Multiple Regression in SPSS worksheet (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.

Multiple Regression practical

In this practical we will look at regressing two different predictor variables individually on a response, followed by a model containing both of them. We will also look at a second approach to doing this. This work builds on the earlier simple linear regression practical.

The dataset we are using is a 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 how household income can be used to predict household expenditure, and whether there is any additional effect according to whether the households' main source of income is from earned income or another source. Although the **maininc** variable is not what we would normally consider to be a continuous variable, the variable only takes two values which means that it can be interpretted meaningfully, as we will see. Both income and expenditure are measured 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 expenditures 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.

We start by running the first linear regression to look at if there is a significant (linear) effect of **income** 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 **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 **OK** button to run the command.

SPSS will produce several tabular outputs but here we will focus on only the model summary and coefficients tables that can be seen below:

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

Here we see some fit statistics for the overall model. The statistic R here takes the value .706 and is equivalent to the Pearson correlation coefficient for a simple linear regression, that is a regression with only one predictor variable. R squared (.499) is simply the value of R squared (R multiplied by itself) and represents the proportion of variance in the response variable, **expenditure** explained by **income**. The table also includes an adjusted R square measure which here takes value .499 and is a version of R squared that is adjusted to take account of the number of predictors (one in the case of this simple linear regression) that are in the model. We next look at the coefficients table which is shown below:

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	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

This table often 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 the value 122.963 and is the predicted value of **expenditure** when **income** takes value 0. This means that our model predicts that an average fixed cost of £123 per week which applies even where there are no residents in the household.

The regression slope, or unstandardised coefficient, (B in SPSS) takes value .575 and is the amount by which we predict that **expenditure** changes for an increase of 1 unit in **income**. In other words, our model predicts that for every extra pound a household has in income, expenditure will also increase by 58 pence per week.

Both coefficients have associated standard errors that can be used to assess their significance. SPSS also reports a standardised coefficient (the Beta) that 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. Here Beta takes the value .706 which represents the predicted change in the number of standard deviations of **expenditure** for an increase of 1 standard deviation in **income**.

To test for the significance of the coefficients we need to form test statistics which are reported under the t column and these are simply B / Std.Error. For the slope on **income** the t statistic is 71.574 and this value can be compared with a t distribution to test the null hypothesis that the slope is 0. We can see the resulting p value for the test under the Sig. column. The p value (quoted under Sig.) is .000 (reported as $p < .001$) which is less than 0.05. We therefore have significant evidence to reject the null hypothesis that the slope coefficient on **income** is zero.

We can also check if the intercept is different from zero though this is often of less interest. For the intercept here the t statistic is 21.349 and the p value (quoted under Sig.) is .000 (reported as $p < .001$) which is less than 0.05. We therefore have significant evidence to reject the null hypothesis that the intercept is zero.

The final two columns give confidence intervals for the coefficients and so a 95 percent confident interval for the intercept takes values between 111.672 and 134.255.

Similarly a 95 percent confidence interval for the slope for **income** takes value between .559 and .591. Here we see the confidence interval does not contain 0 which corresponds to the fact we could reject the null hypothesis that the slope was 0.

We will next run the second linear regression to look at if there is a significant (linear) effect of **maininc** on **expenditure**.

Maininc is a binary variable, it only takes two values;

- 0 when the household's main income source is earnings
- 1 when the household's main income is another source.

A one unit increase in **maininc** can therefore be interpretted as meaning that a household has mainly unearned income (as opposed to mainly earned income).

This is done in SPSS as follows:

1. Select **Linear** from the **Regression** submenu available from the **Analyze** menu.
2. Remove the **income** variable from the **Independent(s)** box.
3. Copy the **Main source of household income (recoded, P425-1)[maininc]** variable into the **Independent(s)** box.
4. The other options will be remembered from last time.
5. Click on the **OK** button to run the command.

The model and coefficients tables for this second model can be seen below:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.404 [a]	.163	.163	267.53203

a. Predictors: (Constant), Main source of household income (recoded, P425-1)

This time we see some fit statistics for the regression with **maininc**. The statistic R here takes the value .404. R squared (.163) represents the proportion of variance in the response variable, **expenditure** explained by **maininc**. This time the adjusted R square measure takes value .163. We next look at the coefficients table which is shown below:

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	585.966	5.019		116.743	.000	576.126	595.806
	Main source of household income (recoded, P425-1)	-237.227	7.501	-.404	-31.624	.000	-251.933	-222.521

This time the coefficients that form the regression equation are as follows: The regression intercept takes value 585.966 while the regression slope takes value -237.227 and is the amount by which we predict that **expenditure** changes for an increase of 1 in **maininc**.

So, by default when **maininc** is zero, which is to say when the household's income comes mainly from earnings, this model predicts that a household will spend £586 per week. By contrast, households with incomes which are mainly from unearned sources are predicted to spend £237 less.

This time under the Beta column the standardised slope takes value -.404 which represents the predicted change in **expenditure** in standard deviation units for an increase of 1 standard deviation in **maininc**.

For the slope cofficient on **maininc** the t statistic is -31.624 and this value can be compared with a t distribution to test the null hypothesis that the slope is 0. The p value (quoted under Sig.) is .000 (reported as $p < .001$) which is less than 0.05. We therefore have significant evidence to reject the null hypothesis that the slope is zero.

For the intercept the t statistic is 116.743 and the p value (quoted under Sig.) is .000 (reported as p < .001) which is less than 0.05. We therefore have significant evidence to reject the null hypothesis that the intercept is zero.

The final two columns give confidence intervals for the coefficients and so a 95 percent confidence interval for the intercept takes values between 576.126 and 595.806.

Similarly a 95 percent confidence interval for the slope for **maininc** takes values between -251.933 and -222.521. Here we see the confidence interval does not contain 0 which corresponds to the fact we could reject the null hypothesis that the slope was 0.

We know therefore that households with lower incomes spend less (regression on page 1), as do households in which the main source of income is not earnings (regression on page 2). Are these two separate affects, or can the differences accounted for by main source of income actually due to the households with earnings are also better off? In order to ascertain this we can include both variables in a single model. This allows us to look at the effect of each variable when the other is taken into account.

We will therefore run the third multiple regression to look at if there are significant (linear) effects of both **income** and **maininc** on **expenditure**. This is done in SPSS as follows:

1. Select **Linear** from the **Regression** submenu available from the **Analyze** menu.
2. Copy the **income** variable into the **Independent(s)** box to join **Main source of household income (recoded, P425-1)[maininc]**.
3. The other options will be remembered from last time.
4. Click on the **OK** button to run the command.

The model and coefficients tables can be seen below:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.707 [a]	.499	.499	206.94980

a. Predictors: (Constant), income, Main source of household income (recoded, P425-1)

This time we see some fit statistics for the multiple regression with both **income** and **maininc**. The statistic R here takes the value .707 . R squared (.499) represents the proportion of variance in the response variable, **expenditure** explained by the multiple regression (both of the predictor variables combined). This time the adjusted R square measure takes value .499 which we can compare with .499 for just **income** and .163 for just **maininc**. An increase in the adjusted R square compared to either one of these implies that the second added variable has increased the explained variance in **expenditure**. We next look at the coefficients table which is shown below:

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	128.783	8.696		14.809	.000	111.735	145.831
	Main source of household income (recoded, P425-1)	-6.261	7.009	-.011	-.893	.372	-20.001	7.480
	income	.570	.010	.700	58.755	.000	.551	.589

This time the coefficients that form the regression equation are as follows: The regression intercept takes value 128.783 while the regression slope for **maininc** takes value -6.261 and the slope for **income** takes value .570. These have changed from -237.227 and .575 respectively when the variables are fitted individually. The coefficient of the income variable is similar to that we have already seen. This time the effect of a household having an income source which is something other than earnings has almost disappeared. Income source only seems to explain a difference of £6 per week once total income is taken into account.

This time there are two standardised slopes with the slope for **maininc** taking value -.011 and the slope for **income** taking value .700.

For **maininc** the slope has t statistic -.893 and the p value (quoted under Sig.) is .372 which is greater than 0.05 and therefore we cannot reject the null hypothesis that the slope on **maininc** is zero.

For **income** the slope has t statistic 58.755 and the p value (quoted under Sig.) is .000 (reported as p < .001) which is less than 0.05. We therefore have significant evidence to reject the null hypothesis that the slope on **income** is zero.

It looks like the effect maininc is not statistically significant if we account for total household income.

For the intercept the t statistic is 14.809 and the p value (quoted under Sig.) is .000 (reported as p < .001) which is less than 0.05. We therefore have significant evidence to reject the null hypothesis that the intercept is zero.

The final two columns give confidence intervals for the coefficients and so a 95 percent confidence interval for the intercept takes values between 111.735 and 145.831.

Similarly a 95 percent confidence interval for the slope for **maininc** takes values between -20.001 and 7.480. Here we see the confidence interval contains 0 which corresponds to the fact we could not reject the null hypothesis that the slope was 0.

Finally a 95 percent confidence interval for the slope for **income** takes values between .551 and .589. Here we see the confidence interval does not contain 0 which corresponds to the fact we could reject the null hypothesis that the slope was 0.

Finally we will show how to run two of the regression models in one go and build up the regression in blocks. This is done in SPSS as follows:

1. Select **Linear** from the **Regression** submenu available from the **Analyze** menu.
2. Remove the **Main source of household income (recoded, P425-1)[maininc]** variable from the **Independent(s)** box to leave just **income**.
3. Click the **Next** button.
4. Copy the **Main source of household income (recoded, P425-1)[maininc]** variable into the now empty **Independent(s)** box.
5. Click on the **Save** button.
6. On the screen appears select the tick for **Standardized** found under **Residuals**.
7. Click on the **Continue** button to return to the main window.
8. Click on the **OK** button to run the command.

The model and coefficients tables can be see below:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.706 [a]	.499	.499	206.94574
2	.707 [b]	.499	.499	206.94980

a. Predictors: (Constant), income

b. Predictors: (Constant), income, Main source of household income (recoded, P425-1)

Here we see the model summaries for the first and third regression models earlier i.e. we fit a model with just **income** and then a second model where we introduce **maininc**.

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients			95.0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	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
2	(Constant)	128.783	8.696		14.809	.000	111.735	145.831
	income	.570	.010	.700	58.755	.000	.551	.589
	Main source of household income (recoded, P425-1)	-6.261	7.009	-.011	-.893	.372	-20.001	7.480

Similarly we have the model coefficients for the first and third models from earlier in one combined table. Having selected standardised residuals we get an additional table, the **Residuals statistics** table.

Residuals Statistics

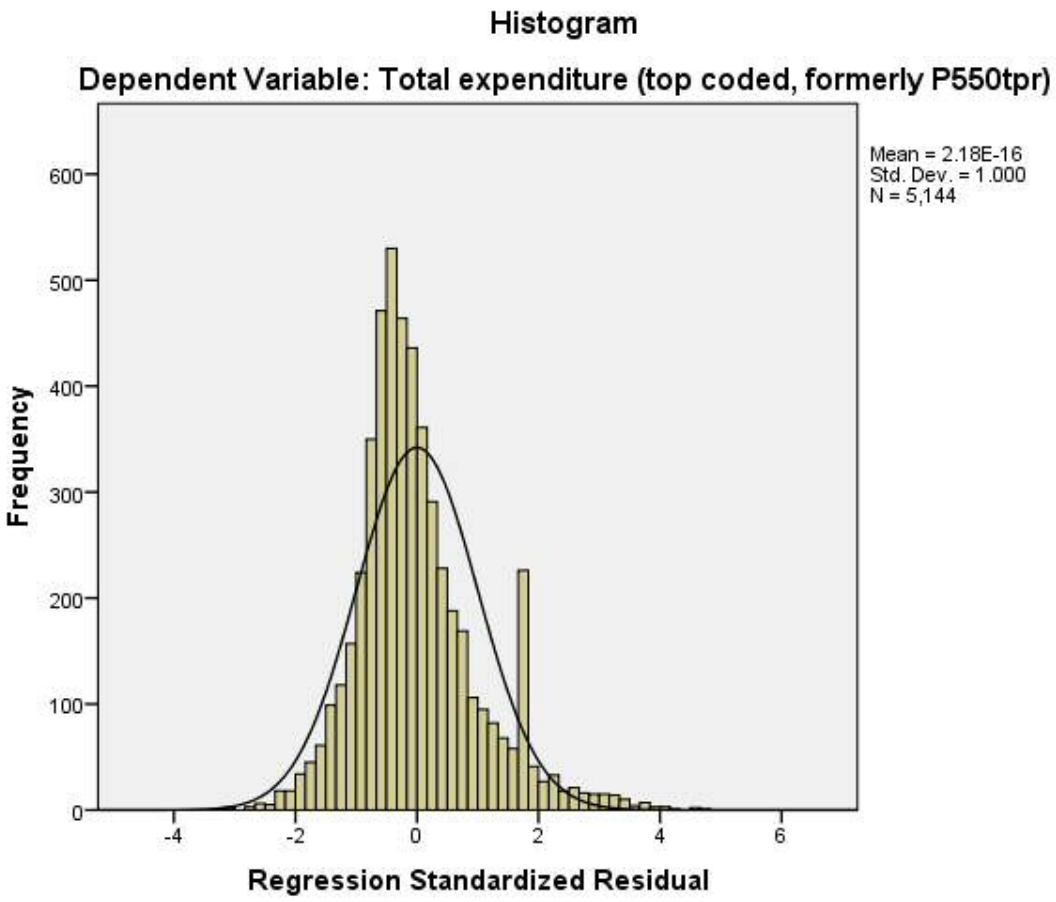
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	122.5221	804.4781	479.7584	206.55716	5144
Residual	-655.16858	985.12305	.00000	206.90956	5144
Std. Predicted Value	-1.729	1.572	.000	1.000	5144
Std. Residual	-3.166	4.760	.000	1.000	5144

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

As we requested that standardized residuals were saved this has resulted in an additional variable being stored in the dataset named **ZRE_1** at the end of the existing variables. We can use this variable to create some residuals plot 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 diagnostics** 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.

This will produce the graph as shown below:

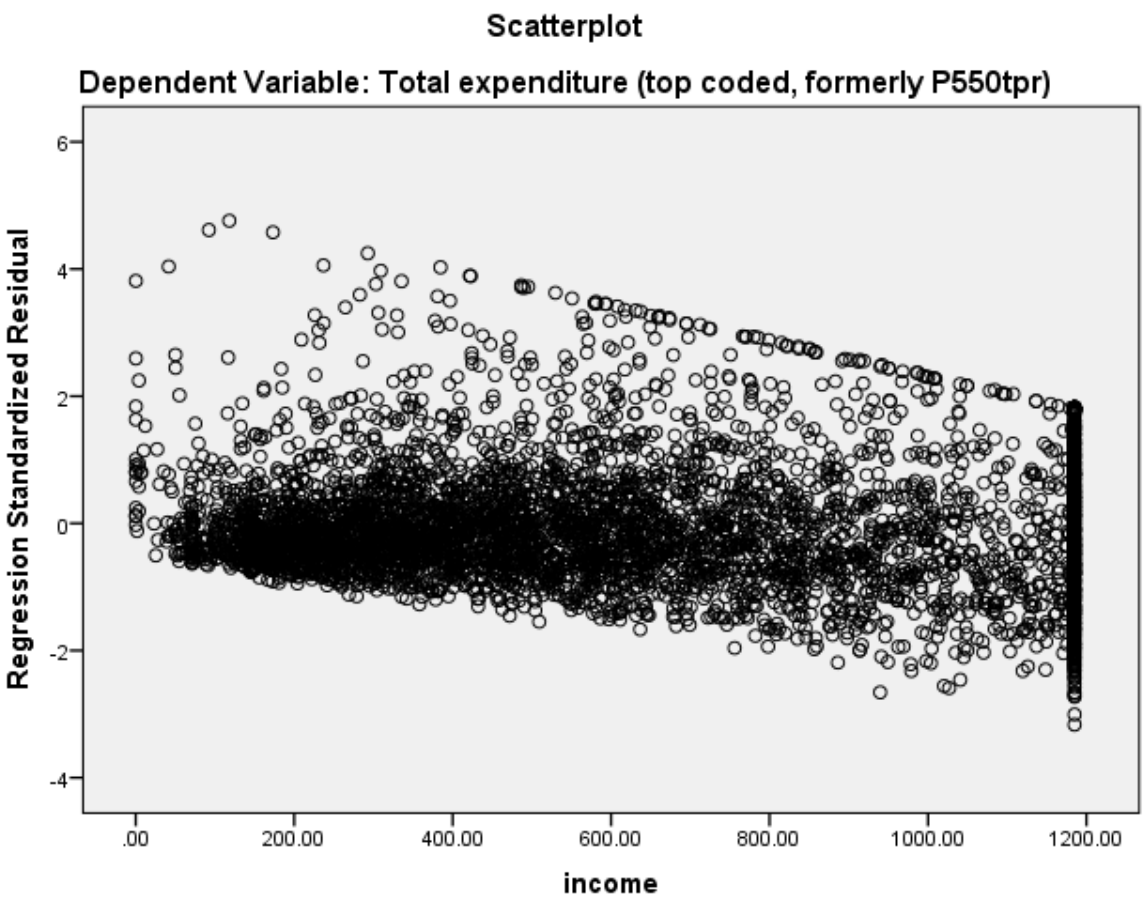


Here we hope to see the histogram of residuals roughly following the shape of the normal curve that is superimposed over them.

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

1. Select **Scatter/Dot** from the **Legacy diagnostics** 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** variable into the **X Axis** box.
5. Click on the **OK** button.

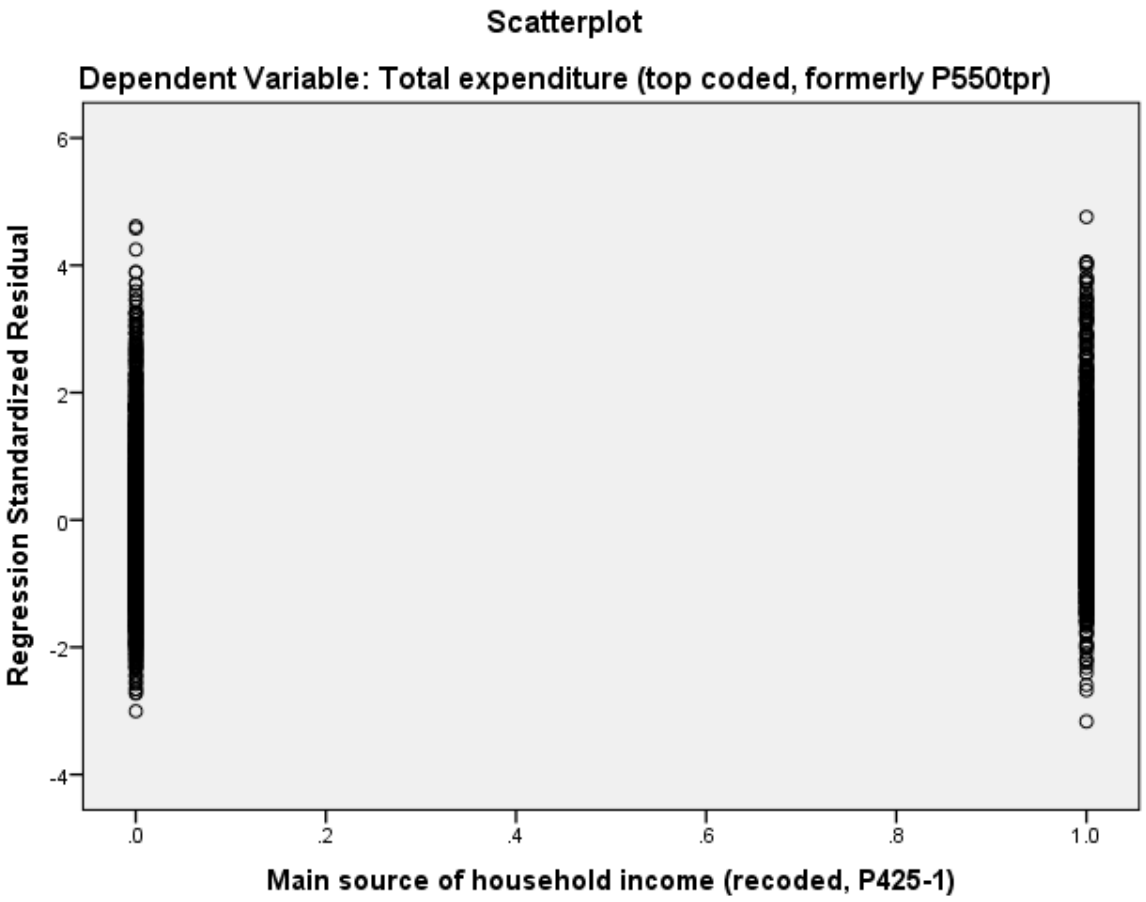
This will produce the graph as shown below:



Here we hope not to see any pattern where there was more variability in the residuals for particular values of **income**. We can repeat this plot for **Main source of household income (recoded, P425-1)[maininc]** as follows:

1. Select **Scatter/Dot** from the **Legacy diagnostics** available from the **Graphs**menu.
2. Select Simple Scatter and click on Define to bring up the Simple Scatterplot window
3. Remove the **income** variable from the **X Axis** box.
4. Copy the **Main source of household income (recoded, P425-1)[maininc]** variable into the **X Axis** box.
5. Click on the **OK** button.

This will produce the graph as shown below:



Note that again we hope not to see any pattern in the residuals.

Copyright and citation

Jo Wathan, Vanessa Higgins, Mark Elliot, William Browne, Chris Charlton, Ana Morales Gomez and Jennifer Buckley (2019) Quantitative methods e-books: Teaching Resources, UK Data Service, NCRM, Centre for Multi-Level Modelling.