Multiple Linear Regression

Mark Tranmer
Mark Elliot

Cathie Marsh Centre for Census and Survey Research

Contents

Section 1: Introduction

1.1 Overview

A multiple linear regression analysis is carried out to predict the values of a dependent variable, *Y*, given a set of p explanatory variables (x_1,x_2,….,$x_{p)}$. In these notes, the necessary theory for multiple linear regression is presented and examples of regression analysis with census data are given to illustrate this theory. This course on multiple linear regression analysis is therefore intended to give a practical outline to the technique. Complicated or tedious algebra will be avoided where possible, and references will be given to more theoretical texts on this technique. Important issues that arise when carrying out a multiple linear regression analysis are discussed in detail including model building, the underlying assumptions, and interpretation of results. However, before we consider multiple linear regression analysis we begin with a brief review of simple linear regression.

1.2 Review of Simple linear regression.

A simple linear regression is carried out to estimate the relationship between a dependent variable, *Y,* and a single explanatory variable, *x,* given a set of data that includes observations for both of these variables for a particular population. For example, for a sample of *n*=17 pupils in a particular school, we might be interested in the relationship of two variables as follows:

- Exam performance at age 16. The dependent variable, *y* (Exam16)
- Exam performance at age 11. The explanatory variable, *x* (Exam11)

(n.b. we would ideally have a bigger sample, but this small sample illustrates the ideas)

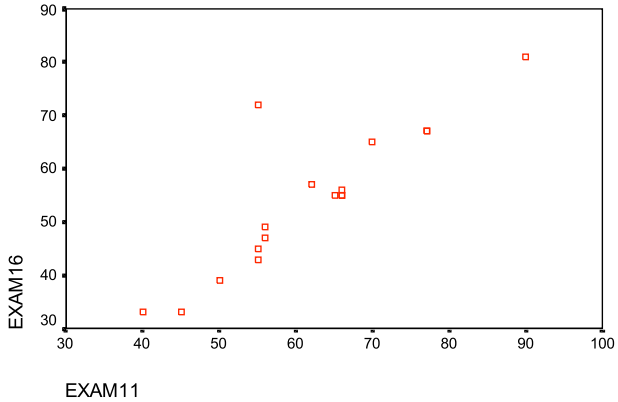
Exam16	Exam11
45	55
67	77
55	66
39	50
72	55
47	56
49	56
81	90
33	40
65	70
57	62
33	45
43	55
55	65
55	66
67	77
56	66

We would carry out a simple linear regression analysis to predict the value of the dependent variable *y*, given the value of the explanatory variable, x. In this example we are trying to predict the value of exam performance at 16 given the exam performance at age 11.

Before we write down any models we would begin such an analysis by plotting the data as follows: Figure 1.1.

Scatterplot of exam performance at 16

against exam performance at 11



We could then calculate a correlation coefficient to get a summary measure of the strength of the relationship. For figure 1.1 we expect the correlation is highly positive (it is 0.87). If we want to fit a straight line to these points, we can perform a simple linear regression analysis. We can write down a model of the following form.

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

Where β_0 the *intercept* and β_1 is the *slope* of the line. We assume that the error terms e_i have a mean value of 0.

The relationship between y and x is then estimated by carrying out a simple linear regression analysis. We will use the least squares criterion to estimate the equations, so that we minimise the sum of squares of the differences between the actual and predicted values for each observation in the sample. That is, we minimise Σe_i^2. Although there are other ways of estimating the parameters in the regression model, the least squares criterion has several desirable statistical properties, most notably, that the estimates are maximum likelihood if the residuals e_i are normally distributed.

For the example above, if we estimate the regression equation we get:

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

where x_i is the value of EXAM11 for the i^{th} student.

We could draw this line on the scatter plot. It is sometimes referred to as the line of *y* on *x,* because we are trying to predict *y* **on** the information provided by *x.*

Predicted values

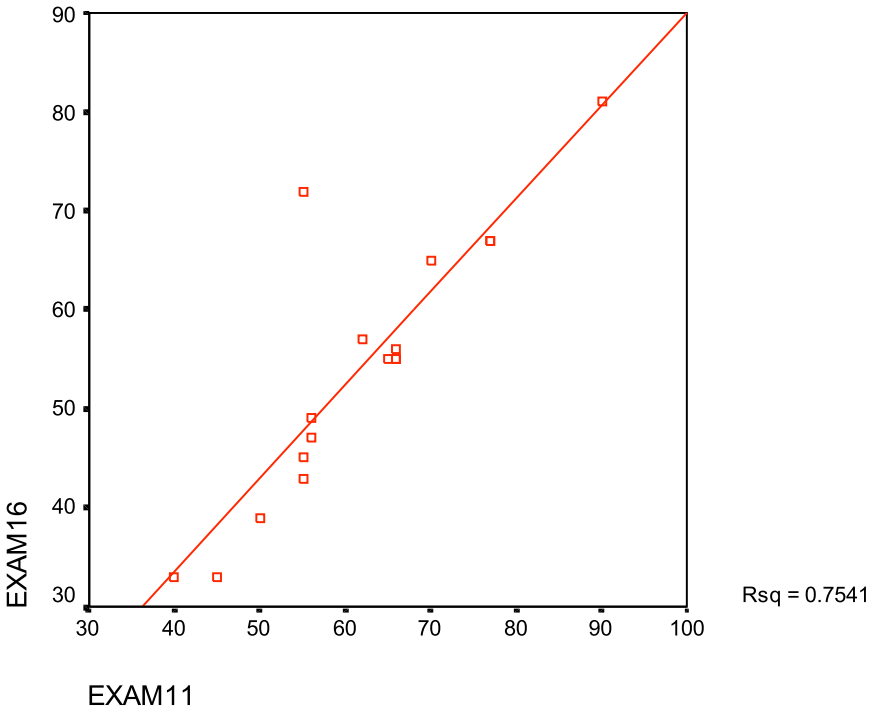
The first student in the sample has a value of 45 for EXAM16 and 55 for exam11.
The *predicted* value of EXAM16 for this student is 47.661.

Residuals

We know that the *actual* value of EXAM16 for the first student is 45, and the predicted value is 47.661, therefore the residual may be calculated as the difference between the actual and predicted values of EXAM16. That is, 45 – 47.661 = -2.661.

Figure 1.2 Scatter plot, including the regression line.

Scatterplot of exam performance at 16 against exam performance at 11



1.3 Theory for multiple linear regression

In multiple linear regression, there are p explanatory variables, and the relationship between the dependent variable and the explanatory variables is represented by the following equation:

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

Where:

β_0 is the constant term and

β_1 to β_p are the coefficients relating the *p* explanatory variables to the variables of interest.

So, multiple linear regression can be thought of an extension of simple linear regression, where there are p explanatory variables, or simple linear regression can be thought of as a special case of multiple linear regression, where p=1. The term 'linear' is used because in multiple linear regression we assume that y is directly related to a linear combination of the explanatory variables.

Examples where multiple linear regression may be used include:

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

As is the case with simple linear regression and correlation, this 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 dependent variable of interest.

In terms of a hypothesis test, for the case of a simple linear regression the null hypothesis, H_0 is that the coefficient relating the explanatory (x) variable to the dependent (y) variable is 0. In other words that there is no relationship between the explanatory variable and the dependent variable. The alternative hypothesis H_1 is that the coefficient relating the x variable to the y variable is not equal to zero. In other words there is some kind of relationship between x and y.

In summary we would write the null and alternative hypotheses as:

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

A1: ***Aside: theory for correlation and simple linear regression***

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:

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

Section 2: Worked Example using SPSS

This document shows how we can use multiple linear regression models with an example where we investigate the nature of area level variations in the percentage of (self reported) limiting long term illness in 1006 wards in the North West of England. The data are from the 2001 UK Census.

We will consider five variables here:

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

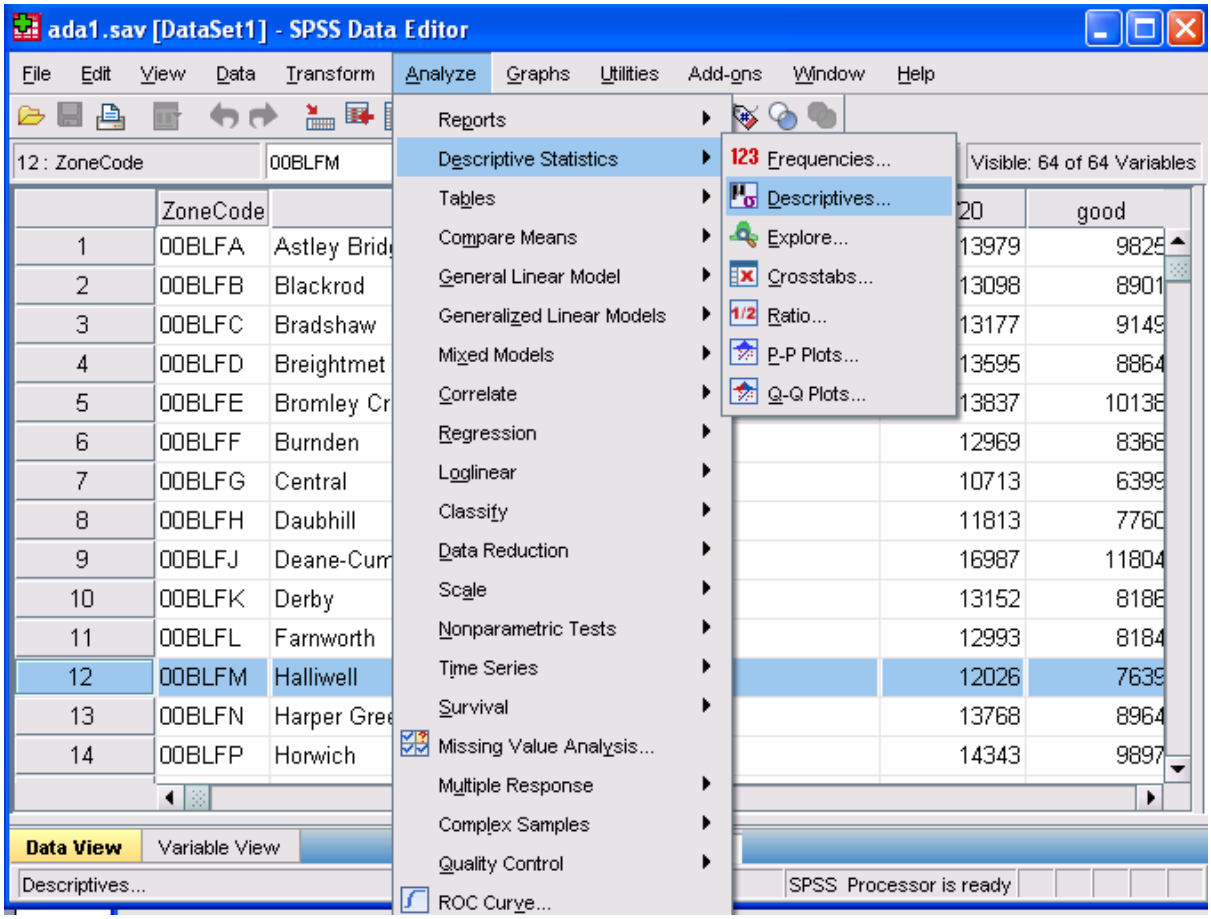
The dependent variable will be LLTI and we will investigate whether we can explain ward level variations in LLTI with A60P, FEMALE, UNEM, SRENT

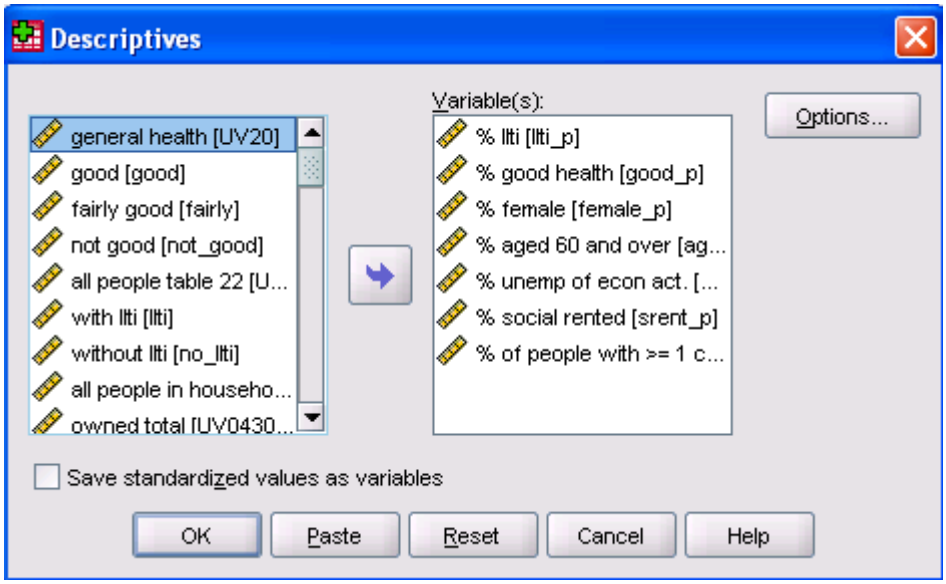
We will consider:

1. Whether this model makes sense substantively
2. Whether the usual assumptions of multiple linear regression analysis are met with these data
3. How much variation in LLTI the four explanatory variables explain
4. Which explanatory variables are most 'important' in this model
5. What is the nature of the relationship between LLTI and the explanatory variables.
6. Are there any wards where there are higher (or lower) than expected levels of LLTI given the explanatory variables we are considering here.

But first we will do some exploratory data analysis (EDA). It is always a good idea to precede a regression analysis with EDA. This may be univariate: descriptives, boxplots, histograms, bivariate: correlations, scatter plots, and occasionally multivariate e.g. principal components analysis.

Univariate EDA - descriptives

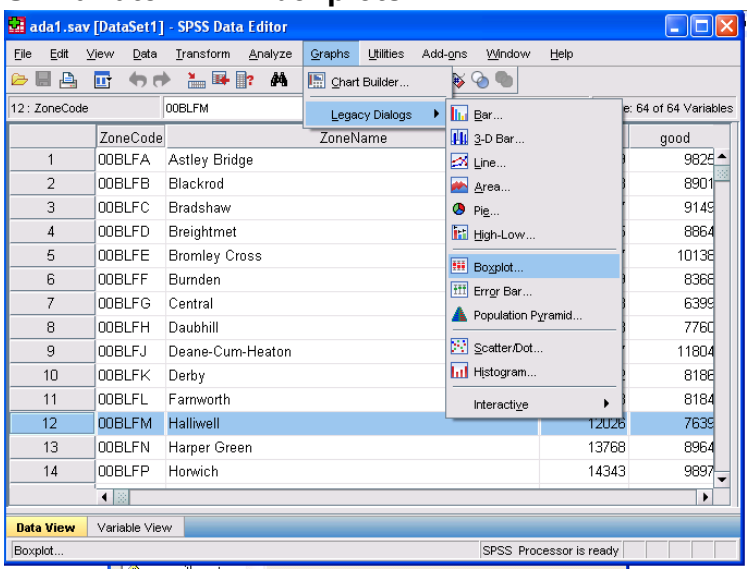


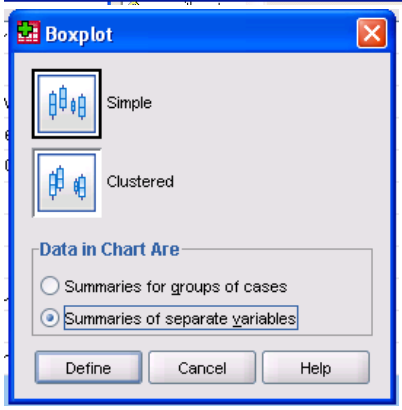


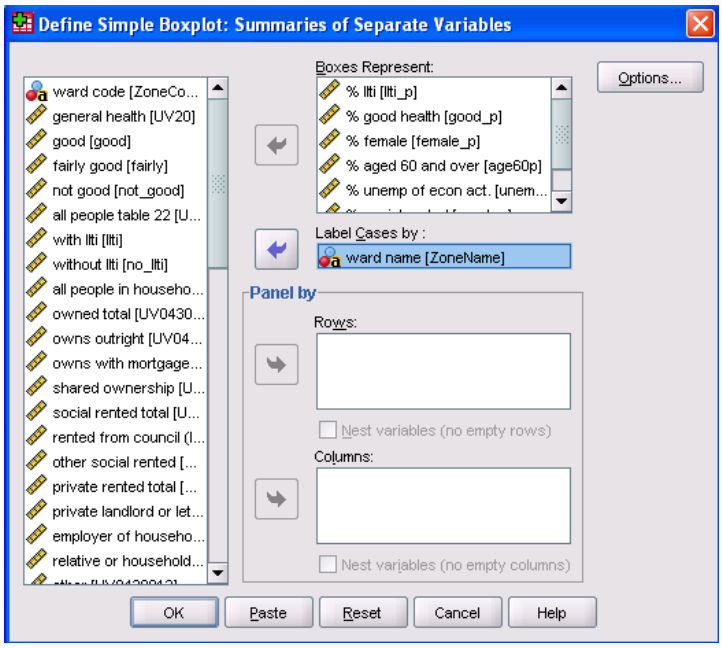
Descriptive Statistics

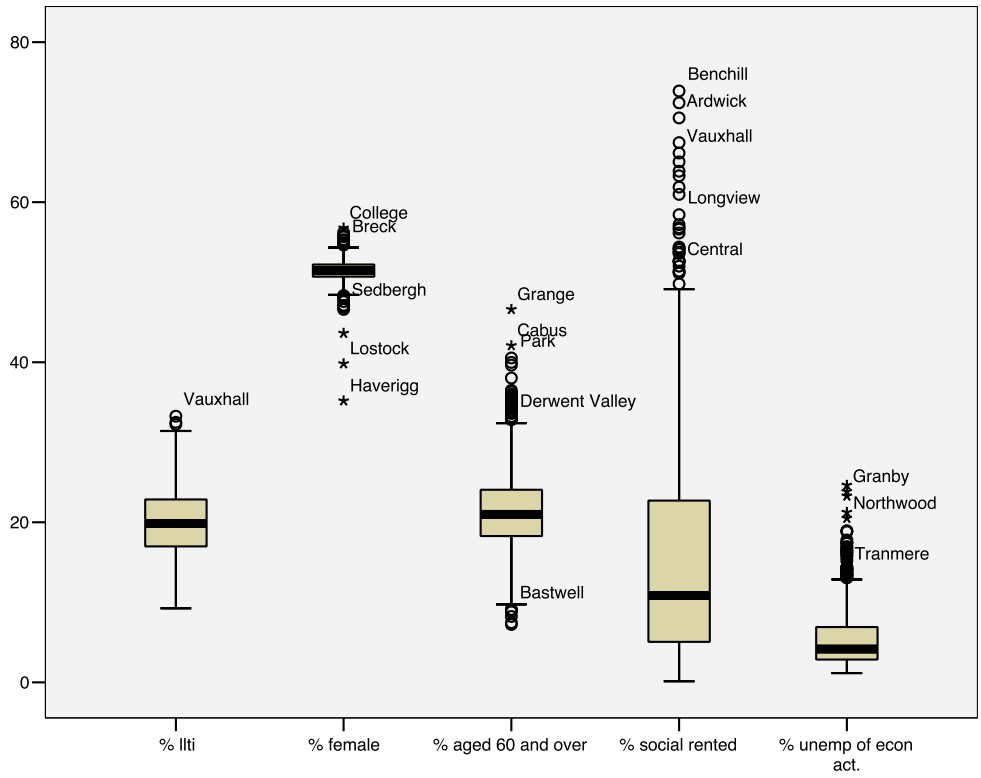
	N	Minimum	Maximum	Mean	Std. Deviation
% llti	1006	9.26	33.26	20.0436	4.13001
% good health	1006	55.70	81.22	67.5984	4.78456
% female	1006	35.18	56.77	51.4180	1.45675
% aged 60 and over	1006	7.24	46.60	21.4374	4.95659
% unemp of econ act.	1006	1.15	24.63	5.3712	3.54237
% social rented	1006	.13	73.89	15.6315	13.90675
% of people with >= 1 car in hh	1006	35.84	99.44	83.4959	12.83261
Valid N (listwise)	1006				

Univariate EDA – boxplots



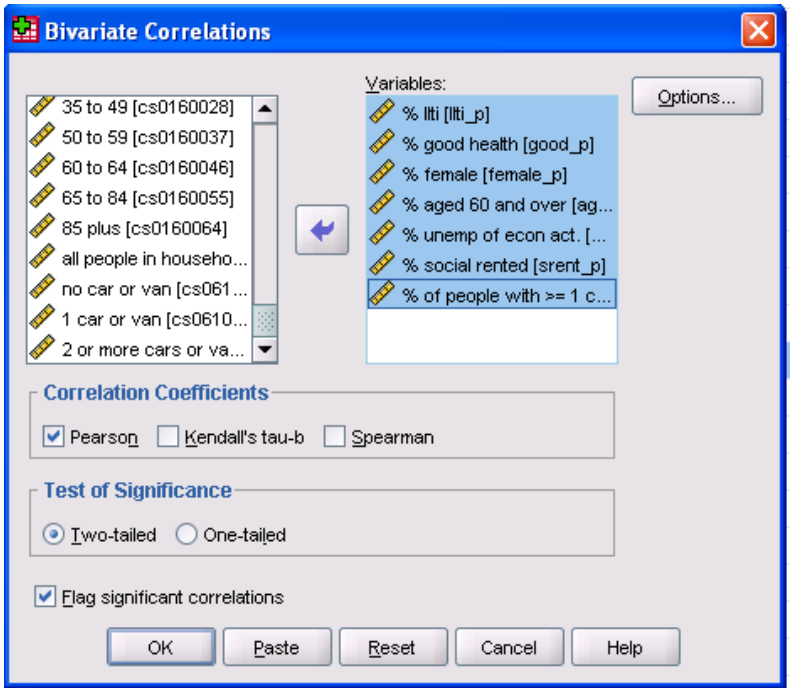






Bivariate EDA - correlations

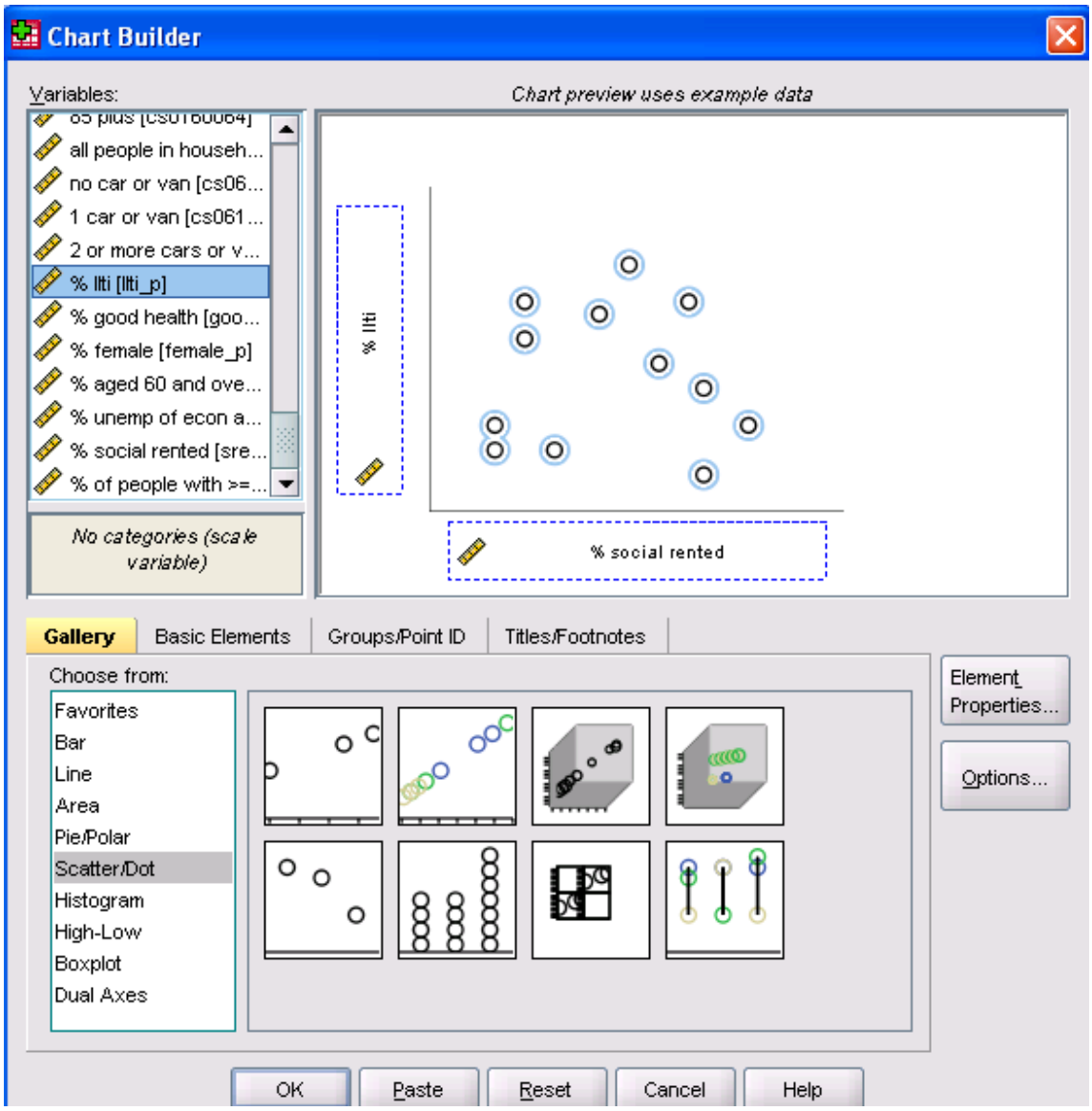
Correlations

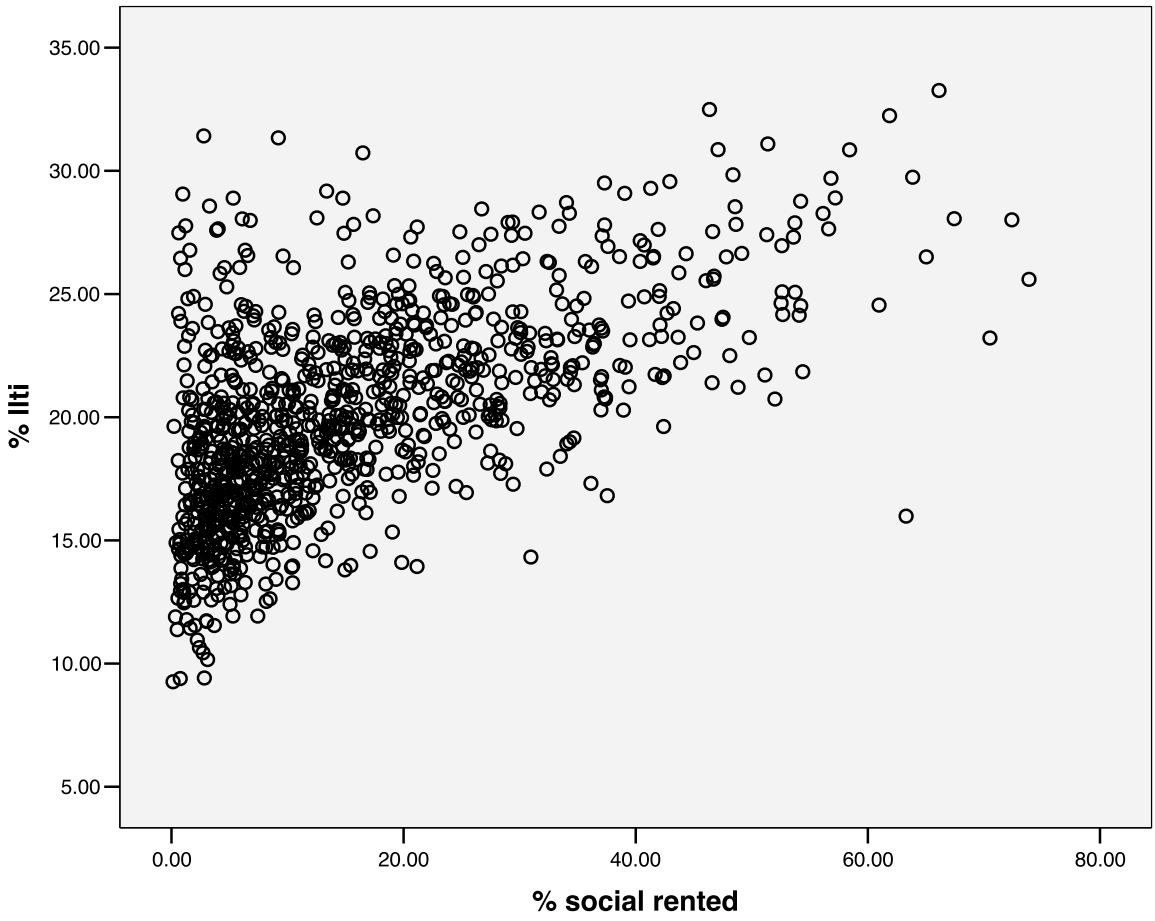


Correlations

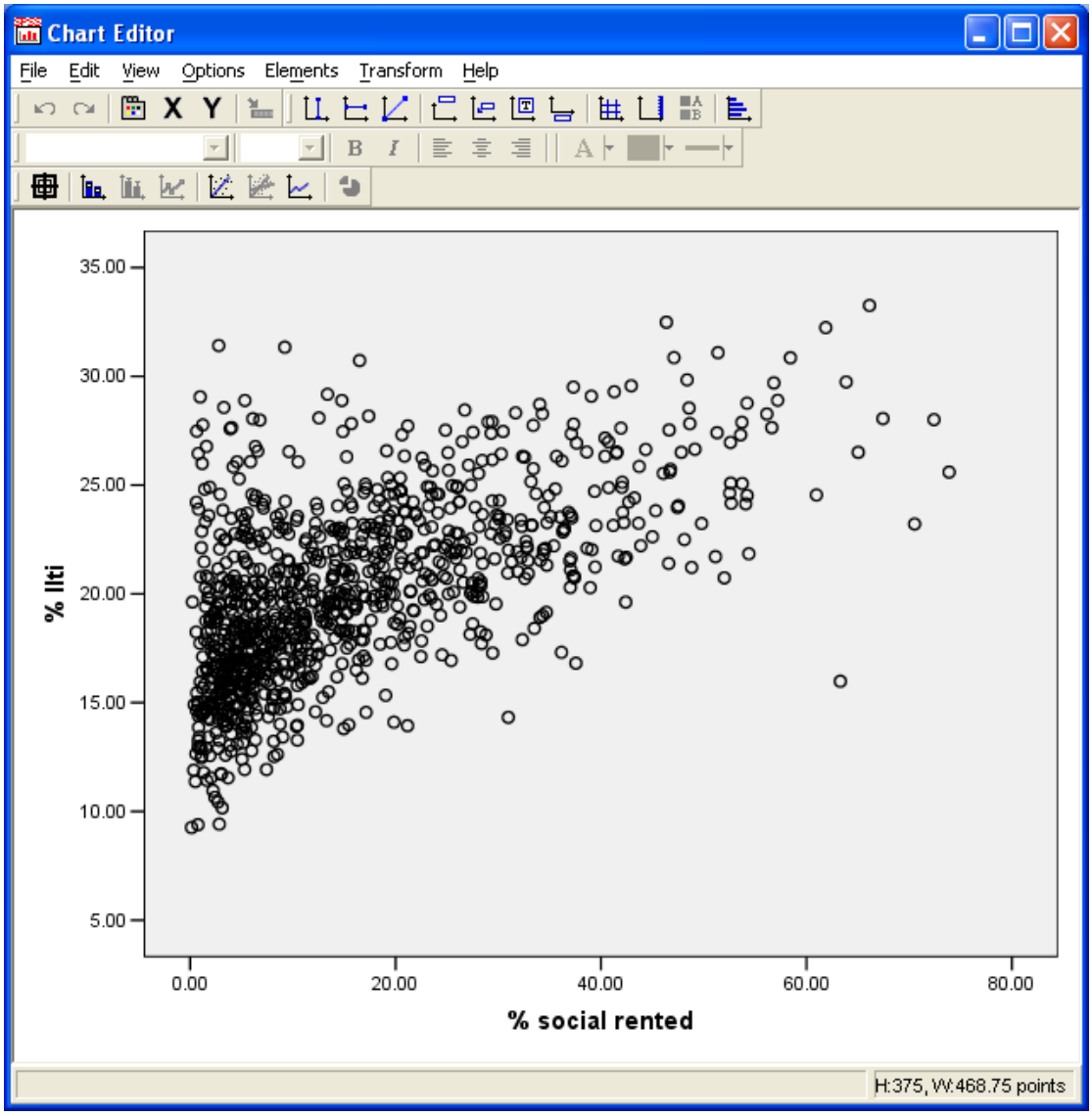
		% llti	% good health	% aged 60 and over	% unemp of econ act.	% female	% social rented	% of people with >= 1 car in hh
% llti	Pearson Correlation	1	-.938	.166	.693	.370	.599	-.723
	Sig. (2-tailed)		.000	.000	.000	.000	.000	.000
	N	1006	1006	1006	1006	1006	1006	1006
% good health	Pearson Correlation	-.938	1	-.063	-.693	-.286	-.637	.769
	Sig. (2-tailed)	.000		.047	.000	.000	.000	.000
	N	1006	1006	1006	1006	1006	1006	1006
% aged 60 and over	Pearson Correlation	.166	-.063	1	-.320	.259	-.321	.346
	Sig. (2-tailed)	.000	.047		.000	.000	.000	.000
	N	1006	1006	1006	1006	1006	1006	1006
% unemp of econ act.	Pearson Correlation	.693	-.693	-.320	1	.162	.797	-.901
	Sig. (2-tailed)	.000	.000	.000		.000	.000	.000
	N	1006	1006	1006	1006	1006	1006	1006
% female	Pearson Correlation	.370	-.286	.259	.162	1	.211	-.200
	Sig. (2-tailed)	.000	.000	.000	.000		.000	.000
	N	1006	1006	1006	1006	1006	1006	1006
% social rented	Pearson Correlation	.599	-.637	-.321	.797	.211	1	-.814
	Sig. (2-tailed)	.000	.000	.000	.000	.000		.000
	N	1006	1006	1006	1006	1006	1006	1006
% of people with >= 1 car in hh	Pearson Correlation	-.723	.769	.346	-.901	-.200	-.814	1
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	
	N	1006	1006	1006	1006	1006	1006	1006

Bivariate EDA - Scatterplot

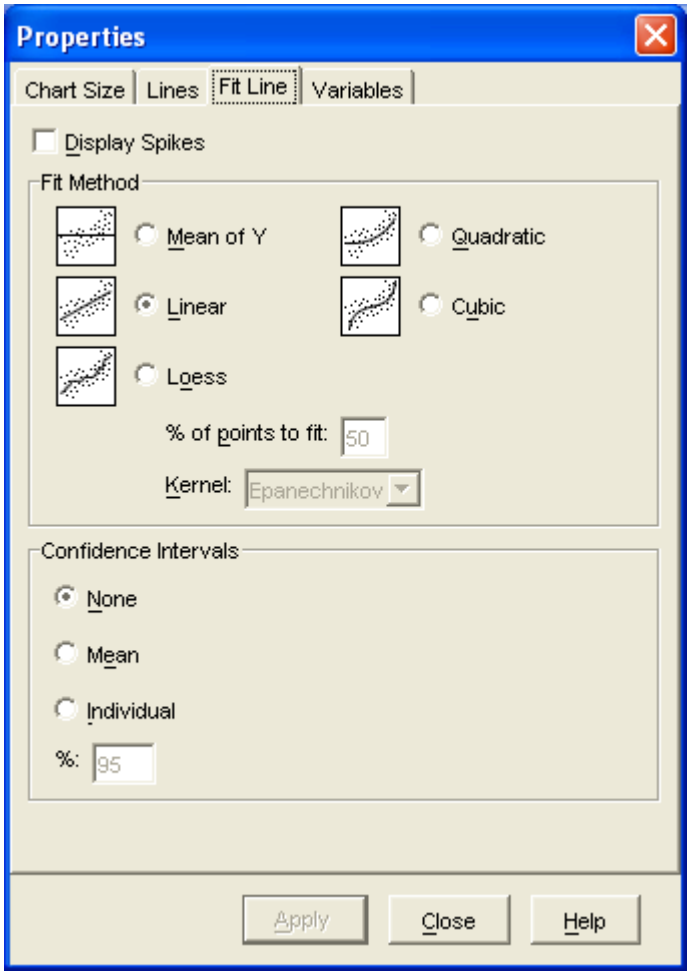


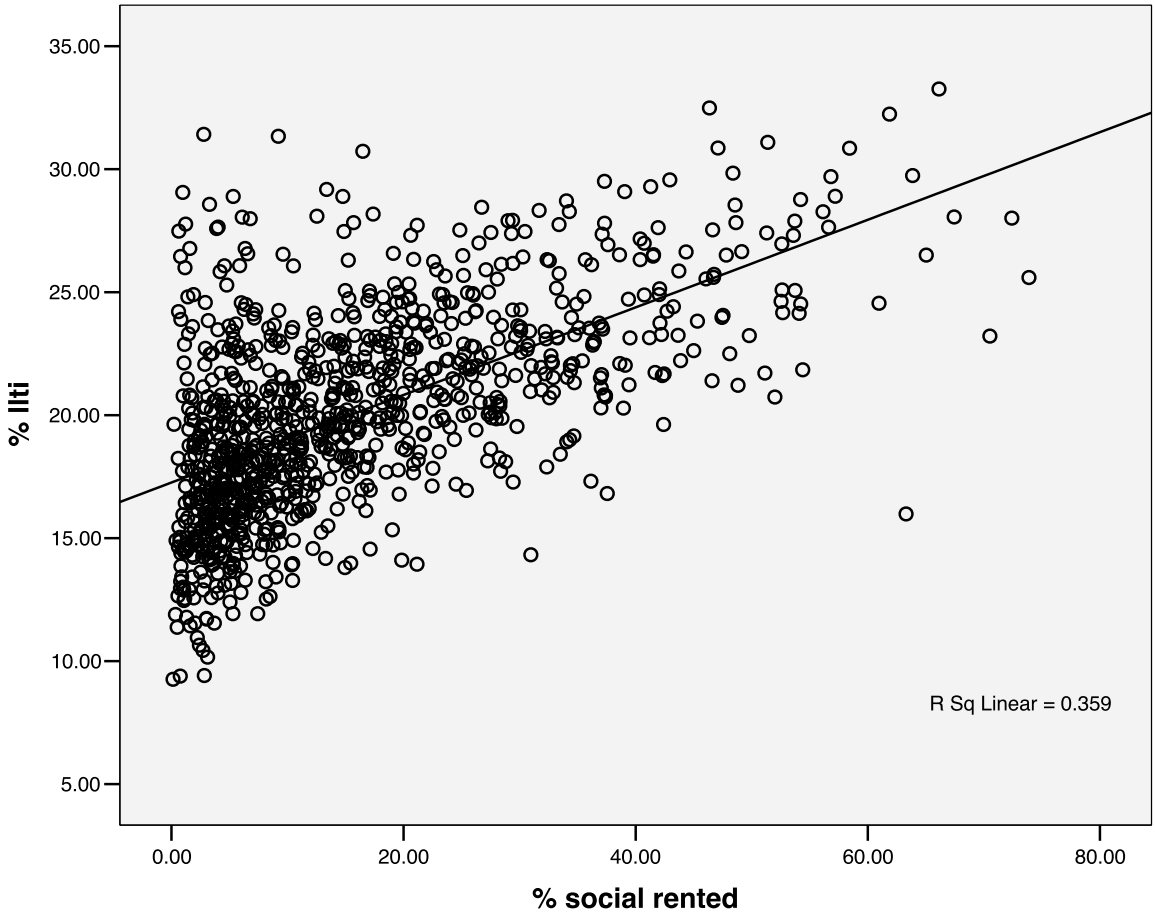


Double click on the graph to go into the graph editor window …



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


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



—

The simple linear regression has an R squared value of 0.359. i.e. it explains 35.9% of the ward level variation in % LLTI

Bivariate Analysis - Simple Linear Regression

Let us continue with the example where the dependent variable is % llti and there is a single explanatory variable, % social rented. Hence we begin with a simple linear regression analysis. We will then add more explanatory variables in a multiple linear regression analysis.

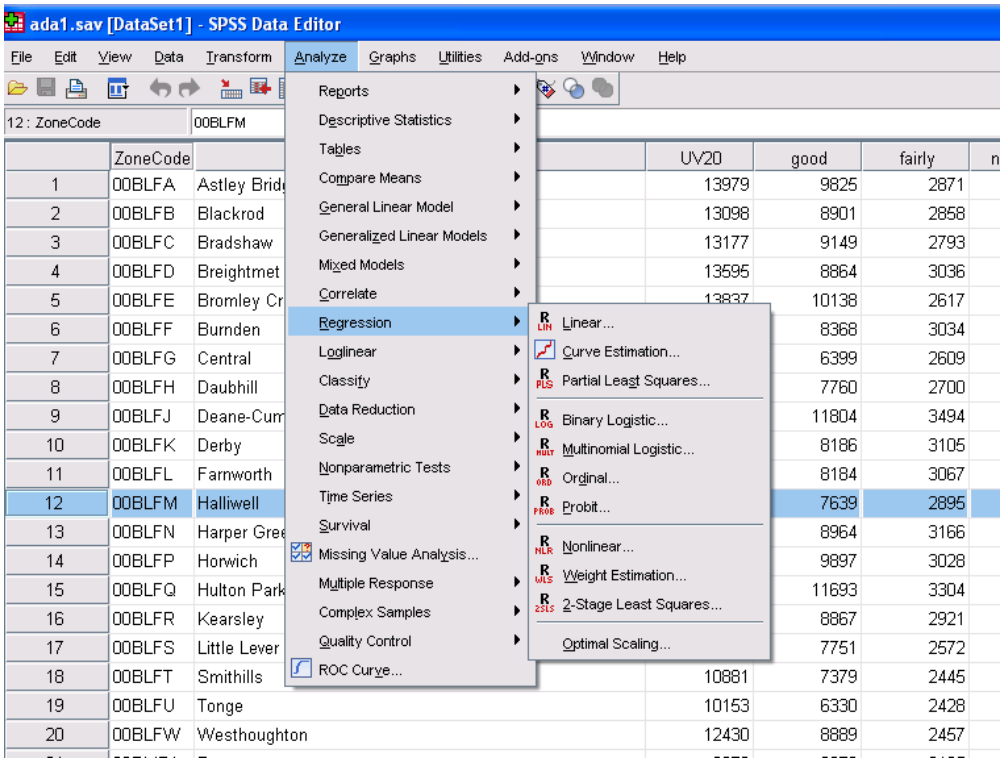
To perform a linear regression analysis, go to the

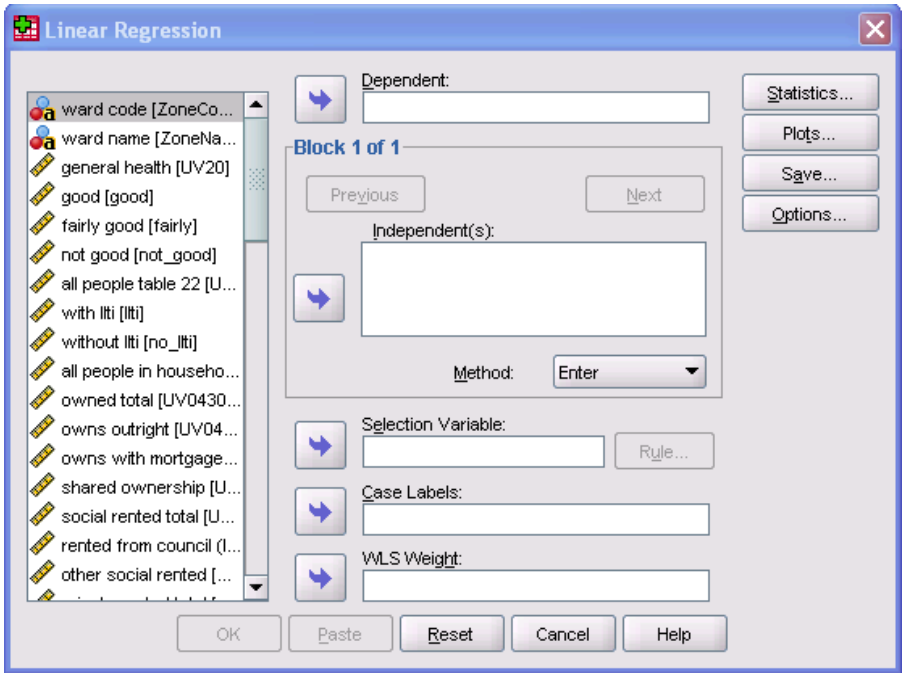
analyze > regression > linear

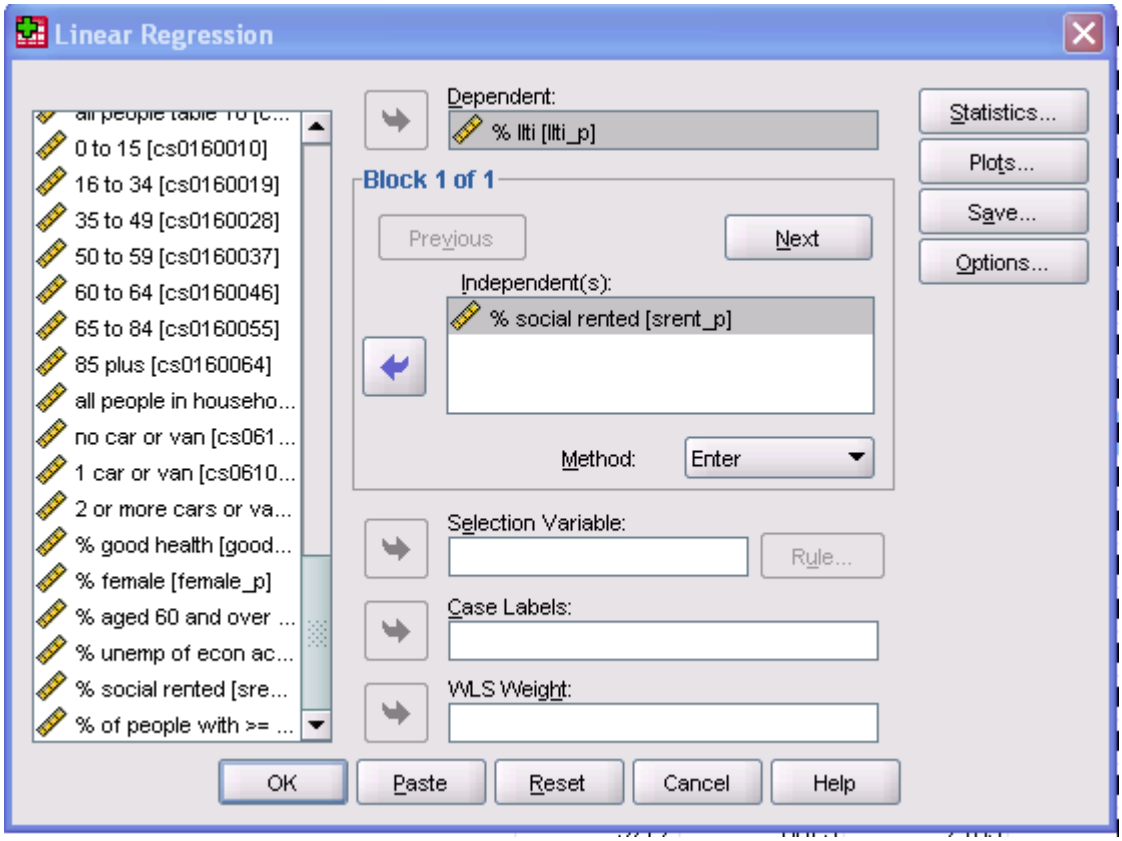
menu options.

Choose the dependent and independent (explanatory) variables you require. The default ‘enter’ method puts all explanatory variables you specify in the model, in the order that you specify them. Note that the order in unimportant in terms of the modeling process.

There are other methods available for model building, based on statistical significance, such as backward elimination or forward selection but when building the model on a substantive basis, the enter method is best: variables are included in the regression equation regardless of whether or not they are statistically significant.


ada1.sav [DataSet1] - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
12 : ZoneCode
00BLFM
Reports
Descriptive Statistics
Tables
Compare Means
General Linear Model
Generalized Linear Models
Mixed Models
Correlate
Regression
Loglinear
Classify
Data Reduction
Scale
Nonparametric Tests
Time Series
Survival
Missing Value Analysis...
Multiple Response
Complex Samples
Quality Control
ROC Curve...
Linear...
Curve Estimation...
Partial Least Squares...
Binary Logistic...
Multinomial Logistic...
Ordinal...
Probit...
Nonlinear...
Weight Estimation...
2-Stage Least Squares...
Optimal Scaling...
ZoneCode UV20 good fairly
1 00BLFA Astley Brid 13979 9825 2871
2 00BLFB Blackrod 13098 8901 2858
3 00BLFC Bradshaw 13177 9149 2793
4 00BLFD Breightmet 13595 8864 3036
5 00BLFE Bromley Cr 10138 2617
6 00BLFF Burnden 8368 3034
7 00BLFG Central 6399 2609
8 00BLFH Daubhill 7760 2700
9 00BLFJ Deane-Cum 11804 3494
10 00BLFK Derby 8186 3105
11 00BLFL Farnworth 8184 3067
12 00BLFM Halliwell 7639 2895
13 00BLFN Harper Gre 8964 3166
14 00BLFP Horwich 9897 3028
15 00BLFQ Hulton Park 11693 3304
16 00BLFR Kearsley 8867 2921
17 00BLFS Little Lever 7751 2572
18 00BLFT Smithills 10881 7379 2445
19 00BLFU Tonge 10153 6330 2428
20 00BLFW Westhoughton 12430 8889 2457


Linear Regression
ward code [ZoneCo...
ward name [ZoneNa...
general health [UV20]
good [good]
fairly good [fairly]
not good [not_good]
all people table 22 [U...
with lti [lti]
without lti [no_lti]
all people in househo...
owned total [UV0430...
owns outright [UV04...
owns with mortgage...
shared ownership [U...
social rented total [U...
rented from council (l...
other social rented [...
Dependent:
Block 1 of 1
Previous
Next
Independent(s):
Method: Enter
Selection Variable:
Rule...
Case Labels:
WLS Weight:
Statistics...
Plots...
Save...
Options...
OK
Paste
Reset
Cancel
Help



Regression

Variables Entered/Removed[b]

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

a. All requested variables entered.

b. Dependent Variable: % llti

the table above confirms that the dependent variable is % llti and the explanatory variable here is % social rented.

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

the table above shows that we have explained about 35.9% of the variation in % llti with the single explanatory variable, % social rented. In general quote the 'adjusted r square' figure. When the sample size, n, is large, r square and adjusted r square will usually be identical or very close. For small n, adjusted r square takes the sample

size (and the number of explanatory variables in the regression equation) into account.

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

the ANOVA table above indicates that the model, as a whole, is a significant fit to the data.

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 coefficients table above shows that:

- the constant, or intercept term for the line of best fit, when x = 0, is 17.261 (%).
- The slope, or coefficient for % social rented, is positive: areas with more social renting tend to be associated with areas with more limiting long term illness.
- The slope coefficient is 0.178 with a standard error of .008.
- The t value = slope coefficient / standard error = 23.733
- This is highly statistically significant ($p << 0.05$) the usual 5% significance level
- The standardized regression coefficient provides a useful way of seeing what the impact of changing the explanatory variable by one standard deviation.
- The standardized coefficient is 0.599 – a one standard deviation change in the explanatory variable results in a 0.599 standard deviation change in the dependent variable % llti.

the theoretical model is

$$LLTI = \beta_0 + \beta_1 SOCIAL\_P + \varepsilon_i$$

or

$$\hat{LLTI} = \hat{\beta}_0 + \hat{\beta}_1 SOCIAL\_P_i$$

where

β_0 is the intercept (constant) term and

β_1 is the slope term; the coefficient that relates the value of SOCIAL_P to the expected value of LLTI.

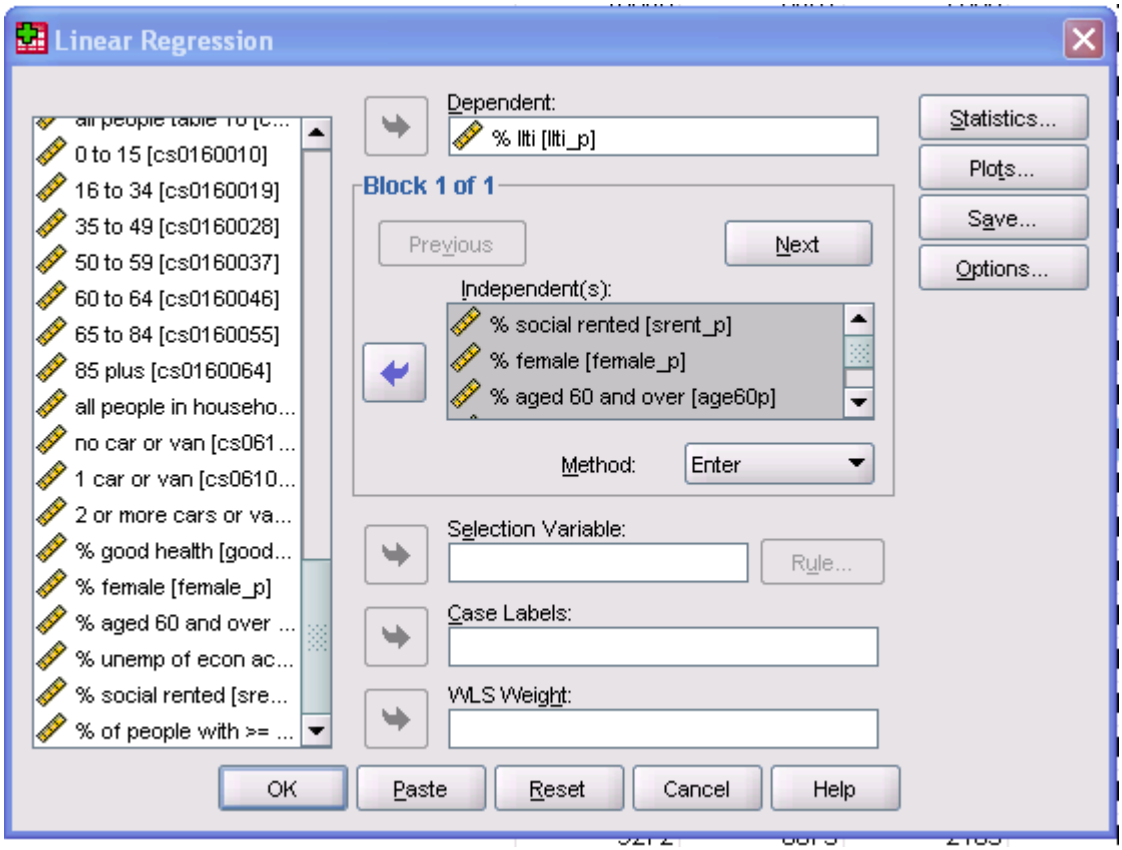
From the results above, our estimated equation is:

$$\hat{LLTI} = 17.261 + 0.178\, SOCIAL\_P_i$$

Multiple linear regression analysis

We will now add some more explanatory variables so that we now have a multiple linear regression mode, which now contains:

Social rented, Age, female and age 60 plus as explanatory variables.



Regression

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

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

This multiple linear regression model, with four explanatory variables, now has an R squared value of 0.675. 67.5 % of the variation in % LLTI can be explained by this model.

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

Once again, the model, as a whole, is a significant fit to the data.

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

From the table above we see that:

- All the explanatory variables are statistically significant.
- All have positive coefficients – for each explanatory variable a greater percentage is associated with a higher level of LLTI
- Taking % aged 60 and over as an example, we see that having controlled for unemp, female and social rented (i.e. holding these variables constant), for every 1% increase in the % of 60 and over, there is an increase of 0.33% in the predicted value of LLTI.

The theoretical model here is:

$$\text{LLTI}_i = \beta_0 + \beta_1 \text{UNEM\_P}_i + \beta_2 \text{FEMALE\_P}_i + \beta_3 \text{SOCIAL\_P}_i + \beta_4 \text{A60P\_P}_i + \varepsilon_i$$

The estimated model here is:

$$\hat{\text{LLTI}}_i = -9.832 + .774\text{UNEM\_P}_i + .344\text{FEMALE\_P}_i + .052\text{SOCIAL\_P}_i + .336\text{A60P\_P}_i$$

We notice that the variables % unemployed (of all economically active) and % social rented are very highly correlated. We can assess the impact of the correlation on the regression results by leaving one of the variables, % unemployed, out of the multiple linear regression analysis.

Regression

Variables Entered/Removed[b]

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

a. All requested variables entered.

b. Dependent Variable: % llti

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

With 3 predictors (+ a constant), we see that we can explain 51.7 % of the variation in % LLTI.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	8881.461	3	2960.487	359.095	.000[a]
	Residual	8260.783	1002	8.244		
	Total	17142.244	1005			

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

b. Dependent Variable: % llti

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
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

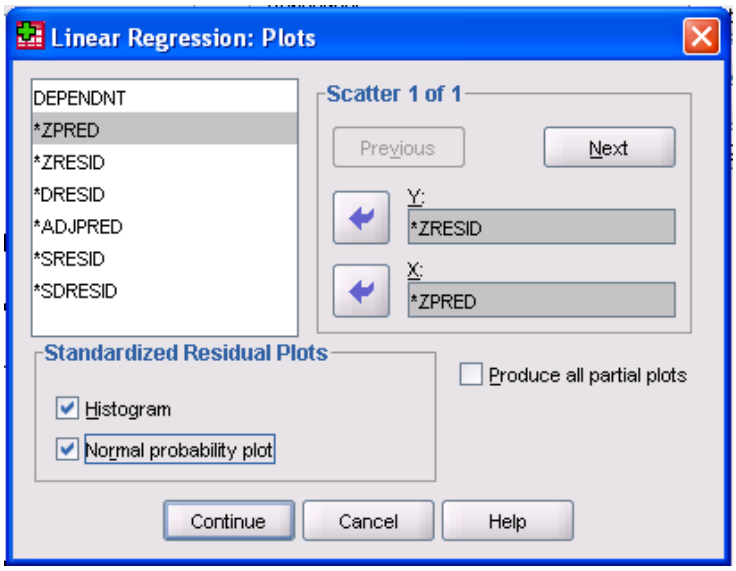
a. Dependent Variable: % llti

The estimated model here is:

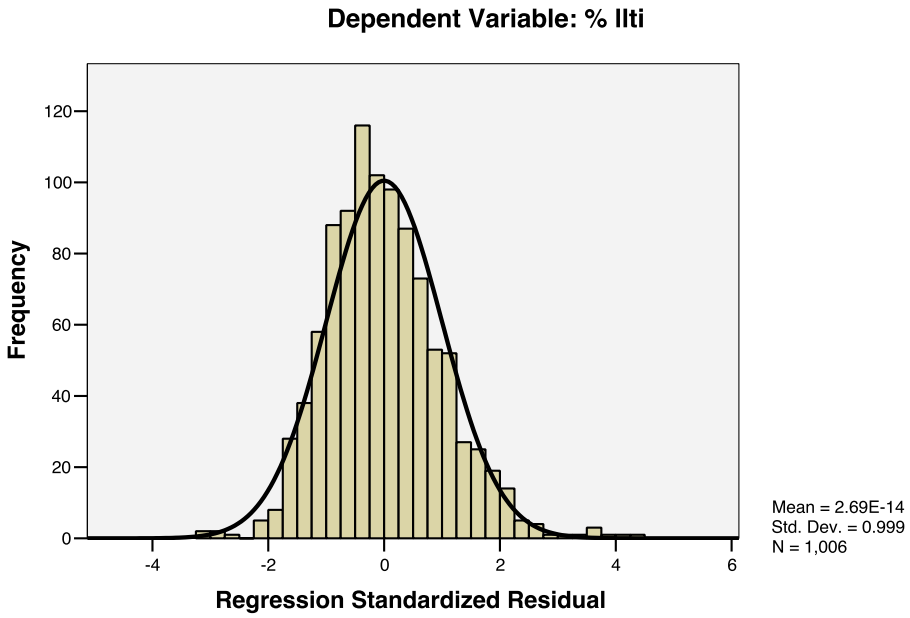
$$\hat{LLTI}_i = -9.127 + .384FEMALE\_P_i + .203SOCIAL\_P_i + .292A60P\_P_i$$

Assumption checking.

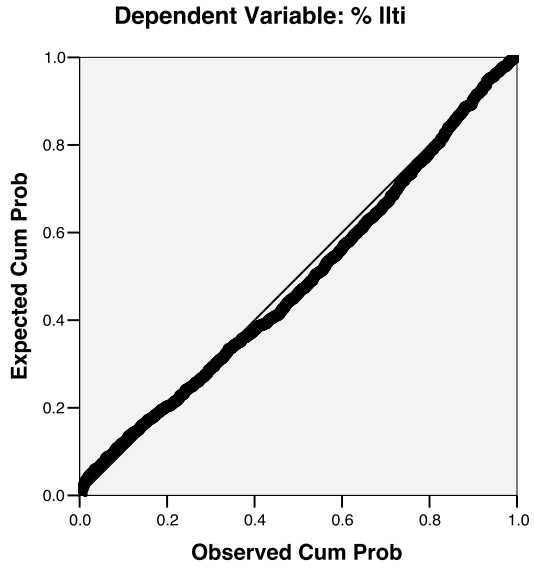
We can check many of the assumptions of the multiple linear regression analysis by producing plots. Based on the results of the last model (with 3 explanatory variables) we can produce plots by clicking the 'plots' button, which appears in the window when we specify the model in analyze > regression > linear. Click 'histogram' and 'normal probability plot' to obtain the full range of plots:



Histogram

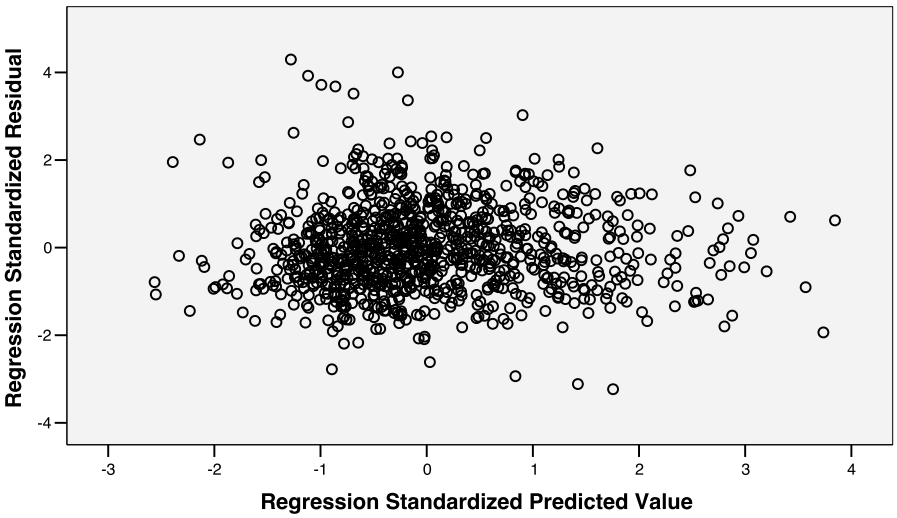

Dependent Variable: % Ilti
Frequency
120
100
80
60
40
20
0
-4
-2
0
2
4
6
Regression Standardized Residual
Mean = 2.69E-14
Std. Dev. = 0.999
N = 1,006

Normal P-P Plot of Regression Standardized Residual

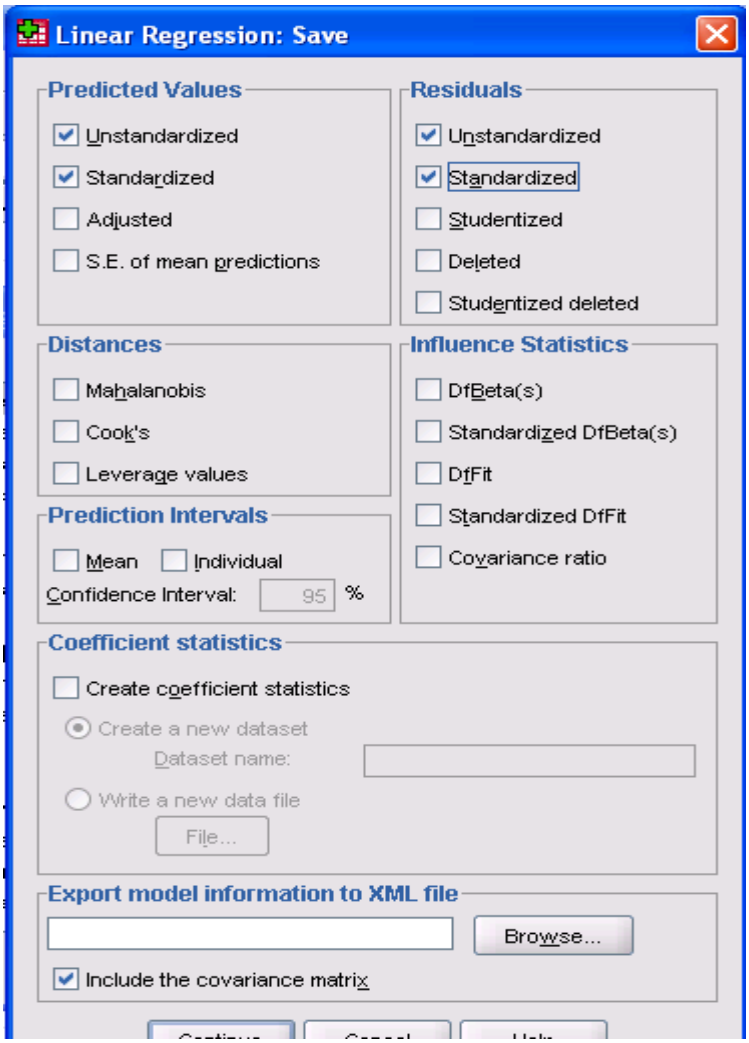


Scatterplot

Dependent Variable: % Ilti

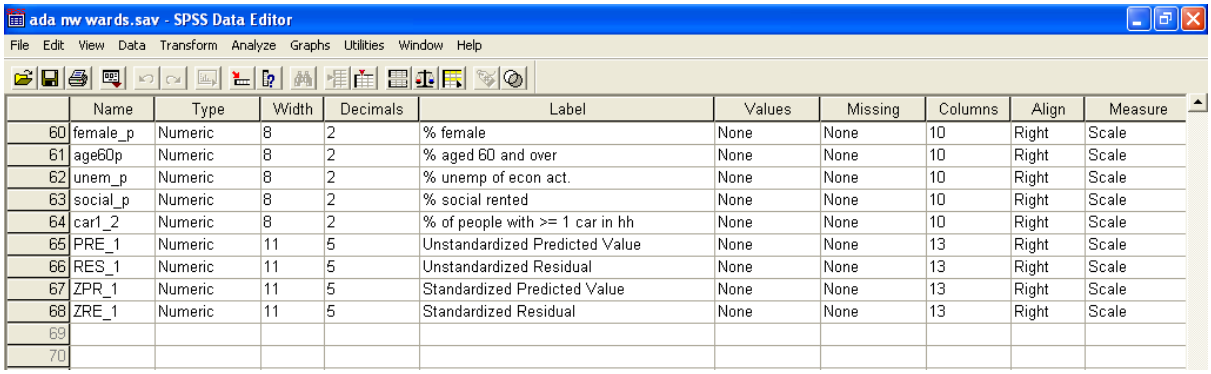


Also in the menu where we specify the regression equation via analyze > regression > linear is a 'save' button, where we can tick values, residuals and measures to be added, as new variables, to the worksheet (i.e. the dataset) we are using. Here we have saved the unstandardised and standardised residuals and predicted values:



new variables are added to the worksheet called

pre_1 = unstandardised predicted
res_1 = unstandardised residual
zpr_1 = standardised predicted
zre_1 = standardised residual



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
60	female_p	Numeric	8	2	% female	None	None	10	Right	Scale
61	age60p	Numeric	8	2	% aged 60 and over	None	None	10	Right	Scale
62	unem_p	Numeric	8	2	% unemp of econ act.	None	None	10	Right	Scale
63	social_p	Numeric	8	2	% social rented	None	None	10	Right	Scale
64	car1_2	Numeric	8	2	% of people with >= 1 car in hh	None	None	10	Right	Scale
65	PRE_1	Numeric	11	5	Unstandardized Predicted Value	None	None	13	Right	Scale
66	RES_1	Numeric	11	5	Unstandardized Residual	None	None	13	Right	Scale
67	ZPR_1	Numeric	11	5	Standardized Predicted Value	None	None	13	Right	Scale
68	ZRE_1	Numeric	11	5	Standardized Residual	None	None	13	Right	Scale
69										
70										

the suffix _1 in the variable names indicates these are the first set of residuals we have saved. If we re-specified the model and saved the residuals, these variable names would have the suffix_2 etc …

A large (positive) standardized residual i.e. > 2 from the model indicates an area where, even when accounting for the explanatory variables in the model, there is still a higher-than-expected level of LLTI in that ward. Conversely a standardized residual < -2 indicated an area that, even when accounting for the explanatory variables, there is still a lower than expected level of LLTI.

Section 3: Further topics

3.1 Checking the assumptions

Most of the underlying assumptions of multiple linear regression can be assessed by examining the residuals, having fitted a model. The various assumptions are listed below. Later, we will see how we can assess whether these assumptions hold by producing the appropriate plots.

The main assumptions are:

1. That the residuals have constant variance, whatever the value of the dependent variable. This is the assumption of **homoscedasticity.** Sometimes textbooks refer to **heteroscedasticity.** This is simply the opposite of homoscedasticity.

2. That there are no very extreme values in the data. That is, that there are no **outliers.**

3. That the residuals are **normally distributed**.

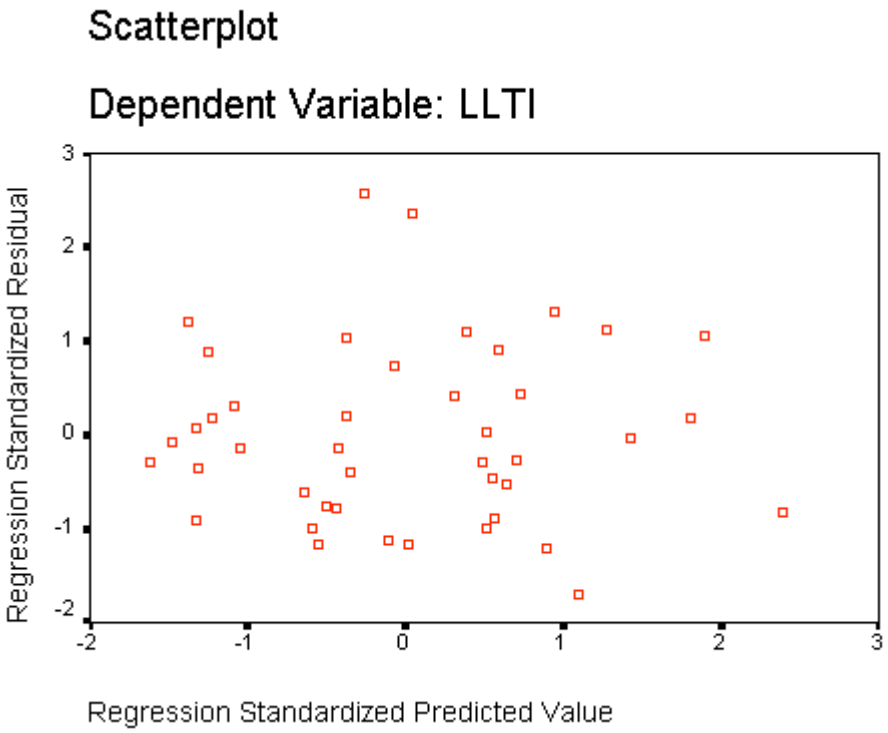
4. That the residuals are **not related to the explanatory variables**.

5. We also assume that the residuals are not correlated with one another.

Residual plots.

1. By plotting the predicted values against the residuals, we can assess the **homoscedasticity** assumption. Often, rather than plotting the unstandardised or raw values, we would plot the *standardised* predicted values against the *standardised* residuals. (Note that a slightly different version of the standardised residual is called the studentized residual, which are residuals standardised by their own standard errors. See Plewis page 15 ff for further discussion of these).

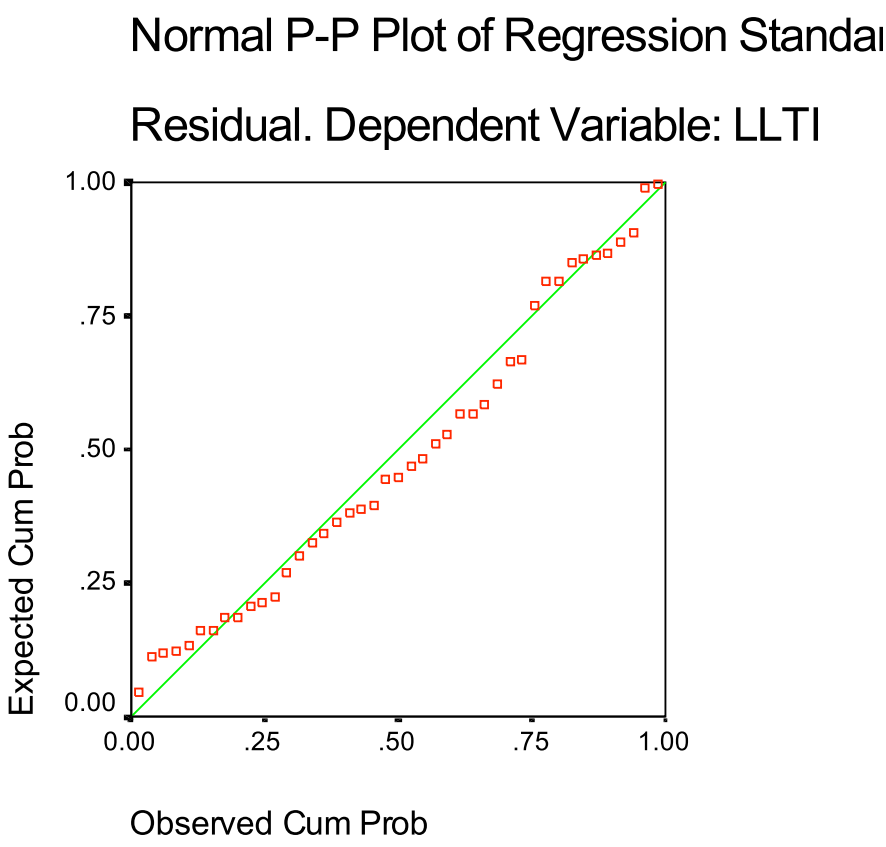
Examples from 1991 census dataset for districts in the North West



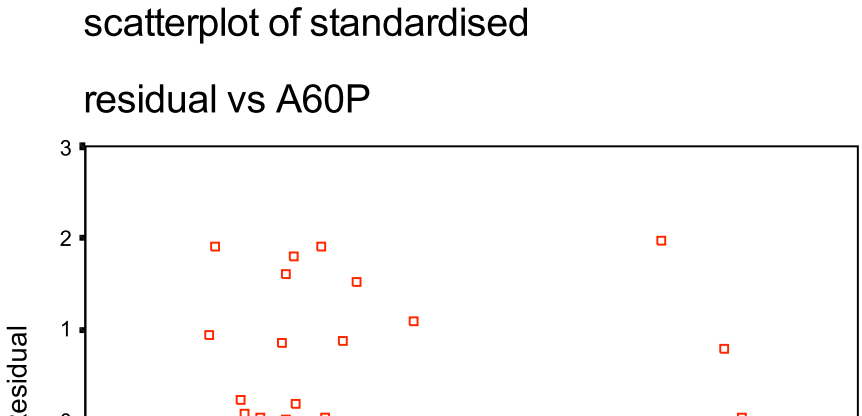
We can also assess the assumption that there are no outliers in our data from the above plot. If there was an extreme value in the standardised predicted values or standardised residuals (say greater/less than +/- 3), we should look at the sample unit (in this case the district) that corresponds to the residual. We should consider the following: is the data atypical of the general pattern for this sample unit? Has the information been recorded/entered into the computer properly for this sample unit? Is there a substantive reason why this outlier occurs: have we left an important explanatory variable out of the regression analysis? In many cases an outlier will affect the general estimate of the regression line, because the least squares approach will try to minimise the distance between the outlier and the regression line. In some cases the extreme point will move the line away from the general pattern of the data. That is, the outlier will have *leverage* on the regression line. In many cases we would consider deleting an outlier from the sample, so that we get a better estimate of the relationship for the general pattern on the data. The above plot suggests that, for our data, there are no outliers.

We can assess the assumption that the residuals are normally distributed by producing a normal probability plot (sometimes called a quantile-quantile or q-q plot).

For this 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. The figure below shows the normal probability plot for our example.



The fourth assumption listed above is that the residuals are not related in some way to the explanatory variables. We could assess this by plotting the standardised residual against the values of each explanatory variable. If a relationship does seem to exist on this plot, we need to consider putting extra terms in the regression equation. For example, there may be a quadratic relationship between the residual and explanatory variable, as indicated by a 'U' or 'n' shaped curve of the points. In order to take into account this quadratic relationship, we would consider adding the square of the explanatory variable to the variables included in the model, so that the model includes a quadratic (x^2) term. E.g. if there appeared to be a quadratic relationship between the residuals and age, we could add age^2 to the model.



The above plot shows a plot of an explanatory variables – AGE60P – against the standardised residual. If the plot had an obvious pattern it would be sensible to consider including further explanatory variables in the model. There does not seem to be an obvious pattern here, but with only 43 observations, it is not easy to tell whether or not a pattern exists.

In general it should be borne in mind that you should have a reasonable size sample to carry out a multiple linear regression analysis when you have a lot of explanatory variables.

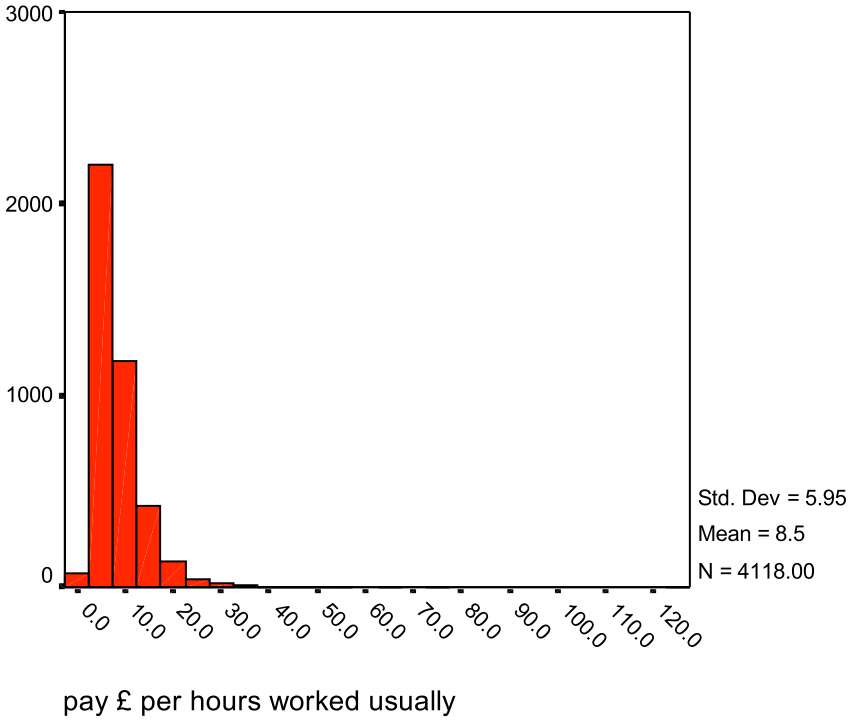
There is no simple answer as to how many observations you need, but in general the bigger the sample, the better.

3.2 Multicollinearity.

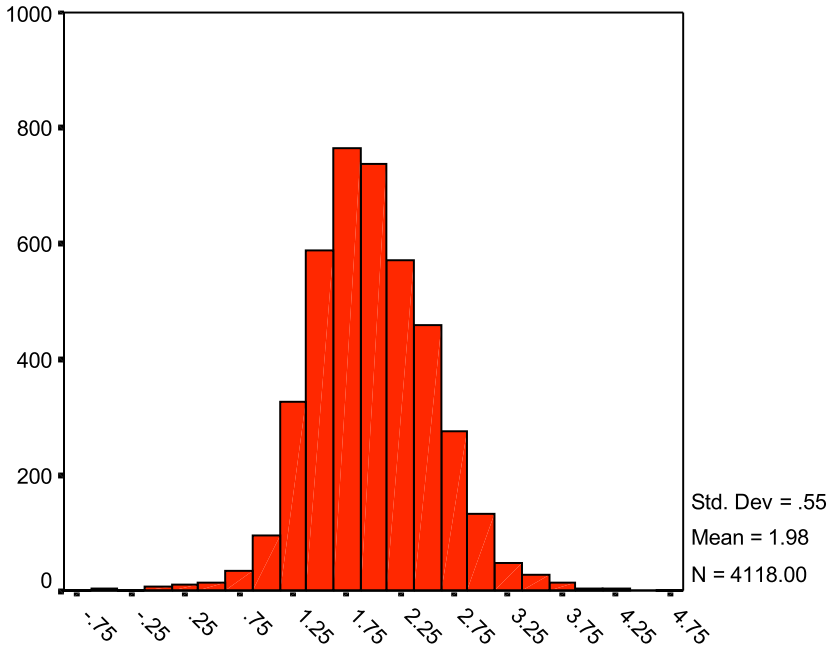
By carrying out a correlation analysis before we fit the regression equations, we can see which, if any, of the explanatory variables are very highly correlated and avoid this problem (or at least this will indicate why estimates of regression coefficients may give values very different from those we might expect). For pairs of explanatory variables with have very high correlations > 0.8 or very low correlations < 0.8 we could consider dropping one of the explanatory variables from the model.

3.3 Transformations:

In some situations the distribution of the dependent variable is not normal, but instead is positively or negatively skewed. For example the distribution of income, and similar variables such as hourly pay, tends to be positively skewed because a few people earn a very high salary. Below is an example of the distribution of hourly pay. As can be seen, it is positively skewed.



If we now take the natural log (LN) of the hourly wage we can see that the resulting distribution is much more 'normal'.



Later we will do some multiple linear regression modelling using log(hourly wage) as the dependent variable.

3.4 Dummy variables

Suppose we were interested in investigating differences, with respect to the y variable (e.g. log(income), in three different ethnic groups. Hence we would have an ethnic group variable with three categories: Afro Caribbean, Pakistani, Indian. We would need to create dummy variables to include this categorical variable in the model

For example we could use this dummy variable scheme, where 'afro-caribbean' is the reference category.

	D1	D2
Afro-caribbean	0	0
Pakistani	1	0
Indian	0	1

$$y_i = \beta_0 + \beta_1 D_{1i} + \beta_2 D_{2i} + e_i$$

Where D_1 is the dummy variable to represent the Pakistani ethnic group and D_2 is the dummy variable to represent the Indian ethnic group

Hence the estimate of the coefficient β_0 gives the average log(income) for the Afro-Caribbean ethnic group. The estimate of β_1 shows how log(income) differs on average for Indian vs Afro-Caribbean ethnic group and the estimate of coefficient β_2 shows how log income differs on average for Pakistani vs Afro Caribbean ethnic group. If we are interested in the way in which log income differs on average for the Indian vs Pakistani ethnic group we can find this out by subtracting the estimate of β_2 from the estimate of β_1.

Dummy variables can be created in SPSS via compute variable or via recode. Both these options appear in the transform menu in SPSS.

3.5 Interactions:

Interactions enable us to assess whether the relationship between the dependent variable and one explanatory variable might change with respect to values of another explanatory variable.
For example, consider a situation where we have a sample of pupils, and the dependent variable is examination performance at 16 (exam16) which we are trying to predict with a previous measure of examination performance based on an exam the pupils took when they were 11 years old (exam11). Suppose we have another explanatory variable, gender.

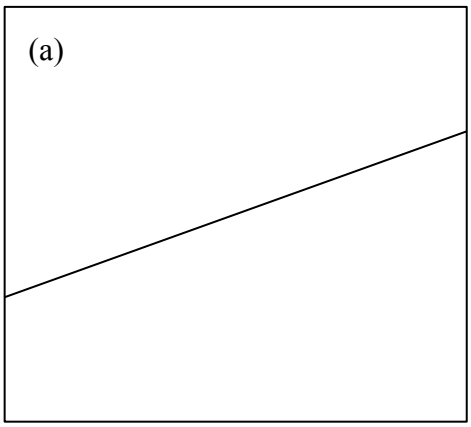
There are three usual things that might be the case for this example (assuming that there is some kind of a relationship between exam16 and exam11).

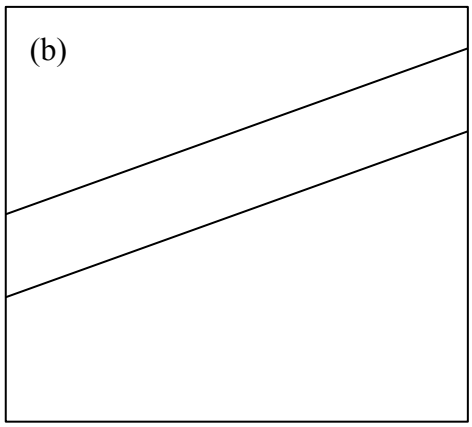
(a) The relationship between exam16 and exam 11 is identical for boys and girls.
(b) The relationship between exam16 and exam11 has a different intercept (overall average) for boys than girls but the nature of the relationship (i.e. the slope) is the same for boys and for girls). In graph (b) below the top line might refer to girls and the bottom line to boys.
(c) The relationship between exam16 and exam11 has a different intercept and a different slope for boys and girls4. In graph (c) below 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.

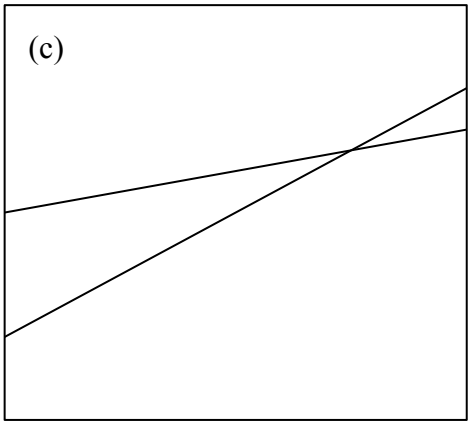
And one other possibility that is less likely to occur in general.

(d) A fourth possibility is that the slope is different for girls and boys but the intercept is identical. In this graph (d, below) one of the lines would refer to girls and the other to boys.

Graphical representations of all four possibilities are shown below:







The simplest model, represented schematically by graph (a) above is one where exam16 and exam11 are positively associated, but there is no difference in this relationship for girls compared with boys. In other words, a single line applies to both genders. The equation for this line is:

$$\text{exam16}_i = \beta_0 + \beta_1 exam11_i + e_i \quad \text{(a)}$$

where exam16 and exam11 are continuous exam scores

If we now consider graph b we might find that there is an overall difference in the level of exam average exam scores but once we have accounted for this overall difference, the relationship between exam16 and exam11 is the same for girls and boys. That is, the lines have the same slope and are therefore parallel.

We can represent this situation via a main effects model where we now have a second explanatory variable. This time it is a categorical (dummy) variable, where gender=0 for boys and gender=1 for girls. Equation (b) is hence a main effects model relating exam11 and gender to exam11.

$$\text{exam16}_i = \beta_0 + \beta_1 exam11_i + \beta_2 gender_i + e_i \quad \text{(b)}$$

Interactions can also be added to the model (this would be appropriate if case (c) applies).

$$\text{exam16}_i = \beta_0 + \beta_1 exam11_i + \beta_2 gender_i + \beta_3 exam11.gender_i + e_i \quad \text{(c)}$$

To create an interaction term such as exam11.gender we simply multiply the two variables exam11 and gender together to create a new variable e.g. ex11gen we then add this to the model as a new explanatory variable. In general you should always leave each of the single variables that make up the interaction term in the model when the interaction term is added.

3.6 Quadratic Relationships.

Sometimes a linear relationship between dependent and explanatory variable may not be appropriate and this is often evident when a scatter plot is produced. For example the linear relationship and quadratic (i.e. curved) relationship for log(hourly wage) vs age are shown below. It seems that although age as a single measure does not explain all the variation in log(hourly wage) it is apparent that the relationship between log(hourly wage) and age is better summarised with a quadratic curve than a straight line.

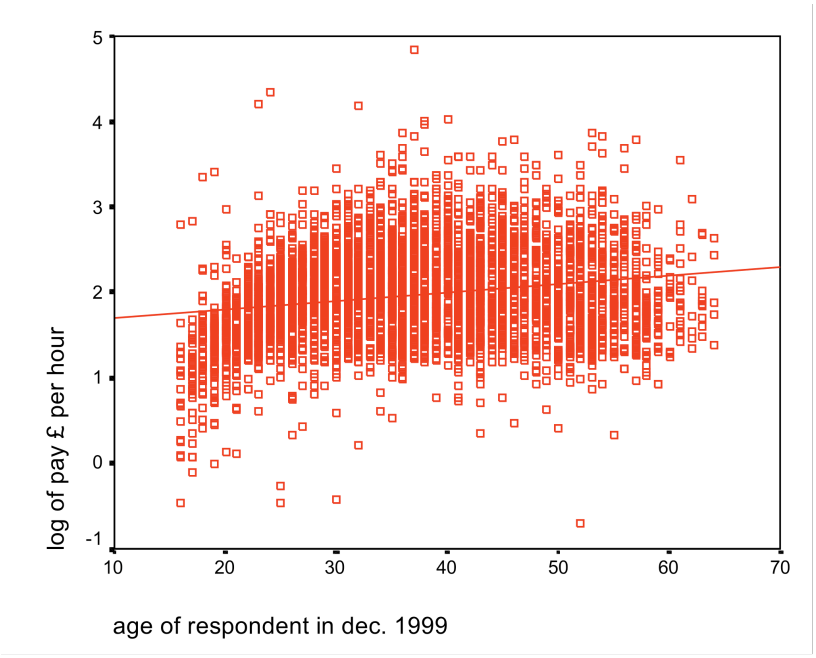


figure (a) above

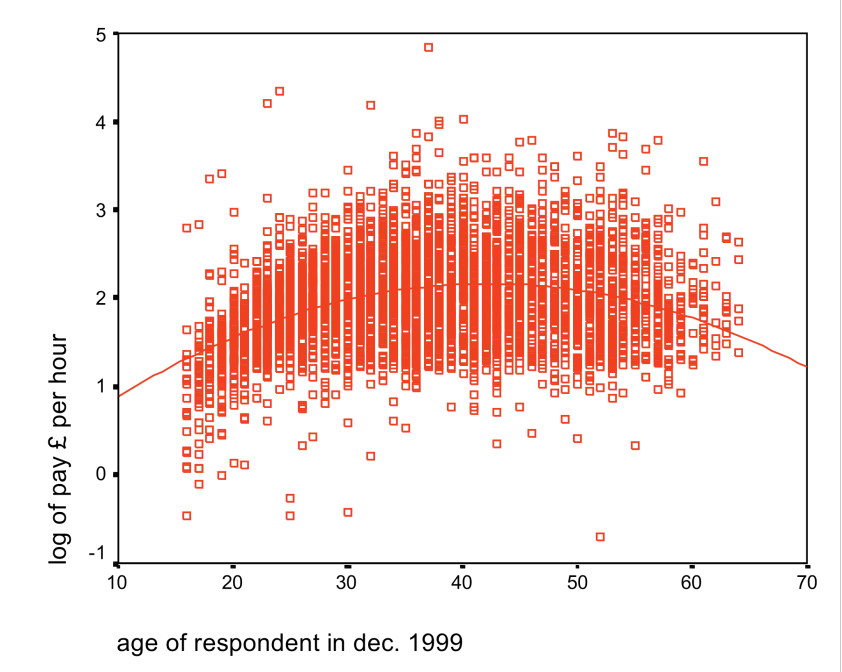


figure (b) above

It is easy to estimate a curve as shown above using SPSS. We first create a new variable: agesq = age^2. We then simply add agesqu into the regression equation as a new explanatory variable.

Hence the equation the straight line shown in figure (a) is:

$$\text{Log(income)}_i = \beta_0 + \beta_1 age_i$$

And the equation for the curve shown in figure (b) is:

$$\text{Log(income)}_i = \beta_0 + \beta_1 age_i + \beta_2 age_i^2$$

Which we could also write equivalently as:

$$\text{Log(income)}_i = \beta_0 + \beta_1 age_i + \beta_2 agesq_i$$

3.6 Model selection methods

In some cases, especially when there are a large number of explanatory variables, we might use statistical criteria to include/exclude explanatory variables, especially if we are interested in the 'best' equation to predict the dependent variable. This is a different fundamental approach to the substantive approach where variables are included on the basis of the research question and this variables are often chosen given the results previous research on the topic and are also influenced by 'common sense' and data availability.

Two examples of selection methods are **backward elimination**, and **stepwise**. The main disadvantage of these methods is that we might miss out important theoretical variables, or interactions. Two selection methods are briefly described below. See Howell page 513 ff for a more detailed description of the methods.

Backward elimination.

Begin with a model that includes all the explanatory variables. Remove the one that is least significant. 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.

Stepwise

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 dependent 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.**

Section 4: BHPS assignment

Using the dataset bhps.sav produce a short report of a multiple regression analysis of the log (hourly wage). The dataset is on the blackboard site.

The report should be between 500-1000 words and might include:

- Appropriate exploratory analysis.
- Appropriate tests of assumptions.
- Dummy variables.
- Interaction terms.
- Squared terms.
- Multiple models (i.e. evidence of a model selection process).

Don't worry about presentational issues for this assignment; we are not after polished pieces of work at this stage. You can cut and paste any relevant SPSS output into appendices. The important point is to the interpretation of the output; the reader should be able to understand the analytical process you have been through. So explain your recodes, dummy variables model selection choices etc.

Reading list

Bryman A and Cramer D (1990) *Quantitative data analysis for social scientists*. Routledge. Chapter 5.

Field, A (2005) Discovering Statistics Using SPSS (Introducing Statistical Methods Second Edition.). Sage Publications

Howell, D (1992*) Statistical methods for psychology*. (3rd Edition) Duxbury. Chapter 15 (and also some of chapter 9).

Plewis, I (1997) *Statistics in Education.* Edward Arnold.

More theoretical:

Draper and Smith (1981) Applied regression analysis (2nd ed). Wiley.[nb: although this book uses the word 'applied' in the title, it is actually more theoretical than the reference above by Howell]