

Regression in SPSS (Quiz)



*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*

Regression practical questions

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. We will find out whether there is a relationship between the size of a household's income (in pounds per week) and the amount that the household spends (also in pounds per week). If there is a relationship, we will also be able to estimate how much more a household spends for every extra pound per week that it receives in income.

The dataset we are using is an excerpt from a cut-down dataset drawn from the Living Costs and Food Survey 2013, available from the UK Data Service: <http://doi.org/10.5255/UKDA-SN-7932-2>. No conditions are required to use the data; however respondents are promised that their data will be kept confidential. As a result high values of some variables such as income and expenditure are grouped together to prevent households being identified by unusually large values. 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.

Begin by looking at some basic summary information about the response variable **expenditure** and the predictor variable **income**.

- Question: What is the range of the predictor?
- Question: What is the mean of the predictor?

Next use a scatter plot to visualise the association between **income** and **expenditure**, and add a regression line to the plot to show the linear relationship.

- Question: What does the chart tell you about the relationship?

Next run the linear regression to formally estimate the effect of **income** on **expenditure**. Make sure that you select the option to save the standardized residuals from the regression so that you can explore these with some plots.

- Question: Looking at the model summary table what proportion of the variance in the response is explained by the predictor?
- Question: Looking at the ANOVA table is the regression significant?
- Question: Looking at the Coefficients table what is the value of the intercept?
- Question: Looking at the Coefficients table what is the value of the slope?
- Question: What is the standardised slope coefficient and what does it mean?
- Question: Is the slope significant?
- Question: What is the 95% CI for the intercept?
- Question: What is the 95% CI for the slope?

Next plot a histogram of the residuals from the regression model.

- Question: Do the residuals look approximately Normal?

Now use a scatterplot to check whether there appears to be any relationship between the residuals and the predictor variable **income**.

- Question: Is there any specific pattern in the plot?

Solutions to Regression practical questions

The instructions for creating the required descriptive statistics in SPSS are 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 **income[income]** 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.

- Question: What is the range of the predictor?
- Question: What is the mean of the predictor?

Solutions: The output from SPSS is as follows:

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Total expenditure (top coded, formerly P550tpr)	5144	30.52	1175.00	479.7584	292.36523
Income	5144	.00	1184.99	620.4336	359.15568
Valid N (listwise)	5144				

The answers are as follows:

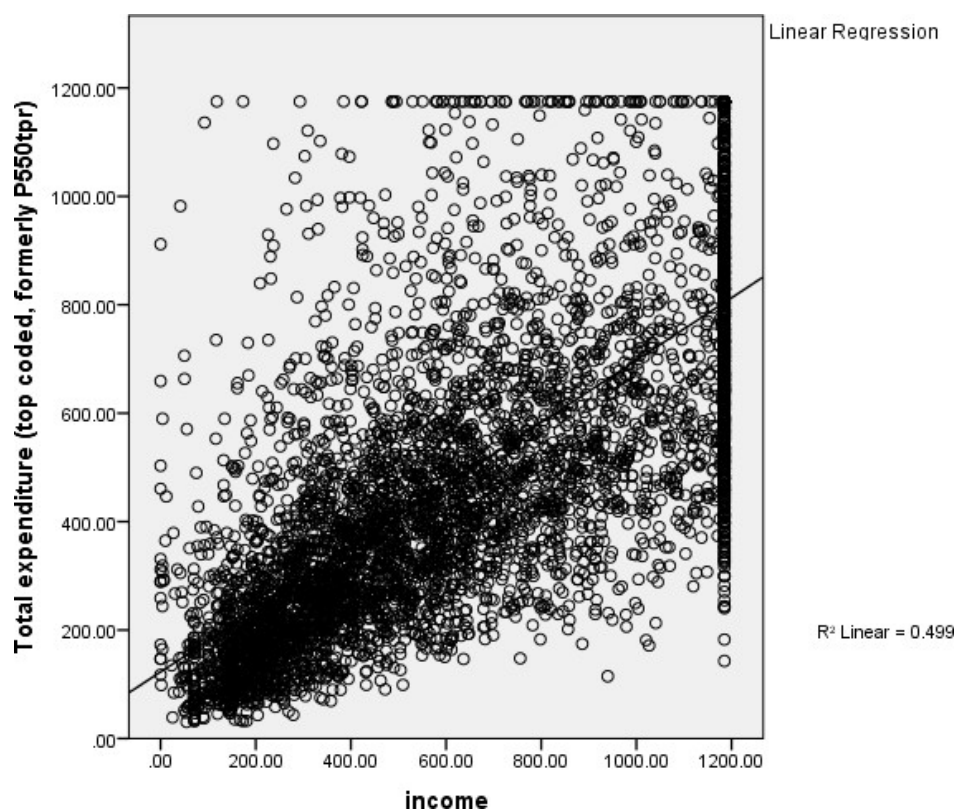
- The predictor takes values between .00 and 1184.99. This upper limit is an artefact of high values being grouped together
- The mean of the predictor is 620.4336

The instructions for creating the required scatterplot in SPSS are 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 **Total expenditure (top coded, formerly P550tpr)[expenditure]** variable into the **Y Axis** box.
4. Copy the **income[income]** variable into the **X Axis** box.
5. Click on the **OK** button.
6. Double click with the left mouse button on the plot and it will pop out into a Chart Editor window.
7. 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)
8. 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.
9. Click on the **Close** button and the line will be added in the Chart Editor window.
10. Finally close the **Chart Editor** window and the graph in the output window will now have the fixed line as shown below:

- Question: What does the chart tell you about the relationship?

Solution: The output from SPSS is as follows:



The scatterplot appears to indicate a significant positive relationship.

The instructions for running the required regression in SPSS are 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 **income** 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.

- Question: Looking at the model summary table what proportion of the variance in the response is explained by the predictor?
- Question: Looking at the ANOVA table is the regression significant?
- Question: Looking at the Coefficients table what is the value of the intercept?
- Question: Looking at the Coefficients table what is the value of the slope?
- Question: What is the standardised slope coefficient and what does it mean?
- Question: Is the slope significant?
- Question: What is the 95% CI for the intercept?
- Question: What is the 95% CI for the slope?

Solutions: The output from SPSS is as follows:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.706 ^a	.499	.499	206.94574

a. Predictors: (Constant), income

ANOVA

Model	Sum of Squares	df	Mean Square	F	Sig.
1 Regression	219396340.515	1	219396340.515	5122.906	.000 ^b
Residual	220214056.285	5142	42826.538		
Total	439610396.799	5143			

b. Predictors: (Constant), income

Coefficients

Model	Unstandardized Coefficients		Standardized Coefficients Beta	t	Sig.	95.0% Confidence Interval for B	
	B	Std. Error				Lower Bound	Upper Bound
1 (Constant)	122.963	5.760		21.349	.000	111.672	134.255
income	.575	.008	.706	71.574	.000	.559	.591

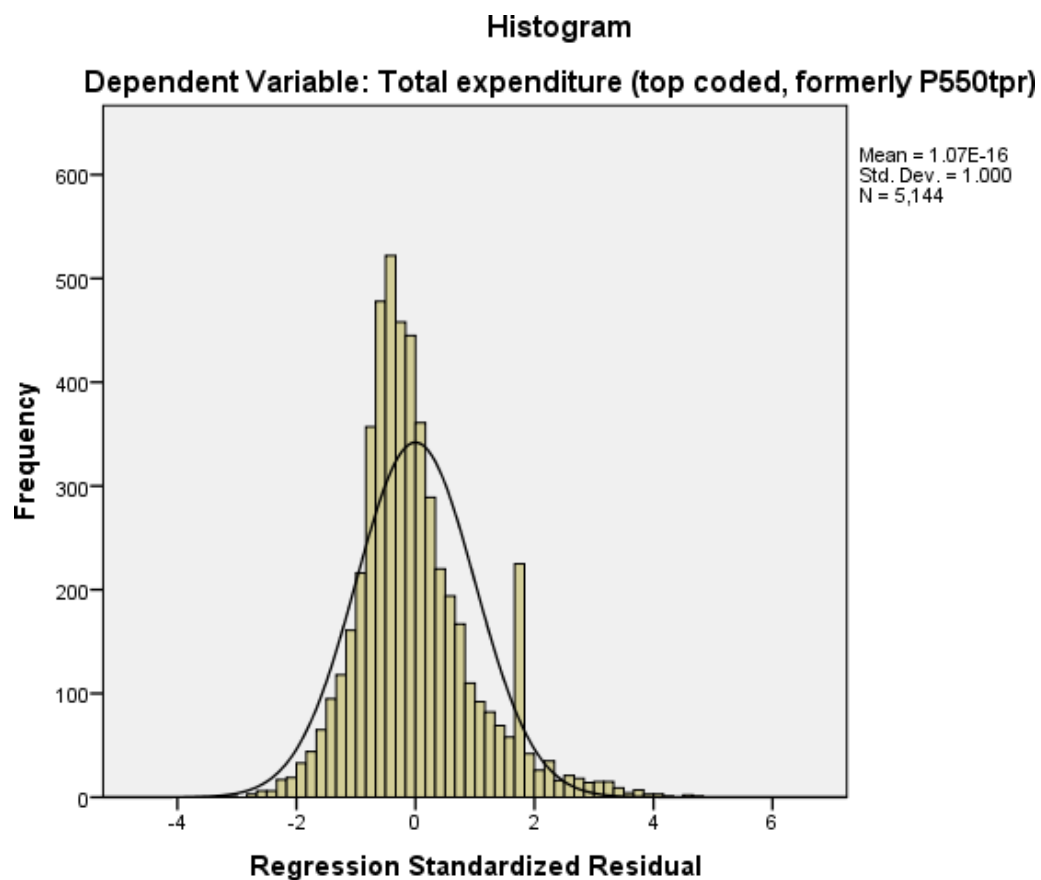
- The R squared provides the proportion of the variance explained by the regression and has the value .499. This suggests that around half of the variance is explained by the regression model.
- Here p is quoted as .000 (reported as $p < .001$) which is less than the conventional .05 level used to judge significance. We therefore have sufficient evidence to reject the null hypothesis that the predictors have no effect.
- The regression intercept takes value 122.963, so the model predicts that if a household had no income it would still spend £123 per week.
- The regression slope takes value .575. The model therefore predicts that for every extra pound in income households have, on average an extra 57 pence will be spent per week.
- The standardised slope coefficient takes value .706 which represents the predicted change in **expenditure** for an increase of 1 standard deviation in **income**.
- 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 **income** is zero. We therefore have significant evidence to reject the null hypothesis that the slope is zero.
- The confidence interval for the intercept is 111.672 to 134.255.
- The confidence interval for the slope is .559 to .591.

The instructions for creating the residual histogram in SPSS are 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.

- Question: Do the residuals look approximately Normal?

Solution: The output from SPSS is as follows:



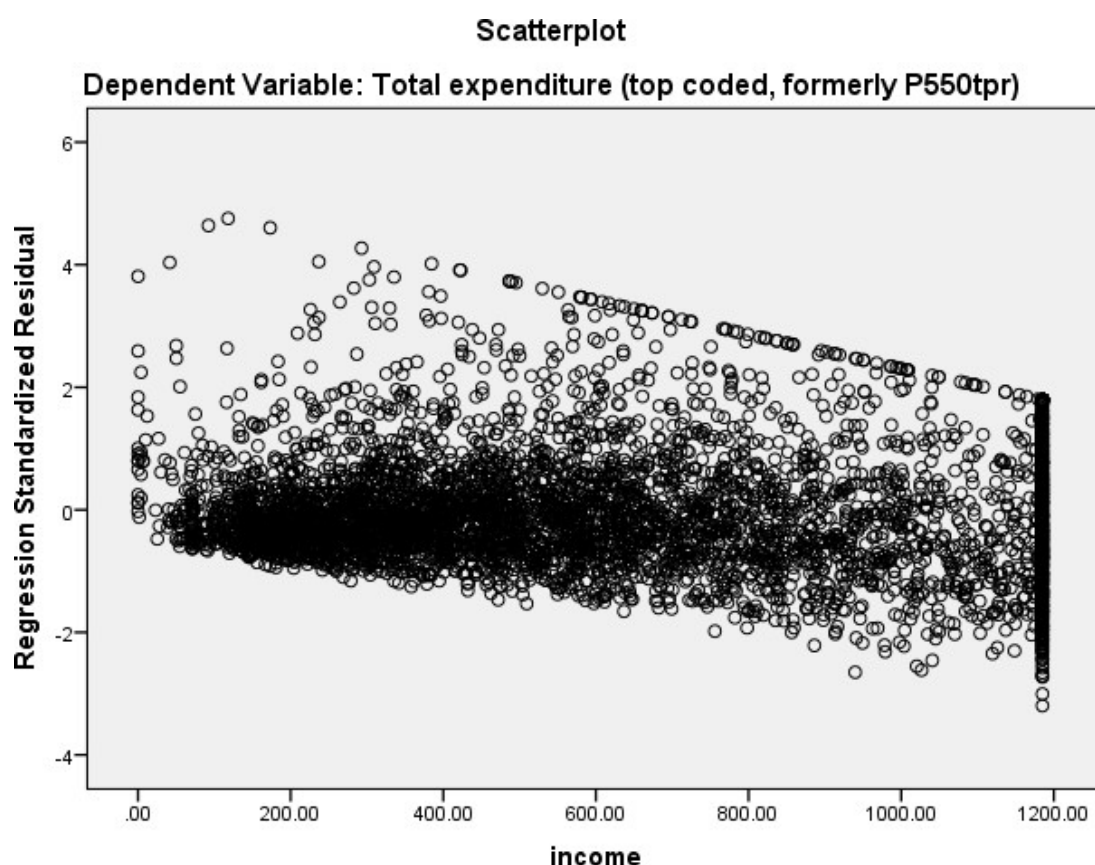
We hope to see the histogram of residuals roughly following the shape of the normal curve that is superimposed over them.

The instructions for creating the required scatterplot in SPSS are 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 **income[income]** variable into the **X Axis** box.
5. Click on the **OK** button and the plot will appear.

- Question: Is there any specific pattern in the plot?

Solution: The output from SPSS is as follows:



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.