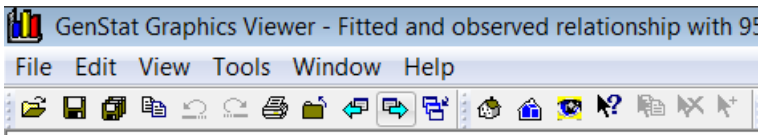
Message: the residuals do not appear to be random; for example, fitted values in the range 222.10 to 246.35 are consistently larger than observed values and fitted values in the range 200.75 to 209.97 are consistently smaller than observed values.

Message: the error variance does not appear to be constant: intermediate responses are less variable than small or large responses.

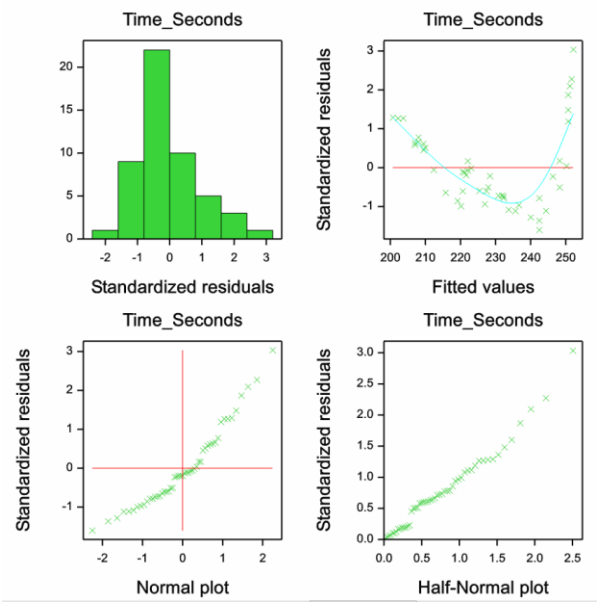
Estimates of parameters

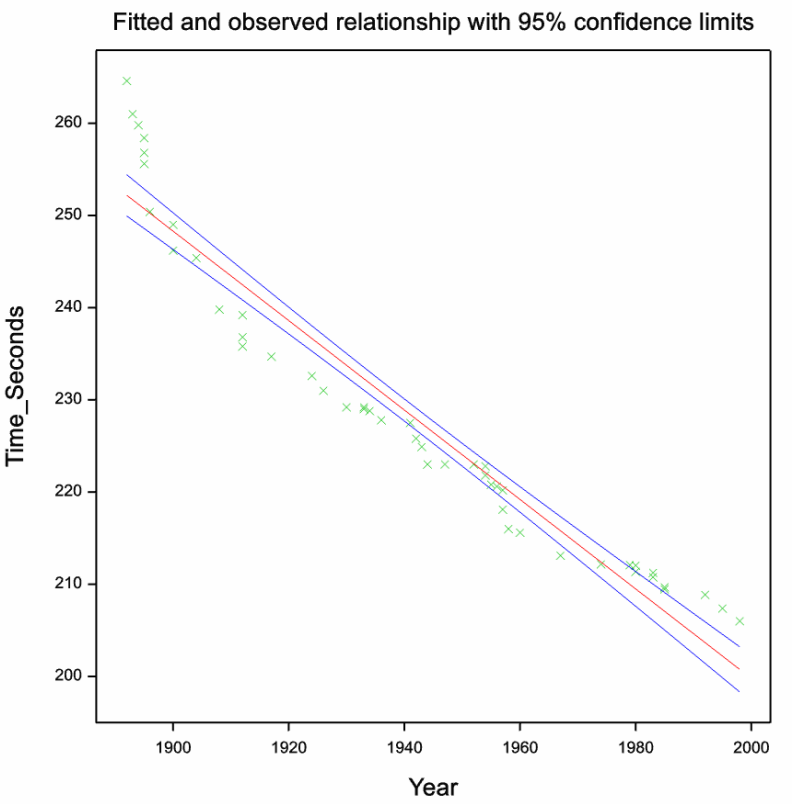
Parameter	estimate	s.e.	t(49)	t pr.
Constant	1170.0	36.8	31.81	<.001
Year	-0.4851	0.0189	-25.61	<.001

And two graphs in the graphs viewer, residual plot and a fitted relationship graph:(note to view the residual plot you will need to use the back button on the toolbar.

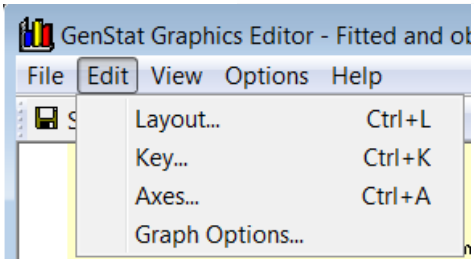


You can see a strong departure of the points from the line at the start of the series.



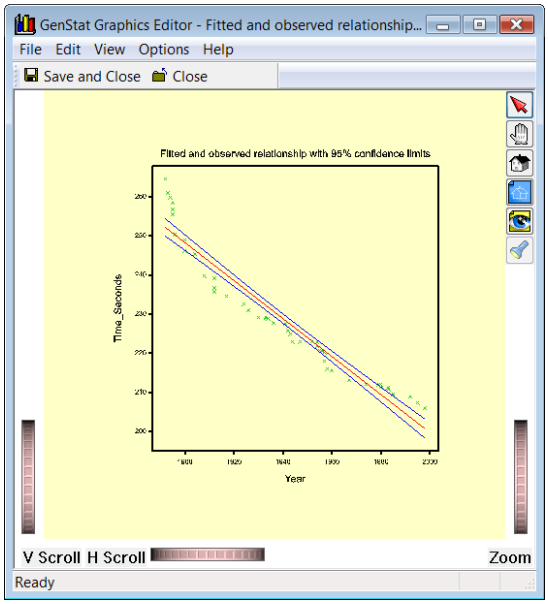


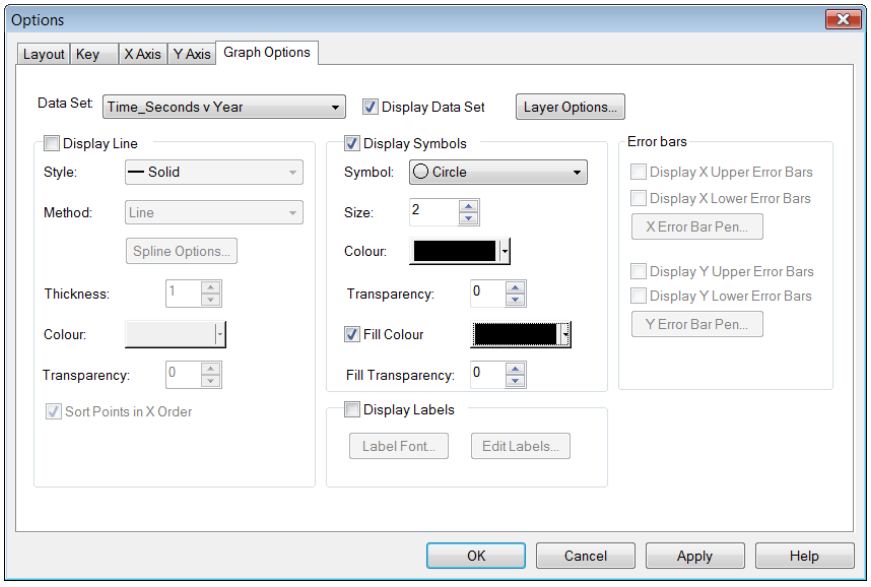
The graph can be put into Word using Edit | Copy (Ctrl+C) and Edit | Paste in Word. The graph can be edited by double clicking the image to bring up the graphics editor.



Using the Edit menu allows you to change the graph:

Use Graph Options to change the points and lines:

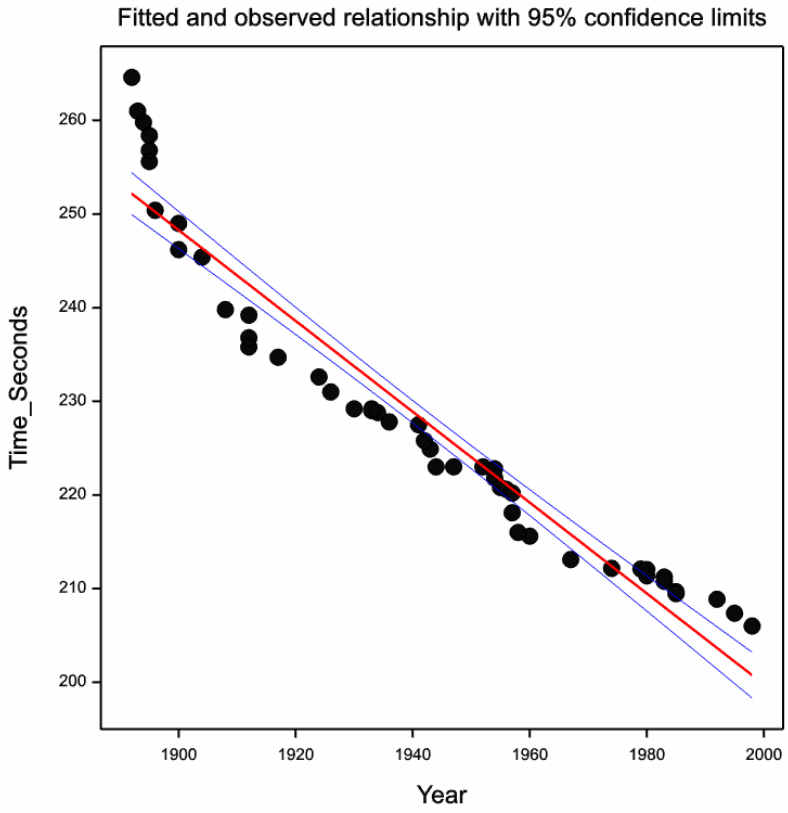




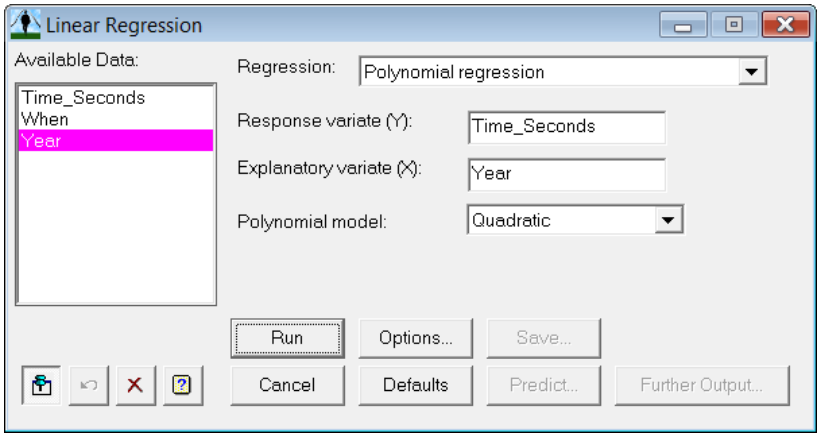
Here the data points are made to be filled black circles 2 twice as large as default.

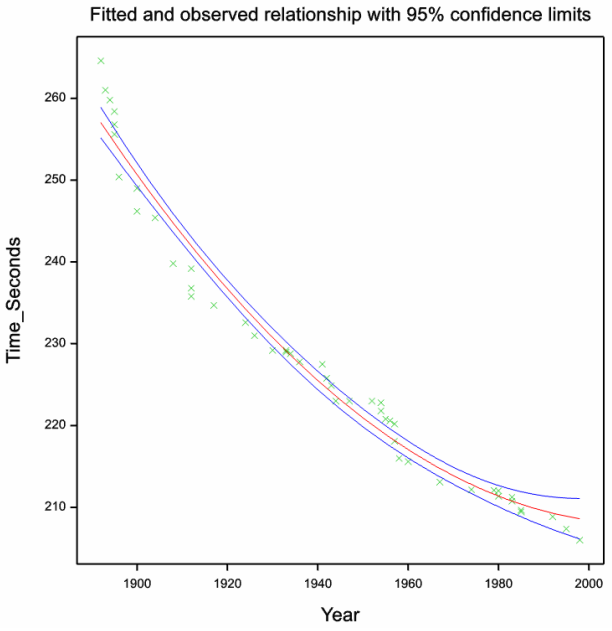
The drop down list Data Set can be used to change the lines and confidence curves.

Modified graph with heavier points and line.

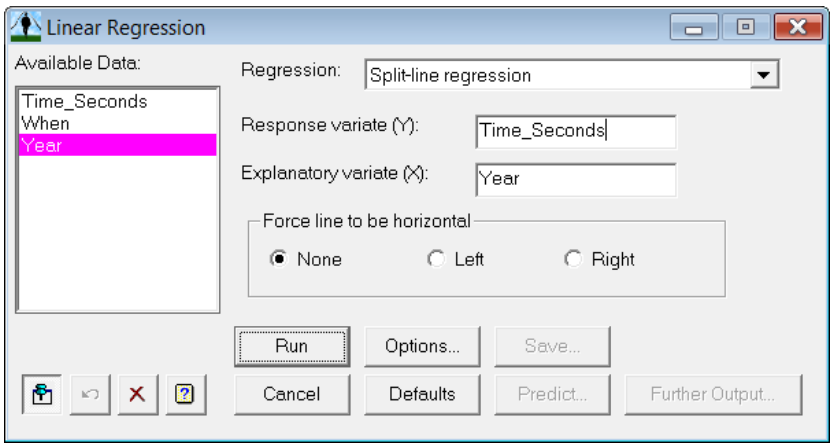


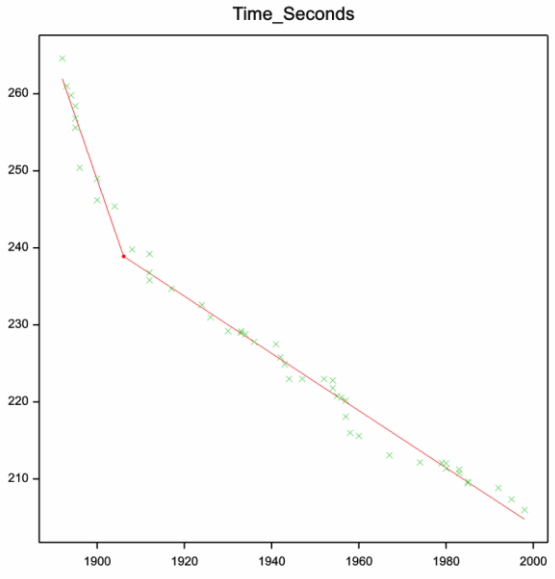
A quadratic model can be fitted to try and take into account the curvature in the data. Dropping down the Regression type to polynomial regression allows us to fit that model:





The quadratic is a better fit but the first few points still depart from the curve.





The output window contains the spli-line model fit.

Fit of two-straight-line model

Nonlinear regression analysis

Response variate: Time_Seconds
Nonlinear parameters: Breakpoint_X
Model calculations: Twolines[1], Twolines[2], Twolines[3]
Fitted terms: Breakpoint_Y, Slope_1, Slope_2

Summary of analysis

Source	d.f.	s.s.	m.s.	v.r.
Regression	4	2664709.9	666177.487	250453.17
Residual	47	125.0	2.660	
Total	51	2664835.0	52251.666	

Percentage variance accounted for 99.0
Standard error of observations is estimated to be 1.63.

Message: the following units have large standardized residuals.

Unit	Response	Residual
7	250.40	-3.05

Estimates of parameters

Parameter	estimate	s.e.
Breakpoint_X	1906.07	1.26
* Linear		
Breakpoint_Y	238.888	0.835
Slope_1	-1.637	0.145
Slope_2	-0.3713	0.0104

X value at intersection of lines

X value 1906.069, approximate s.e. 1.258
95% confidence interval (1902, 1910)

A two-part model with separate lines on either side of a break point can be fitted with the Split-Line Regression menu. This gives a much better fit to the data, and the break point is estimated as 1906, very close to the date of 1912 when the record keeping was taken over by the IAAF.