Multiple Linear Regression

(2nd Edition)

Mark Tranmer

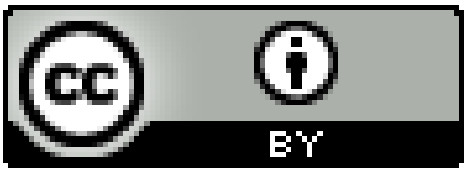
Jen Murphy

Mark Elliot

Maria Pampaka

January 2020

License and attribution



When citing please use the following (or equivalent):

Tranmer, M., Murphy, J., Elliot, M., and Pampaka, M. (2020) Multiple Linear Regression (2nd Edition); *Cathie Marsh Institute Working Paper 2020-01. https://hummedia.manchester.ac.uk/institutes/cmist/archive-publications/working-papers/2020/2020-1-multiple-linear-regression.pdf*

CONTENTS

1 THE BASICS – UNDERSTANDING LINEAR REGRESSION

Linear regression is a modelling technique for analysing data to make predictions. In simple linear regression, a bivariate model is built to predict a response variable (y) from an explanatory variable (x)[1]. In multiple linear regression the model is extended to include more than one explanatory variable ($x_1, x_2, \ldots, x_p$) producing a *multivariate* model.

This primer presents the necessary theory and gives a practical outline of the technique for bivariate and multivariate linear regression models. We discuss model building, assumptions for regression modelling and interpreting the results to gain meaningful understanding from data. Complex algebra is avoided as far as is possible and we have provided a reading list for more in-depth learning and reference.

1.1 SIMPLE LINEAR REGRESSION – ESTIMATING A BIVARIATE MODEL

A simple linear regression estimates the relationship between a response variable y, and a single explanatory variable x, given a set of data that includes observations for both of these variables for a particular sample.

For example, we might be interested to know if exam performance at age 16 – the response variable – can be predicted from exam results at age 11 – the explanatory variable.

Table 1 Sample of exam results at ages 11 and 16 (n = 17)

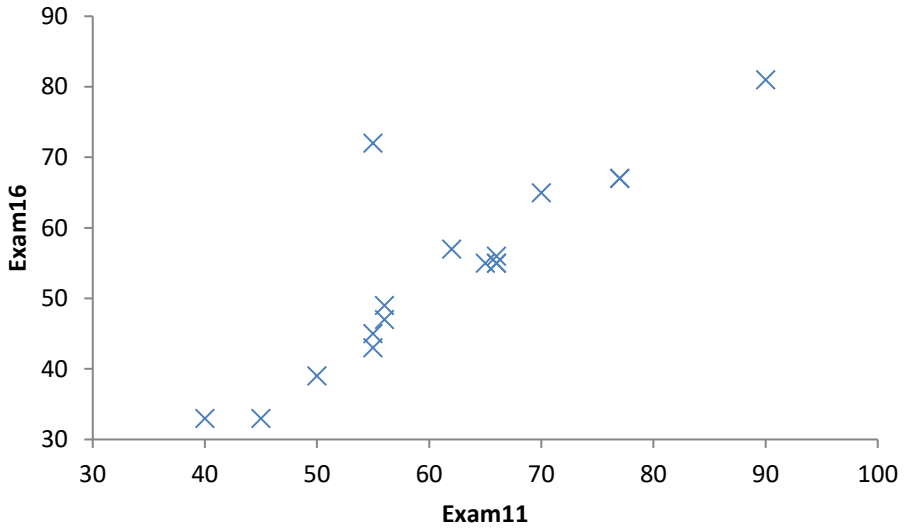
Results at age 16 (Variable name: Exam16)	Results at age 11 (Variable name: Exam11)
45	55
67	77
55	66
39	50
72	55
47	56
49	56
81	90

[1] The terms response and explanatory variables are the general terms to describe predictive relationships. You will also see the terms dependent and independent used. Formally, this latter pair only applies to experimental designs but are sometimes used more generally. Some statistical software (e.g. SPSS) uses dependent/independent by default.

33	40
65	70
57	62
33	45
43	55
55	65
55	66
67	77
56	66

Table 1 contains exam results at ages 11 and 16 for a sample of 17 students. Before we use linear regression to predict a student's result at 16 from the age 11 score, we can plot the data (Figure 1).

Figure 1 Scatterplot of exam score at age 16, against score at age 11



We are interested in the relationship between age 11 and age 16 scores – or how they are correlated. In this case, the correlation coefficient is 0.87 – demonstrating that the two variables are indeed highly positively correlated.

To fit a straight line to the points on this scatterplot, we use linear regression – the equation of this line, is what we use to make predictions. The equation for the line in regression modelling takes the form:

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

We refer to this as our model. For the mathematical theory underlying the estimation and calculation of correlation coefficients, see Appendix A.

β_0 is the *intercept* also called the *constant*– this is where the line crosses the y axis of the graph. For this example, this would be the predicted age 16 score, for someone who has scored nil in their age 11 exam.

β_1 is the *slope* of the line – this is how much the value of y increases, for a one-unit increase in x, or for each additional mark gained in the age 11 exam, how much the student scores in the age 16 exam.

e_i is the error term for the i^{th}student. The error is the amount by which the predicted value is different to the actual value. In linear regression we assume that if we calculate the error terms for every person in the sample, and take the mean, the mean value will be zero. The error term is also referred to as the residual (see 1.3 for more detail on residuals).

1.2 HYPOTHESIS TESTING

Our hypothesis is that the age 16 score can be predicted from the age 11 score that is to say that there is an association between the two. We can write this out as null and alternative hypotheses:

$$H_0\colon\ \beta_1 = 0$$
$$H_1\colon\ \beta_1 \neq 0$$

The null hypothesis is that there is no association – it doesn't matter what the age 11 score is for a student when predicting their age 16 score, so the slope of the line, denoted β_1, would be zero.

If there is a relationship, then the slope is not zero – our alternative hypothesis.

The relationship between *x* and *y* is then estimated by carrying out a simple linear regression analysis. SPSS estimates the equation of the line of best fit by minimising the sum of the squares of the differences between the actual values, and the values predicted by the equation (the residuals) for each observation. This method is often referred to as the ordinary least squares approach; there are other methods for estimating parameters but the technical details of this are beyond this primer.

For this example:

β_0 = -3.984

β_1 = 0.939

This gives us a regression equation of:

$$\hat{y}_i = -3.984 + 0.939x_i$$

where x_i is the value of EXAM11 for the ith student. The ^ symbol over the y_i is used to show that this is a predicted value.

So, if a student has an EXAM11 score of 55 we can predict the EXAM16 score as follows:

$$Predicted\ EXAM16\ score = -3.984 + (0.939 \times 55)$$
$$= 47.7$$

If we draw this line on the scatter plot, as shown in Figure 2, it is referred to as the line of best fit of *y* on *x,* because we are trying to predict *y* using the information provided by *x.*

1.3 RESIDUALS

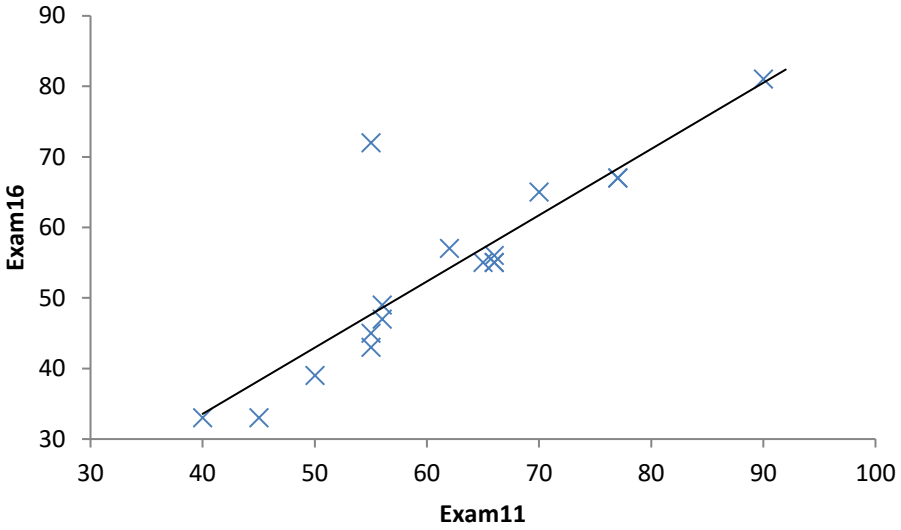
The predicted EXAM16 score of the student with an EXAM11 score of 55 is 47.7;; however, if we refer to the original data, we can see that the first student in the table scored 55 at age 11, but their actual score at age 16 was 45. The difference between the actual or observed value, and the predicted value is called the error or residual.

$$e_i = y_i - \hat{y}_i$$

Remember that $\hat{y}$ means predicted, and y means actual or observed.

The residual for the first student is therefore 45 – 47.7 = -2.7. The residual is the distance of each data point away from the regression line. In Figure 2 the prediction equation is plotted on the scatter plot of exam scores. We can see that very few if any of the actual values fall on the prediction line.

Figure 2 Plotting the regression line for age 11 and age 16 exam scores



If we calculate the predicted value using the regression equation for every student in the sample, we can then calculate all the residuals. For a model which meets the *assumptions for linear regression*, the mean of these residuals is zero. More about assumptions and testing data to make sure they are suitable for modelling using linear regression later!

Our model has allowed us to predict the values of EXAM16, however it is important to distinguish between correlation and causation. The EXAM11 score value, has not caused the EXAM16 score value, they are simply correlated – there may be other variables through which the relationship is mediated: base intellect, educational environment, parental support, student effort and so on and these could be causing the score, rather than the explanatory variable itself. To illustrate this further, statistically speaking, we would have just as good a model if we used EXAM16 to predict the values of EXAM11. Clearly one would not expect a student's EXAM scores at age 16 to be causing in any sense their exam scores at age 11! So a good model does not mean a causal relationship.

Our analysis has investigated how an explanatory variable is associated with a response variable of interest, but the equation itself is not grounds for causal inference.

1.4 MULTIPLE LINEAR REGRESSION – A MULTIVARIATE MODEL

Multiple linear regression extends simple linear regression to include more than one explanatory variable. In both cases, we still use the term 'linear' because we assume that the response variable is directly related to a linear combination of the explanatory variables.

The equation for multiple linear regression has the same form as that for simple linear regression but has more terms:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + e_i$$

As for the simple case, β_0 is the constant – which will be the predicted value of y when all explanatory variables are 0. In a model with p explanatory variables, each explanatory variable has its own β_coefficient.

Again, the analysis does not allow us to make causal inferences, but it does allow us to investigate how a set of explanatory variables is associated with a response variable of interest.

2 BASIC ANALYSIS USING SPSS

Multiple linear regression is a widely used method within social sciences research and practice. Examples of suitable problems to which this method could be applied include:

- Prediction of an individual's income given several socio-economic characteristics.
- Prediction of the overall examination performance of pupils in 'A' levels, given the values of a set of exam scores at age 16.
- Estimation of systolic or diastolic blood pressure, given a variety of socio-economic and behavioural characteristics (occupation, drinking smoking, age etc.).

This section shows how to use the IBM program SPSS to build a multiple linear regression model to investigate the variation between different areas in the percentage of residents reporting a life limiting long-term illness.

The data are taken from the 2001 UK Census and are restricted to the council wards in the North West of England (n = 1006).

2.1 VARIABLES IN THE ANALYSIS

We will consider five variables in this analysis (See Table 2).

Table 2 Variables in the analysis

Variable Name	Description
Response variable	
% LLTI	The percentage of people in each ward who consider themselves to have a limiting long-term illness
Explanatory variables	
A60P	The percentage of people in each ward that are aged 60 and over
FEMALE	The percentage of people in each ward that are female
UNEM	The percentage of people in each ward that are unemployed (of those Economically active)
% Social Rented	The percentage of people in each ward that are 'social renters' (i.e. rent from the local authority)

In this example, we need to consider:

- Does the model make sense in real world terms?
- Are the assumptions of linear regression met?
- How well do these four explanatory variables explain the variation in the outcome variable?

- Which explanatory variables make the most difference to the outcome variable?
- Are there any areas that have higher or lower than expected values for the outcome?

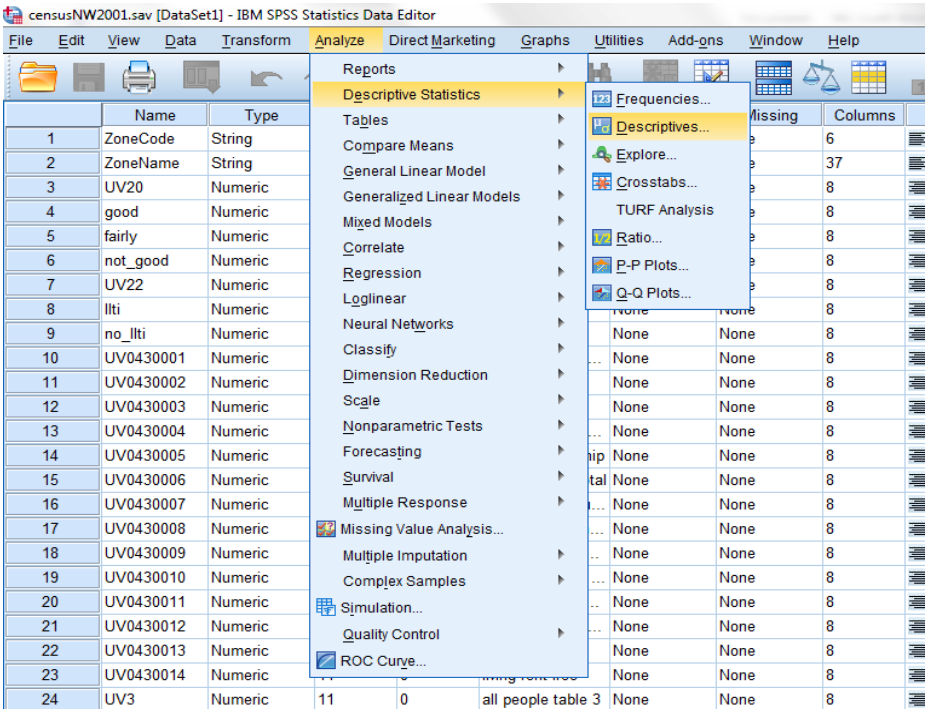
2.2 EXPLORATORY DATA ANALYSIS

The first task in any data analysis is to explore and understand the data using descriptive statistics and useful visualisations. This has two purposes:

1. It will help you to get a feel for the data you are working with;
2. It will inform decisions you make when you carry out more complex analyses (such as regression modelling).

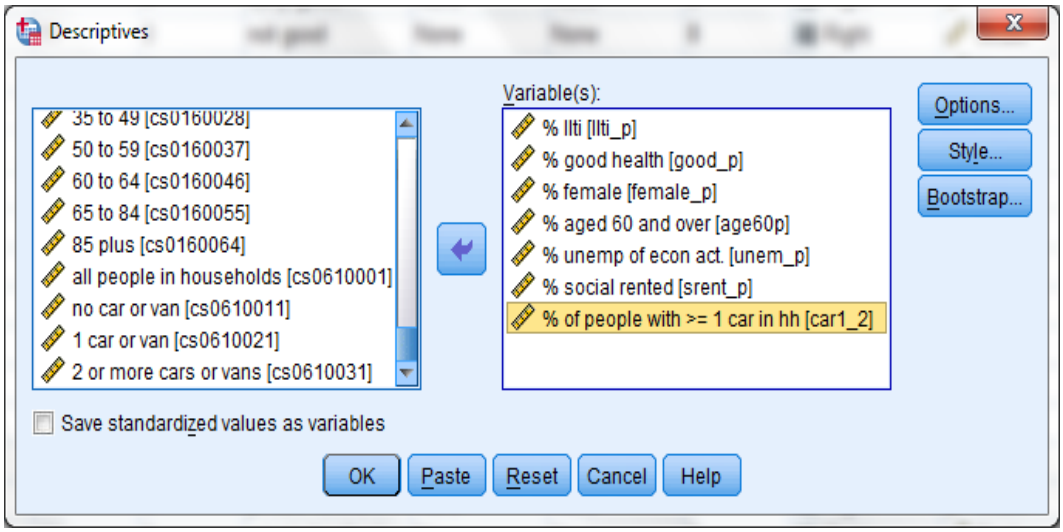
2.2.1 DESCRIPTIVE STATISTICS

SPSS uses a point and click menu-based interface to allow the user to explore the data. These screen shots show the menu selections required and are followed by outputs to show what to expect from an exploratory analysis within SPSS[2]. In the first example, we want descriptive statistics for the variables we are going to use in our model.



[2] Here we are using SPSS version 23. If you are using a different version then the look and feel may be a little different.

This selection opens the following dioalog box.



Clicking on **OK** at this dialog box will prompt SPSS to open an output window in which the following output will appear (Table 3).[3]

Table 3 An example of descriptive statistics output in SPSS

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
% llti	1006	9.26	33.26	20.0436	4.13001
% aged 60 and over	1006	7.24	46.60	21.4374	4.95659
% female	1006	35.18	56.77	51.4180	1.45675
% unemp of econ act.	1006	1.15	24.63	5.3712	3.54237
% social rented	1006	.13	73.89	15.6315	13.90675
Valid N (listwise)	1006				

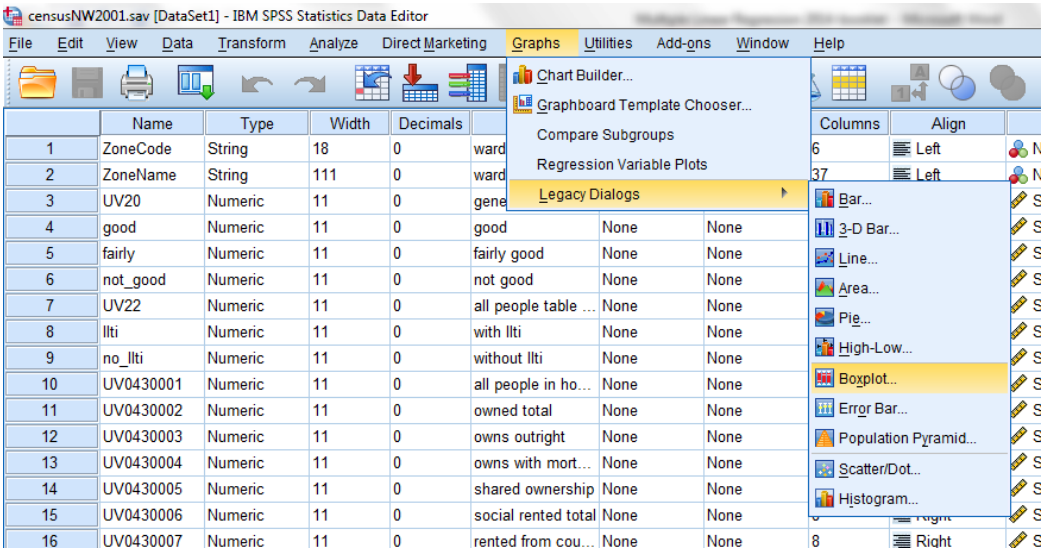
For the purposes of decision-making, we expect to find a reasonable amount of variability in both our explanatory and response variables. A response variable with a low standard deviation would mean there is little to explain; an explanatory variable with little variability

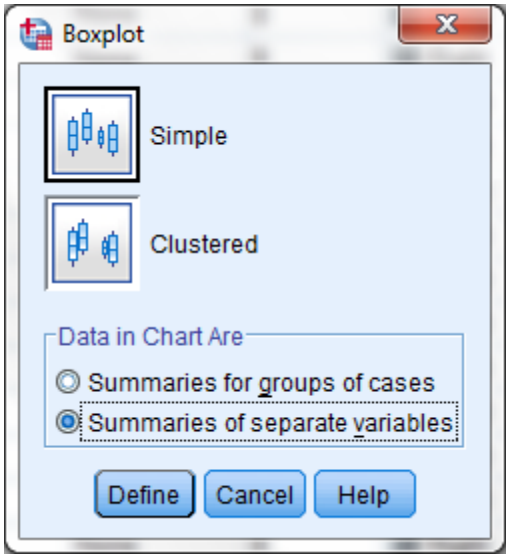
[3] Note that using the Paste button in a dialog box above allows the syntax to be pasted into a script window from which it can be directly edited, saved and run again later. There are numerous online sources for SPSS syntax and it is not intended that this primer covers the writing of syntax.

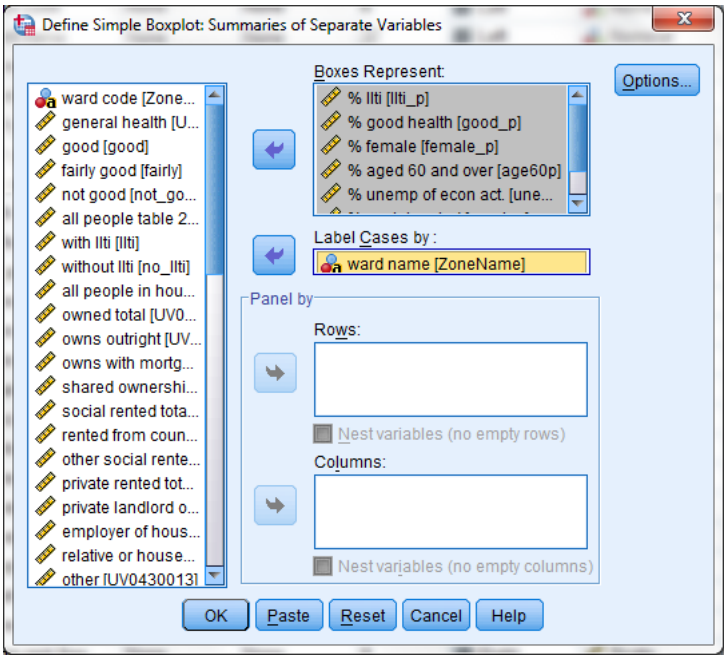
is unlikely to add value to a model. In this case, the variables all look to have sufficient variability with the possible exception of the *%female* variable.

2.2.2 PRODUCING UNIVARIATE BOX PLOTS

A box-plot can be a useful tool for visualising the distribution of a number of variables side by side. To produce these, the simplest approach is as shown below:



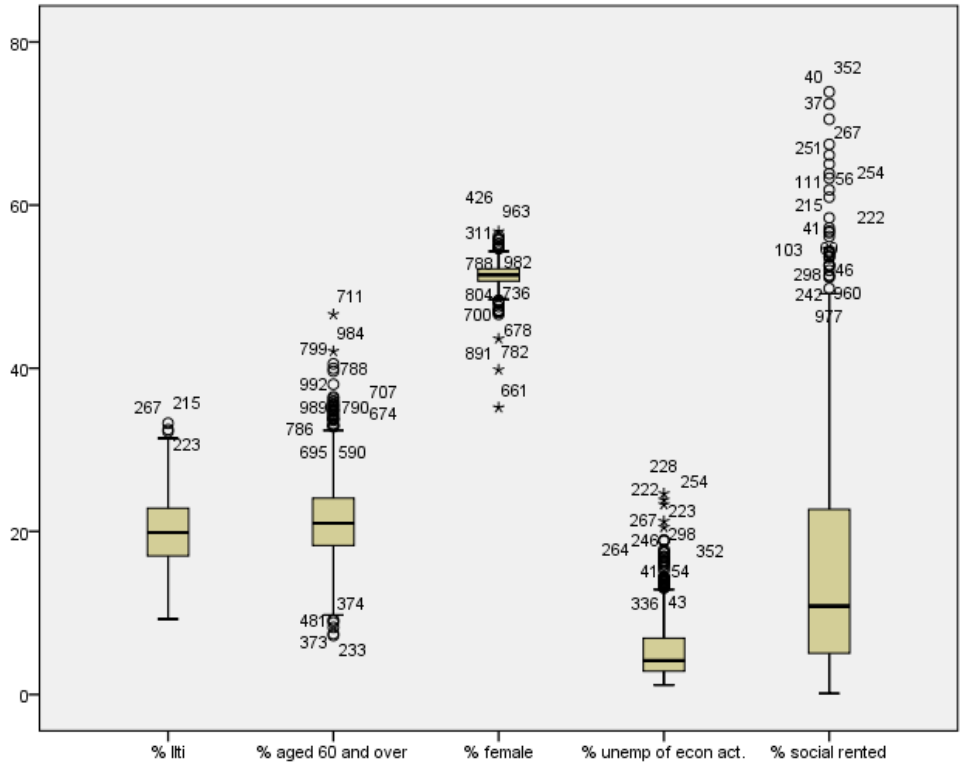




Here we plot the values for each variable. You can see in

Figure 3 that the distribution for each variable is quite different – for example, there are much greater differences between the wards in the *%social renters*, than in *%females*. This is in line with our expectations – we would expect most wards to have a similar gender split, but that poorer areas would have a much higher incidence of social renting.

Figure 3 Box plot of univariate distributions



2.2.3 BIVARIATE CORRELATIONS

SPSS will calculate the Pearson correlation for all pairs of specified variables. Select **Analyze > Correlate > Bivariate** to reach the dialogue box:

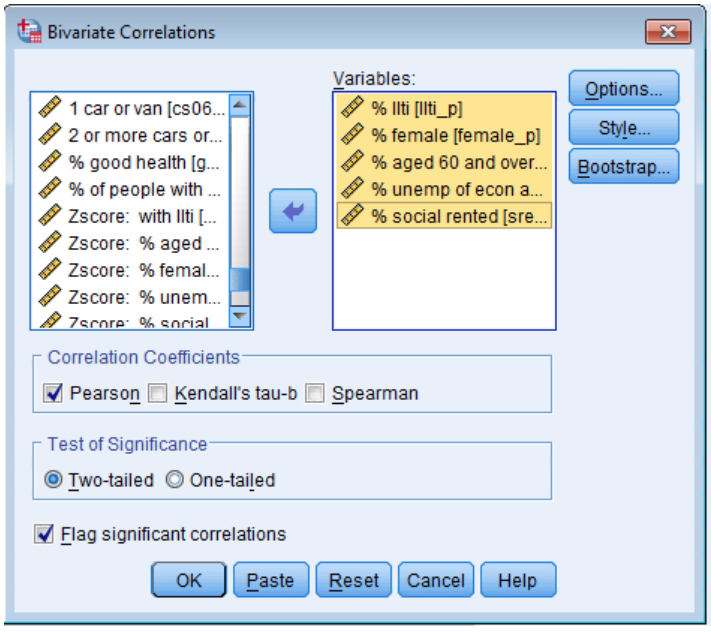


Table 4 shows the SPSS output where the five variables above are selected. The output shows that N = 1006 for all correlations. This tells us that the data are complete and there

are no missing values – in a real life data scenario it is likely that N will differ for each calculated correlation as not all cases will have complete values for every field. *Missing data* is an area for research within itself and there are many methods for dealing with missing data such that a sample remains *representative* and/or any results are *unbiased*. For the purposes of this example, all cases with missing data have been excluded – a somewhat heavy-handed approach but which works well for a worked example and may indeed be appropriate in many analyses.

The two values of the bivariate correlation table:

1. The correlations between your hypothesised explanatory variables and your response variables should be reasonable sized (as a rule of thumb, ignoring the sign of the correlation, they should be >0.15) and statistically significant.
2. The correlations between your explanatory variable should not be too high. We cover this more detail in section 3.5.

Table 4 Pearson Correlations

Correlations

		% llti	% female	% aged 60 and over	% unemp of econ act.	% social rented
% llti	Pearson Correlation	1	.370**	.166**	.693**	.599**
	Sig. (2-tailed)		.000	.000	.000	.000
	N	1006	1006	1006	1006	1006
% female	Pearson Correlation	.370**	1	.259**	.162**	.211**
	Sig. (2-tailed)	.000		.000	.000	.000
	N	1006	1006	1006	1006	1006
% aged 60 and over	Pearson Correlation	.166**	.259**	1	-.320**	-.321**
	Sig. (2-tailed)	.000	.000		.000	.000
	N	1006	1006	1006	1006	1006
% unemp of econ act.	Pearson Correlation	.693**	.162**	-.320**	1	.797**
	Sig. (2-tailed)	.000	.000	.000		.000
	N	1006	1006	1006	1006	1006
% social rented	Pearson Correlation	.599**	.211**	-.321**	.797**	1
	Sig. (2-tailed)	.000	.000	.000	.000	
	N	1006	1006	1006	1006	1006

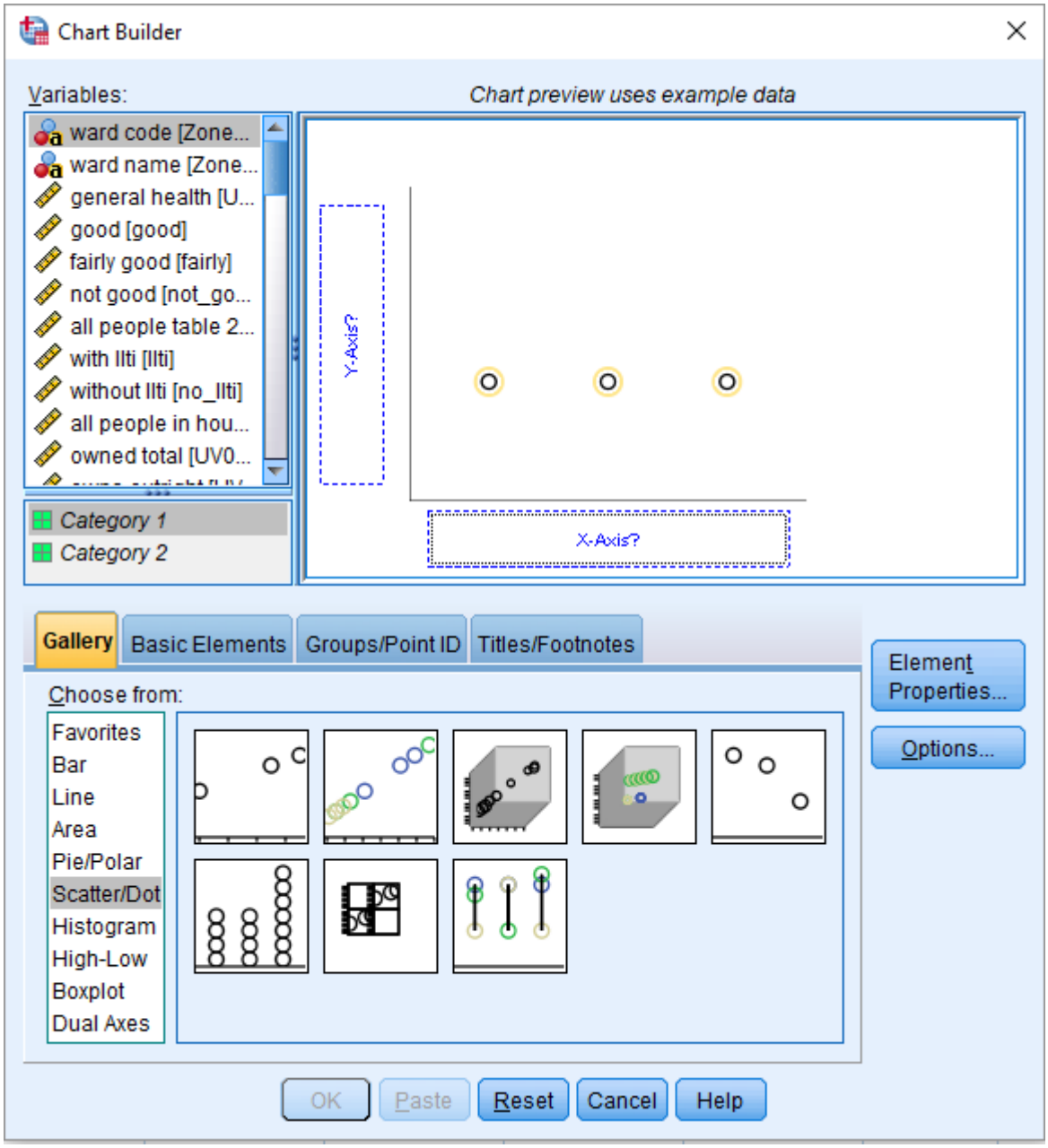
**. Correlation is significant at the 0.01 level (2-tailed).

In this case, the correlations of the explanatory variables with the response variable, apart from age 60 look good enough (according to the criteria set above). We will leave this in consideration now, but will watch out for issues with this variable later. Similarly, the correlation between social rented and unemployment is quite high but not high enough for rejection at this stage.

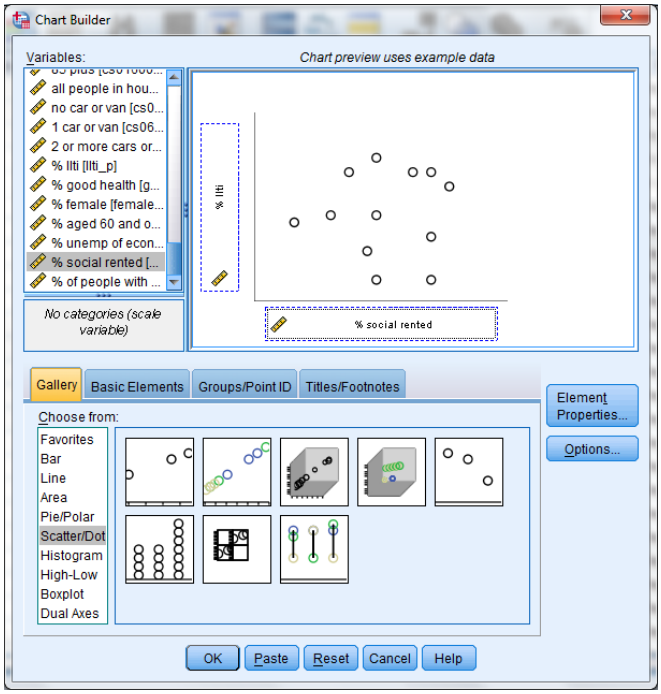
2.2.4 PRODUCING SCATTERPLOTS (IN SPSS)

SPSS will produce scatterplots for pairs of variables. This example shows a scatter plot of the percentage of residents reporting a life limiting illness, against the percentage of residents residing in rented social housing (for example housing association or local authority homes). Use the **Graphs > Chartbuilder** menu path to access the chart builder dialogue box. You may see a warning about setting the measurement level – in this example all of our variables are continuous – that is to say they are numerical and can take any value. Dealing with nominal or categorical variables will be discussed in section 4.

The dialog box is shown below. Select Scatter/Dot and then the top left hand option (simple scatter).

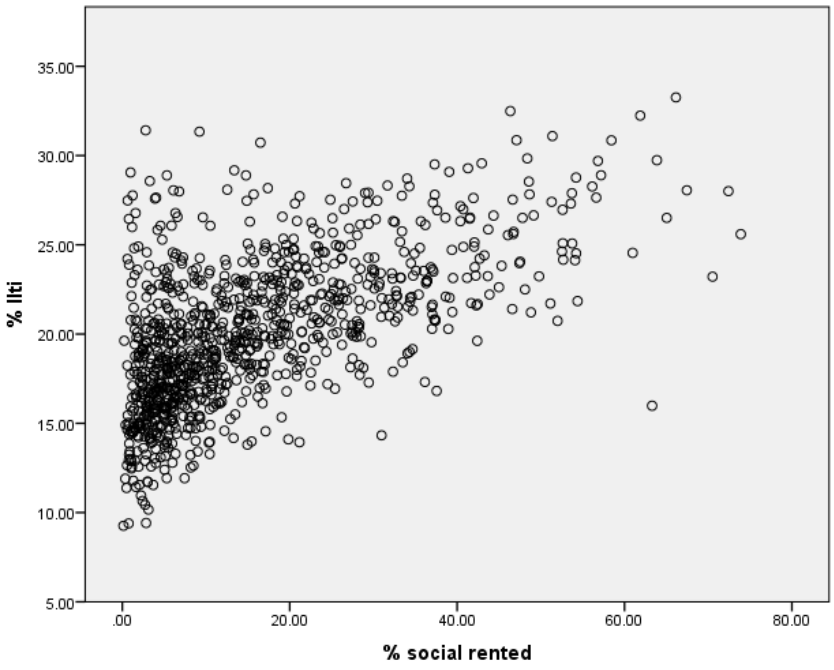


To generate the graph you need to drag the variable names from the list on the left onto the pane on the right and then click **OK**:



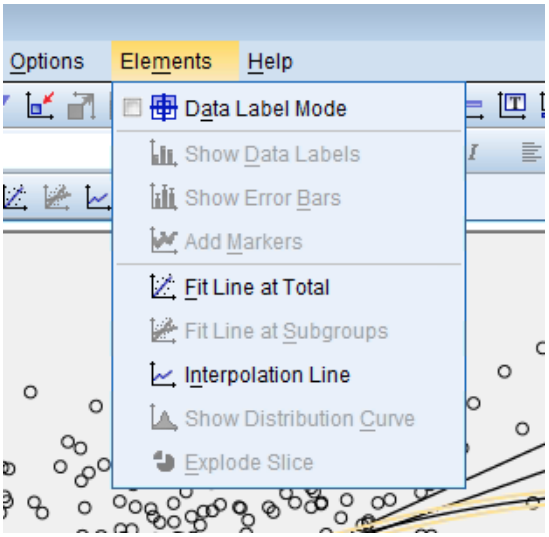
The output should look like Figure 4.

Figure 4 Scatter plot of % llti against % social rented



Double clicking on the graph from the output page will open the graph editor and allow a straight line to be fitted and plotted on the scatterplot as shown in Figure 5.

Choose – **Elements, Fit line, Linear** to fit a simple linear regression line of *% LLTI* on *% social rented.*


Options
Elements
Help
Data Label Mode
Show Data Labels
Show Error Bars
Add Markers
Fit Line at Total
Fit Line at Subgroups
Interpolation Line
Show Distribution Curve
Explode Slice

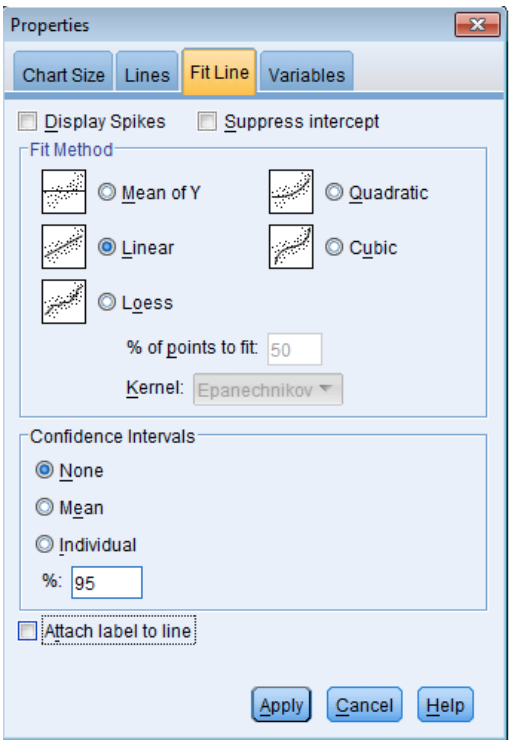
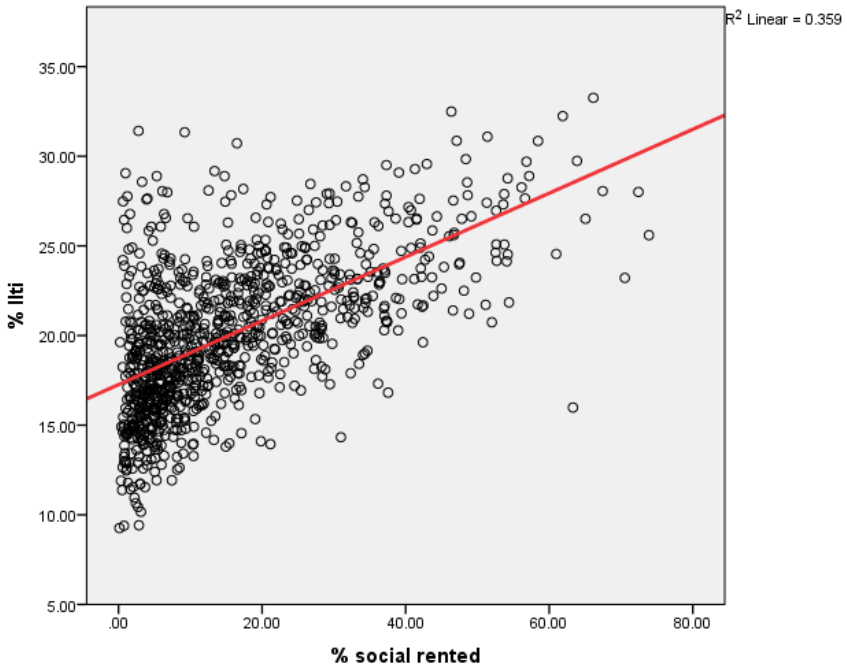

Properties
Chart Size
Lines
Fit Line
Variables
Display Spikes
Suppress intercept
Fit Method
Mean of Y
Quadratic
Linear
Cubic
Loess
% of points to fit:
50
Kernel:
Epanechnikov
Confidence Intervals
None
Mean
Individual
%:
95
Attach label to line
Apply
Cancel
Help

Figure 5 Simple linear regression of %llti by % social rented using graph editor

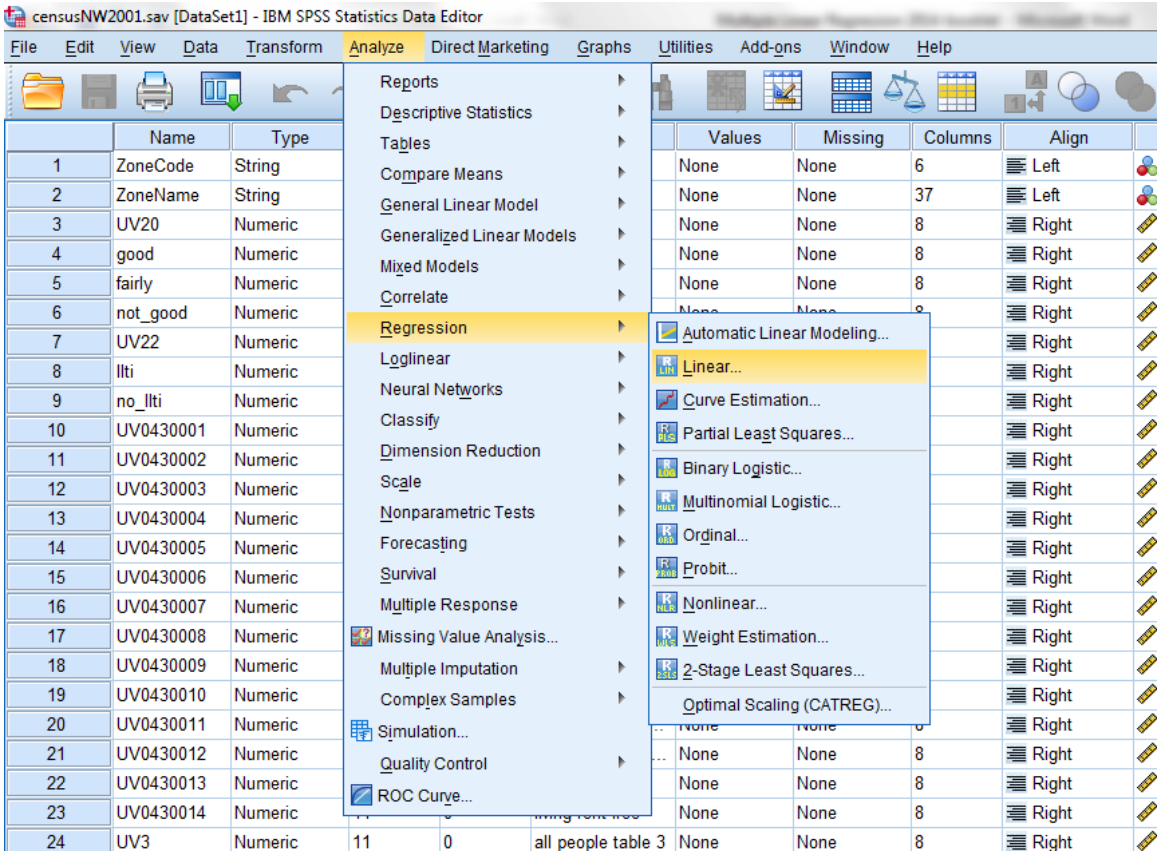


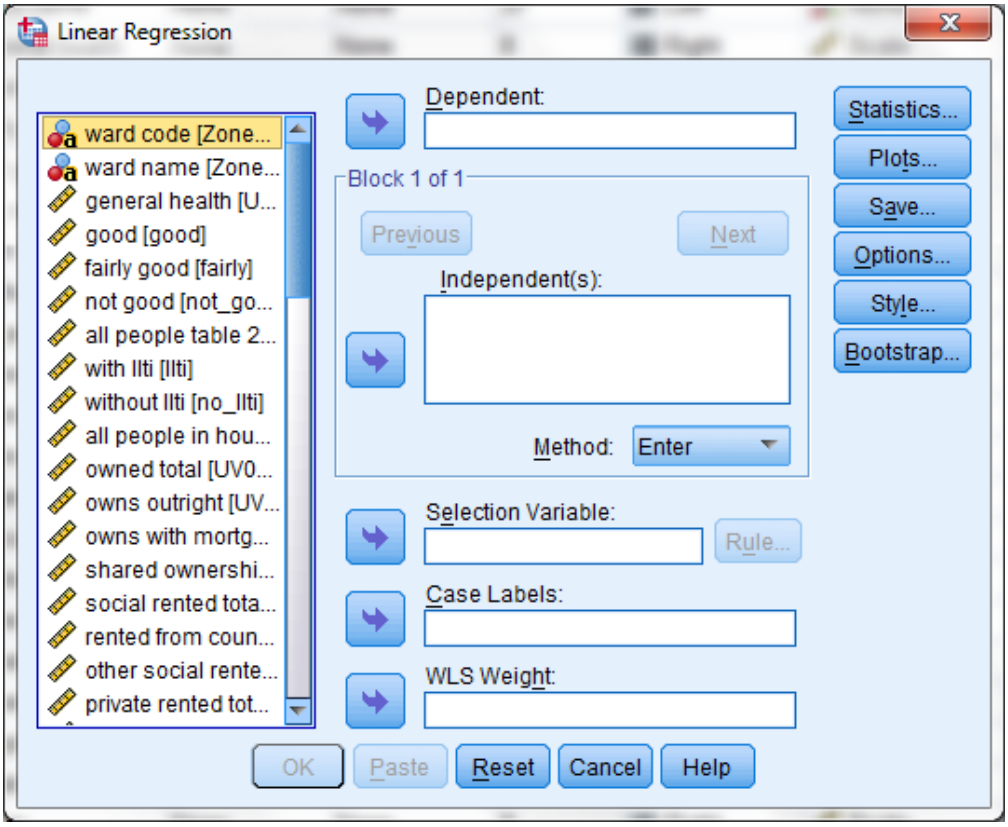
The simple linear regression line plot in Figure 5 shows an R^2 value of 0.359 at the top right hand side of the plot. This means that the variable *% social rented* explains 35.9% of the ward level variation in *% LLTI.* This is a measure of how well our model fits the data – we can use R^2 to compare models, the more variance a model explains, the higher the R^2 value.

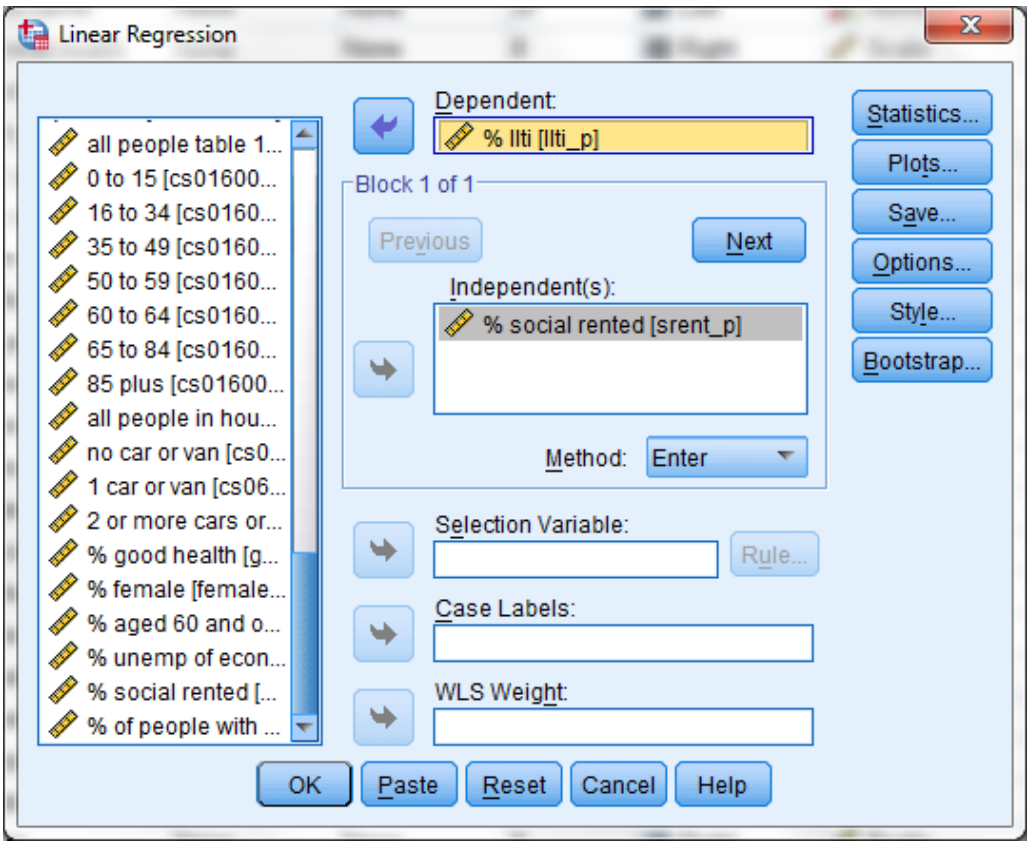
2.3 SIMPLE LINEAR REGRESSION

The linear regression line plotted in Figure 5 through the graph editor interface can be specified as a model.

Our response variable is *%llti* and for a simple linear regression we specify one explanatory variable, *% social rented*. These are selected using the **Analyze > Regression > Linear menu path.**


censusNW2001.sav [DataSet1] - IBM SPSS Statistics Data Editor
File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Neural Networks
Classify
Dimension Reduction
Scale
Nonparametric Tests
Forecasting
Survival
Multiple Response
Missing Value Analysis...
Multiple Imputation
Complex Samples
Simulation...
Quality Control
ROC Curve...
Automatic Linear Modeling...
Linear...
Curve Estimation...
Partial Least Squares...
Binary Logistic...
Multinomial Logistic...
Ordinal...
Probit...
Nonlinear...
Weight Estimation...
2-Stage Least Squares...
Optimal Scaling (CATREG)...
Name Type Values Missing Columns Align
1 ZoneCode String None None 6 Left
2 ZoneName String None None 37 Left
3 UV20 Numeric None None 8 Right
4 good Numeric None None 8 Right
5 fairly Numeric None None 8 Right
6 not_good Numeric
7 UV22 Numeric
8 llti Numeric
9 no_llti Numeric
10 UV0430001 Numeric
11 UV0430002 Numeric
12 UV0430003 Numeric
13 UV0430004 Numeric
14 UV0430005 Numeric
15 UV0430006 Numeric
16 UV0430007 Numeric
17 UV0430008 Numeric
18 UV0430009 Numeric
19 UV0430010 Numeric
20 UV0430011 Numeric
21 UV0430012 Numeric None None 8 Right
22 UV0430013 Numeric None None 8 Right
23 UV0430014 Numeric None None 8 Right
24 UV3 Numeric 11 0 all people table 3 None None 8 Right


Linear Regression
ward code [Zone...
ward name [Zone...
general health [U...
good [good]
fairly good [fairly]
not good [not_go...
all people table 2...
with llti [llti]
without llti [no_llti]
all people in hou...
owned total [UV0...
owns outright [UV...
owns with mortg...
shared ownershi...
social rented tota...
rented from coun...
other social rente...
private rented tot...
Dependent:
Block 1 of 1
Previous
Next
Independent(s):
Method: Enter
Selection Variable:
Rule...
Case Labels:
WLS Weight:
Statistics...
Plots...
Save...
Options...
Style...
Bootstrap...
OK
Paste
Reset
Cancel
Help



2.3.1 REGRESSION OUTPUTS

The output for a model within SPSS contains four tables. These are shown as separate Tables here with an explanation of the content for this example.

Table 5 Variables entered

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	% social rented[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: % llti

Table 5 confirms that the response variable is *% llti* and the explanatory variable here is % social rented. The model selection 'method' is stated as 'Enter'. This is the default and is most appropriate here. More about "methods" later!

Table 6 Model Summary

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.599[a]	.359	.359	3.30724

a. Predictors: (Constant), % social rented

Table 6 is a summary of the model fit details. The adjusted R^2 figure [4]is 0.359 – the same as we saw in Figure 5 showing that the model explains 35.9% of the variance in the % of life limiting illness reported at a ward level.

Table 7 ANOVA table

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6160.641	1	6160.641	563.240	.000[a]
	Residual	10981.604	1004	10.938		
	Total	17142.244	1005			

a. Predictors: (Constant), % social rented

b. Dependent Variable: % llti

ANOVA stands for Analysis of Variance; SPSS produces an ANOVA table as part of the regression output as shown in Table 7. The variance in the data is divided into a set of components. The technical background to an ANOVA table is beyond the scope of this primer. We look mainly at the Sig. column, which tells us the p-value for the R^2 statistic. If this is greater than 0.05 then the whole model is not statistically significant and we need to stop our analysis here. The value here is below 0.05 and so we can say that the fit of the model as a whole is statistically significant.

[4] In SPSS, both R^2 and "adjusted" R^2 are quoted. For large sample sizes, these two figures are usually very close. For small values of n, the figure is adjusted to take account of the small sample size and the number of explanatory variables and so there may be a difference. The technical details of the adjustment are beyond the scope of this primer. The adjusted figure should be used in all instances.

Table 8 Model parameters

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	17.261	.157		109.999	.000
	% social rented	.178	.008	.599	23.733	.000

a. Dependent Variable: % llti

The estimated model parameters are shown in the Coefficients table (Table 8). The B column gives us the β coefficients for the prediction equation.

To best understand this table it helps to write out the model equation. Remember:

$$y_i = \beta_0 + \beta_1 x_i + e_i$$

Substituting the variables and results of our regression analysis gives:

$$\widehat{\%\ llti} = \beta_0 + \beta_1(\%\ social\ rented)$$

So:

$$\widehat{\%\ llti} = 17.261 + 0.178(\%\ social\ rented)$$

The ^ over the *%lltii* indicates that this is a predicted value rather than the actual value (and therefore we don't need the error term).

2.3.1.1 INTERPRETING THE RESULTS

In our example, for every 1% increase in the percentage of people living in social rented housing in a ward, we expect a 0.178% increase in the percentage of people living with a life limiting illness in that same ward. The relationship is positive – areas with more social tenants have greater levels of long-term illness.

For a ward with no social tenants, we expect 17.261% illness as this is the intercept – where the line of best fit crosses the y-axis.

Again, we must be careful to remember that this statistically significant model describes a relationship but does not tell us that living in socially rented accommodation, causes life limiting illnesses. In fact, those people reporting illness in each ward may not even be the same people who report living in social housing as the data are held at a ward, rather than person level. Instead, an increase in social tenants may indicate that a ward has higher levels of people with lower incomes and higher levels of poverty. There is a significant body of literature that links poverty with illness, so this does make substantive sense.

2.3.2 STANDARDISED COEFFICIENTS

The unstandardised coefficients shown in Table 8 can be substituted straight into the theoretical model. The issue with these is that they are dependent on the scale of measurement of the explanatory variables and therefore cannot be used for comparison – bigger does not necessarily mean more important. The standardised coefficients get round this problem and relate to a version of the model where the variables have been standardised to fit a normal distribution with a mean of zero and a standard deviation of 1. We interpret the standardised coefficients in terms of standard deviations.

For this model, for one standard deviation change in the *% of social renters* in a ward, there is a 0.599 standard deviation change in the % of people reporting a life limiting illness.

The descriptives table we produced in SPSS (Table 3) tells us that the standard deviation of social tenancy is 13.9% and the standard deviation of the outcome variable is 4.13%. So for a 13.9% change in social tenancy, there is a (4.13*0.599) change in illness – 2.47%. This is the same as a change of 0.178% for a 1% increase in social tenancy[5].

2.3.3 STATISTICAL SIGNIFICANCE

The table of the coefficients (Table 8) shows that both intercept and slope (β_0 and β_1) are *statistically significant*.

The parameters are estimates drawn from a distribution of possible values generated by SPSS when computing the model – the true value for each parameter could in fact fall anywhere within its distribution. The standard error of the estimate shows us the spread of this distribution, and the Sig. column tells us whether or not these values are statistically different from zero.

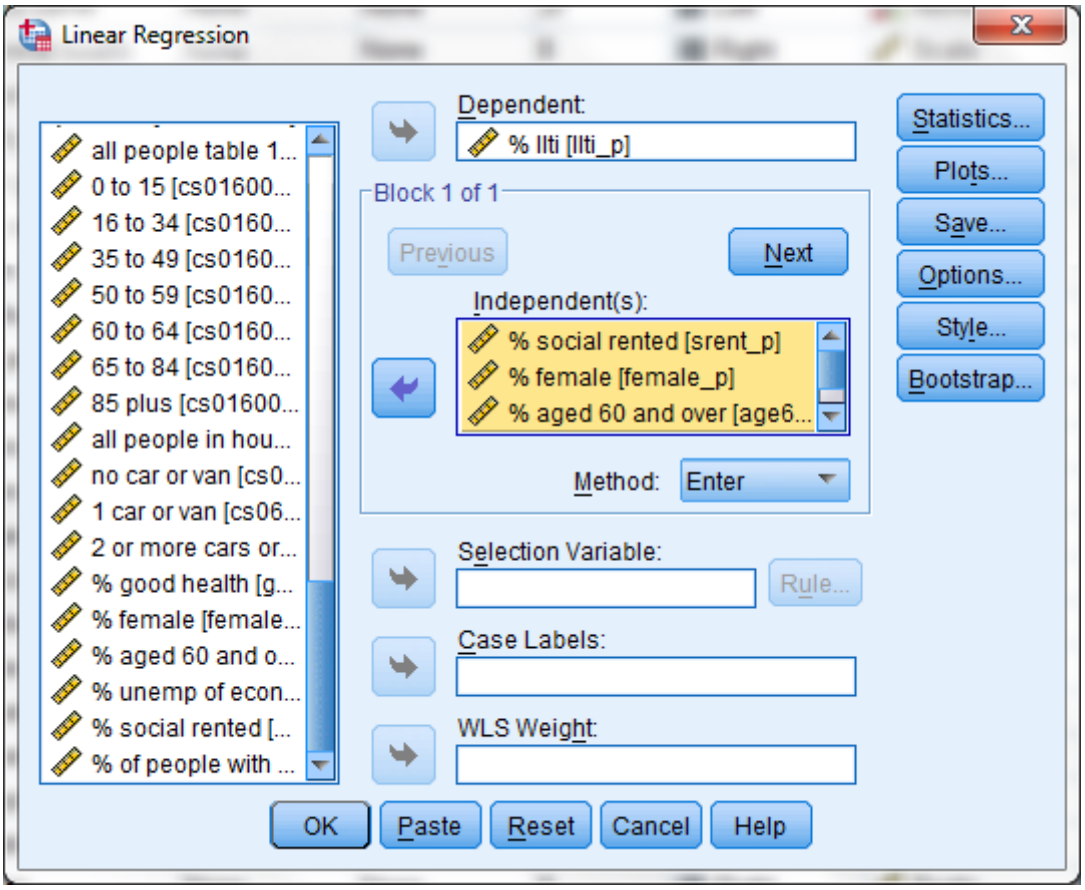
If these values are not statistically different from zero, then the true value sits within a distribution which includes zero within the 95% confidence bounds. If the estimate for the parameter could be zero, then it could be that there is in fact no relationship – a zero coefficient and a flat line of best fit.

A value which is not statistically significant is indicated by a p-value greater than 0.05 (the Sig. column). For this model, p <0.05 and so we can say that the estimates of the parameters are statistically significant and we can infer that there is an association between the variables.

[5] 2.47% / 13.9 % = 0.178, the unstandardised value for β_1

2.4 MULTIPLE LINEAR REGRESSION ANALYSIS

Adding additional explanatory variables to a simple linear regression model builds a multiple linear regression model. The process is identical within SPSS – including additional variables in the specification stages. This example includes the percentage of females, the percentage of over 60s and the percentage of unemployed economically active residents as additional explanatory variables, over the simple regression using just the percentage of social tenants.



2.4.1 MORE ON METHODS – 'ENTER'

This worked example is a case of a deductive model. A deductive model is one that is built on real world understanding of the problem to be modelled and is grounded in theory – often drawn from existing understanding or published literature.

Here we are interested in the levels of life limiting illness in different areas. We have a theory that poverty is linked with life limiting illnesses, and that differences in age and gender may play a part. We have a dataset that contains variables which are related to this theory and so we build a model that reflects our theory.

For the default method is 'Enter', the order of the explanatory variables is not important. The method uses all the specified explanatory variables, regardless of whether or not they turn out to be statistically significant.

Other methods are covered later in this primer.

2.4.2 REGRESSION OUTPUTS

Including the extra variables has increased the adjusted R^2 value from 0.395 to 0.675. This means 67.5 % of the variation in percentage LLTI is now explained by the model – a large improvement. The ANOVA table (Table 11) shows that the model is a statistically significant fit to the data.

Table 9 Variables

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	% aged 60 and over, % female, % unemp of econ act., % social rented[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: % llti

Table 10 Model Summary

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.823[a]	.677	.675	2.35344

a. Predictors: (Constant), % aged 60 and over, % female, % unemp of econ act., % social rented

Table 11 ANOVA

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	11598.023	4	2899.506	523.501	.000[a]
	Residual	5544.221	1001	5.539		
	Total	17142.244	1005			

a. Predictors: (Constant), % aged 60 and over, % female, % unemp of econ act., % social rented

b. Dependent Variable: % llti

Table 12 Tables of coefficients (sometimes called Model parameter values)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-9.832	2.734		-3.596	.000
	% unemp of econ act.	.774	.035	.664	22.147	.000
	% female	.344	.056	.121	6.176	.000
	% social rented	.052	.009	.175	5.728	.000
	% aged 60 and over	.336	.017	.404	19.762	.000

a. Dependent Variable: % llti

2.4.3 INTERPRETING THE RESULTS

From Table 12 we can see that all of the explanatory variables are statistically significant. So our theory that these variables are related to long-term limiting illness rates is supported by the evidence.

All the β coefficients are positive – which tells us that an increase in the value of any of the variables leads to an increase in long term limiting illness rates.

From the information in Table 12, we can now make a prediction of the long term limiting illness rates for a hypothetical ward, where we know the values of the explanatory variables but don't know the long term limiting illness rate.

Say that in our hypothetical ward that the unemployment rate is 18%, females are 45% of the population, social tenancy is at 20%, and 20% of the population are aged 60 and over. The general form of the model is:

$$\begin{aligned} \%\, llti = \; & \beta_0 + \beta_1(\%\, unemployed) \\ & + \beta_2(\%\, female) \\ & + \beta_3(\%\, social\ rented) \\ & + \beta_4(\%\, age\ 60\ and\ over) \\ & + \varepsilon_i \end{aligned}$$

Substituting the values from Table 12 gives us:

$$\begin{aligned} \widehat{\%llti} = \; & -9.832 + 0.774 \times (\%\, unemployed) \\ & + 0.344 \times (\%\, female) \\ & + 0.052 \times (\%\, social\ rented) \\ & + 0.336 \times (\%\, age\ 60\ and\ over) \end{aligned}$$

This would give a predicted value for our hypothetical ward of 27.3%:

$$\begin{aligned} \widehat{\%llti} = \; & -9.832 + 0.774 \times 18 \\ & + 0.344 \times 45 \\ & + 0.052 \times 20 \\ & + 0.336 \times 20 = 27.3 \end{aligned}$$

We can also use Table 12 to examine the impact of an older population in a ward as a single variable. If we leave all other variables the same (sometimes called "holding all other variables constant"), then we can see that an increase of 1% in the proportion of the population that is over 60 leads to a 0.336% increase in the predicted value of long term limiting illness rate (i.e. the precise value of the B coefficient). Another way of saying this is to say this is "controlling for employment, gender and social tenancy rates, a 1 unit increase in the percentage of people over sixty leads to 0.336 unit increase in long term limiting illness rates". This simple interpretability is one of the strengths of linear regression.

3 THE ASSUMPTIONS OF LINEAR REGRESSION

OK so we have just shown the basics of linear regression and how it is implemented in SPSS. Now we are going to go a bit deeper. In this section we will consider some of the assumptions of linear regression and how they affect the models that you might produce.

To interpret a model and its limitations, it is important to understand the underlying assumptions of the method and how these affect the treatment of the data and modelling choices made.

When we use linear regression to build a model, we assume that:

- The response variable is continuous and the explanatory variables are either continuous or binary.
- The relationship between outcome and explanatory variables is linear
- The residuals are homoscedastic

- The residuals are normally distributed
- There is no more than limited multicollinearity
- There are no external variables – that is variable that are not included in the model that have strong relationships with the response variable (after controlling for the variables that are in the model).
- Independent errors
- Independent observations.

For most of these assumptions, if they are violated then it does not necessarily mean we cannot use a linear regression method, simply that we may need to acknowledge some limitations, adapt the interpretation or transform the data to make it more suitable for modelling.

3.1 ASSUMPTION 1: VARIABLE TYPES

The most basic assumption of a linear regression is that the response variable is continuous. The normal definition of continuous is that it can take any value between its minimum and its maximum. Two useful tests for continuity are:

1. Can you perform meaningful arithmetic on the numbers on the scale?
2. Can you meaningfully continuously subdivide the numbers on the scale into infinitely small parts?

In many cases these two tests are clear cut but there is a certain class of variables called count variables which pass test 1 but the result of test 2 is ambiguous and depends in part on the meaning of the variable. For example, number of cigarettes smoked is usually OK to treat as continuous whereas number of cars in a household is not.

Binary variables are indicators of whether feature is present or whether something is true or false not they are usually coded as 1 – the feature is present/true and 0 the feature absent/false.

Variables which are not binary or continuous can be used in a regression model if there are first converted into Dummy variables (see section 4.1)

3.2 ASSUMPTION 2: LINEARITY

Linear regression modelling assumes that the relationship between outcome and each of the explanatory variables is linear[6], however this may not always be the case.

3.2.1 CHECKING FOR NON-LINEAR RELATIONSHIPS

Non-linear relationships can be difficult to spot. If there are just two variables, then a curve in the data when looking at a two-way scatter plot may indicate a non-linear relationship. However, non-linear relationships can be hidden, perhaps because of complex dependencies in the data; a curve or even a cubic shape, in the scatter plot of residuals may also indicate that there are non-linear effects.

3.2.2 MODELLING A NON-LINEAR RELATIONSHIP, USING LINEAR REGRESSION

We can take account of a non-linear relationship into a linear regression model through a neat trick. By transforming the explanatory variable into something that does have a linear relationship with the outcome and entering that transformed variable into our model we can maintain the assumption of linearity.[7]

For example, there may be a curve in the data, which is better represented by a quadratic rather than a linear relationship.

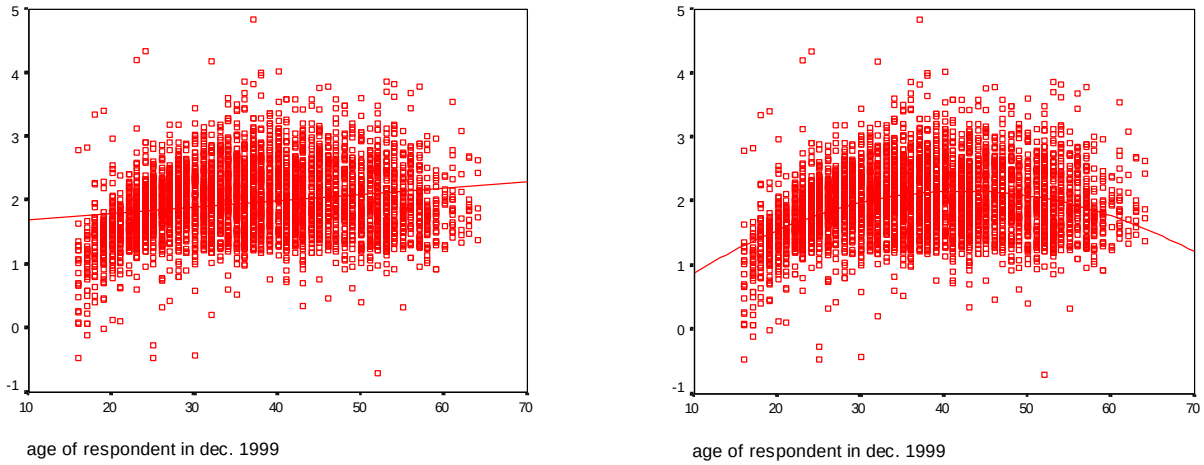
Figure 6 shows the log of hourly wage by age for a sample of respondents. In the left hand plot a straight line of best fit is plotted. In the right hand plot, we can see that a curved line looks to the naked eye to be a much more sensible fit. We, therefore, propose that there is a quadratic relationship between the log of pay per hour, and age. This means that the log of pay per hour and age squared are linearly related.

---

[6] i.e. in the sense that it conforms to a straight line. It might seem slightly odd as a curve is also a line but when statisticians refer to "linear", they mean straight, everything else is "non-linear". See https://study.com/academy/lesson/how-to-recognize-linear-functions-vs-non-linear-functions.html for further discussion.

[7] This may seem a little confusing; since we have added in non-linear predictors why is the model still referred to as a *linear* regression model? The reason is that the linearity here refers to the model not the data. The term linear regression denotes an equation in which the effect of each parameter in the model is simply additive (but the parameters themselves could represent non-linear relationships in the data). See: https://blog.minitab.com/blog/adventures-in-statistics-2/what-is-the-difference-between-linear-and-nonlinear-equations-in-regression-analysis for more details.

Figure 6 Scatterplots of log of hourly wage, by age



To account for this non-linear relationship in our linear model, we need to compute a new variable – the square of age (here called agesq where agesq = age^2). If there is a statistically significant quadratic relationship between hourly wage and age, then the model should contain a statistically significant linear coefficient for age squared which we can then use to make better predictions.

The general form of model for the linear relationship would be:

$$Ln(Hourly\ Wage)_i = \beta_0 + \beta_1 age_i + \varepsilon_i$$

The model for the quadratic relationship would be:

$$Ln(Hourly\ Wage)_i = \beta_0 + \beta_1 age_i + \beta_2 agesq_i + \varepsilon_i$$

Note that we have retained the linear component in the model. This is generally regarded as best practice regardless of the significance of the linear component. In this case the left hand graph in Figure 6 does indicate that there is a linear component.

3.3 ASSUMPTION 3: NORMAL DISTRIBUTION OF RESIDUALS

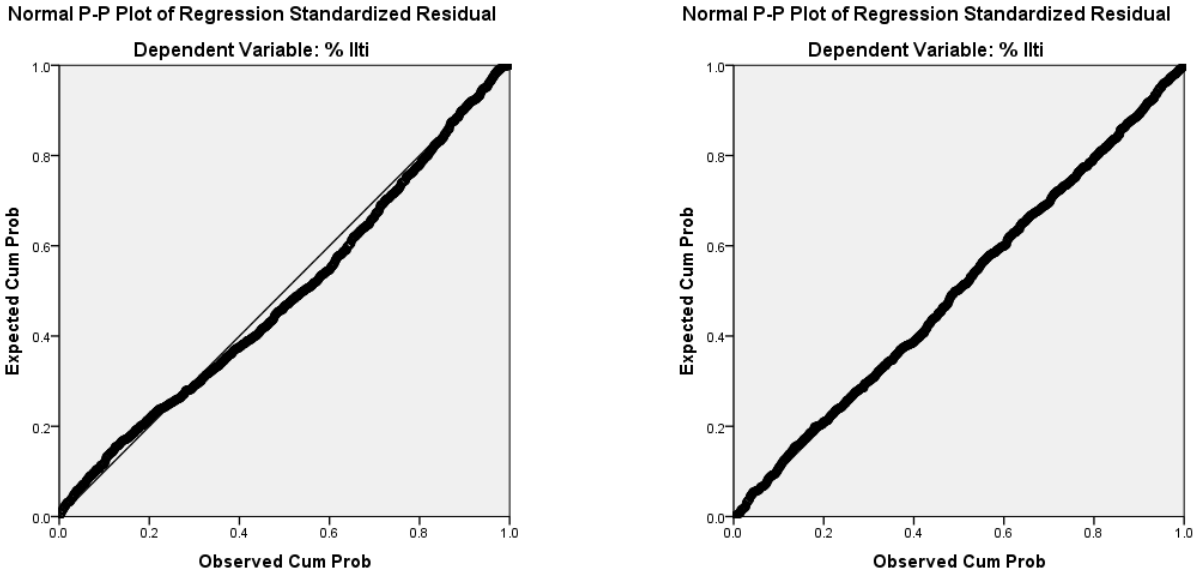
3.3.1 P-P PLOTS

We can assess the assumption that the residuals are normally distributed by producing a P-P plot[8]through the regression dialogue box.

[8] This is sometimes referred to as a normal probability plot or a quantile-quantile or q-q plot.

The ordered values of the standardised residuals are plotted against the expected values from the standard normal distribution. If the residuals are normally distributed, they should lie, approximately, on the diagonal.

Figure 7 P-P plots for the simple linear regression (left – Table 8) and multiple linear regression (right Table 12) examples



In Figure 7, the left hand example shows the plot for the simple linear regression and the right hand plot shows the multiple linear regression. We can see that the line deviates from the diagonal on the left plot, whereas in the right hand example the line stays more closely to the diagonal.

This makes substantive sense – our multiple linear regression example explains much more of the variance and therefore there are no substantively interesting patterns left within the residuals and they are normally distributed. In our simple linear regression, we are missing some important explanatory variables – there is unexplained variance and this shows in the residuals where the distribution deviates from normal.[9]

3.3.2 HISTOGRAMS OF RESIDUALS

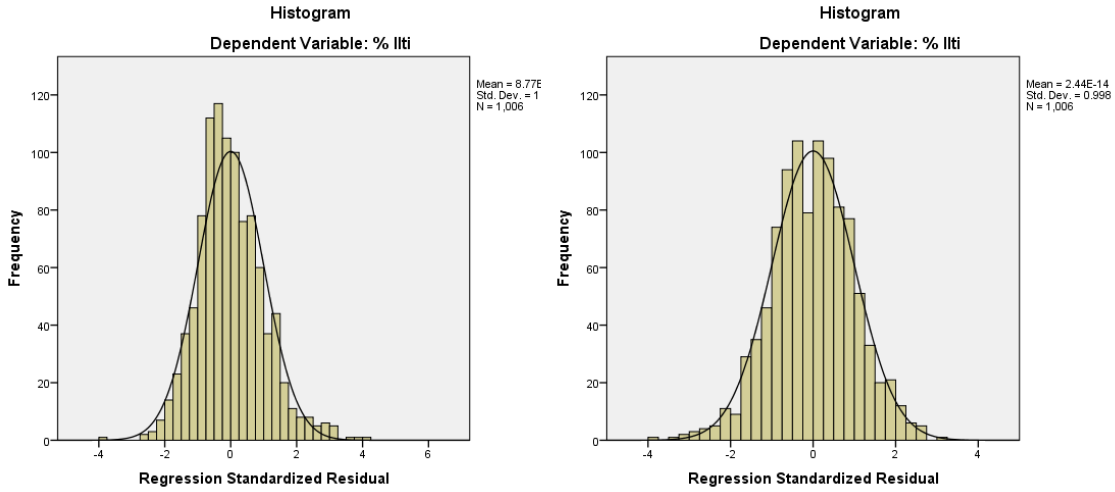
If we plot the standardised residuals for our two regression examples with histograms we can see that both examples follow approximately a normal distribution (Figure 8). The left

[9] Note that the reverse is not necessarily true. Normally distributed residuals does not imply that you have no missing (or extraneous) variables.

hand example is our simple linear regression and the right hand example is the multiple linear regression. The multiple linear regression example here has residuals that follow the normal distribution more closely.

We could use technical tests for normality such as the Shapiro-Wilk or Kolmogorov-Smirov statistics; however, these are beyond the scope of this primer.[10]

Figure 8 Histogram of standardised residuals for simple regression (left, Table 8) and multiple regression (right, Table 12)



3.4 ASSUMPTION 4: HOMOSCEDASTICITY

Homoscedasticity refers to the distribution of the residuals or error terms. If this assumption holds then the error terms have constant variance – in other words, the error for each observation does not depend on any variable within the model. Another way of saying this is that the standard deviation of the error terms are constant and do not depend on the explanatory variable values.

3.4.1 CHECKING FOR HOMOSCEDASTICITY OF THE RESIDUALS

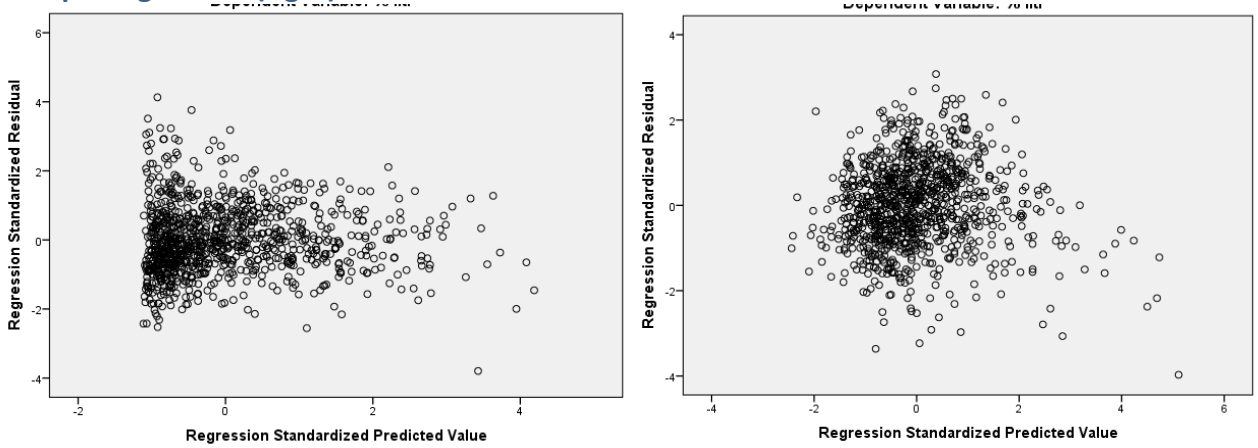
Plotting the residuals against the explanatory variables is a useful method for visually checking whether or not the residuals are homoscedastic. The scatter plot should look like random noise – no patterns should be visible.

[10] See Field (2017) for more details.

Figure 9 shows a plot of the standardised residuals against the standardised predicted values for the response variable measuring long term illness from our ward level multiple linear regression (right) and the simple linear regression (left) examples.

Also, we should plot the saved residuals against any of the other variables in the analysis to assess on a variable-by-variable basis wherever there is any dependency in the residuals on the variables in the analysis (there should not be).

Figure 9 Plotting residuals to check homoscedasticity for simple regression (left) and multiple regression (right)



The left hand plot shows a clear cone shape typical of heteroscedasticity. The right hand plot shows a more random noise type pattern, indicating homoscedastic residuals.

In this case, the left hand plot refers to a simple linear regression with only one explanatory variable. There are still patterns in the variance which have not been explained and this is seen in the residuals.[11]

The right hand plot includes more variables and there are no discernible patterns within the variance: these residuals look to be meeting the assumption of homoscedasticity.

3.4.2 WHAT TO DO IF THE RESIDUALS ARE NOT HOMOSCEDASTIC AND WHY DOES IT MATTER

[11] Another way to think about this is that the model is only addressing part of the distribution of the response variable.

Some models are more prone to displaying heteroscedasticity, for example if a data set has extreme values. A model of data collected over a period of time can often have heteroscedasticity if there is a significant change in the outcome variable from the beginning to the end of the collection period.

Heteroscedasticity therefore arises in two forms. The model may be correct, but there is a feature of the data that causes the error terms to have non-constant variance such as a large range in values. Alternatively, the model may be incorrectly specified so there is some unexplained variance due to the omission of an important explanatory variable and this variance is being included in the error terms.

When the problem is the underlying data, the β coefficients will be less precise as a result and the model fit may be overstated.

For an incorrectly specified model, introducing additional explanatory variables may solve the problem. For an underlying data issue, removing outliers may help, or it may be appropriate to transform the outcome variable – possibly using a standardised form of the variable to reduce the range of possible values.

3.5 ASSUMPTION 5: MULTICOLINEARITY

When two of the explanatory variables in a model are highly correlated (and could therefore be used to predict one another), we say that they are *collinear*.

In our model, it may be that these variables are actually representing the same societal factors which influence rates of illness - we can investigate this by removing one of the variables and producing an alternative model.

When there are collinear variables, the model can become unstable – this is often indicated by the standard error around the estimation of the β coefficients being large and the coefficients being subject to large changes when variables are added or deleted form the model. The model cannot distinguish between the strength of the different effects and one of the assumptions of linear regression is violated.

Signs that there is multicollinearity include:

- β coefficients which are not significant, even though the explanatory variable is highly correlated with the outcome variable.
- β coefficients which change radically when you add or remove a variable from the model.
- β coefficients which are in the opposite direction to your expectation based on theory –a negative coefficient when you expect a positive relationship.

- High pairwise correlation between explanatory variables.

3.5.1 TESTING FOR COLINEARITY - CORRELATIONS

By carrying out a correlation analysis before we fit the regression equations, we can see which, if any, of the explanatory variables are very highly correlated and identify any potential problems with collinearity.

If we refer back to the Pearson correlations that we produced in **Error! Reference source not found.** we note that the unemployment and social tenancy variables were correlated with a Pearson coefficient of 0.797. What is meant by a "high level of correlation" is somewhat subjective, here we apply a rule of thumb that any correlation over $|0.7|$ is considered high. Where a pair of variables are highly correlated, it may be worth considering removing one of them from the analysis.

We can remove one of the variables and investigate the effect. Using the same example, we remove the unemployment variable and check the model fit.

Table 12: Model summary

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.720[a]	.518	.517	2.87129

a. Predictors: (Constant), % aged 60 and over, % female, % social rented

Removing the unemployment variable produces a model that explains 51.5% of the variance in illness rates. This is 16% less than for when the variable is included so we can conclude that this variable is useful for the model – despite being highly correlated with social tenancy. The parameters of the model are given in Table 13.

Table 13: Model parameters

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-9.127	3.336		-2.736	.006
	% female	.384	.068	.135	5.648	.000
	% social rented	.203	.007	.683	27.952	.000
	% aged 60 and over	.292	.021	.350	14.165	.000

a. Dependent Variable: % llti

3.5.2 TESTING FOR COLLINEARITY – VARIANCE INFLATION FACTOR

The variance inflation factor (or VIF) of a linear regression gives us an idea of how much the variance of the regression estimates has been increased because of multicollinearity. This is easily calculated in SPSS as part of the model outputs. As a rule of thumb, if the VIF values are greater than 10, then multicollinearity may be a problem,

The VIF values can be generated as part of the regression output in the coefficients table – see section 3.6.2

3.5.3 COLLINEARITY – WHAT TO DO

The simplest method for dealing with collinearity is to remove the variable in question from the model as in the example in 3.5.1.

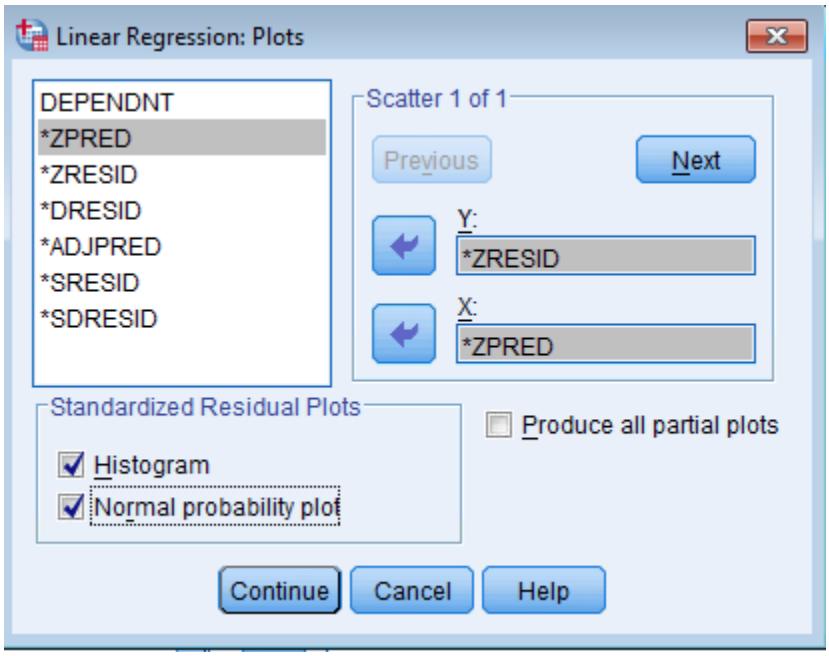
Alternatively, the variables of interest can be reduced in dimensionality by using a technique such as principal component analysis.[12]

3.6 CHECKING THE ASSUMPTIONS OF LINEAR REGRESSION WITH SPSS

3.6.1 REQUESTING PLOTS

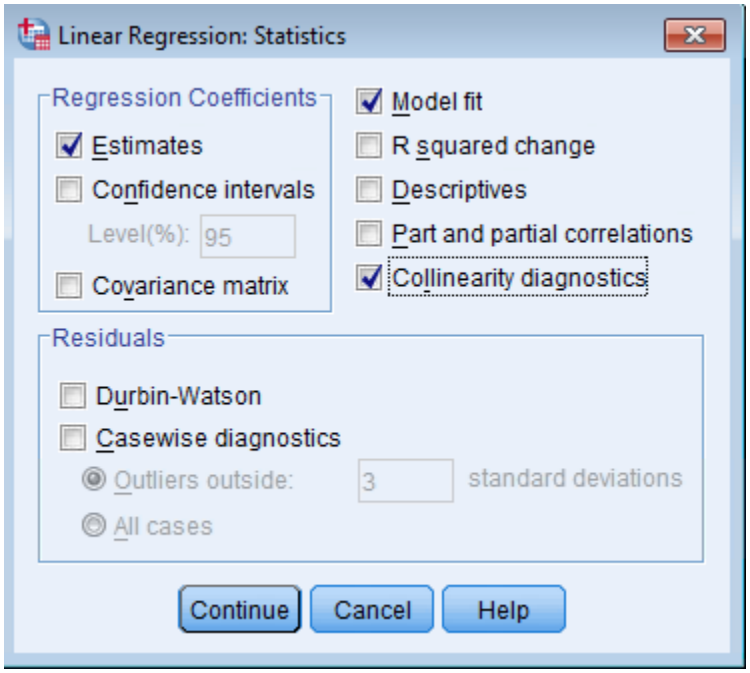
From the regression dialogue box, select Plots. From here, requesting a scatter plot of predicted values by variance, and the standardised residual plots will provide the three key visualisations used to assess the assumptions of linear regression as part of the regression output.

[12] See for example Field (2017) or Hair et al (2010) for discussion of this method.



3.6.2 CALCULATING VARIANCE INFLATION FACTORS

From the regression dialogue box select **Statistics** to open the dialogue for requiring VIF. Tick the Collinearity diagnostics checkbox and exit.



When the regression analysis is run, the VIFs form part of the output. Table shows our example multiple linear regression output with the additional information. We can see in this example that there are no variables which cause concern.

Table 14: Coefficients of a model with variance inflation factors

Coefficients[a]

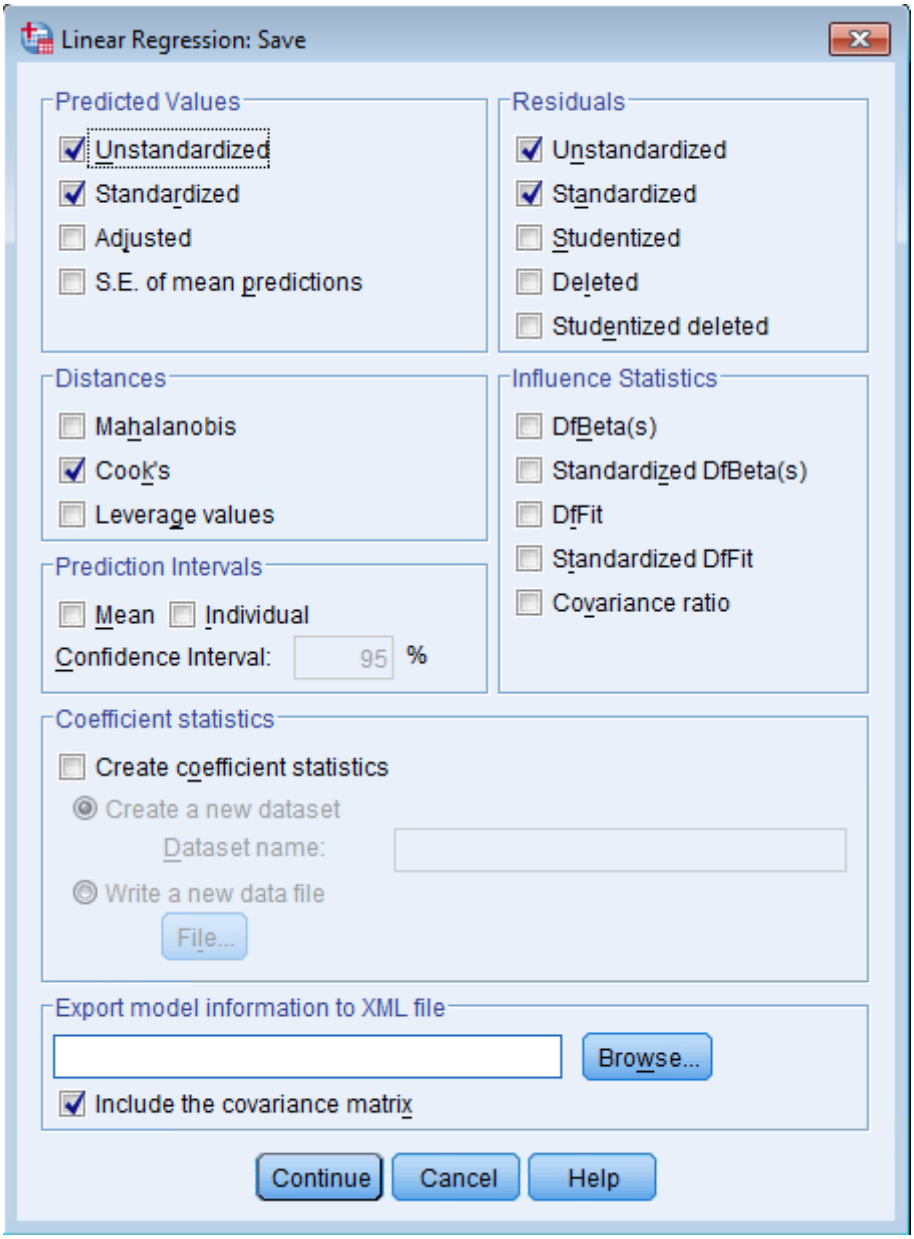
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	-9.832	2.734		-3.596	.000		
	% social rented	.052	.009	.175	5.728	.000	.348	2.876
	% female	.344	.056	.121	6.176	.000	.836	1.196
	% aged 60 and over	.336	.017	.404	19.762	.000	.775	1.291
	% unemp of econ act.	.774	.035	.664	22.147	.000	.360	2.778

a. Dependent Variable: % llti

3.7 SAVING REGRESSION VALUES

Select Save from the regression dialogue box. Here we can request that predicted values and residuals are saved as new variables to the dataset. We can also save the Cook's distance for each observation.

In this example, we have saved the unstandardised and standardised residuals and predicted values, and the Cook's distance.



New variables are added to the dataset:

pre_1 = unstandardised predicted

res_1 = unstandardised residual

zpr_1 = standardised predicted

zre_1 = standardised residual

coo_1 = Cook's Distance

Further model specifications save as separate variables with the suffice _2 and so on.

3.8 EXTREME VALUES

A large residual means that the actual value and that predicted by the regression model are very different.

Extreme values seen on a scatter plot of residuals suggests that there is a sample unit which needs to be checked, as a rule of thumb, a standardised residual of magnitude 3 or greater should be investigated.

When this occurs it is worth considering:

- Is the data atypical of the general pattern for this sample unit?
- Is there a data entry error?
- Is there a substantive reason why this outlier occurs?
- Has an important explanatory variable been omitted from the model?

Some times in a regression analysis it is sensible to remove such outliers from the data before refining the model. An outlier will have a disproportionate effect on the estimations of the β parameters because the least squares method minimises the squared error terms – and this places more weight on minimising the distance of outliers from the line of best fit. This in turn can move the line of best fit away from the general pattern of the data.

When an outlier has an influence like this, it is described as having *leverage* on the regression line. In this example, in the simple model there are many residuals that have a magnitude greater than 3. This is further evidence that important explanatory variables have been omitted. In the multiple regression model there are very few points of concern and all of those are only just over the threshold, so no need to examine any of the wards for removal from the analysis.

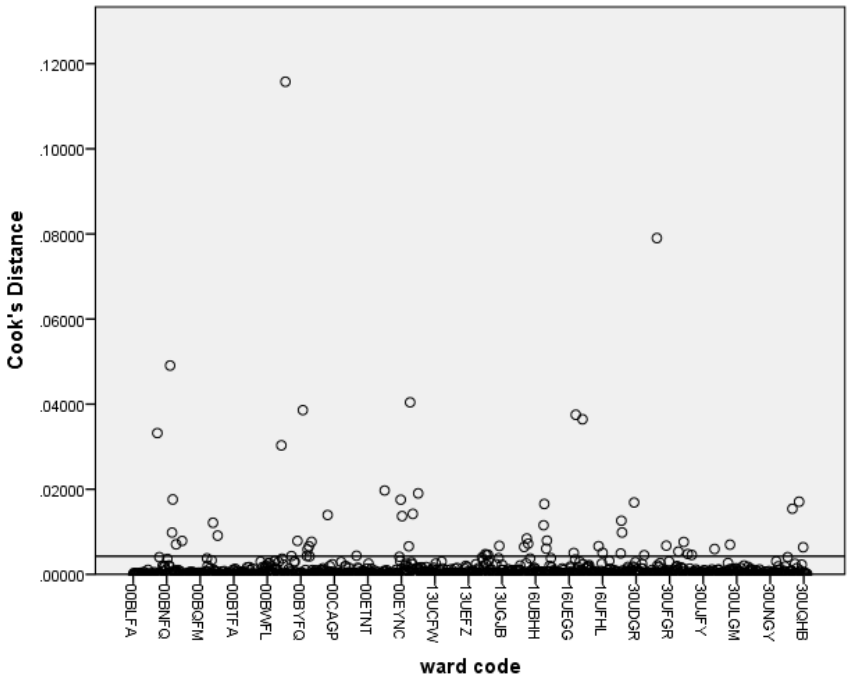
3.8.1 COOK'S DISTANCE

To assess outliers we can plot a scatter plot for inspection or another approach is to calculate the Cook's distance for each observation. This can be specified as part of the regression and is saved as an extra variable in the dataset. The Cook's distance is a measure of the change in the predicted values, if the observation is removed. Any value with a distance of larger than three times the mean Cook's distance might be an outlier.

In our multiple regression model example, if we save the Cook's distances and visualise them by ward (as in Figure 10Figure 10 Cook's distance by ward code) we can see that there are several values that breach the threshold (which is typically 3 times the mean of the Cook's distances in this case marked as a horizontal line at around y= 0.004).Two cases in particular have very high Cook's distances; these may be worth investigating as outliers.[13]

[13] The Breusch-Pagan test is a further analysis where the outcome variable is the squared residual. The explanatory variables are the same as for the model in question. This regression generates a test statistic for a

Figure 10 Cook's distance by ward code



4 MOVING TO A MORE COMPLEX MODEL

4.1 NOMINAL VARIABLES

Up to this point, our models have included only *continuous variables*. A continuous variable is numeric, and can take any value. In our examples, the value has had a minimum of zero but actually, mathematically, it wouldn't have mattered if the values extended into negative numbers – although this would not have made sense in the real world.

A *nominal* or *unordered categorical variable* is one where the possible values are separate categories but are not in any order.

Consider a survey that asks for a participant's gender, and codes the answers as follows:

1. Male
2. Female

χ^2 test where the null hypothesis is homoscedasticity. This test is not available through the menu interface in SPSS but can be run using a readily available macro. The technical details of the test and the method for executing it through SPSS are beyond the scope of this primer. Note that a function exists within both python and R for automating the test.

3. Trans gender
4. Non binary

Each case within the data would have a numerical value for gender. If we were to use this number within a linear regression model, it would treat the value for gender of a non-binary respondent as four times the value for gender of a male. This doesn't make sense and we could have listed the answers in any order resulting in them being assigned a different number within the dataset; the numerical codes are arbitrary.

The variable is not continuous but our theory may still be that the outcome variable is affected by gender so we want to include it in the model. To do this we construct a series of *dummy variables*. Dummy variables are binary variables constructed out of particular values of a nominal variable.

We need (n-1) dummy variables where n is the number of possible responses/categories. In this example, we are using the 'Male' response as our reference category and therefore all of the dummy coefficients are interpreted as comparisons to the male case. We have 4 possible responses so need 3 dummy variables.

This means that when the value of all of the dummy variables is zero, the prediction we make using the regression equation is for a male. Table 13 shows the values for the three new dummy variables against the original question for gender.

If D_female = 1, and all other dummies are zero, then we are predicting for a female. If D_trans = 1 and all other dummies are zero, we are predicting for a transgender person and so on.

Table 13 Creating dummy variables

Gender response	D_female	D_trans	D_nb
Male	0	0	0
Female	1	0	0
Transgender	0	1	0
Non-binary	0	0	1

Remembering the earlier model for exam results, if we had a theory that gender could also be used to predict the age 16 results, we might include it as follows:

$$exam16_i = \beta_0 + \beta_1(D_{female}) + \beta_2(D_{trans}) + \beta_3(D_{nb}) + \beta_4(exam11) + \varepsilon_i$$

For a male, the equation collapses to:

$$exam16_i = \beta_0 + \beta_4(exam11) + \varepsilon_i$$

because all of the dummy variables take a value of zero.

For a female:

$$exam16_i = \beta_0 + \beta_1 + \beta_4(exam11) + \varepsilon_i$$

because D_nb and D_trans are equal to zero, and D_female is equal to 1.

4.2 INTERACTION EFFECTS

An interaction effect is when the relationship between an outcome variable and an explanatory variable, changes, based on another explanatory variable.

Going back to our sample of exam results, let's say that we know the sex of the students. For this example, we will assume that sex is binary and we have only males and females in the sample.

We are trying to predict the age 16 scores, using the age 11 scores and the sex of the student. There are four possible outcomes for our modelling work.

4.2.1 SCENARIO A: SAME SLOPE, SAME INTERCEPT

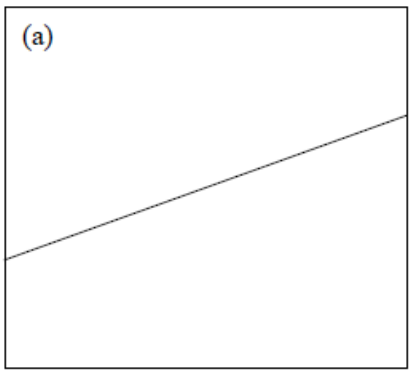
The relationship between exam16 and exam11 is identical for boys and girls – sex is not significant, and there is no interaction effect.

Our model for scenario A is the same as in the earlier section on simple linear regression:

$$exam16_i = \beta_0 + \beta_1 exam11_i + e_i$$

There is no difference between boys and girls so there is no term for sex in the equation.

Figure 11 Same slope, same intercept



4.2.2 SCENARIO B: DIFFERENT INTERCEPT, SAME SLOPE

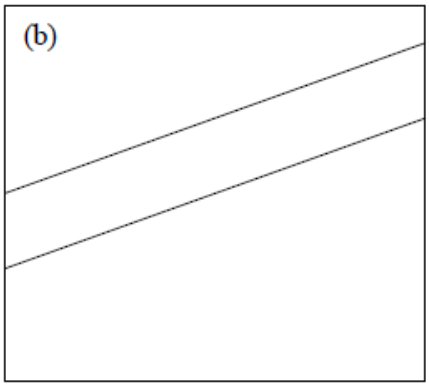
Here the relationship between exam16 and exam11 has a different intercept for boys than girls but the nature of the relationship (the slope) is the same for boys and for girls. This means that boys on average do differently to girls at age 11 and age 16, but the change in the scores between the two ages is the same regardless of sex.

In scenario (b) the slopes are the same but there is an overall difference in the average exam scores. We need a dummy variable to represent sex – let's say that if sex = 0 for a male and sex = 1 for a female.

$$exam16_i = \beta_0 + \beta_1 exam11_i + \beta_2 Sex_i + e_i$$

There are two separate lines for girls and boys, but they are parallel.

Figure 12 Same slope, different intercept



4.2.3 SCENARIO C: DIFFERENT INTERCEPT, DIFFERENT SLOPES

The relationship between exam16 and exam11 has a different intercept and a different slope for boys and girls. The line with the lower intercept but steeper slope might refer to boys and the line with the higher intercept and shallower slope to girls.

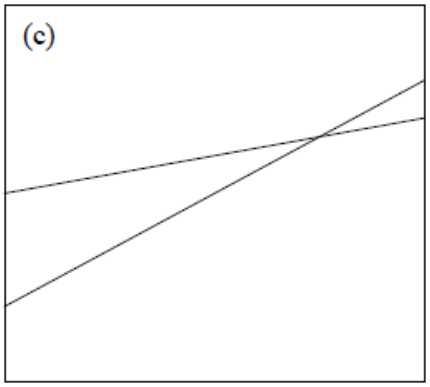
This is an interaction effect.

The difference in intercept is modelled by including a term for sex as in scenario b. The difference in slope is modelled by inclusion of an interaction term. To do this we simply create an extra variable that is the product of the two variables we wish to interact, and including that new variable in the model.

For every case, we multiply the exam11 score by the Sex dummy variable and compute this into a new variable, here called exam11Sex.

$$exam16_i = \beta_0 + \beta_1 exam11_i + \beta_2 Sex_i + \beta_3 exam11Sex_i + e_i$$

Figure 13 Different slope, different intercept



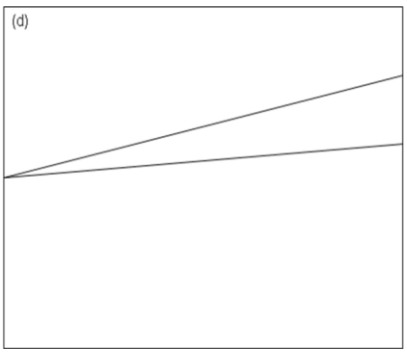
4.2.4 SCENARIO D: DIFFERENT SLOPE, SAME INTERCEPT

The slope is different for girls and boys but the intercept is identical. In this graph one of the lines would refer to girls, and the other line to boys. This means that boys and girls do the same (on average) at age 11 but that the different sexes progress differently between age 11 and age 16.

In scenario (d) we have different slopes, but the same intercept for the two sexes[14]. The equation for the line is the same as scenario c, but β_2 is zero so the model equation collapses to:

$$exam16_i = \beta_0 + \beta_1 exam11_i + \beta_3 exam11Sex_i + e_i$$

Figure 14 Same intercept, different slope



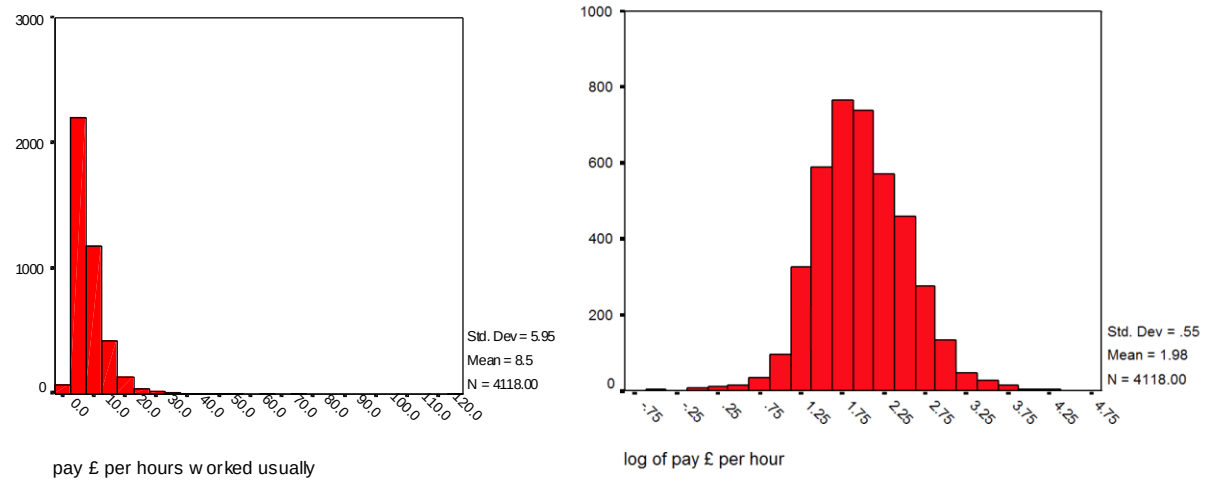
[14] Note that this is a theoretical possibility. In practice, this will rarely happen and when building models one should by default include all the main effects for all of the variables in an interaction term as this improves model stability.

4.3 TRANSFORMING A VARIABLE

Variables do not need to be normally distributed to be used within a linear regression; however, the assumptions of linear regression are sometimes more easily met when the response variable conforms to a normal, or near normal distribution.

The distribution of income is often subject to significant skew and is bounded at zero. This is because a few people earn a very high salary and it is not possible to have a negative wage. Figure 15 shows a histogram of hourly pay with significant positive skew on the left hand side, and the result of taking the log of this variable as a histogram on the right hand side. We can see that by taking the natural log of the hourly wage, the distribution becomes closer to normal.

Figure 15 Histograms of hourly pay (left) and log of hourly pay (right)



Another common transformation is to standardise the data. To standardise a variable we subtract the mean, and divide by the standard deviation. This gives a distribution with a mean of zero and a standard deviation of 1.

The SPSS menu and dialogue boxes for transforming variables are shown in section 4.5.

4.4 MORE MODEL SELECTION METHODS – BEYOND THE DEFAULT

The default method within SPSS linear regression is the **enter** method.

In the enter method, a substantive theory based model is built, including all explanatory variables considered relevant based on the research question, previous research, real-world understanding and the availability of data.

When there are a large number of explanatory variables, we might use statistical criteria to decide which variables to include in the model and produce the "best" equation to predict the response variable.

Two examples of such selection methods are discussed here; *backwards elimination* and *stepwise selection*. Even with these automatic methods, inclusion of many variables without a robust theory underlying why we think they may be related risks building spurious relationships into our model. We may build a good predictive model, but if this is based upon spurious correlations, we do not learn anything about the problem our research is trying to address.

4.4.1 BACKWARDS ELIMINATION

Begin with a model that includes all the explanatory variables. Remove the one that has the highest p-value. Refit the model, having removed the least significant explanatory variable, remove the least significant explanatory variable from the remaining set, refit the model, and so on, until some 'stopping' criterion is met: usually that all the explanatory variables that are included in the model are significant.

4.4.2 STEPWISE

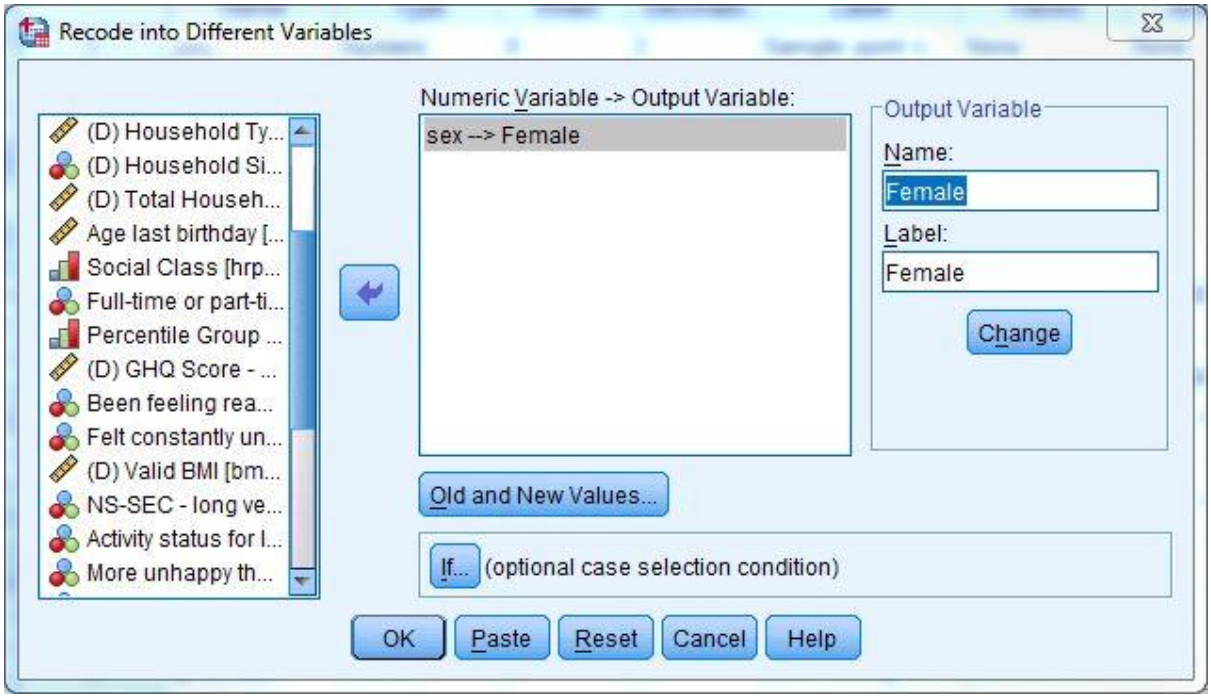
This is more or less the reverse of backward elimination, in that we start with no explanatory variables in the model, and then build the model up, step-by-step. We begin by including the variable most highly correlated to the response variable in the model. Then include the next most correlated variable, allowing for the first explanatory variable in the model, and keep adding explanatory variables until no further variables are significant. In this approach, it is possible to delete a variable that has been included at an earlier step but is no longer significant, given the explanatory variables that were added later. If we ignore this possibility, and do not allow any variables that have already been added to the model to be deleted, this model building procedure is called forward selection.

4.5 SPSS SKILLS FOR MORE ADVANCED MODELLING

4.5.1 RECODING INTO A DUMMY VARIABLE

Use the **Transform > Recode into different variables** menu path to open the recode dialogue box.

Here you can select the variable to recode, and specify the name and label of the new 'output variable'. Then click on **Change** to see variable within the **Numeric Variable -> Output Variable** box.

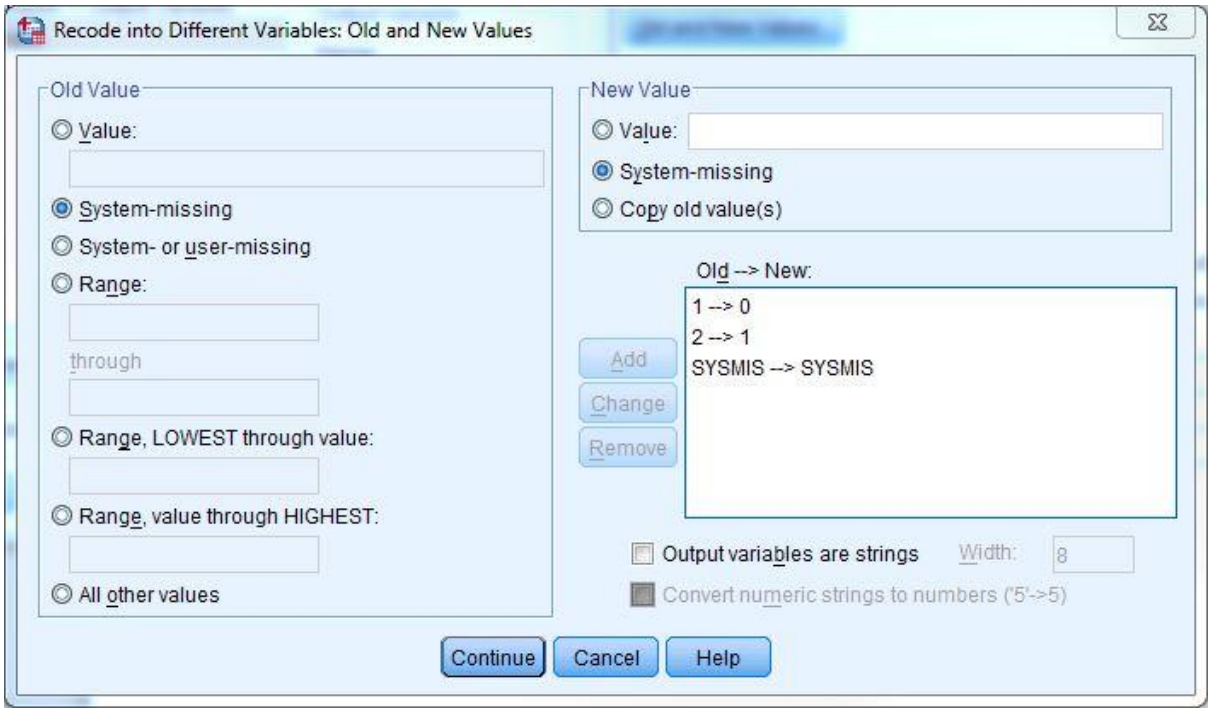


Click on **Old and new Values** to open the next dialogue box for specifying the recode. In this example, we have selected sex and are recoding into a dummy variable called "Female". The previous and new codings are shown in Table 15.

Table 14 Recoding sex to a dummy variable

Previous code (Sex)	New code (Female)	Meaning
1	0	Male
2	1	Female

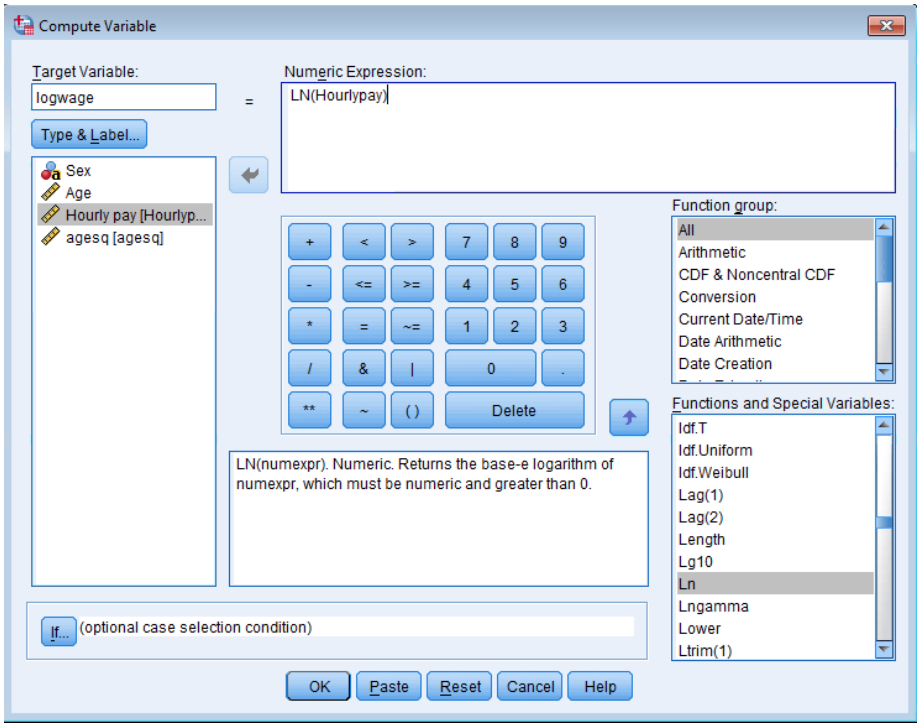
Specify each old and new value and then click **Add** to generate the list of recodings. In this dataset, the variable sex was binary and so only a few lines of recoding are needed (see below) but a variable with more categories would need many values recoding to zero, and multiple dummies. Also adding a recode of System Missing to System Missing ensures that values coded as missing within the data retain that coding.



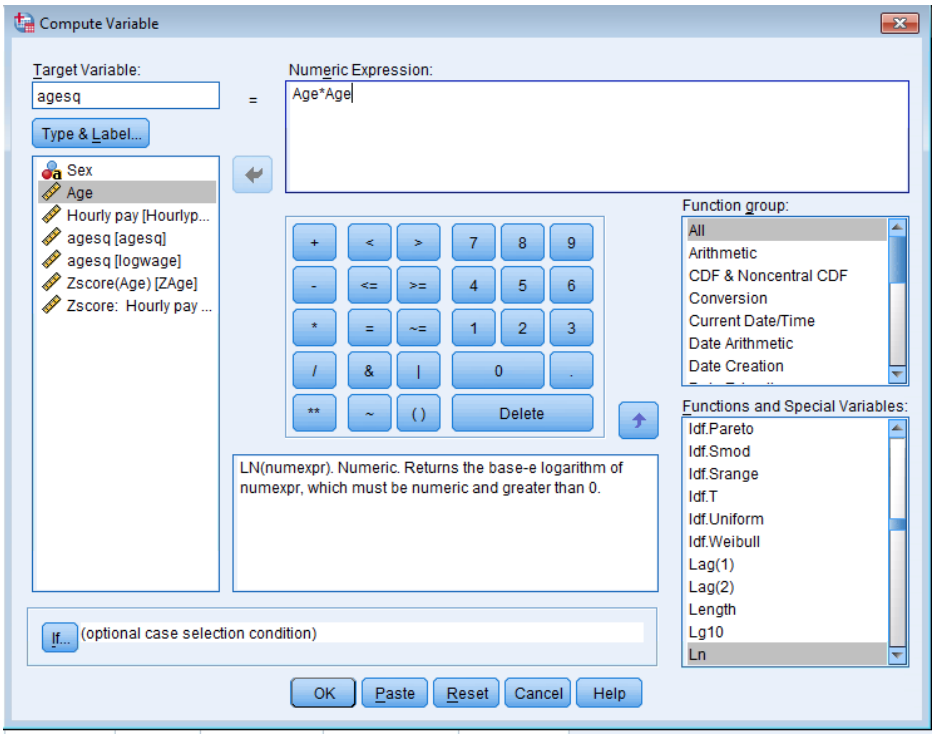
4.5.2 COMPUTING A NEW VARIABLE

New variables can be computed via the **Transform > Compute Variable**… menu path.

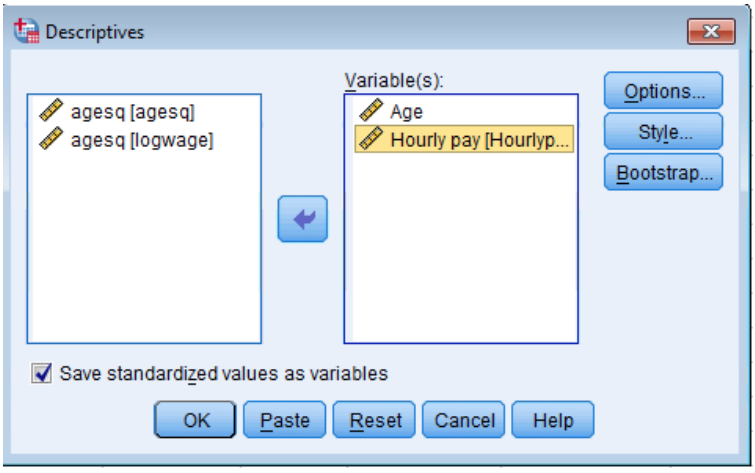
To compute the natural log of pay in this example dataset:



To compute a quadratic term – here age squared:



To save standardised versions of a variable, go to Descriptives and select the check box.



The resulting dataset will look like this – we now have three original variables and four computed variables displayed in the Variables viewer.

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	Sex	String	6	0		None	None	6	Left	Nominal	Input
2	Age	Numeric	11	0		None	None	11	Right	Scale	Input
3	Hourlypay	Numeric	11	2	Hourly pay	None	None	11	Right	Scale	Input
4	agesq	Numeric	8	2	agesq	None	None	10	Right	Scale	Input
5	logwage	Numeric	8	2	agesq	None	None	10	Right	Scale	Input
6	ZAge	Numeric	11	5	Zscore(Age)	None	None	13	Right	Scale	Input
7	ZHourlypay	Numeric	11	5	Zscore: Hourly...	None	None	13	Right	Scale	Input

5 FURTHER READING

A number of excellent texts have been written with significantly more technical detail and worked examples, a selection of which are listed below. Field is available in both SPSS and also a version in R (a free to use open source data analysis program widely used in academia and the public and private sectors).

Bryman, A., Cramer, D., 1994. Quantitative Data Analysis for Social Scientists. Routledge.

Dobson, A.J., 2010. An Introduction to Generalized Linear Models, Second Edition. Taylor & Francis.

Field, A., 2017. Discovering Statistics Using IBM SPSS Statistics. SAGE.

Hair, J.F., Anderson, R.E., Babin, B.J. and Black, W.C., 2010. Multivariate data analysis: A global perspective (Vol. 7).

Howell, D.C., 2012. Statistical Methods for Psychology. Cengage Learning.

Hutcheson, G.D., 1999. The Multivariate Social Scientist: Introductory Statistics Using Generalized Linear Models. SAGE.

Linneman, T.J., 2011. Social Statistics: The Basics and Beyond. Taylor & Francis.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, Second Edition. CRC Press.

Plewis, I., Everitt, B., 1997. Statistics in Education. Arnold.

6 APPENDIX A: CORRELATION, COVARIANCE AND PARAMETER ESTIMATION

The correlation coefficient, r, is calculated using:

$$r = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

Where,

$$Var(X) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Is the variance of *x* from the sample, which is of size *n.*

$$Var(Y) = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$$

Is the variance of *y*, and,

$$Var(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Is the covariance of *x* and *y.*

Notice that the correlation coefficient is a function of the variances of the two variables of interest, and their covariance.

In a simple linear regression analysis, we estimate the intercept, β_0, and slope of the line, β_1 as:

$$\hat{\beta}_1 = \frac{Cov(X,Y)}{Var(X)} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}_i$$

7 GLOSSARY

categorical	A variable where the responses are categories. For example, ethnicity.
collinear	When one variable can be used to predict another. When two variables are closely linearly associated.
continuous	A continuous variable is a variable which takes a numeric form and can take any value. For example, distance in miles to the nearest shop.
Cook's distance	A measure of whether or not observations are outliers. The threshold for further consideration is three times the mean of the Cook's distance. The creator of the measure defines any point as having a Cook's distance of 1 to be of concern. Used to assess whether or not an observation within a dataset should be removed to improve the fit of the model.
correlated	Two continuous variables are said to be correlated if a change in one variable results in a measurable change in the other. The correlation coefficient is a measure of the strength of this association or relationship.
response variable	The outcome we want to predict. The value of this variable is predicted to be dependent on the other terms in the model. Sometimes referred to as the dependent variable or the Y variable.
error term	See residual
explanatory variable	The variables which we use to predict the outcome variable. These variables are also referred to as independent or the X variable(s).
homoscedastic	One of the key assumptions for a linear regression model. If residuals are homoscedastic, they have constant variance regardless of any explanatory variables.
linear regression	A method where a line of best fit is estimated by minimising the sum of the square of the differences between the actual and predicted observations.
multicolinearity	When two or more variables are closely linearly associated or can be used to predict each other.
multiple linear regression	Linear regression with more than one explanatory variable.
negative correlation	
ordinalordinal variable	A variable where the responses are categories, which can be put in an order. For example, the highest level of education achieved by a respondent. Remember that the possible

	responses may not be evenly spaced.
Pearson's coefficient	a measure of correlation.
population	The whole group we are interested in.
positive correlation	A situation where if one variable increases in value, another variable also tends to increase in value.
R^2	A measure of model fit. The percentage of variance explained by the model.
representative	When a sample is representative, it has the same statistical properties as the population as a whole. This means that when we get results of a statistical analysis of the sample, we can infer that the same results are true for the population. To be representative a sample needs to be of sufficient size and the correct composition to reflect the means of groups within the underlying population.
residual	The difference between the predicted value from the model, and the actual value of the observation. When texts refer to the residuals, it means the data that is generated if we calculate the residual for every observation in the dataset.
sample	The sub section of the population which we are studying. A smaller number of units, drawn from the population. For example we might be interested in menu choices in a school canteen. Our population of interest is everyone in the school. We could then take a survey of 5 students from each year group. This would be our sample.
simple linear regression	Linear regression with one explanatory variable
skew	Measures the symmetry of a distribution. A symmetrical distribution has a skew of 0. Positive skew means more of the values are at the lower end of the distribution, negative skew means that more of the values are at the higher end of the distribution.
statistically significant	When a result is statistically significant, we mean that it meets our criteria for the hypothesis test. Statistically significant is not the same as "important" or "interesting" and has a specific technical meaning.