Chapter 15

Regression Models – Examples

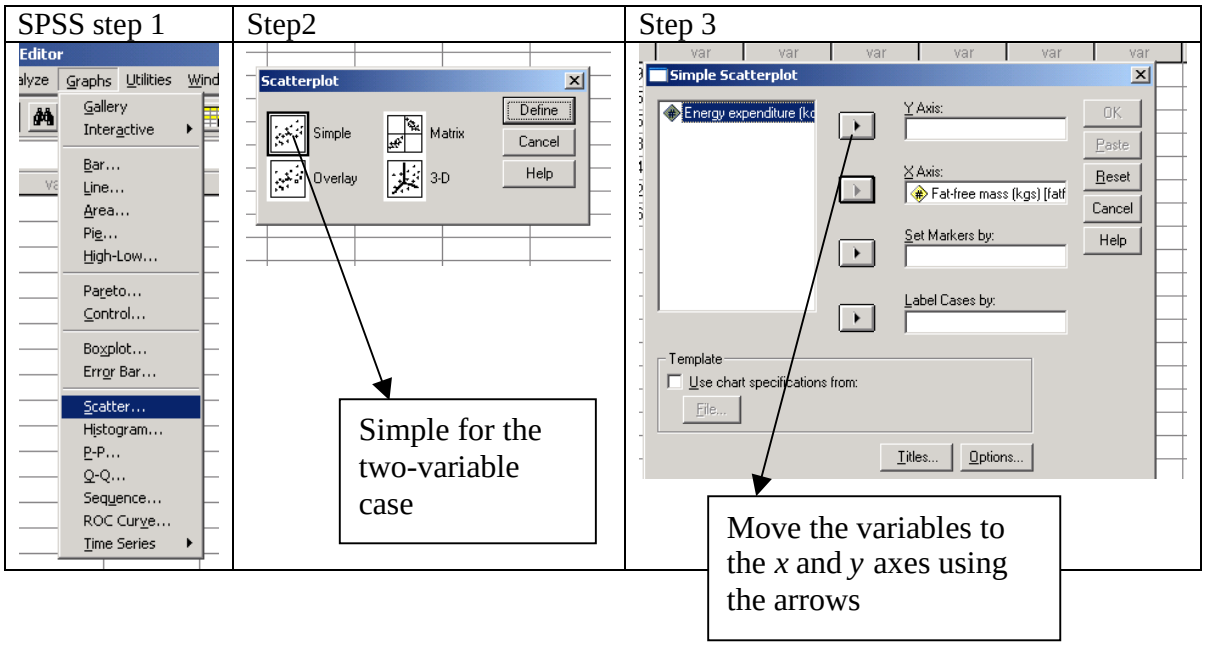
15.1 Simple Linear Regression Example:

A study investigated the relationship between energy expenditure and body build. For a random sample of seven adult men, underwater weighing techniques were used to determine the fat-free body mass in kg. of each of them. The total 24-hour energy expenditure was also measured. The data are as follows:

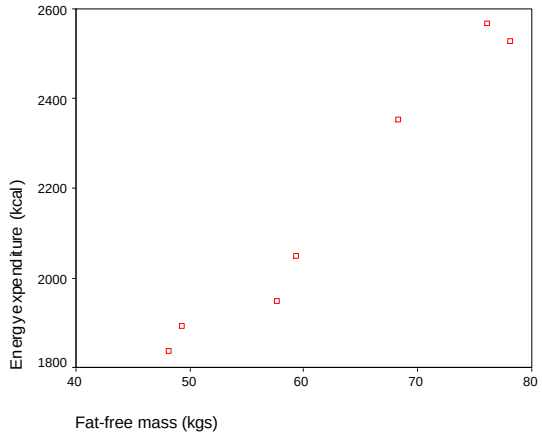
Participant		1	2	3	4	5	6	7
Fat-free mass kg	*x*	49.3	59.3	68.3	48.1	57.6	78.1	76.1
Energy kcal	*y*	1894	2050	2353	1838	1948	2528	2568

15.1.1.The Scatter Diagram:

A plot of the data provides a scatter diagram showing the relationship between *x* and y. The response variable *y* is plotted on the *y*-axis, the explanatory variable is plotted on the *x*-axis. If the relationship between the two variables is **linear**, then the scatter plot should show this straight-line relationship (albeit with a certain amount of random scatter)



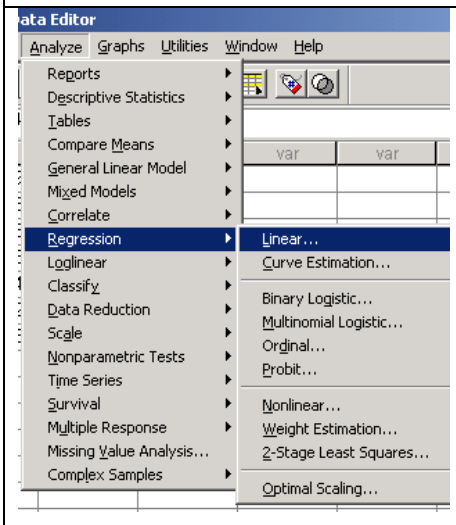
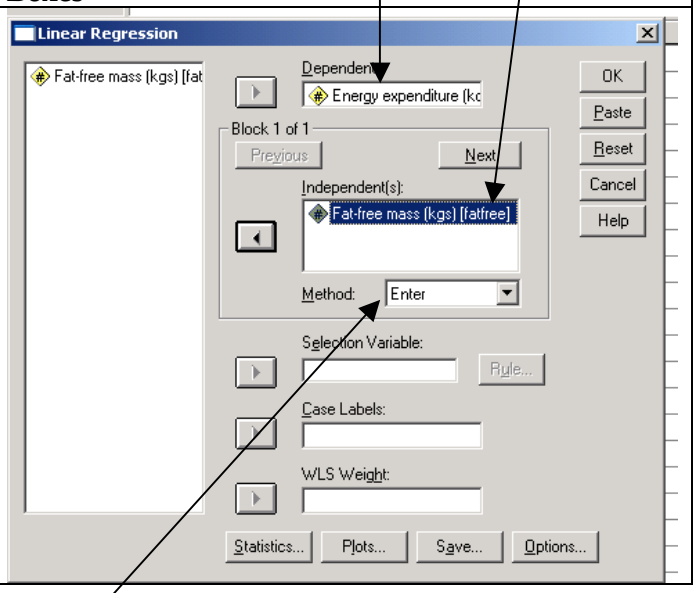
Example (continued): Scatter Plot.



It is assumed that energy expenditure depends on fat-free body mass; hence energy expenditure is the response variable (or dependent variable) y and fat-free mass is the explanatory variable (or independent variable x).

This plot is approximately linear; as the fat-free mass increases, the energy expenditure increases.

15.1.2 Least Square Estimation for the simple linear regression line

Analyze – Regression - Linear	**Move variables to the Dependent and Independent Boxes**
	

The **Method** is left at the default '**Enter**'

Resulting Output

15.1.3 Simple linear regression: inference using SPSS output

a) Model fitting

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	Fat-free mass (kgs)[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: Energy expenditure (kcal)

Variation explained by the model

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.981[a]	.963	.956	64.848

a. Predictors: (Constant), Fat-free mass (kgs)

The R Square (R^2) measures the % of variation explained by the model. For example in the above model 96.3 % of the variation in the *y*-variable (energy expenditure) was explained by the model.

The least squares estimates of the coefficients

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	607.703	138.765		4.379	.007
	Fat-free mass (kgs)	25.012	2.189	.981	11.427	.000

a. Dependent Variable: Energy expenditure (kcal)

From the Table, the estimated regression line is

Energy Expenditure = 607.703 + 25.012 fat-free mass

b) Inference about β_1 the slope.

Confidence Interval about β_1: a 100(1-α)% confidence interval for β_1 is given by:

$$\hat{\beta}_1 \pm t_{n-2,\alpha/2}(std.error.\hat{\beta}_1)$$

Example: $\hat{\beta}_1$**= 25.012** with Std. Error = **2.189.** There are 7 observations (n=7) and two parameters have been estimated, giving n-2 df=5df. $t_{5,0.025} = 2.571$

Hence 95% CI for β_1 is given by **25.012 ± 2.189*2.571 = (19.389, 30.645)**

Notice that this interval does not contain zero.

Hypothesis test about β_1:

$H_0: \beta_1 = 0$ No linear relationship between x and y
$H_1: \beta_1 \neq 0$ There is a linear relationship

Test Statistic: $T = \dfrac{\hat{\beta}_1}{std.Error(\hat{\beta}_1)}$

Reject H_o if $T > t_{n-2,\alpha/2}$ or $T < -t_{n-2,\alpha/2}$

Otherwise accept the null hypothesis

Example

$H_o: \beta_1 = 0$ No relationship between fatfree mass and energy expenditure
$H_i: \beta_1 \neq 0$ There is a linear relationship
Observed T = 11.427, p = 0.000.

Hence very strong evidence to reject the null hypothesis and conclude that there is a linear relationship between fat free mass and energy expenditure.

b) Similarly, β_0 is significantly different from zero in this model since T = 4.379 and p= 0.007.

c) **Test of the significance of the regression using the F-distribution.**

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	549097.6	1	549097.619	130.575	.000[a]
	Residual	21026.096	5	4205.219		
	Total	570123.7	6			

a. Predictors: (Constant), Fat-free mass (kgs)

b. Dependent Variable: Energy expenditure (kcal)

If the regression is not significant, then *y* does not depend on *x*. The hypotheses may be written:

$H_o: \beta_1 = 0$ (y does not depend on x) model : $y_i = \beta_o + \varepsilon_i$
$H_i: \beta_1 = 0$ (y does depend on x) model: $y_i = \beta_o + \beta_1 x_i + \varepsilon_i$

Test Statistic $F = \dfrac{MS(regression)}{MS(residual)}$

Reject H_o if observed $F > F_{1,n-2,\alpha}$. Conclude that *y* does depend on *x*.
Otherwise accept the null hypothesis and conclude that *y* does not depend on *x*.

Example:

$H_o : \beta_1 = 0$ energy expenditure does not depend on fat - free mass, model : $y_i = \beta_o + \varepsilon_i$
$H_i : \beta_1 = 0$ energy expenditure depends linearly on fat - free mass model : $y_i = \beta_o + \beta_1 x_i + \varepsilon_i$

Observed F = 130.575 and p =0.000 < 0.001, Very, very strong evidence to reject the null hypothesis and accept the alternative. Conclude that the energy expenditure depends linearly on fat-free mass.

Here the **best** linear equation is given by y = 607.7 + 25.012 x

Prediction

The predicting equation is given by:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

Substituting specific values for *x* will give the predicted value of y.

***Predicting equation*:** $\hat{y}$ = 607.7 +25.012x
when x = 60 kg, $\hat{y}$ = 607.7 +25.012*60 = 2108 kcals

Comments:

i) Predicting outside the range of values of *x* on which the equation was estimated, should be done with caution. The seemingly linear relationship may not persist over all values of *x*

ii) **Interpretation of the coefficients:**
The *slope* represents the amount by which *y* changes for every unit change in *x*. The *intercept* represents the value of *y* when *x* is zero.

Example: The value of 25.012 for the gradient implies that, on average, with every increase in one kg of fat-free mass, the energy expenditure increases by 25.01kcals. The constant term of 607.7 could be interpreted as the energy expenditure when the fat-free mass was zero.

15.1.4 Assumptions of the Model:

1. For any given *x*, *y* is a random variable with a probability distribution.
2. The errors are random errors with mean zero and variance σ^2. The errors are independent.
3. The errors are assumed to have constant variance, σ^2.
4. In order to make tests of significance and construct confidence intervals, it is necessary to assume that the errors are Normally distributed.

Thus the model is $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, $i = 1,2,\ldots n.$ where the ε_i are independent, $N(0, \sigma^2)$ for all i = 1,2,…n.

The Residuals

After fitting the model, the assumptions of the model are validated using residual analysis.
The residuals, e_i provide an estimate of the errors, ε_i.

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

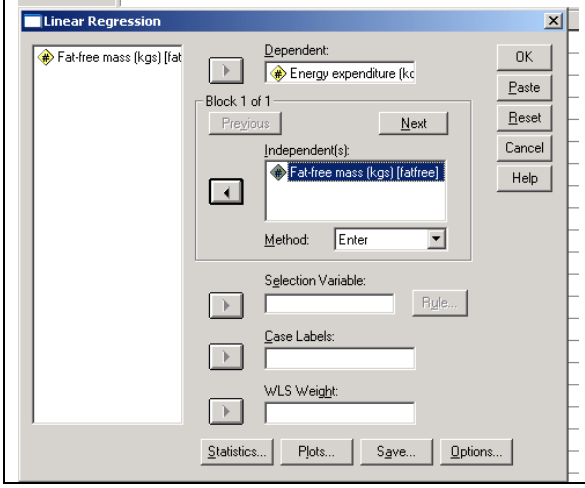
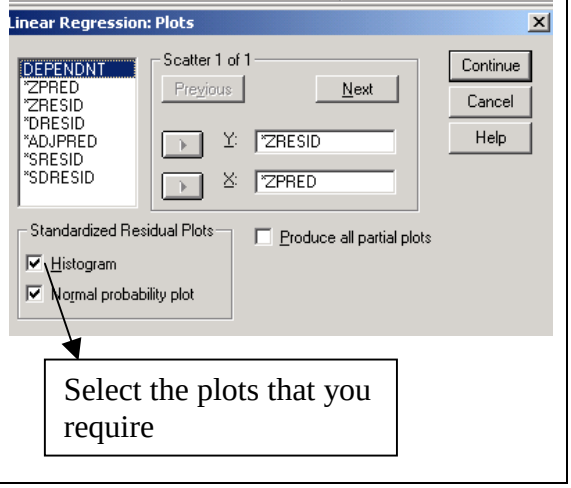
Residual Analysis to validate the assumptions of the model.

The residuals are given by: $e_i = y_i - \hat{y}_i$

The assumptions are:

i) Errors are independent
ii) Mean zero, constant variance
iii) Normally distributed.

Using SPSS to validate the assumptions

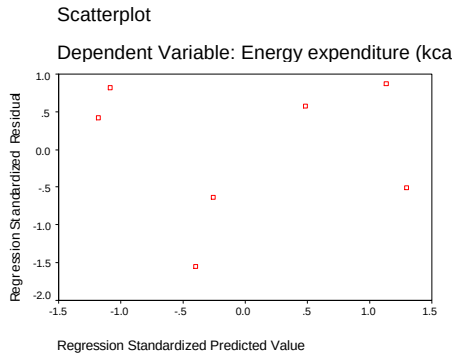
Within the Regression dialog box,	**Click on Plots**
	

SPSS Output:

Plot of residuals v fitted values

Should be randomly scattered about zero with fairly constant 'spread' if assumption of independence and homogeneity are valid.

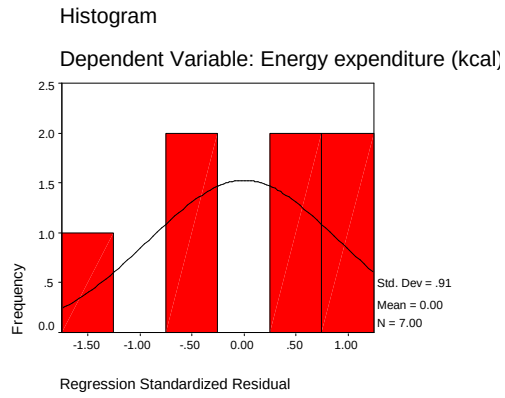
If the model fitted is inadequate this can also be noticed from a distinctive pattern to this plot.



Comments: This appears to be random but the sample size is very small

Histogram

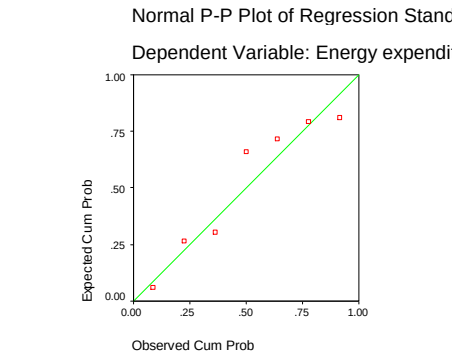
Should be symmetric about zero and bell shaped, if the Normality assumption is valid.



The smallness of the sample size makes it difficult to see any definite shape to the histogram.

Normal Probability Plot

Should be a straight line if the normality assumption is valid



The sample size again makes this difficult to interpret with any degree of certainty.

15.2 Multiple linear regression

15.2.1 Introduction

This is applicable when the data are multivariate. A multiple linear regression model relates a **response** variable *Y* to **more** than one explanatory variable.
The main purpose of the multiple regression analysis is to find which explanatory variables contribute to the variation of the response variable. We are usually looking for the 'best' **subset** of the explanatory variables.

The Model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} + \varepsilon_i , \qquad i = 1,2,...n.$$

where: k is the number of explanatory variables,

β_o , $\beta_1, \cdots \beta_k$ are the parameters of the model,

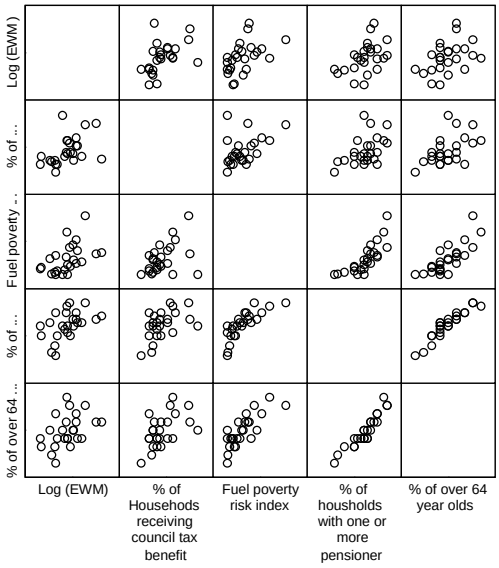
ε_i is a random error term.

15.2.2 Example: Data were collected by Rudge (2004) on excess winter morbidity in the period 1993 – 1996 in 25 grouped enumeration districts in Newham, London.

The explanatory variables are as follows:
CTB = % households receiving council tax benefit
FPR = Fuel poverty risk index
HHI = % of households with one or more pensioners
PERS = % of over 65 year olds

The basic response variable is the excess winter morbidity EWM in the period 1993-1996 and the analysis uses log(EWM).

The Multiple Scatter Diagrams:
The response variable *y* and the all the *x* continuous variables are plotted against each other.



From the scatter plots it is not clear which of the explanatory variables has the 'best' single linear relationship with the response variable. However, there are some very strong relationships between the explanatory variables. This may cause a problem.

Correlation Matrix

The relationships are confirmed by the output from the correlation matrix below.

Correlations

		Log (EWM)	% of Households receiving council tax benefit	Fuel poverty risk index	% of housholds with one or more pensioner	% of over 64 year olds
Log (EWM)	Pearson Correlation	1	.473*	.497*	.485*	.436*
	Sig. (2-tailed)		.017	.012	.014	.029
	N	25	25	25	25	25
% of Households receiving council tax benefit	Pearson Correlation	.473*	1	.430*	.495*	.523**
	Sig. (2-tailed)	.017		.032	.012	.007
	N	25	25	25	25	25
Fuel poverty risk index	Pearson Correlation	.497*	.430*	1	.796**	.797**
	Sig. (2-tailed)	.012	.032		.000	.000
	N	25	25	25	25	25
% of housholds with one or more pensioner	Pearson Correlation	.485*	.495*	.796**	1	.940**
	Sig. (2-tailed)	.014	.012	.000		.000
	N	25	25	25	25	25
% of over 64 year olds	Pearson Correlation	.436*	.523**	.797**	.940**	1
	Sig. (2-tailed)	.029	.007	.000	.000	
	N	25	25	25	25	25

*. Correlation is significant at the 0.05 level (2-tailed).

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

The response variable is significantly correlated with each of the explanatory variables and the Fuel Poverty Risk Index is the single variable that 'best' describes log(EWM).
(% households with one or more pensioners is highly correlated with % of over 64 year olds – unsurprisingly and the other explanatory variables show significant relationships amongst themselves too).

Model 1: Best Simple Linear Regression Model

The response variable is log EWM. The best simple linear regression model will be the model relating log EWM to fuel poverty risk index. The output is given below.

Output from Simple Linear Regression (using ENTER).

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	Fuel poverty risk index[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: Log (EWM)

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.497[a]	.247	.214	.1414701

a. Predictors: (Constant), Fuel poverty risk index

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.151	1	.151	7.538	.012[a]
	Residual	.460	23	.020		
	Total	.611	24			

a. Predictors: (Constant), Fuel poverty risk index

b. Dependent Variable: Log (EWM)

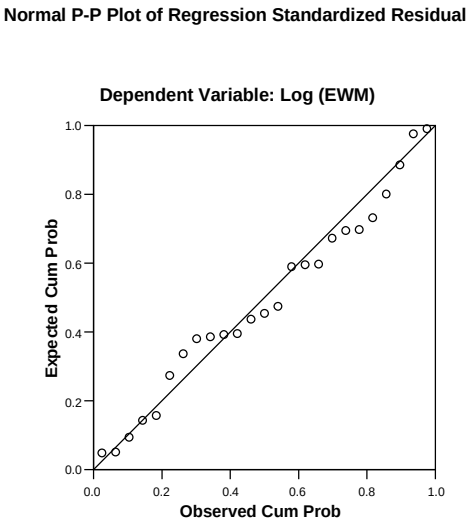
Coefficients[a]

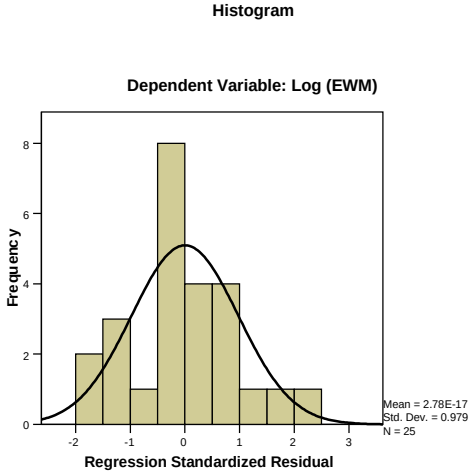
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.213	.042		5.066	.000
	Fuel poverty risk index	5.982E-05	.000	.497	2.746	.012

a. Dependent Variable: Log (EWM)

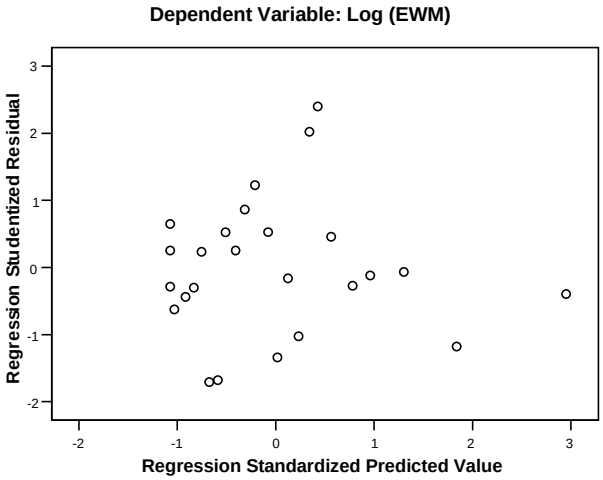
Comment: From the ANOVA Table and from the Table of Coefficients, it can be seen that there is a significant, linear relationship between log(EWM) and the Fuel Poverty Risk Index since the sig. values are less than 0.05.

Write down the assumptions of the model and use the residual analysis to comment on whether they are likely to be valid



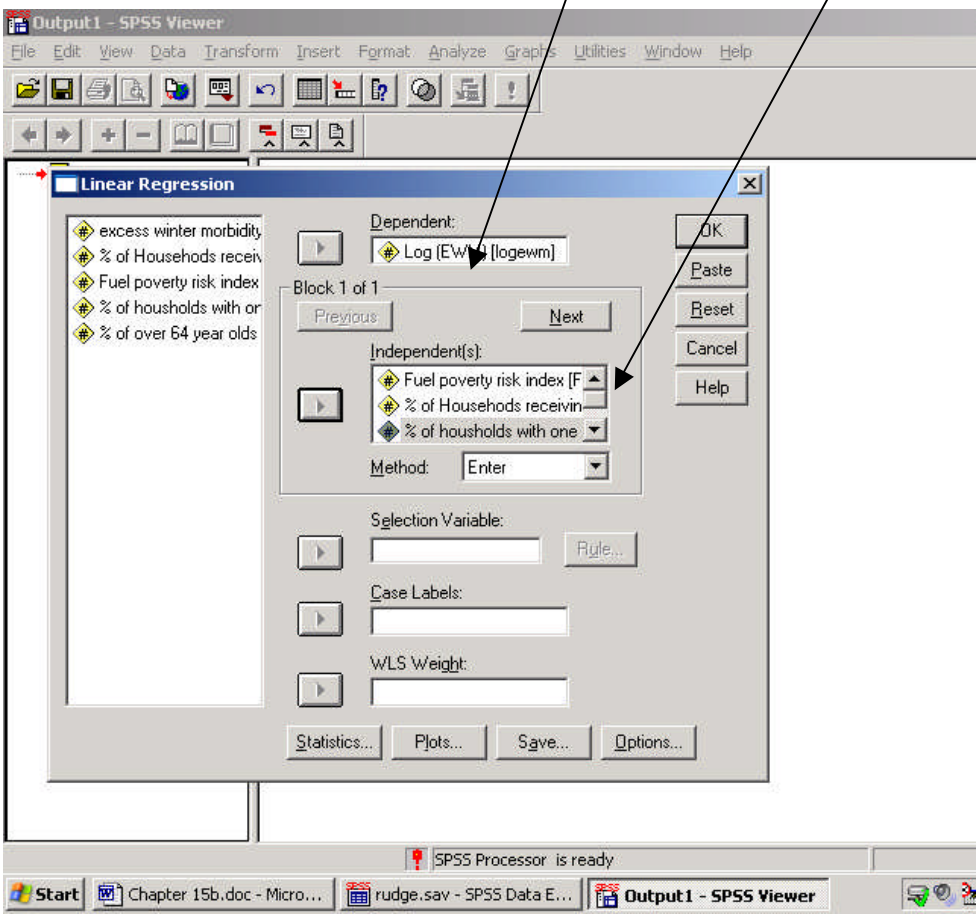


Scatterplot



15.2.3 Multiple Regression Analysis

This example is a small example. In the Regression dialog box, all or some of the explanatory variables of choice can be moved to the **Independent(s)** Box. The **Method** can be selected from a list including ENTER, STEPWISE, REMOVE, FORWARD, BACKWARD.



The main ones to use are ENTER, REMOVE and STEPWISE.
There are 4 explanatory variables to choose from. Fuel Poverty Risk Index (FPR) was entered first. We now can think about adding another variable to the list of independent variables and achieving a better model that explains more of the variation.

% of households with one or more pensioners (HHI) is the next most correlated variable with log (EWM). However this variable is also very highly correlated with the FPR (their correlation is 0.796 with sig = 0.000). A better choice may be to include % Households receiving Council Tax Benefit (CTB) as this is less correlated with the already included explanatory variable FPR but significantly correlated with the response variable log(EWM)

Model 2: Multiple Regression Model including FPR and CTB

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.574[a]	.329	.268	.1365235

a. Predictors: (Constant), % of Households receiving council tax benefit, Fuel poverty risk index

b. Dependent Variable: Log (EWM)

R Square Model 2 = 0.329, compared to R Square Model 1 = 0.247
Adjusted R Square Model 2 = 0.268 compared to Adjusted R Square Model 1 = 0.214

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.201	2	.101	5.396	.012[a]
	Residual	.410	22	.019		
	Total	.611	24			

a. Predictors: (Constant), % of Households receiving council tax benefit, Fuel poverty risk index

b. Dependent Variable: Log (EWM)

Model 2 is significant as the Sig = 0.012 < 0.05. Hence FPR and CTB are jointly significant in explaining log(EWM).

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.034	.116		.298	.769
	Fuel poverty risk index	4.337E-05	.000	.360	1.862	.076
	% of Households receiving council tax benefit	.005	.003	.318	1.642	.115

a. Dependent Variable: Log (EWM)

The Sig values here indicate that **after including FPR** in the model, CTB does **not** add significantly to the model and hence should **not** be included, 0.115>0.05.

Model 3: Multiple Regression Model including FPR and HHI

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.518[a]	.269	.202	.1425511

a. Predictors: (Constant), % of housholds with one or more pensioner, Fuel poverty risk index

b. Dependent Variable: Log (EWM)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.164	2	.082	4.038	.032[a]
	Residual	.447	22	.020		
	Total	.611	24			

a. Predictors: (Constant), % of housholds with one or more pensioner, Fuel poverty risk index

b. Dependent Variable: Log (EWM)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-.019	.290		-.065	.949
	Fuel poverty risk index	3.650E-05	.000	.303	1.006	.325
	% of housholds with one or more pensioner	.009	.012	.243	.808	.428

a. Dependent Variable: Log (EWM)

In all respects this model is worse than model 2. And model 1 was better than model 2.

It can be seen, perhaps, that finding the best subset of the explanatory variables to explain the behaviour of the response variable is not easy.

The Method: Stepwise can often provide useful guidance.

15.2.4 Stepwise Regression

There are many ways to construct a 'best' regression equation from a large set of x-variables.

Backward elimination: We begin with a model that includes all the predictors and we try to eliminate the ones that contributed the least to the model.

Forward selection: We start with the constant and add only significant variables.

Stepwise selection: Add one variable at the time in the models as in forward but also check whether existing variables can be removed.

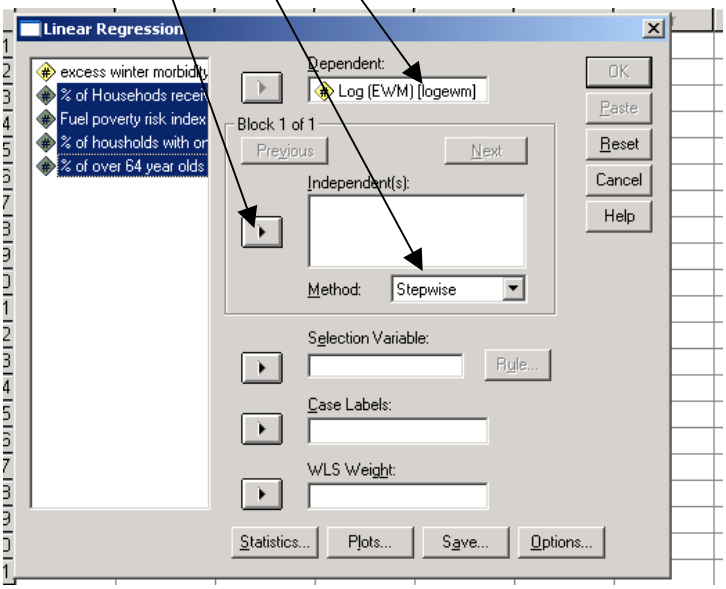
The following output uses the option STEPWISE.

In stepwise regression, all the explanatory variables are usually included in the independent(s) list. Using Stepwise procedure on the Rudge example provides the following output

Stepwise Output on the Rudge Example:
Move **log(EWM)** to the **dependent** box

Move **ALL** the possible explanatory variables to the **Independent** Box

Set the Method to **Stepwise**



STEPWISE Output

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Fuel poverty risk index	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

a. Dependent Variable: Log (EWM)

This confirms that the best model contains Fuel Poverty Risk Index only

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.497[a]	.247	.214	.1414701

a. Predictors: (Constant), Fuel poverty risk index

b. Dependent Variable: Log (EWM)

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.151	1	.151	7.538	.012[a]
	Residual	.460	23	.020		
	Total	.611	24			

a. Predictors: (Constant), Fuel poverty risk index

b. Dependent Variable: Log (EWM)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	.213	.042		5.066	.000
	Fuel poverty risk index	5.982E-05	.000	.497	2.746	.012

a. Dependent Variable: Log (EWM)

Excluded Variables[b]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
1	% of Households receiving council tax benefit	.318[a]	1.642	.115	.330	.815
	% of housholds with one or more pensioner	.243[a]	.808	.428	.170	.366
	% of over 64 year olds	.109[a]	.357	.724	.076	.365

a. Predictors in the Model: (Constant), Fuel poverty risk index

b. Dependent Variable: Log (EWM)

This table lists the excluded variables.

The output confirms the decision that the best model relates log(EWM) to the Fuel Poverty Risk Index. When FPR is included, the other explanatory variables do not add significantly to the model.

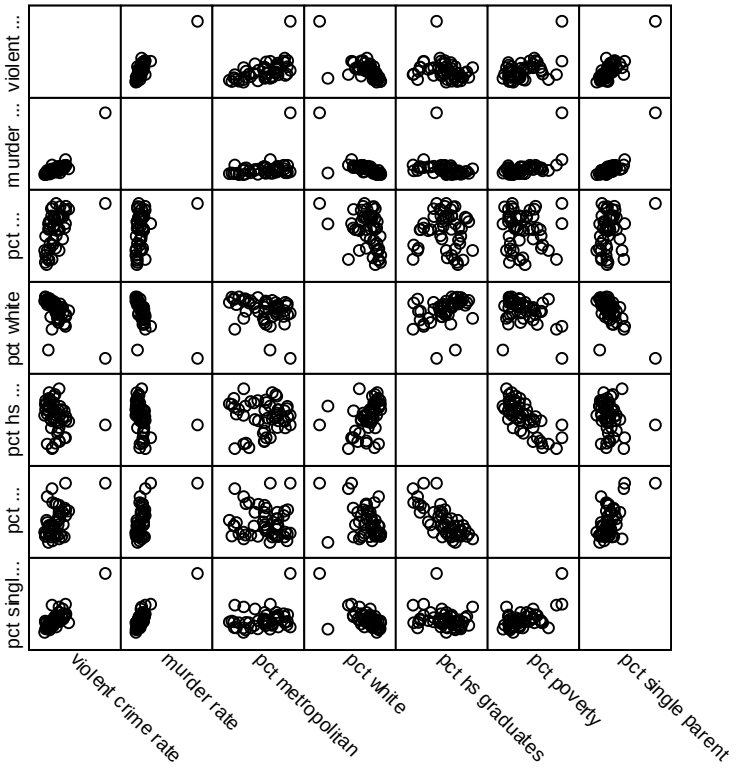
15.3 Stepwise Regression

15.3.1 Example

These data are from *Statistical Methods for the Social Sciences, Third Edition* by A. Agresti and B. Finlay (Prentice Hall, 1977). The variables are as follows:

Crime: violent crimes per 100,000 population
Murder: murders per 1,000,000 population
Pctmetro: % population living in metropolitan areas
Pctwhite: % population that is white
Pcths: % population with high school education or above
Poverty: % population living under the poverty line
Single: % population that are single parents

The data are collected from the states of the USA. The variable of interest is **crime.**



Comment: Notice that in most of the plots there is one unusual observation. It may be worth investigating (at some point) the effect of removing this case.

Correlations

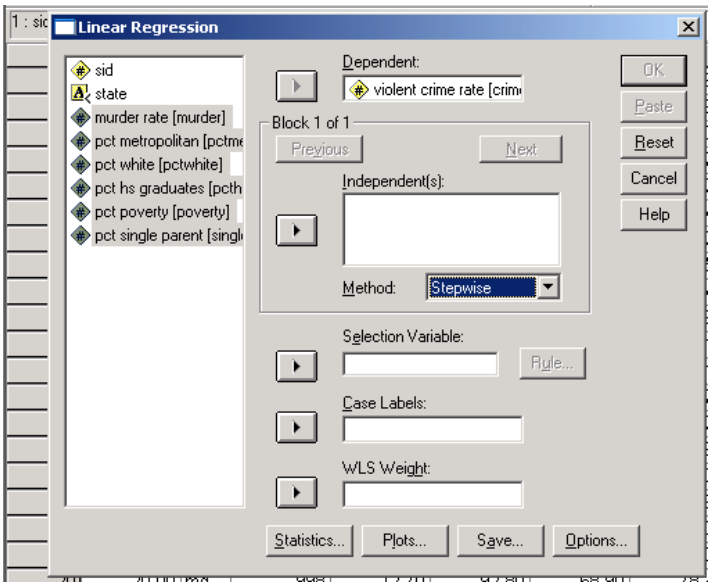
		violent crime rate	murder rate	pct metropolitan	pct white	pct hs graduates	pct poverty	pct single parent
violent crime rate	Pearson Correlation	1	.886**	.544**	-.677**	-.256	.510**	.839**
	Sig. (2-tailed)		.000	.000	.000	.070	.000	.000
	N	51	51	51	51	51	51	51
murder rate	Pearson Correlation	.886**	1	.316*	-.706**	-.286*	.566**	.859**
	Sig. (2-tailed)	.000		.024	.000	.042	.000	.000
	N	51	51	51	51	51	51	51
pct metropolitan	Pearson Correlation	.544**	.316*	1	-.337*	-.004	-.061	.260
	Sig. (2-tailed)	.000	.024		.016	.978	.673	.066
	N	51	51	51	51	51	51	51
pct white	Pearson Correlation	-.677**	-.706**	-.337*	1	.339*	-.389**	-.656**
	Sig. (2-tailed)	.000	.000	.016		.015	.005	.000
	N	51	51	51	51	51	51	51
pct hs graduates	Pearson Correlation	-.256	-.286*	-.004	.339*	1	-.744**	-.220
	Sig. (2-tailed)	.070	.042	.978	.015		.000	.121
	N	51	51	51	51	51	51	51
pct poverty	Pearson Correlation	.510**	.566**	-.061	-.389**	-.744**	1	.549**
	Sig. (2-tailed)	.000	.000	.673	.005	.000		.000
	N	51	51	51	51	51	51	51
pct single parent	Pearson Correlation	.839**	.859**	.260	-.656**	-.220	.549**	1
	Sig. (2-tailed)	.000	.000	.066	.000	.121	.000	
	N	51	51	51	51	51	51	51

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

*. Correlation is significant at the 0.05 level (2-tailed).

The correlations show that, unsurprisingly, the murder rate is highly correlated with the crime rate.

SPSS Stepwise:
The response variable **crime** goes in the **Dependent** box
All the scale explanatory variables go in the **Independent** box (do not include categorical variables here)
Stepwise is the chosen method



You should investigate the options offered by **Statistics, Plots, Save, Options.**
The plots will be required when we have selected a model and wish to test the validity of the assumptions.

Stepwise Output

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	murder rate	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
2	pct metropolitan	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
3	pct single parent	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

a. Dependent Variable: violent crime rate

Murder Rate goes in first (highest correlation)
Pct metropolitan goes in second
Pct single parent goes in third.

The following variables are not included:
Pctwhite: % population that is white
Pcths: % population with high school education or above
Poverty: % population living under the poverty line

Model Summary[d]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.886[a]	.785	.781	206.441
2	.929[b]	.863	.857	166.804
3	.942[c]	.888	.881	152.369

a. Predictors: (Constant), murder rate

b. Predictors: (Constant), murder rate, pct metropolitan

c. Predictors: (Constant), murder rate, pct metropolitan, pct single parent

d. Dependent Variable: violent crime rate

This provides diagnostics for the three models that have been fitted. Notice how the adjusted R Square increases to 0.881.

ANOVA[d]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	7640199	1	7640198.858	179.272	.000[a]
	Residual	2088276	49	42617.875		
	Total	9728475	50			
2	Regression	8392939	2	4196469.484	150.824	.000[b]
	Residual	1335536	48	27823.662		
	Total	9728475	50			
3	Regression	8637307	3	2879102.319	124.012	.000[c]
	Residual	1091168	47	23216.336		
	Total	9728475	50			

a. Predictors: (Constant), murder rate

b. Predictors: (Constant), murder rate, pct metropolitan

c. Predictors: (Constant), murder rate, pct metropolitan, pct single parent

d. Dependent Variable: violent crime rate

The ANOVA shows that the three fitted models are all highly significant. (The sig values are all less than 0.001).

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	294.527	37.428		7.869	.000
	murder rate	36.473	2.724	.886	13.389	.000
2	(Constant)	-69.117	76.174		-.907	.369
	murder rate	32.658	2.320	.794	14.077	.000
	pct metropolitan	5.890	1.132	.293	5.201	.000
3	(Constant)	-707.561	208.727		-3.390	.001
	murder rate	21.663	3.997	.526	5.420	.000
	pct metropolitan	5.971	1.035	.297	5.771	.000
	pct single parent	64.364	19.839	.310	3.244	.002

a. Dependent Variable: violent crime rate

This table gives the coefficients for the three models that were fitted.
Notice that the higher the pct metropolitan, the higher the crime rate
The higher the pct single parent, the higher the crime rate

For the excluded variables, at each stage, diagnostics are presented below:

Excluded Variables[d]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics
						Tolerance
1	pct metropolitan	.293[a]	5.201	.000	.600	.900
	pct white	-.102[a]	-1.098	.278	-.157	.501
	pct hs graduates	-.003[a]	-.040	.969	-.006	.918
	pct poverty	.012[a]	.146	.885	.021	.680
	pct single parent	.296[a]	2.402	.020	.328	.262
2	pct white	-.037[b]	-.477	.636	-.069	.487
	pct hs graduates	-.031[b]	-.543	.590	-.079	.910
	pct poverty	.127[b]	1.915	.062	.269	.616
	pct single parent	.310[b]	3.244	.002	.428	.262
3	pct white	-.004[c]	-.060	.953	-.009	.477
	pct hs graduates	-.040[c]	-.776	.442	-.114	.907
	pct poverty	.099[c]	1.604	.116	.230	.603

a. Predictors in the Model: (Constant), murder rate

b. Predictors in the Model: (Constant), murder rate, pct metropolitan

c. Predictors in the Model: (Constant), murder rate, pct metropolitan, pct single parent

d. Dependent Variable: violent crime rate

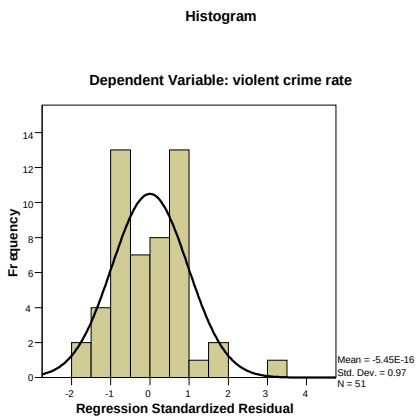
Looking at the table above, when model 1 is fitted, you can see that the next most important variable is **pct metropolitan (t = 5.201)** and that is the variable next included.
At stage 2 (murder, pct metropolitan included) **pct single parent** will be included next (t = 3.244).
At stage 3, none of the remaining variables are significant (pct poverty has a t value of 1.604 which is non- significant).

Diagnostics:
The assumptions of the model are that the errors are

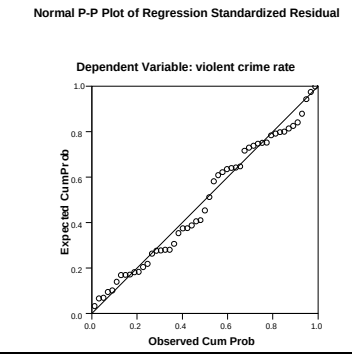
1. Random errors with mean zero and variance σ^2. .
2. The errors are assumed to have constant variance, σ^2.
3. The errors are independent
4. In order to make tests of significance and construct confidence intervals, it is necessary to assume that the errors are Normally distributed.

The Residuals
After fitting the model, the assumptions of the model are validated using residual analysis. The residuals, e_i provide an estimate of the errors, ε_i.

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

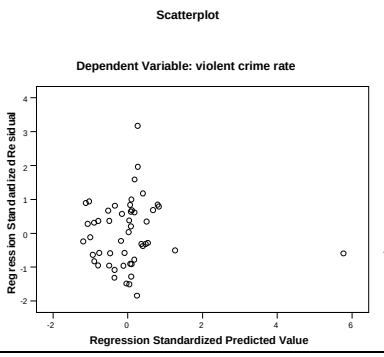


The histogram evaluates the normality assumption. This looks reasonably symmetric.



The points should lie on a straight line if the errors are normally distributed.

This seems to show that there may be further work required before the normality assumption is valid.



Residuals v fitted values should show random scatter.

Notice here the outlier. The next step should be to remove this and see what effect that has.

Practical 15: Multiple Linear Regression

The data set in the file SHARED (K):\SCTMS\SOM\MA2010\REGRESSION\RUDGE1.SAV comprises data collected by Rudge (2004) on excess winter morbidity in the period 1993 – 1996 in 25 grouped enumeration districts in Newham, London.

The explanatory variables are as follows:
CTB = % households receiving council tax benefit
FPR = Fuel poverty risk index
HHI = % of households with one or more pensioners
PERS = % of over 65 year olds

The basic response variable is the excess winter morbidity EWM in the period 1993-1996 and the analysis uses ln(EWM).

a) Open the data file in SPSS and calculate ln(EWM).

b) Reproduce the results from the multiple linear regression example in your course notes (section 15.2.2 and 15.2.3)

i) Use Graph>Scatter/Dot>Matrix Scatter to produce multiple plots of ln(EWM) and the explanatory variables.

ii) Use Analyze>Correlate>Bivariate to obtain a correlation matrix of ln(EWM) and the explanatory variables.

iii) Fit the best linear regression model relating ln(EWM) to fuel poverty risk index, FPR (Model 1) and use the option PLOTS to obtain residual plots for the model.

Write down the estimated regression equation for the best simple linear model fitted. Calculate the predicted value of EWM for FPR = 2500.

Write down the assumptions of the model and use the residual analysis to comment on whether they are likely to be valid.

iv) Now fit the multiple regression model relating ln(EWM) to FPR and CTB (Model 2) and the model relating ln(EWM) to FPR and HHI (Model 3).

For each of the models, write down the estimated regression equation and calculate the predicted value of EWM for CTB = 50, FPR = 2500, HHI = 40 and PERS = 30.