Residuals

Relating to page 27 of *Level 3 Bivariate Data Learning Workbook*

Residuals

In a scattergraph of bivariate data, each point is a certain vertical distance from the trend line. The difference between the actual *recorded value* of the response variable, and the *value predicted by the trend line* is called a **residual**.

Residual = observed value – trend (predicted) value

If a point lies above the trend line, the value of its residual is positive; if a point lies below the trend line, the value of its residual is negative.

Calculating and graphing residuals by spreadsheet

The following explanation uses the spreadsheet for Oregon settlement population versus annual waste (1998), with figures rounded to the nearest whole number.
(The data for this file is available on the ESA website RESOURCES www.esa.co.nz/pages/esa-online.)

Finding trend (predicted) values on Excel

In cell C1 type: Trend values

In cell C2 type: tons

In cell C3 type: =0.8784*A3 (or click on the cell A3) – 4763.6

Press **Enter** (the number 9 906 should appear). This is the annual number of tons of waste predicted for a settlement of 16 700 people, found by using the regression equation).

Click on cell C3 and click on the small black square at the bottom right-hand corner (a cross should appear).

Left-click on the cross, and holding the mouse key down, drag the cross down to cell C37. When you release the mouse key, cells C3–C37 should be filled with predictions from the regression equation.

Calculating residuals on Excel

In cell D1 type Residuals

In cell D2 type tons

In cell D3 type =B3-C3

Press **Enter** (the number 2470 should appear)

Click on cell D3, hold and drag down (as done for column C) to D37

Cells D3–D37 should be filled with the values of the residuals.

Graphing the residuals

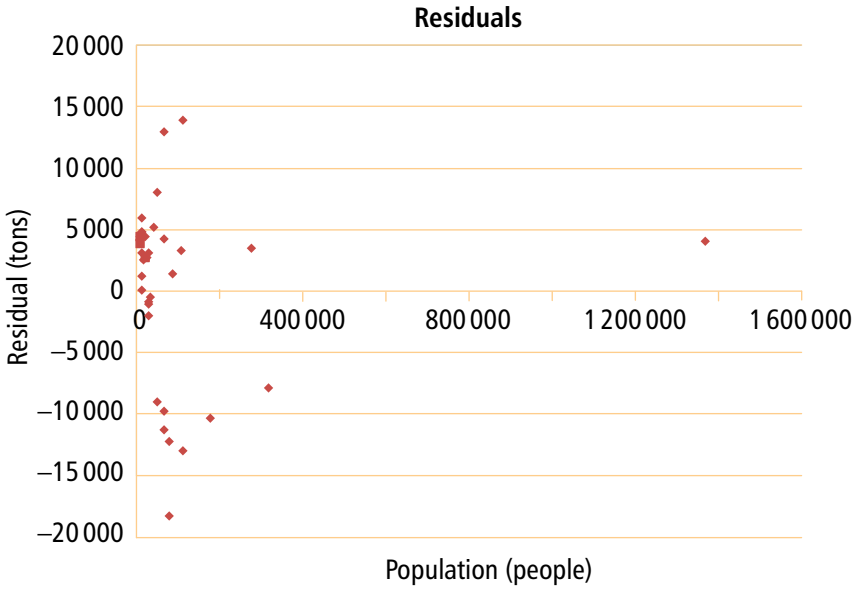
Click on the **A** at the top of column A (column A should now be highlighted)

Press **Ctrl** on the keyboard and click on the **D** at the top of column D

Press **Insert** then **Scatter** selecting the scatter plot with unconnected points (top left).

As before, insert a title for the graph, and label the axes.

Your graph should look like this:



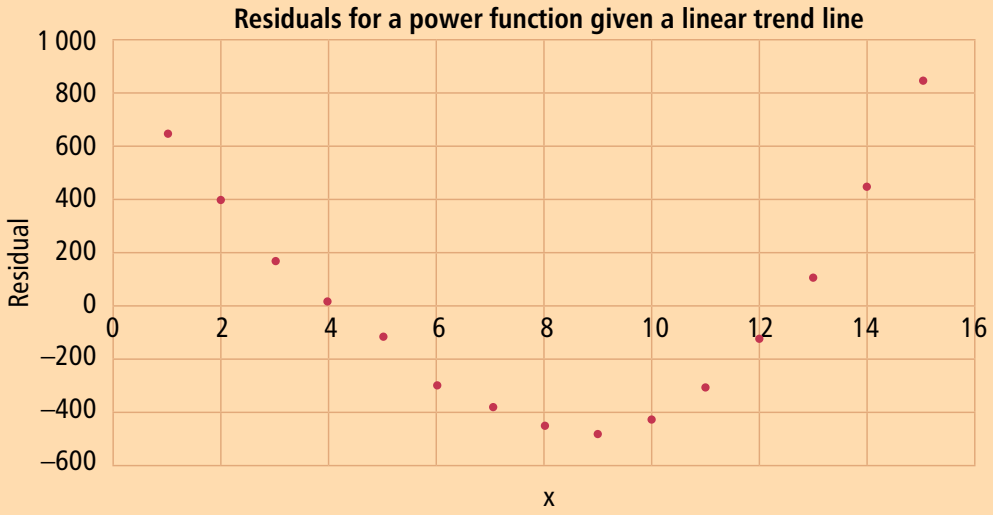
Interpreting the residuals

The graph of the residuals can be used to decide whether a linear trend line is an appropriate model for the relationship between the variables.

- If the residuals appear *randomly scattered* above and below the *x*-axis, then a linear relationship between the variables you are investigating would be appropriate.
- If the residuals form a *pattern* when graphed, then a linear relationship between the variables you are investigating would not be appropriate.

Example

In the graph of residuals shown below, the residuals form a curved pattern.



This means that a linear trend line is not appropriate for the data.

Exercise: Residuals

Graph the residuals for the following data sets (available online at ESA RESOURCES www.esa.co.nz/pages/esa-online). The equations of the trend lines (found in a previous exercise) are included alongside each question. Interpret the residuals.

1. Columbia River ($V = -0.116d + 1.68$)

2. Wrist/neck measurement ($n = 1.3517w + 10.623$)

3. Rainbow Glacier retreat ($p = -16.75t + 37.51$)

4. Sacramento Valley drainage ($R = 0.618r - 9.791$)

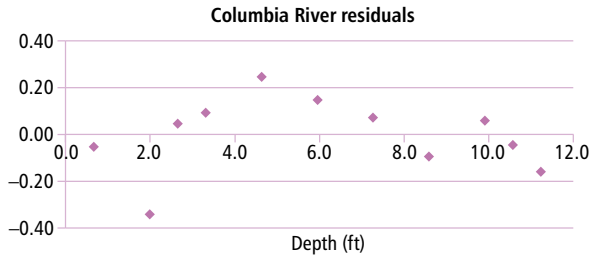
5. Zinc v lead concentrations ($z = 28.72l + 19.54$)

6. California power plants ($C = 0.0584h + 50\ 306$)

Answers

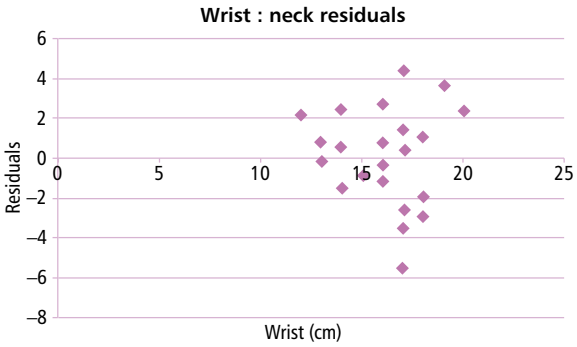
Residuals

1.



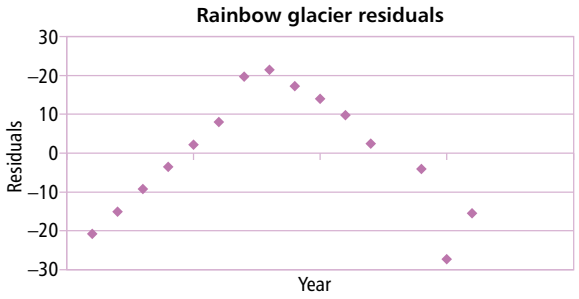
One feature of the graph is that the residuals are all positive between depths of 2.6 m and 7.3 m, so the linear model is tending to underestimate velocity through this range of depths. Otherwise the linear model is appropriate for the relationship between the depth and velocity of the Columbia River.

2.



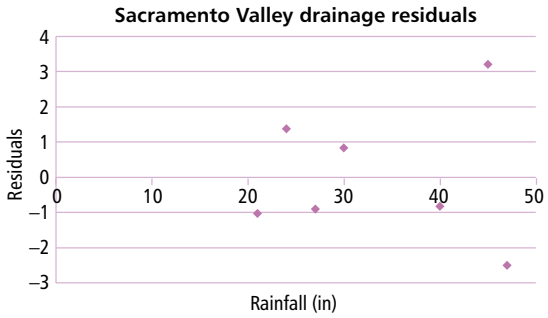
As there is no apparent pattern in the residuals this is further evidence that a linear model is appropriate for the relationship between wrist and neck circumferences.

3.



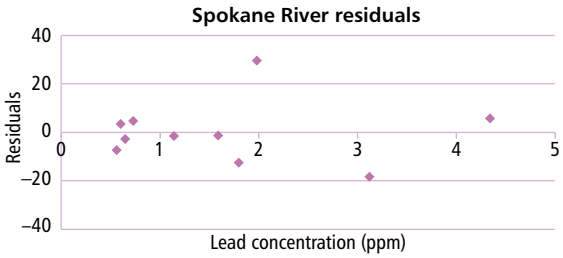
There is a pattern in the residuals, so this is evidence that a linear model is not appropriate for the relationship between the years since 1958 and the retreat of the Rainbow Glacier.

4.



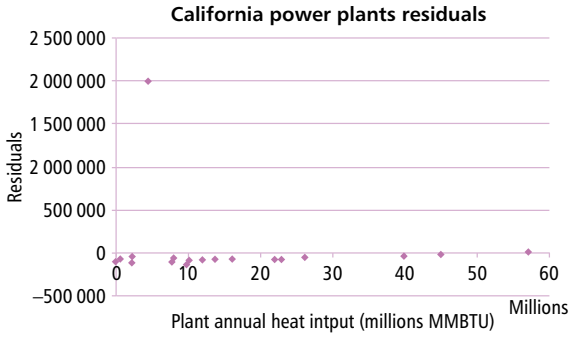
There is no apparent pattern in the residuals, so this is further evidence that a linear model is appropriate for the relationship between the runoff and rainfall in the Sacramento Valley.

5.



As there is no apparent pattern in the residuals this is further evidence that a linear model is appropriate for the relationship between lead and zinc concentrations in fish in the Spokane River.

6.



The large number of points on the horizontal axis show that the linear model is appropriate as the points lie on or very close to the trend line. One outlier is evident when the plant annual heat input is 4 487 554 MMBTU. (This outlier has raised the trend line so that it consistently overestimated annual emissions, so most residuals are negative.)