Regression in SPSS (Practical)

The development of this E-Book has been supported by the British Academy
This implementation is by National Centre for Research Methods and UK Data Service

Note: Weights have not been applied to the analyses. You can find out more about weighting survey data on the UK Data Service website.

Regression practical

In this practical we will look at regressing one variable on another variable to explore the relationship between them. This work builds on the concept of correlation that we have looked at in earlier practicals but here we specify one variable as a response (or dependent) variable and look at the effect of another predictor (or independent) variable upon it.

The dataset we are using is an excerpt from a cut-down dataset drawn from the Living Costs and Food Survey, available from the UK Data Service: http://doi.org/10.5255/UKDA-SN-7932-2, and we will be exploring the characteristics of two variables; household size and total household expenditure (in pounds per week). No conditions are required to use the data; however respondents are promised that their data will be kept confidential. As a result high values are grouped together to prevent households being identified by their large household sizes or unusually high expenditure. This protects respondents but it also affects the quality of the results produced in this workbook. Users who wish to use better quality data are encouraged to explore the full data from the Living Costs and Food Survey, which is available through the UK Data Service (http://doi.org/10.5255/UKDA-SN-7702-1), for which users need to register and adhere to some conditions of use.

In this example **expenditure** is the response variable and **hhsize** is the predictor variable. This will allow us to understand whether larger households spend more than smaller households. If so, we will find out how much more, on average, a household spends for every additional resident in the household.

To begin with we will simply look at some basic summary information about the variables and plot them in a scatterplot in SPSS which is done as follows:

1. Select **Descriptives** from the **Descriptive Statistics** submenu available from the **Analyze** menu.
2. Copy the **Total expenditure (top coded, formerly P550tpr)[expenditure]** and **Household size, number of people in household (recoded) formerly A049r[hhsize]** variables into the **Variable(s)** box.
3. Click on the **Options** button.
4. Ensure that the **Mean, Std. deviation, Minimum** and **Maximum** options are selected only.
5. Click on the **Continue** button to return to the main window.
6. Click on the **OK** button to run the command.

The descriptive statistics will then appear as shown below:

Descriptive Statistics

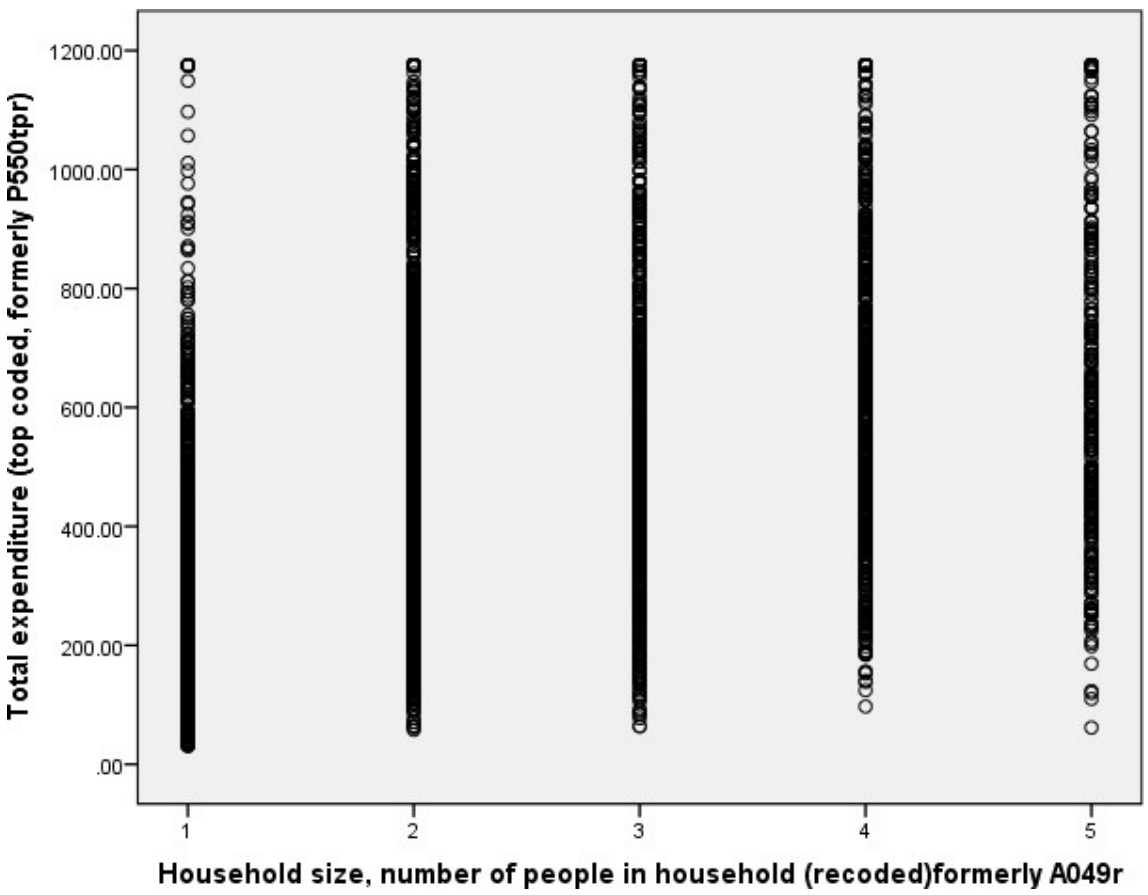
	N	Minimum	Maximum	Mean	Std. Deviation
Total expenditure (top coded, formerly P550tpr)	5144	30.52	1175.00	479.7584	292.36523
Household size, number of people in household (recoded) formerly A049r	5144	1	5	2.33	1.196
Valid N (listwise)	5144				

Here we see a row in the table for each variable. **expenditure** is the response variable and takes values between 30.52 and 1175.00 with a mean of 479.7584. **hhsize** is the predictor variable and takes values between 1 and 5 with a mean of 2.33. The maximum values are artificially low as high values of each variable have been grouped together.

We can next plot these variables against each other following instructions below:

1. Select **Scatter/Dot** from the **Legacy Dialogs** available from the **Graphs** menu.
2. Select **Simple Scatter** and click on Define to bring up the Simple Scatterplot window.
3. Copy the **Total expenditure (top coded, formerly P550tpr)[expenditure]** variable into the **Y Axis** box.
4. Copy the **Household size, number of people in household (recoded)formerly A049r[hhsize]** variable into the **X Axis** box.
5. Click on the **OK** button.

SPSS will then draw a scatterplot of the two variables which can be seen below:



The scatterplot gives a general idea of the relationship between the two variables and we can look by eye to see if a linear relationship is suitable. In this example there appears to be a positive relationship as there are more points in the bottom-left and top-right quarters of the plot than in the top-left and bottom-right corners.

We now need to actually run the linear regression to look at if there is a significant (linear) effect of **hhsize** on **expenditure**. This is done in SPSS as follows:

1. Select **Linear** from the **Regression** submenu available from the **Analyze** menu.
2. Copy the **Total expenditure (top coded, formerly P550tpr)[expenditure]** variable into the **Dependent** box.
3. Copy the **Household size, number of people in household (recoded) formerly A049r [hhsize]** variable into the **Independent(s)** box.
4. Click on the **Statistics** button.
5. On the screen appears add the tick for **Confidence Interval** to those for **Estimates** and **Model fit**.
6. Click on the **Continue** button to return to the main window.
7. Click on the **Save** button.
8. On the screen appears select the tick for **Standardized** found under **Residuals**.
9. Click on the **Continue** button to return to the main window.
10. Click on the **OK** button to run the command.

The command will run and five output tables will be presented. The first of which is the **Variables Entered/Removed** table but as it is only useful when we do multiple regression we will not show it.

The next table is the **Model Summary** table.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.449 [a]	.202	.202	261.20871

a. Predictors: (Constant), Household size, number of people in household (recoded)formerly A049r

Here we see some fit statistics for the overall model. The statistic R here takes the value .449 and is equivalent to the Pearson correlation coefficient for a simple linear regression, that is, a regression with only one predictor variable. R square (.202) is simply the value of R squared (R multiplied by itself) and represents the proportion of variance in the response variable, **expenditure** explained by the predictor variables. The table also includes an adjusted R square measure which here takes value .202 and is a version of R squared that is adjusted to take account of the number of predictors (one in the case of a simple linear regression) that are in the model.

The next table is the **ANOVA** table.

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	88771774.744	1	88771774.744	1301.067	.000 [b]
	Residual	350838622.056	5142	68229.993		
	Total	439610396.799	5143			

b. Predictors: (Constant), Household size, number of people in household (recoded)formerly A049r

The ANOVA (Analysis of Variance) table is used to look at how well the predictors as a whole can account for differences in the response variable. It is used in SPSS for many models to show how the variability in the response variable is partitioned between different explanatory components (which it measures in terms of sums of squares).

Here we see in the sum of squares column that the total sum of squares, which captures the total variance in **expenditure** is 439610396.800. This can be split into two parts. The regression sum of squares (88771774.744) captures the variability in the values of **expenditure** that would be predicted on the basis of the predictor variables alone, so here on the basis of **hhsize**. The residual sum of squares (350838622.056) is the variation in the dependent variable that remains unexplained by the predictor variables. The R-squared that we saw in the earlier table is related to the sum of squares - it expresses the regression (or explained) SS as a fraction of the total SS i.e. 88771774.744/439610396.800=0.202.

These sums of squares have associated degrees of freedom (df). For the total sum of squares the df is one less than the number of observations (N - 1, here 5143) due to fitting a mean to the data. The regression sum of squares has df = 1 to account for the 1 predictor in the model. The residual df is then the difference between the total df and the regression df = 5142. The next column is the mean squares (sums of squares adjusted for dfs) which are used to construct a test statistic, F, shown in the fifth column. Here we see that F takes value 1301.067 and can be used to test the null hypothesis that there is no linear relationship between the predictors and the response variable. To do this the value of F needs to be compared with an F distribution with 1 and 5142 degrees of freedom. This test results in a p value that is given in the Sig. column. Here p is quoted as .000 (reported as p < .001) which is less than the conventional .05 level used to judge significance. We therefore have sufficient evidence to reject the null hypothesis that the predictors have no effect.

The next table is the **Coefficients** table.

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	223.903	7.974		28.081	.000	208.271	239.534
	Household size, number of people in household (recoded) formerly A049r	109.878	3.046	.449	36.070	.000	103.906	115.850

This table gives the most interesting information about the regression model. We begin with the coefficients that form the regression equation. The regression intercept (labelled Constant in SPSS) takes value 223.903 and is the value of the regression line when **hhsize** takes value 0. The regression slope, the B, takes value 109.878 and is the amount by which we predict that **expenditure** changes for an increase in **hhsize** of one unit. In other words, the model predicts that the fixed household cost before any residents are accounted for is £224 per week. For every extra person in the household a predicted extra £110 per week is spent by the household.

Both coefficients have associated standard errors that can be used to assess their significance and also in the case of the slope, to construct a standardised coefficient. This can be seen under the Beta column and takes value .449 which represents the predicted change in **expenditure** in standard deviation units for an increase of one standard deviation in **hhsize**. The standardised coefficient beta can be interpreted as a "unit-free" measure of effect size, one that can be used to compare the magnitude of effects of predictors measured in different units.

To test for the significance of the coefficients we need to form test statistics which are reported under the t column and are simply B / Std. Error. For the slope the t statistic is 36.070 and this value can be compared with a t distribution to test the null hypothesis that B = 0. We can see the resulting p value for the test under the Sig. column. Here p is quoted as .000 (reported as p < .001) so we have sufficient evidence to reject the null hypothesis that the slope coefficient on **hhsize** is zero. We therefore have significant evidence to reject the null hypothesis that the slope is zero.

We can also check if the intercept is different from zero though this is often of less interest. For the intercept the t statistic is 28.081 and this value can be compared with a t distribution to test the null hypothesis that the intercept B = 0. We can see the resulting p value for the test under the Sig. column. Here p is quoted as .000 (reported as p < .001) so we have sufficient evidence to reject the null hypothesis that the intercept is zero .We therefore have significant evidence to reject the null hypothesis that the intercept is zero.

The final two columns give 95% confidence intervals for the coefficients, which show a lower bound of 208.271 and an upper bound of 239.534 for our estimate of the intercept.

Similarly the 95% confidence interval bounds for the **hhsize** B coefficient are 103.906 and 115.850. Zero does not lie between these values, consistent with our rejection of the null hypothesis that B = 0.

The final table is the **Residuals statistics** table.

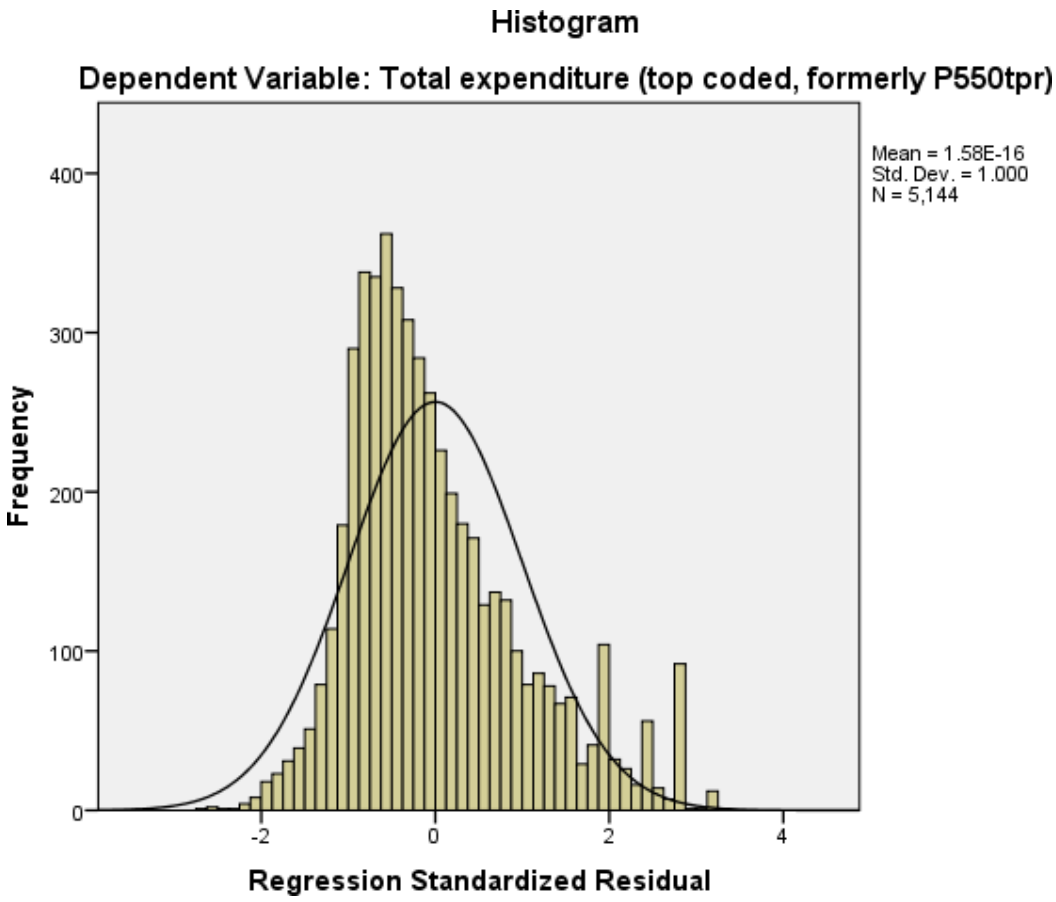
Residuals Statistics

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	333.7809	773.2941	479.7584	131.37998	5144
Residual	-711.60413	841.21912	.00000	261.18332	5144
Std. Predicted Value	-1.111	2.234	.000	1.000	5144
Std. Residual	-2.724	3.220	.000	1.000	5144

This table just summarises the predictions and residuals that come out of the regression and it is perhaps easier to look at these via plots.

As we ticked the box to request that standardised residuals were saved this has resulted in an additional variable, named **ZRE_1**, being stored in the dataset at the end of the existing variables (You can see this by viewing the dataset in the SPSS Data Editor Window). We can use this variable to create some residual plots to assess the fit of the model. We will firstly plot a histogram of the residuals to check their normality which can be done in SPSS as follows:

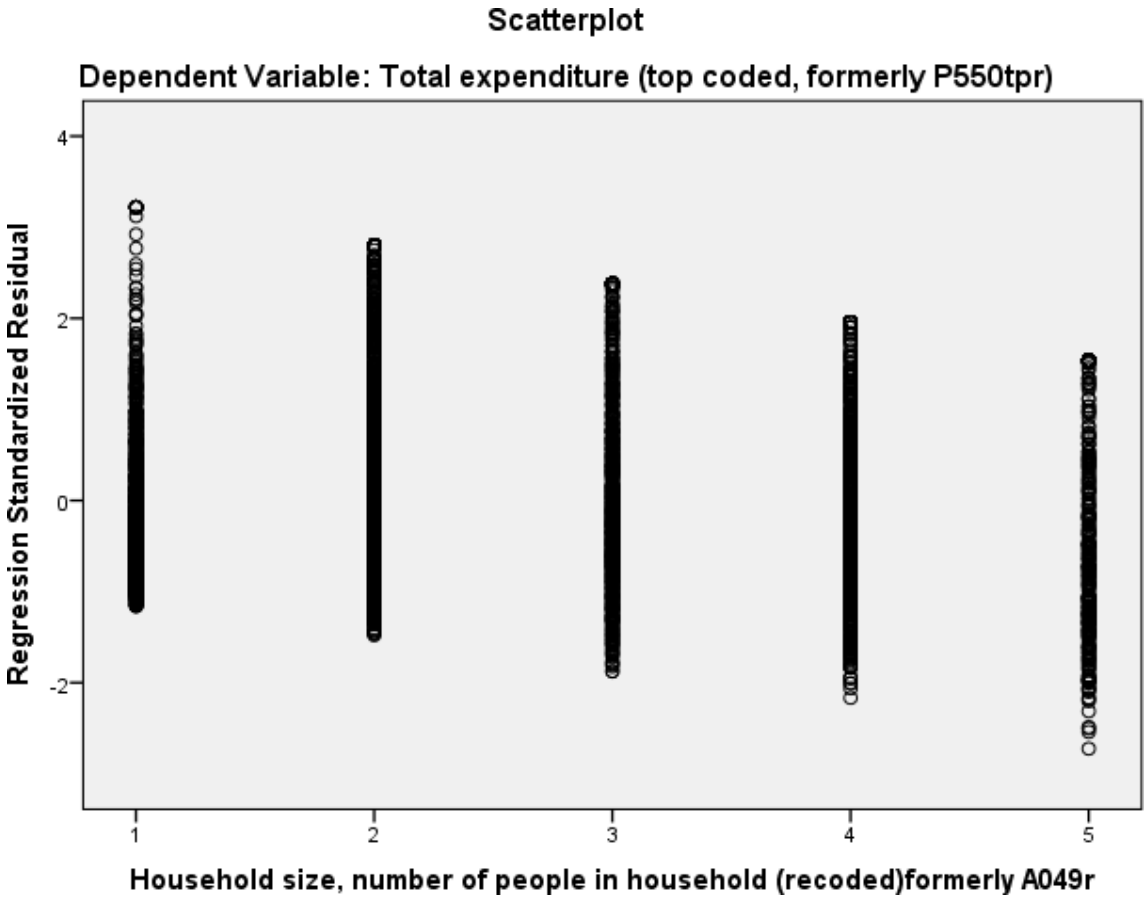
1. Select **Histogram** from the **Legacy Dialogs** available from the **Graphs** menu.
2. Copy the **Standardized Residual [ZRE_1]** variable into the **Variable** box.
3. Click on the **Display normal curve** tick box.
4. Click on the **OK** button to produce the graph.



Here we hope to see the histogram of residuals roughly following the shape of the normal curve that is superimposed over them.

We can also look at how the distribution of the residuals interacts with the predictor variable to check there is no relationship. We do this via a scatterplot which can be produced in SPSS as follows:

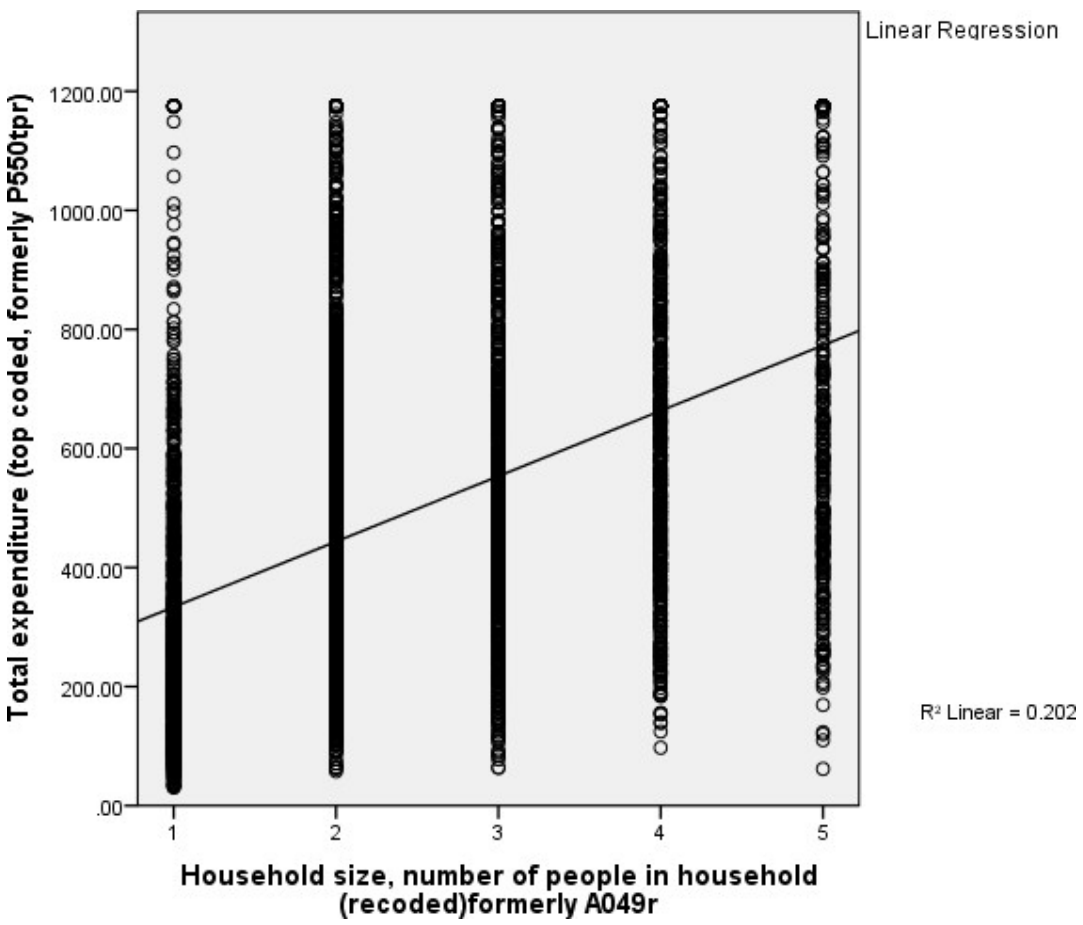
1. Select **Scatter/Dot** from the **Legacy Dialogs** available from the **Graphs** menu.
2. Select Simple Scatter and click on Define to bring up the Simple Scatterplot window.
3. Copy the **Standardized Residual [ZRE_1]** variable into the **Y Axis** box.
4. Copy the **Household size, number of people in household (recoded) formerly A049r[hhsize]** variable into the **X Axis** box.
5. Click on the **OK** button and the plot will appear.



Here we hope that the residuals show a random scatter when plotted against the predictor variable and also that their variability is constant across different values of the predictor variable.

Finally we might like to superimpose the regression line onto the scatterplot we drew earlier of the response against the predictor, so that the strength of the linear relationship is clear. To do this we will need to use the **Chart Editor** in SPSS so follow the following instructions:

1. Locate the earlier scatterplot in the SPSS output window noting you may need to scroll up to find it.
2. Double click with the left mouse button on the plot and it will pop out into a Chart Editor window.
3. On the window click on the 5th button from the left on the bottom row of icons (It will say **Add Fit Line at Total** if you hover the mouse over it)
4. On the **Properties** window that appears remove the tick next to **Attach label to line** as otherwise the equation is superimposed on the plot which looks untidy.
5. Click on the **Close** button and the line will be added in the Chart Editor window.
6. Finally close the **Chart Editor** window and the graph in the output window will now have the fixed line as shown below:



Note that the scatterplot now also contains the R-squared value which corresponds to the R-squared value we saw in the regression fit earlier.