Task 5: Estimating values of one measurement from another using a regression method

In Task 4 we have observed that there is strong linear correlation between the measurements. Therefore, in this task we utilize this linear correlation to estimate values of one measurement from values of another measurement, using a regression method.

Because the correlation observed in the North Island sample is not different from the correlation in the South Island sample, we put both samples together and treat them as a single sample for the regression method.

(In an advance regression method, we can also estimate the differences between the two samples if any. You will learn such advanced regression methods at university, typically in a second year-level Statistics course.)

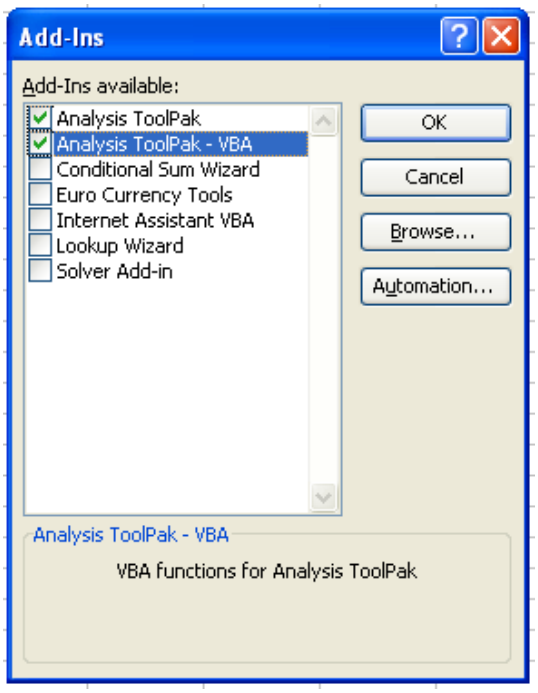
(1) Insert a new worksheet, and name it as "Regression".

(2) We first estimate values of X2 using the values of X1. Copy and paste the values of X1 and X2 from the "Dolphins data" worksheet to the "Regression" worksheet.

The regression method is available in Excel only through its add-in package **Data Analysis**. If this package has already been added to Excel, you can see **Data Analysis…** option in the drop list when clicking on the **Tools** on the menu. If not, load the package by following the instruction below. Otherwise, go to step (3).

Click on the **Tools** on the menu and select **Add-Ins…** option.

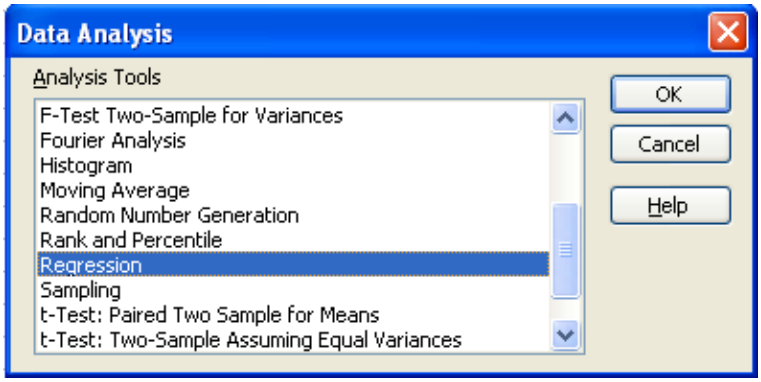
In the Add-Ins dialog box, select **Analysis ToolPak** and **Analysis ToolPak – VBA** as shown, and click **OK.**



Now the add-in package is loaded into Excel.

(3) Click on an empty cell on the "Regression" worksheet and then, click the **Tools** menu, and select **Data Analysis…** option.

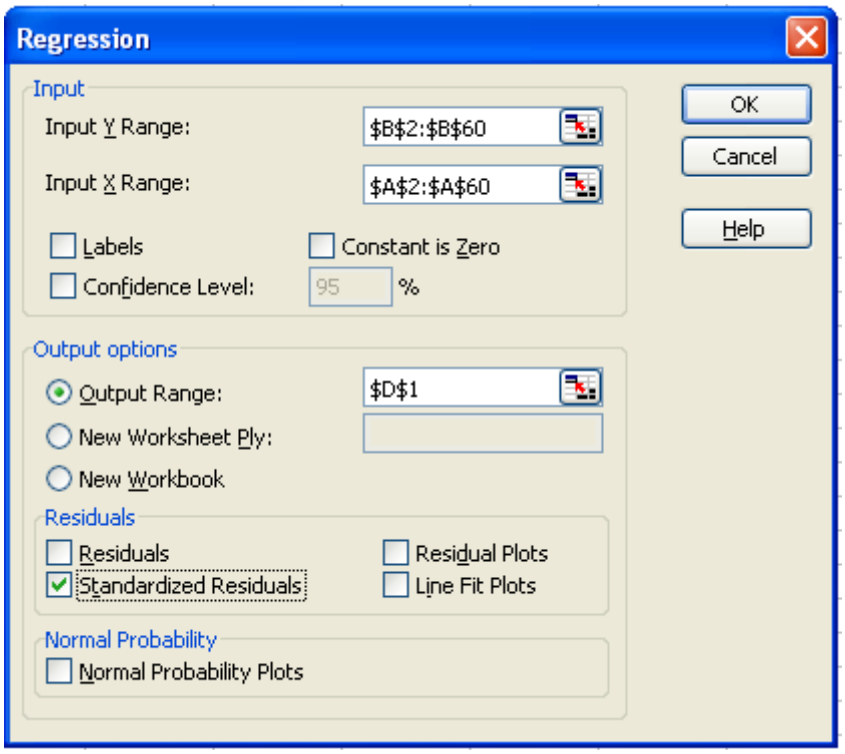
(4) In the **Data Analysis** window, select the **Regression** option in the **Analysis Tools** box. Click **OK.**



(5) In the following window, in the **Input** area, select all the 59 X2 values for the **Input Y Range** box, and the 59 X1 values for the **Input X Range** box.

In the **Output options** area, click on the **Output Range** option and then, select an empty cell on the "Regression" worksheet in the **Output Range** box.

Click also on the **Standardized Residuals** option in the **Residuals** area, as shown below. Then, click **OK**.



The following results now should appear.

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.820217225							
R Square	0.672756297							
Adjusted R Sq	0.667015179							
Standard Error	2.863372098							
Observations	59							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	960.7644078	960.7644078	117.182114	1.89884E-15			
Residual	57	467.3372871	8.198899774					
Total	58	1428.101695						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-31.07086195	7.748666784	-4.009833281	0.00017871	-46.58729942	-15.5544245	-46.58729942	-15.55442448
X Variable 1	0.292271085	0.026999467	10.82506877	1.8988E-15	0.238205585	0.346336584	0.238205585	0.346336584

In the *Regression Statistics* table, the R Square value of 0.673 indicates that 67.3% of the variation observed in the X2 values are explained by the X2's linear relationship with X1.

In the ANOVA table, the *Significance F* value of 1.899E-15 (= 1.899×10^{-15} = 0.000000000000001899 < 0.05) also indicates that X2 values change along with X1 values, and hence, indeed we are able to estimate values of X2 using the values of X1.

The *Coefficients* values in the third table are the estimated values of the regression equation, i.e. $X2 = -31.071 + 0.292 \times X1$.
This means that if the value of X1 is, say, 300, we can estimate the corresponding X2 value as $-31.071 + 0.292 \times 300 = 56.529 \cong 56.53$.

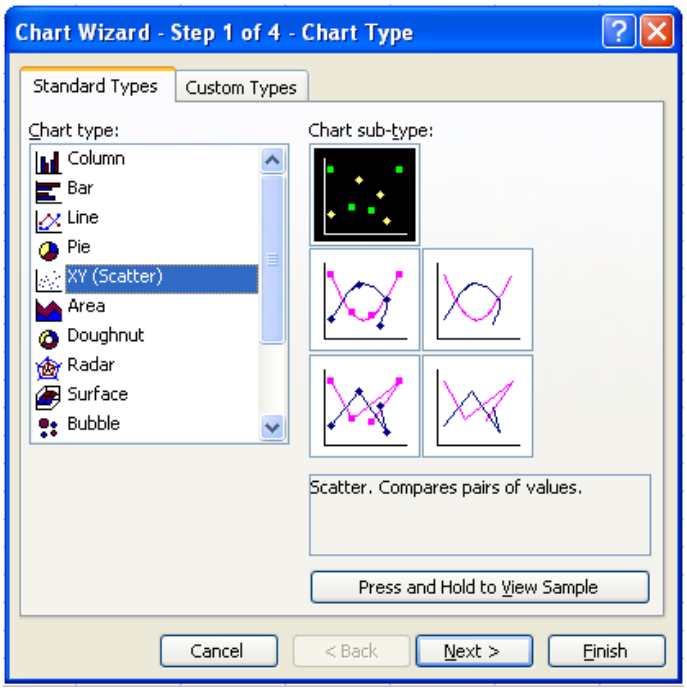
The RESIDUAL OUTPUT section of the results on the "Regression" worksheet shows the differences between the estimated X2 values from the above equation and the actually observed X2 values for the 59 dolphin skulls. It is always very important for us to check these differences, using a scatterplot. The following steps show how to draw the scatterplot.

(6) Copy and paste the X1 variable (variable name and the 59 values) into a new column on the "Regression" worksheet.

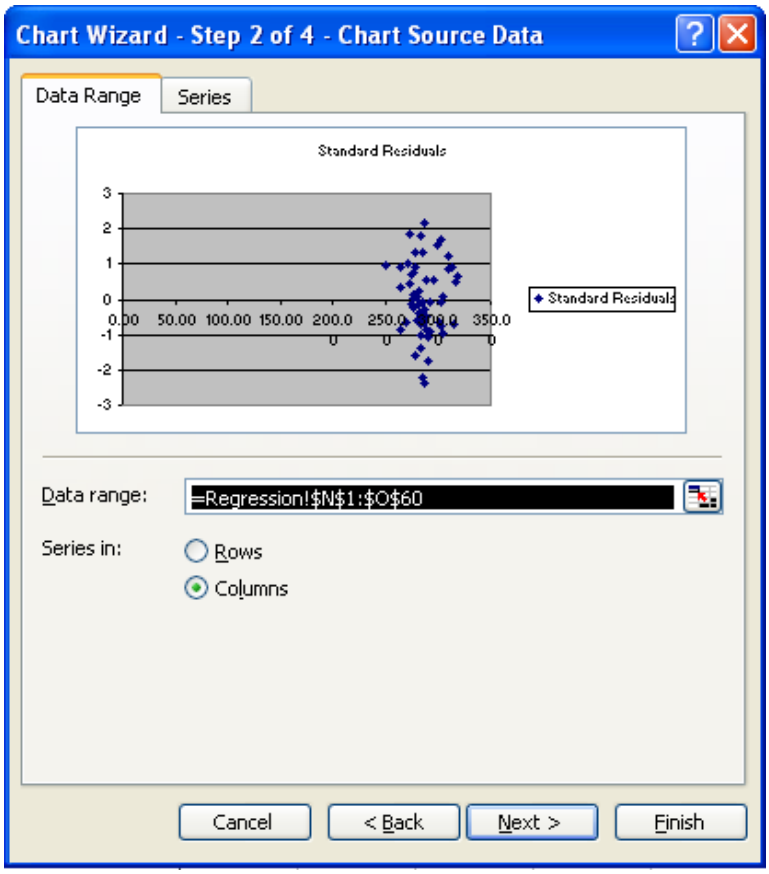
(7) In the next column, copy and paste the *Standard Residuals* (variable name and the 59 values) from the RESIDUAL OUTPUT section.

(8) Click on an empty cell on the "Regression" worksheet and then, click the **Chart Wizard** icon in the Tool bar.

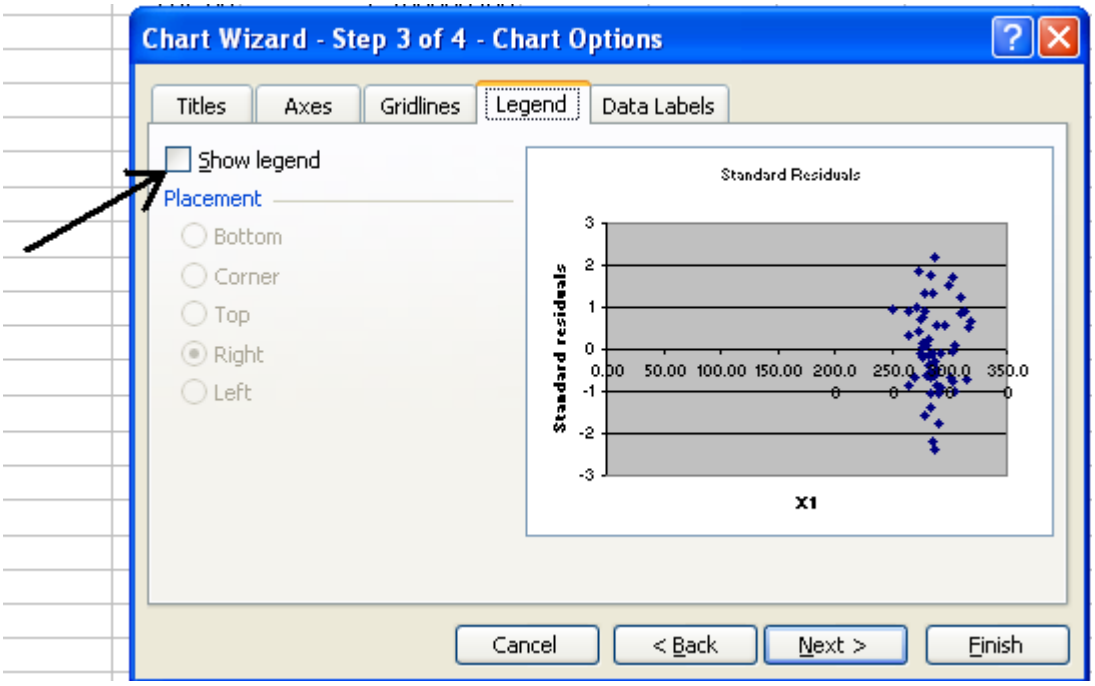
(9) Select the **XY (Scatter)** option in the **Standard Types**, then the first option in the **Chart sub-type**. Click **Next >**.



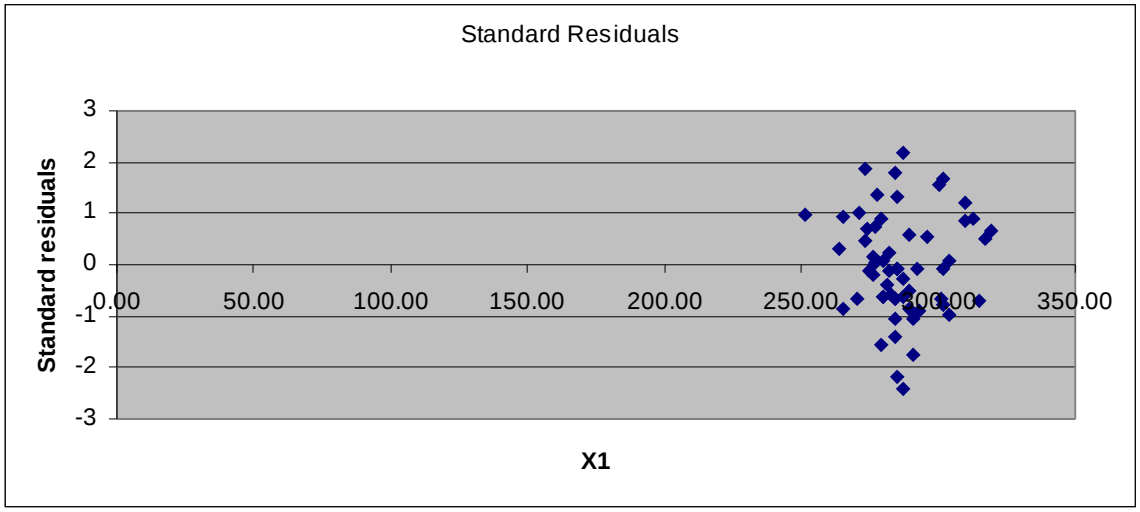
(10) Select the X1 and Standard Residuals columns in the **Data range** box and click on the **Columns** option. Click **Next >**.



(11) Enter an appropriate title and axis labels on the **Titles** tab**.** Then, click on the **Legend** tab and click once on the **Show legend** check box to uncheck.



Click **Finish.** Then, we have the following scatterplot.



In the scatterplot, we need to check the following:

- The residuals are randomly scattered around the mean of 0
- No systematic pattern is observed in the spread of the residuals across the values of X1
- The great majority of the residuals are within the range between –2 and + 2

(If any of the three points is not true, it indicates that there is a problem with the estimated regression equation. Although there are ways to solve such problems, unfortunately they are beyond Year 13 level. You can learn these and a lot more about advanced regression methods at university, typically in a second year Statistics course.)

(12) Repeat the steps (1) ~ (11) for another pair of the measurements, for example, try estimating X5 values using the X3 values.

Congratulations! You have successfully completed this task.

Do not forget to save all the worksheets you created, so that we can use them later for further analysis.