NUM-NUM Analysis

Student Worksheet

T³
Teachers Teaching with Technology™
Professional Development from Texas Instruments

TI-Nspire CAS   Investigation   Student   90 min

NUM-NUM analysis

If we have data from two numeric variables (i.e. 'NUM–NUM'), we may be interested in the statistical relationship between them – it may be useful as a way of predicting values of the **response variable** (y), from the values of the **explanatory variable** (x).

From your previous study of bivariate data analysis, you may recall the following:

- Produce a scatter plot, and observe the plot, and look for evidence of an association between the variables.
- If the scatter plot suggests an association, we may look at Pearson's correlation coefficient (r) or the coefficient of determination (r^2) to confirm our observation.
- If these values are strong enough, we may try to fit a line of the form $y = a + bx$, and perhaps use this to help predict other values of y from given x values.

In Year 12 Further Mathematics, we look at some new ways of evaluating whether a linear model is appropriate.

Residual analysis: should we reject a linear model?

A linear model of the form $y = a + bx$ might be used to describe the relationship between two numerical statistical variables.

For instance, in the plot shown right, the least-squares regression line appears to fit the data well, and the r^2 value is close to 1.

However, look carefully at the scatter plot – the data values are above the line at the left and right ends of the plot, but below the line in the middle of the plot. This gives the impression that there might be some curvature in the scatter plot and hence a linear model may not be appropriate. Hence predictions based on such a model may be unreliable.
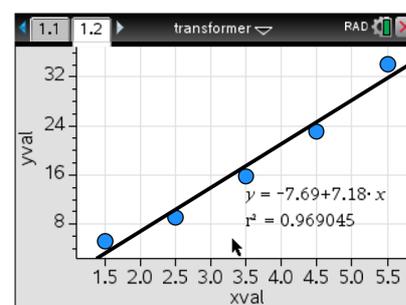
One way to check the suitability of a model is, for each x-value, to calculate the difference between the y-value $\left(y_{data}\right)$, and the y-value predicted from the model $\left(y_{predicted}\right)$.

This difference is called the **residual value,** and is calculated as:

$$residual\ value = y_{data} - y_{predicted}$$

Author: D. Tynan

TEXAS INSTRUMENTS

A **residual plot** is a plot of these residual values (against each *x* value). For example the residual plot (underneath the scatter plot) highlights whether the regression rule *overpredicts* or *underpredicts* for each *x*-value, and by how much. For example, it shows that the data point (4.5, 23.1) on the scatter plot has a residual value of −1.520, which means that the predicted *y*-value is 1.520 units higher than the actual data *y*-value (i.e. the model *overpredicts* that value of *y*).
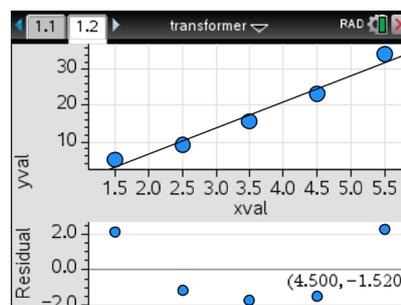
Further, if a linear model was appropriate, the residual values would be randomly positioned around zero. So the curvature in the residual plot suggests a linear model does not explain the relationship well.

Choosing a non-linear model

If it is clear from a residual plot that a linear model (i.e. $y = a + bx$) is not suitable, it is possible that a non-linear model (for example $y = a + bx^2$) may be suitable.

To help decide, we consider whether a change (**transformation**) to either the explanatory variable (*x*), or the response variable (*y*) results in a better model.

The possible transformations we consider in Year 12 Further Mathematics are:

Transformations of the *x*-variable	Transformations of the *y*-variable
$y = a + bx^2$	$y^2 = a + bx$
$y = a + b\log_{10}(x)$	$\log_{10}(y) = a + bx$
$y = a + \dfrac{b}{x}$	$\dfrac{1}{y} = a + bx$

In evaluating the suitability of a particular transformation, we need to consider what happens as a result of the transformation:

- does the transformation 'straighten' the scatter plot (make it more 'line–like')?
- do the new residual values appear to be relatively small and randomly scattered around zero
- Is the coefficient of determination improved (i.e. is r^2 closer to 1?)

The circle of transformations

The diagram at right is very helpful in narrowing the choice of suitable non-linear models. Check the curvature of the original plot. For example, the scatter plot considered earlier suggests one of the following three transformations may be suitable.

- A 'log-*y*' transformation (model : $\log_{10}(y) = a + bx$)

- A '1/*y*' transformation (model : $\dfrac{1}{y} = a + bx$)

- An 'x^2' transformation (model : $y = a + bx^2$)
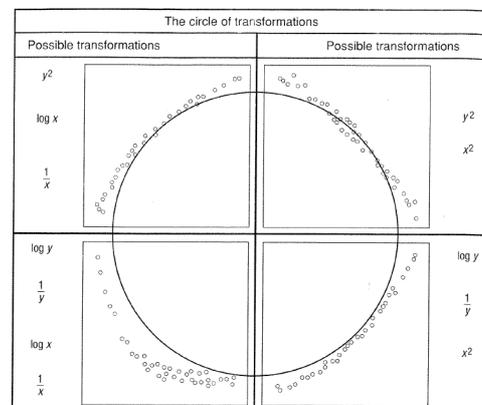
OK so what next? We will build a TI-Nspire template file called 'transformer', which makes it much easier to decide which model (linear or non-linear model) is appropriate.

Author: D. Tynan

TEXAS INSTRUMENTS

Building and testing a 'Transformer' template file

Step 1. Create & saving the template file

- Press (⌂) and then (1) to create a new document.
- Press **4: Add Lists & Spreadsheet**
- Press (ctrl)(S) to save the document as "transformer"
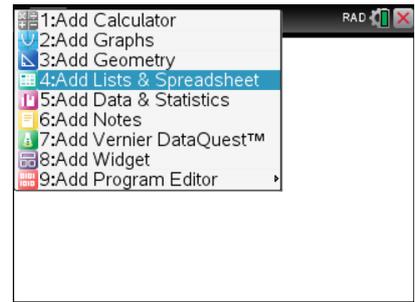
Note: The TI-Nspire should be set into "APPROX" calculation mode.
Press (⌂) **> Settings > Document Settings** to check.

Step 2. Construct the spreadsheet

In the top row, enter the following statistical variable names

	A xval	B yval	C xsqu	D xlog	E xrec	F ysqu	G ylog	H yrec	I

In the second row, type in the following formulas for each of the transformed variables (as shown).

For **xsqu**, type "=xval^2"   For **xlog**, type "=log(**xval**,10)"   For **xrec**, type "=1/**xval**"
For **ysqu**, type "=yval^2"   For **ylog**, type "=log(**yval**,10)"   For **yrec**, type "=1/**yval**"

	A xval	B yval	C xsqu	D xlog	E xrec	F ysqu	G ylog	H yrec	I
=			=xval^2	=log(xval,10)	=1/xval	=yval^2	=log(yval,10)	=1/yval	

Step 3. Enter new data

Enter the following sample data into **xval** and **yval**. It should appear similar to the screen shown below.

| xval	yval
1.5	5.2
2.5	9.1
3.5	15.7
4.5	23.1
5.5	34.1

*[Note for future use: If there is data already in these columns, press ▲ until an entire column has been selected, then press (menu) **Data > Clear Data** to clear the data from this variable.]*

Step 4. Construct the scatter and residual plots

**a.** Press (ctrl)(doc▾) then select **Add Data & Statistics page** (as per screen shown).
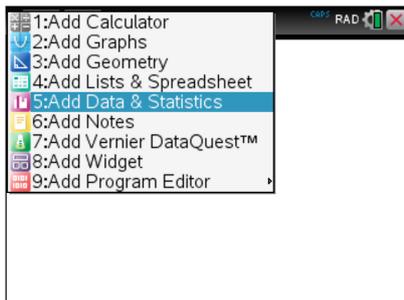
b. Press (tab), then select **xval** as *explanatory* variable
Press (tab), then select **yval** as *response* variable

Author: D. Tynan

TEXAS INSTRUMENTS
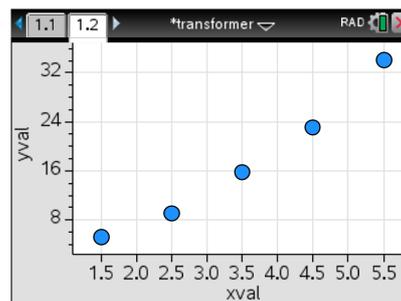
c. Press (menu) **Analyse > Regression > Show Linear (a+bx)**
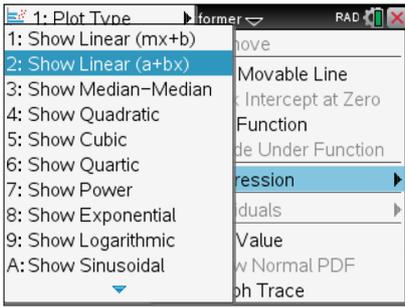
d. This will be the resulting line with equation shown

$y = -7.69 + 7.18 \cdot x$
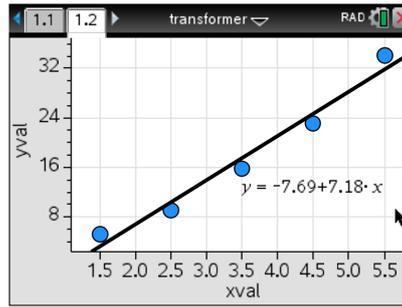
e. Press (menu) **Analyse > Residuals > Show Residual Plot**
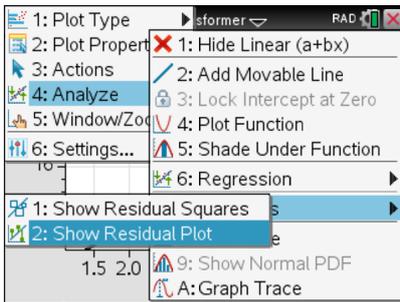
f. This places the residual plot under the scatter plot.

$y = -7.69 + 7.18 \cdot x$
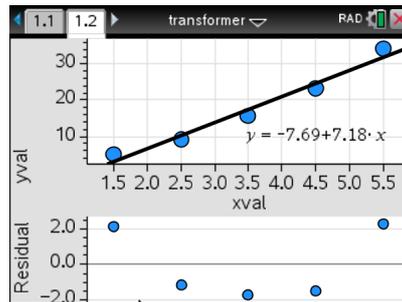
Step 5. Transform variables and check the coefficient of determination

g. To perform a transformation, click on either the **xval** or **yval** and change it.
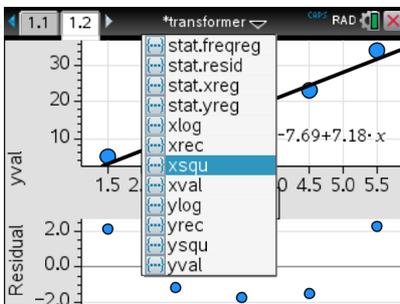
h. The plots are updated.
[Note: To get a better view of the residuals, press (menu) *> **Window/Zoom > Zoom–Data**]*

$y = 2.80548 + 1.02698 \cdot x$
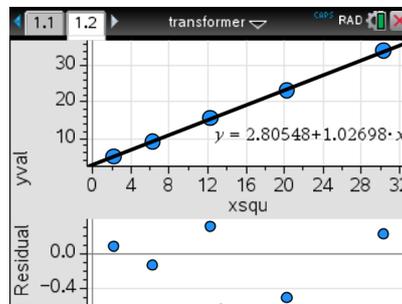
i. Press (menu) **> Settings** and click 'Diagnostics' so that the r^2 value diagnostic is displayed.
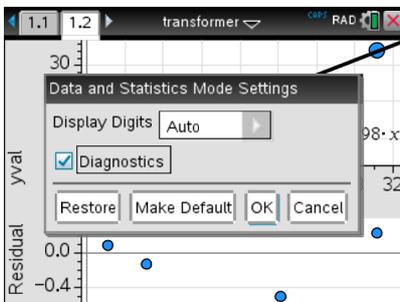
j. Note that the r^2 value is now displayed under the regression equation.

$y = 2.80548 + 1.02698 \cdot x$
$r^2 = 0.999201$
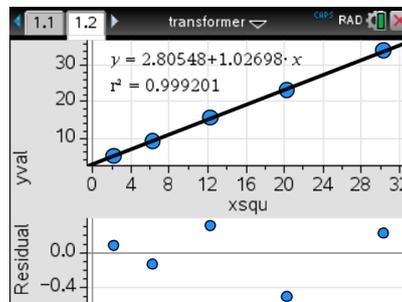
k. Save your template file again with, and it is ready to use for future analyses!

Author: D. Tynan

TEXAS INSTRUMENTS

What we just did …

As the '**xsqu**' (*x*-squared) transformation has straightened the scatter plot, and the residual values seem to be small and randomly scattered about zero, we can say that an '*x*-squared' model is reasonable.

This means that the equation $y = a + bx^2$ or more specifically $y = 2.805 + 1.027x^2$ (correct to 3 decimal places) is a better model to describe the relationship. The r^2 value (0.999) is also higher, which means that 99.9% of the variation in the response variable (*y*) can be explained by variation in the explanatory variable (*x*).

A sample analysis …

The number of hours spent doing homework each week (*Homework hours*) and the number of hours spent watching television each week (*TV hours*) were recorded for a group of 20 Year 12 students.

| TV hours | 6 | 28 | 14 | 6 | 9 | 30 | 10 | 3 | 3 | 18 | 7 | 9 | 4 | 25 | 8 | 24 | 21 | 5 | 15 | 10 |
| Homework hours | 17 | 16 | 9 | 21 | 17 | 8 | 15 | 48 | 37 | 9 | 21 | 18 | 24 | 14 | 14 | 10 | 14 | 32 | 6 | 13 |

Question 1.

Enter the data into the 'transformer' template, using *TV hours* as the explanatory variable, and view the scatter plot and residual plot. Comment on the suitability of a linear model for this relationship. *[Note: You may need to use* (menu) *> **Window > Zoom > Zoom–Data** to get a better view of the plots.]*
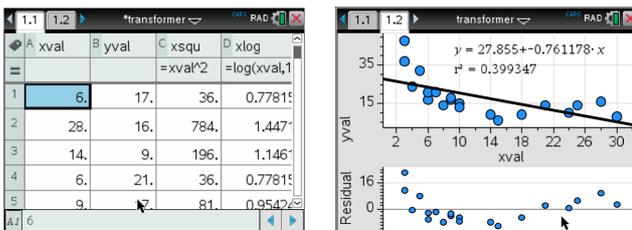
Question 2.

Using the *Circle of Transformations* diagram presented earlier, suggest four possible non-linear models that might explain the relationship. In each case, record the regression equation (to 3 decimal places), and the coefficient of determination (to 3 decimal places).
Record also whether the residuals appear to have any pattern.

Model	Plots	Equation	r^2	Residuals

Author: D. Tynan

TEXAS INSTRUMENTS

Model	Plots	Equation	r^2	Residuals

Question 3.   On the basis of these results, propose the 'best' of these four models, giving reasons.

Question 4.   If a student spends 10 hours per week watching television, use your chosen 'best' model to predict the number of hours that the student spends doing homework. Give your answer correct to the nearest hour.

TEXAS
INSTRUMENTS