
6.5. Model I: Simple linear regression

In this section we consider the calculations in two BOXES. The first (BOX 6.3.) describes how to calculate a regression line which allows you to model the association or make predictions. In BOX 6.4. the significance of this association is tested. Both calculations are illustrated in relation to Example 6.2.

EXAMPLE 6.2. Heavy metal contamination of soil under electricity pylons

BOX 6.3. How to carry out a Model I: Simple linear regression: drawing a regression line

BOX 6.4. How to carry out a Model I: Simple linear regression: testing the significance of the association

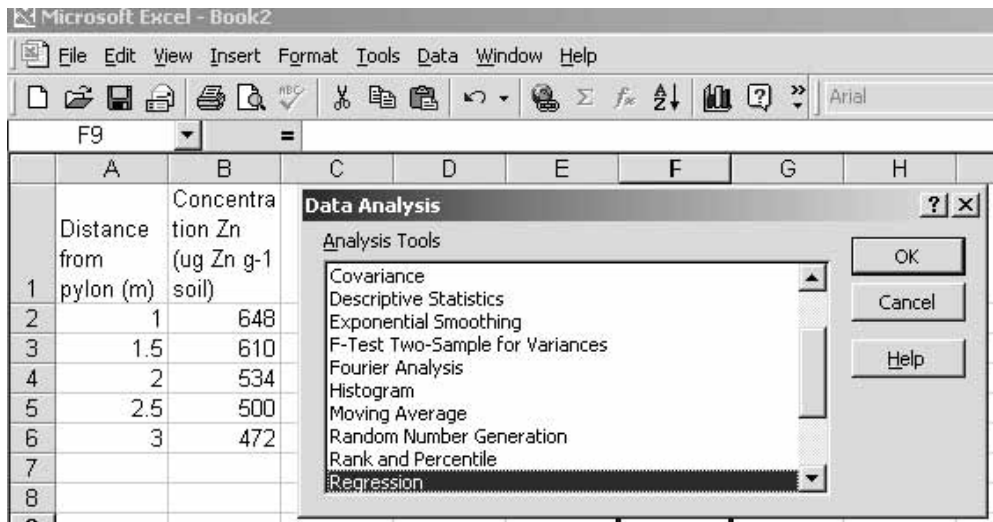
For Excel these two aspects will be tackled together as Excel completes all the necessary calculations at once.

In linear regression, you are seeking to describe the association between two data sets in terms of the equation for a straight line: $y = a + bx$.

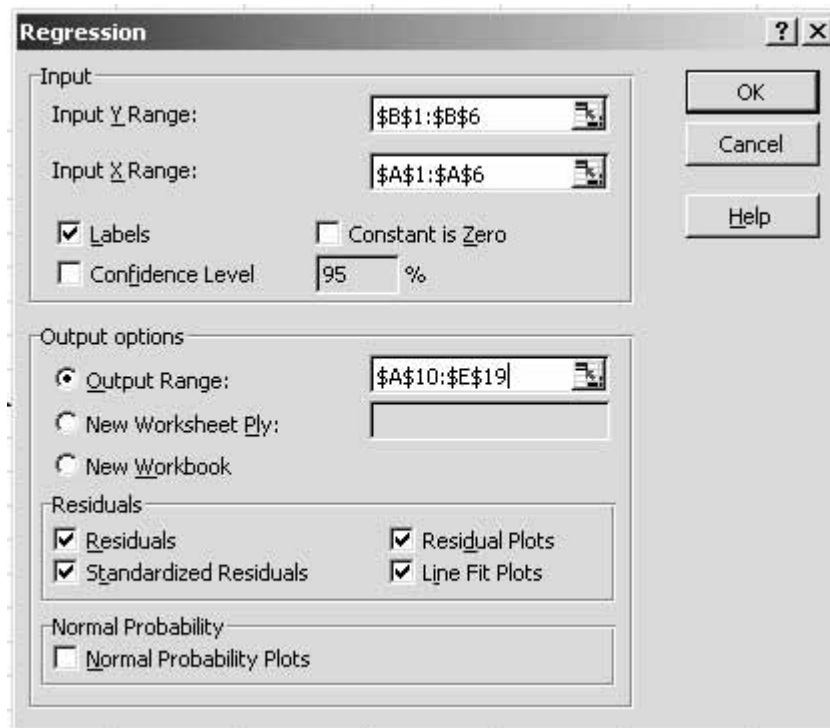
The Data Analysis Tool Kit has a comprehensive approach to linear regression and returns much useful, statistical information including plotting the data as a scatter plot, so you do not have to do this independently.

Step 1. The data are entered onto the spreadsheet using appropriate row and column headings.

Step 2. Select 'Tools' from the tool bar, then 'Data Analysis'.



From the dialogue box, select Regression and click OK.



Step 3. In the dialogue box which opens enter the input information.

You must determine which set of data is to be plotted on the x -axis (the independent variable) and which on the y -axis (the dependent variable). In this example, 'distance from the pylon' is the independent variable and so the cell locations for these data, including the data label, must be entered in the space labelled 'Input X Range'. In the Input box bring the cursor to the box marked 'Input X Range' by clicking in it. Then click on the first cell for that data set, A1, and holding down the mouse button drag down the data set. Next click in the box, 'Input Y Range', then click on the first cell for that data set, B1, and holding down the mouse button drag down the data set.

In the dialogue box there are a number of other options that can be selected. Ensure that you check the box for labels so that Excel reads the first cell from each data set as a label and not as data.

There is an option to select 'Constant is Zero'. This will assign the value of zero to the constant value a in the equation $y = a + bx$ and force the linear regression line through the origin (point 0,0). For some data sets this property is very useful, but it is not appropriate in this example, as we anticipate a negative association between these data sets.

In the 'Input' box you can also select a confidence level and this defaults to 95%, so normally would not need to be changed.

Step 4. There are a number of selectable output options. First select where your results will be returned. If you select 'Output Range', remember to click on the button and then also click in the adjacent box to bring the cursor to that location.

In the 'Residuals' box you can select all items by clicking in each box: Click on 'Residuals' to include residual values in the residuals output table, Click on 'Standardised Residuals' to include standardized residual values in the residuals output table. Click on 'Residual Plots' for a chart (automatically generated) showing each independent variable versus the residual value.

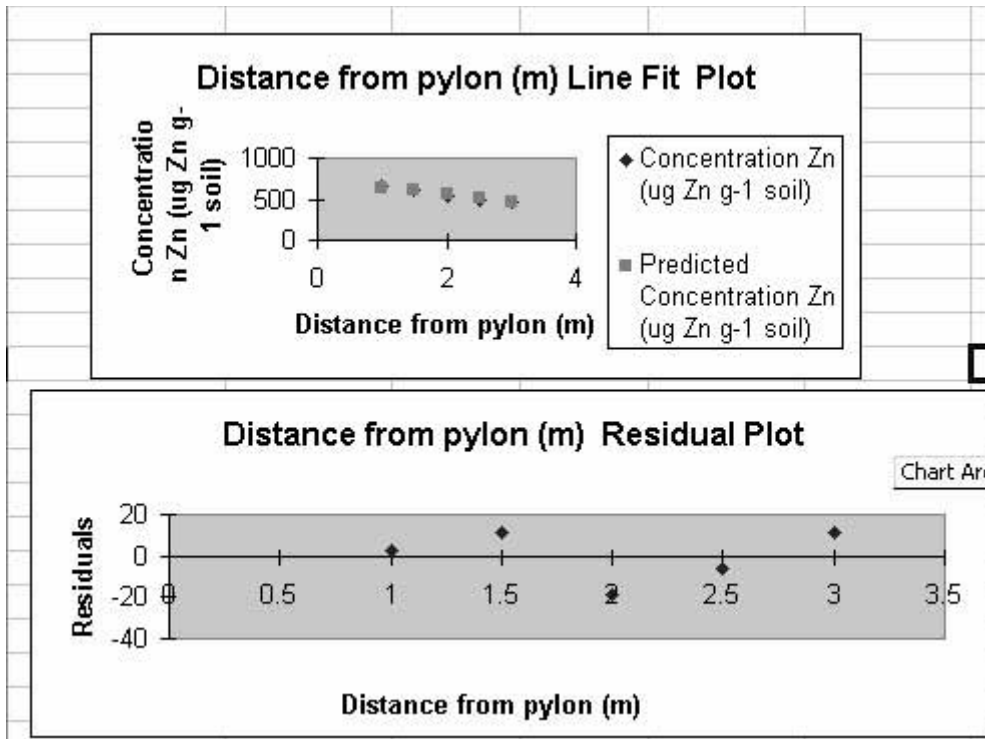
Click on 'Line Fit Plots' for a chart (automatically generated) showing predicted values versus the observed values. Having selected all the desired options, click on 'OK'.

10	SUMMARY OUTPUT									
11										
12	<i>Regression Statistics</i>									
13	Multiple R	0.984879874								
14	R Square	0.969988366								
15	Adjusted R Square	0.959984488								
16	Standard Error	14.83689096								
17	Observations	5								
18										
19	ANOVA									
20		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>				
21	Regression	1	21344.4	21344.4	96.9612356	0.002226791				
22	Residual	3	660.4	220.1333						
23	Total	4	22004.8							
24										
25		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>	
26	Intercept	737.6	19.90578	37.05457	4.3232E-05	674.2508707	800.9491	674.2509	800.9491	
27	Distance from pylc	-92.4	9.383674	-9.84689	0.00222679	-122.263066	-62.5369	-122.263	-62.5369	
28										
29										
30										
31	RESIDUAL OUTPUT									
32										
33	<i>Observation</i>	<i>Concentration Zn</i>	<i>Residuals</i>	<i>Standard Residuals</i>						
34	1	645.2	2.8	0.217914						
35	2	599	11	0.856089						
36	3	552.8	-18.8	-1.46313						
37	4	506.6	-6.6	-0.51365						
38	5	460.4	11.6	0.902785						
39										

Step 5. Interpreting the output data.

The first box gives a summary output returns values for R (the product moment correlation coefficient) and R^2 (the coefficient of determination). Adjusted R^2 adjusts the value in relation to the number of observations.

An ANOVA table is given. The ANOVA tests the significance of the regression line. A value for F is given (96.96) and its significance ($p = 0.002$), therefore the regression line is statistically significant.



Step 6. In the following table, ‘Intercept’ gives the value for ‘*a*’ in the equation: $y = a + bx$. $a = 737.5$

‘Distance from pylon’ gives the value of ‘*b*’ in the equation and is the gradient of the regression line (i.e. gradient of the slope). Notice that the value is negative since as one variable increases the other decreases. $b = -92.4$

For each coefficient, values are assigned for the standard error. *t* statistics are also generated, which test the significance of the coefficients. The probability for these statistics are given and note that both are significant ($p < 0.005$ for the gradient (actual value $p = 0.002$) and $p < 0.00005$ for the intercept actual value, $p = 4.32 \times 10^{-5}$).

Step 7. Residual values are given: these indicate the distance of the actual data points from the regression line. These values can be plotted above and below a horizontal line representing the regression line in the residual plot chart, which facilitates a visualization of the spread of values round the line.

The charts are returned ‘stacked’. It is necessary to drag them apart and edit the charts as required – particularly expanding the plot area (see below for line fit plot).

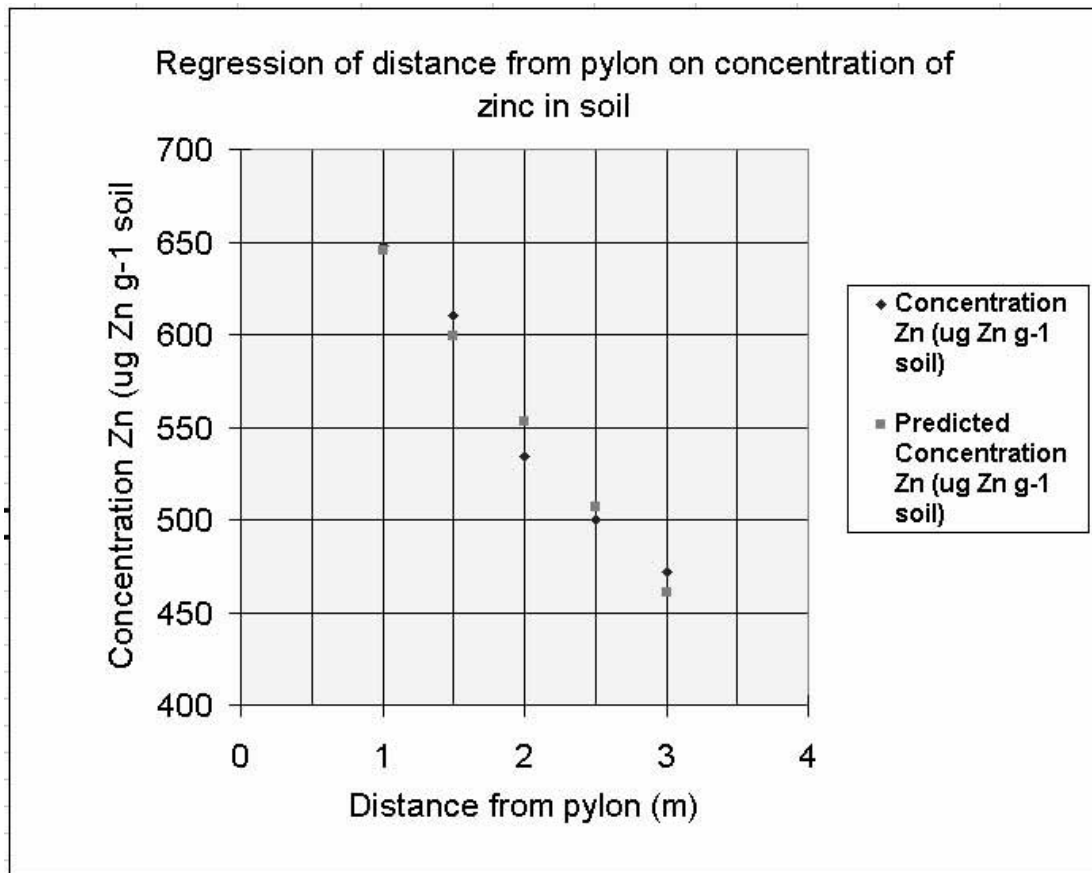
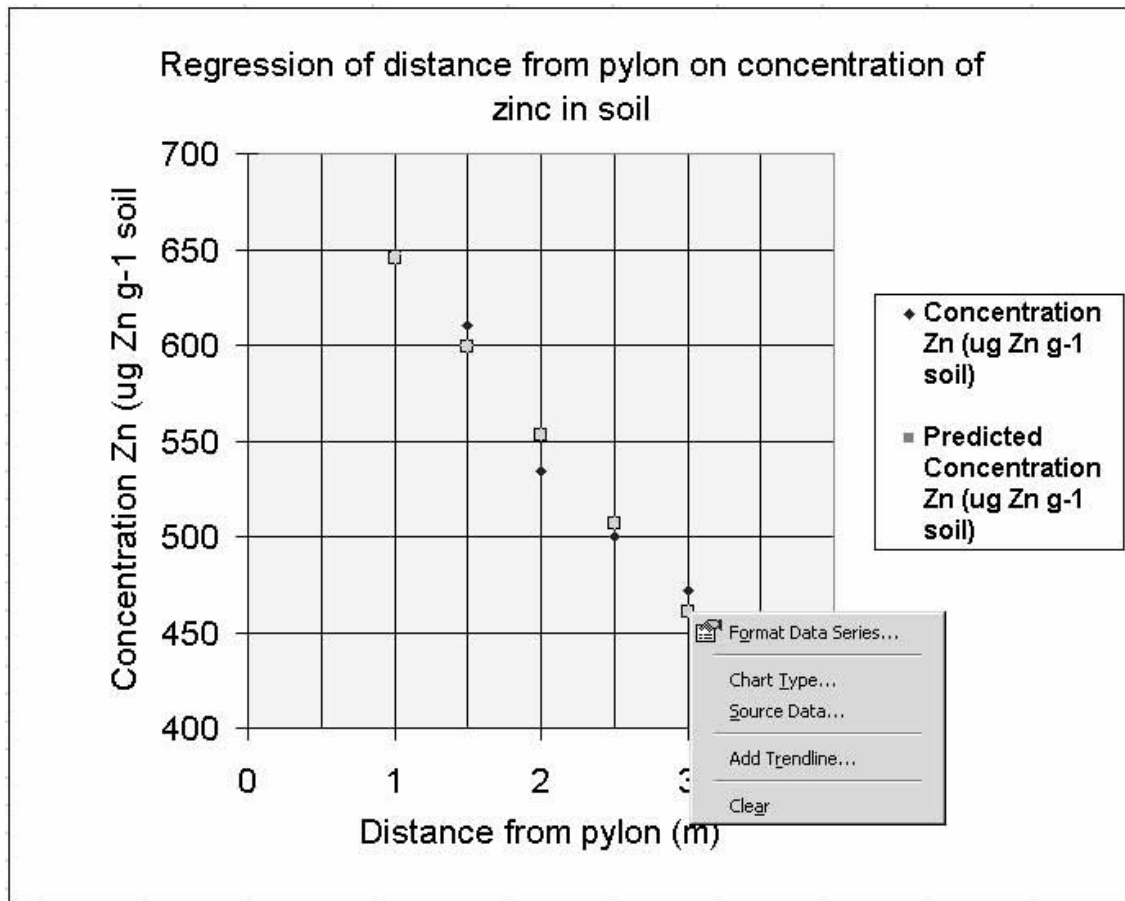
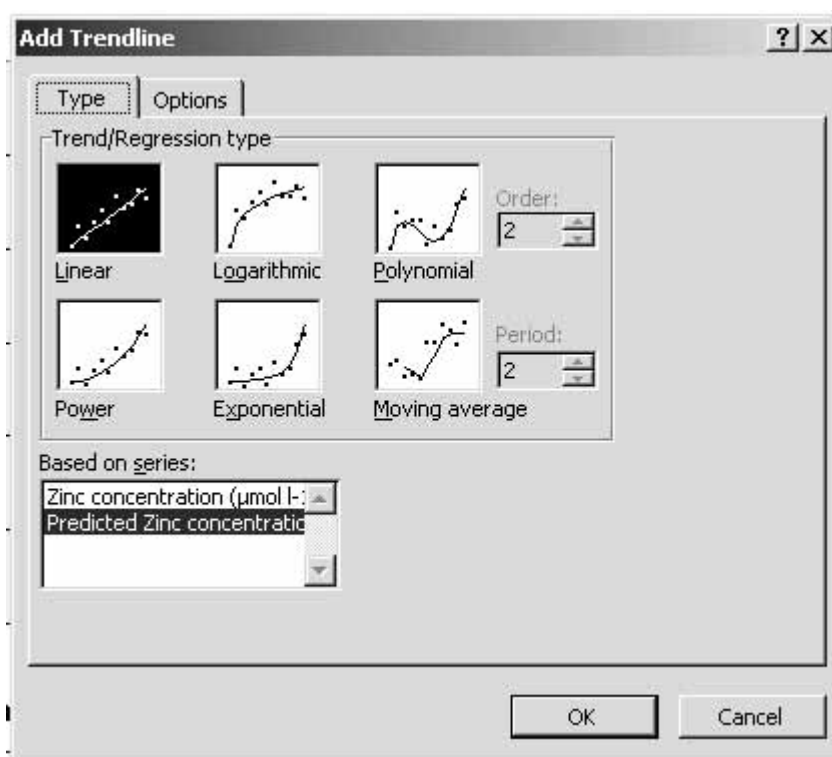


Fig. W6.5.1. Line fit plot of levels of zinc in soil and distance from a pylon

Fig. W6.5.1. shows the expanded plot area with reduced sized text. Note that the actual data are plotted as blue diamonds and the 'predicted' data points which are positioned for the linear regression line are given as pink squares. In order to add the regression line, move the cursor over one of the pink squares and click with the right mouse button. This will highlight all the pink squares (they now appear green) and a pop-up menu will appear.



From the menu, click on 'Add Trendline'. A new box will open.



On the 'Type' tab you can select the type of line required and for regression analysis select 'Linear' by clicking on the box. Then select the 'Options' tab.

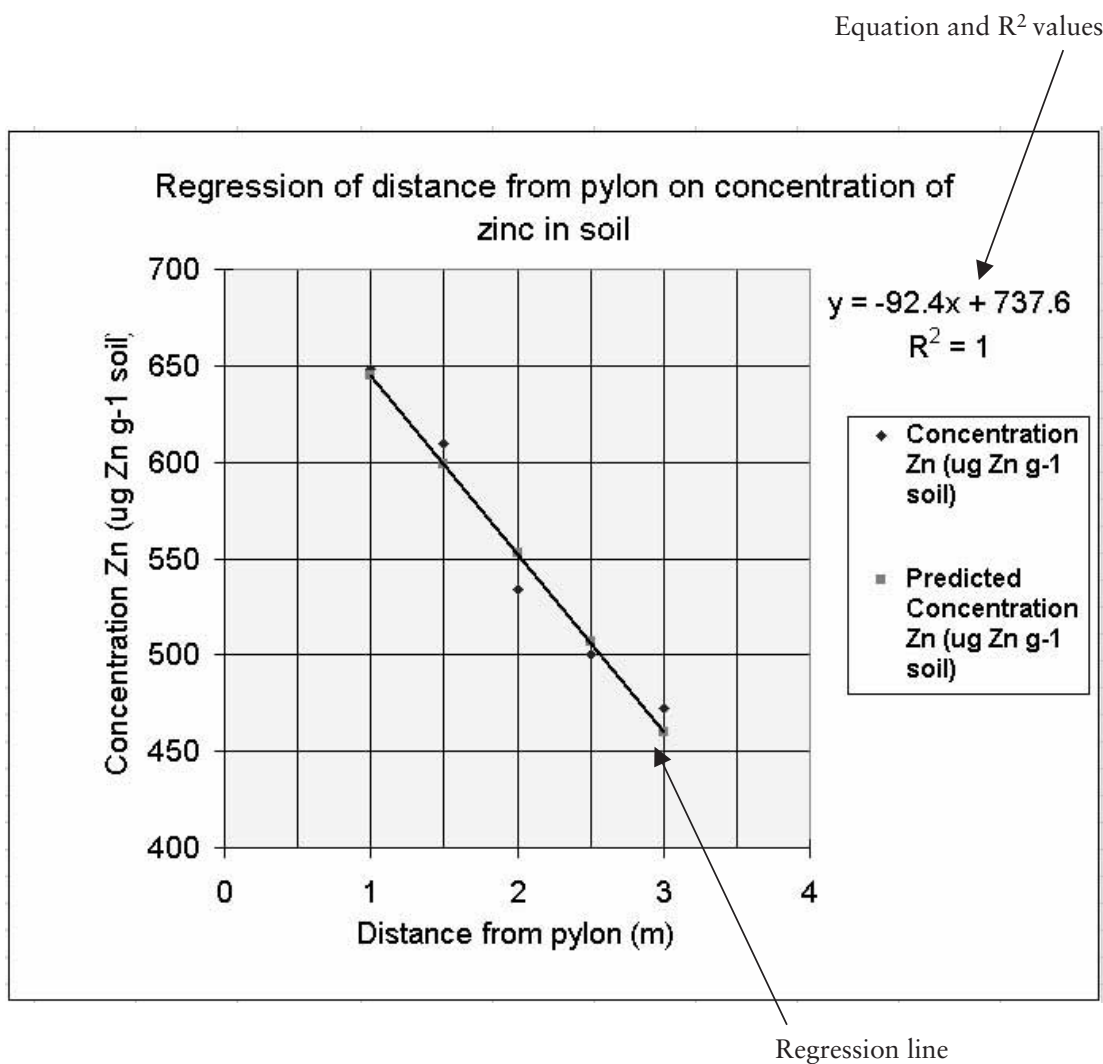
The screenshot shows the 'Add Trendline' dialog box with the following settings:

- Tab:** Options
- Trendline name:** Automatic: Linear (Predicted Zinc concentration (µmol l-1 Zn g-1 soil))
- Forecast:** Forward: 0 Units; Backward: 0 Units
- Set intercept =:** 0
- Display equation on chart:**
- Display R-squared value on chart:**

Here you have a number of options that can be selected. You can add a trendline name by clicking on 'Custom' and entering text in the space provided.

The 'Forecast' box enables the fitted regression line to be extended either side of the regression line. This is a useful facility in some cases, but note that on most occasions with biological data that this must not be done as it would be erroneous. The next option is to set the intercept to zero or another chosen value. This is not required in this example.

Click in the next box, which will allow the linear regression equation to be displayed on the chart. Also click in the final box, which allows the R^2 value to be displayed on the chart (this is the coefficient of determination which has a value of the square of the correlation coefficient). Finally click on 'OK' to obtain the edited chart.



The residuals plot (Fig. W6.5.2.) can also be expanded in the same way as was the line fit plot.

Here the relationship of the actual data points to the regression line is shown. The regression line is shown horizontally and the data points lie above and below this line where the values are greater or less than the regression line respectively. The greater the deviation of the points from the regression line, the smaller would be the value for R^2 .

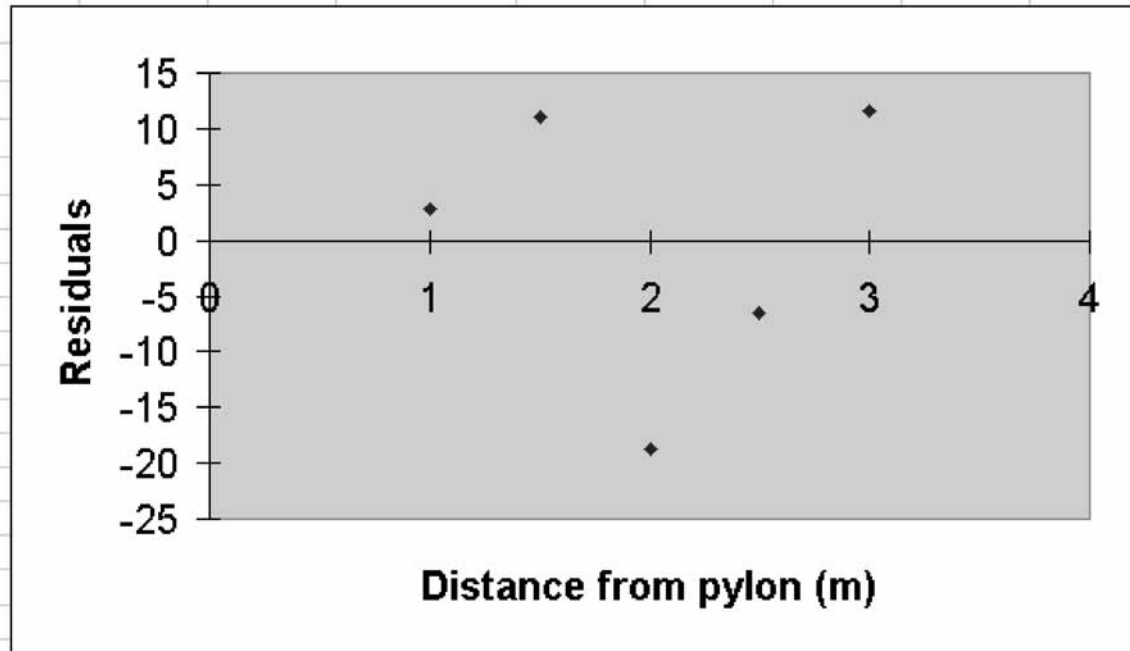


Fig. 6.5.2. Residual plot of the association between zinc concentration in the soil and distance from the pylon (m)