

# Session 7

## Log-linear Models

	<i>page</i>
Multi-way Tables	7-6
Example 1	7-8
Interpretation of Parameters	7-12
Continuous Covariates	7-13
SPSS commands for Log-Linear Models	7-14
Practical Session 7: Log-Linear Models	7-15

## Session 7: Log-Linear Models

The analysis of multi-way contingency tables is based on log-linear models. In order to develop this theory, consider the simpler situation of a two-way tables as produced by a cross-tabulation of **SEX** by **LIFE** (GSS91 data).

Respondent's Sex \* Is Life Exciting or Dull Crosstabulation

			Is Life Exciting or Dull			Total
			Exciting	Routine	Dull	
Respondent's Sex	Male	Count	213	200	12	425
		Expected Count	188.2	219.0	17.8	425.0
		% within Respondent's Sex	50.1%	47.1%	2.8%	100.0%
	Female	Count	221	305	29	555
		Expected Count	245.8	286.0	23.2	555.0
		% within Respondent's Sex	39.8%	55.0%	5.2%	100.0%
Total	Count	434	505	41	980	
	Expected Count	434.0	505.0	41.0	980.0	
	% within Respondent's Sex	44.3%	51.5%	4.2%	100.0%	

### Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	11.994 <sup>a</sup>	2	.002
Likelihood Ratio	12.109	2	.002
Linear-by-Linear Association	11.973	1	.001
N of Valid Cases	980		

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 17.78.

We might ask how were these results derived?

Firstly, we make the assumption that the two variables are **independent**. This means that we are assuming that the probability of a response to **LIFE** is the same for both sexes, so that the probability of being in a particular cell, e.g. SEX=1 and LIFE=2 can be found by multiplying together the probability that SEX=1 and the probability that LIFE=2 (for both sexes). Knowing the total sample size, we can then work out the expected cell frequency counts. We measure the discrepancy between the observed and expected cell counts using the general principle of Deviance based on the Likelihood. The statistical distribution of the cell counts is the Multinomial distribution, which is a generalisation of the Binomial distribution. It turns out that, mathematically, we get exactly the same results if we assume the cell counts have the Poisson distribution. This Deviance measure is tested using the Chi-squared distribution following general theory. A modified version due to Karl Pearson is often referred to as the chi-squared goodness-of-fit statistic.

How do we express this as a statistical model?

Consider the model for the cell probabilities:

$$p(\text{SEX}=i, \text{LIFE}=j) = p(\text{SEX}=i) \times p(\text{LIFE}=j)$$

So, for  $p(\text{SEX}=1 \text{ (male)}, \text{LIFE}=2 \text{ (routine)})$  we need  $p(\text{SEX}=1)$  and  $p(\text{LIFE}=2)$ .

On the basis of the sample we see that we can estimate  $p(\text{SEX}=1)$  by  $425/980=0.434$  as the proportion of males in the sample. Similarly, we can estimate  $p(\text{LIFE}=2)$  by  $505/980=0.515$ . This is based on the assumption that it is the same for males and females, i.e. that it is **independent** of SEX.

Thus, we can now estimate  $p(\text{SEX}=1, \text{LIFE}=2)$  by  $0.434 \times 0.515 = 0.2235$ .

The expected count in the cell would then be  $0.2235 \times 980 = 219.0$  (see SPSS output).

This is carried out for each cell of the table and the observed counts are compared to these expected counts using some definition of Loss. The Deviance loss function is referred to as **Likelihood Ratio**. The form of this loss function is derived from the assumption that the counts have the **Multinomial** distribution. Following the general theory this has approximate distribution given by the Chi-squared distribution. Testing the value of this Loss is a test of our assumption of independence. The modified version of this, due to Pearson, is also given. The advantage of the Pearson version is that the approximate Chi-squared assumption holds more accurately when expected counts are small. The approximation can become unreliable when expected counts become very small (e.g. close 1). We may be forced to amalgamate categories to increase these small expected counts.

We can see that the model for  $p(\text{SEX}=i, \text{LIFE}=j)$  is a simple multiplicative model. We can turn it into a simple additive model by taking logs:

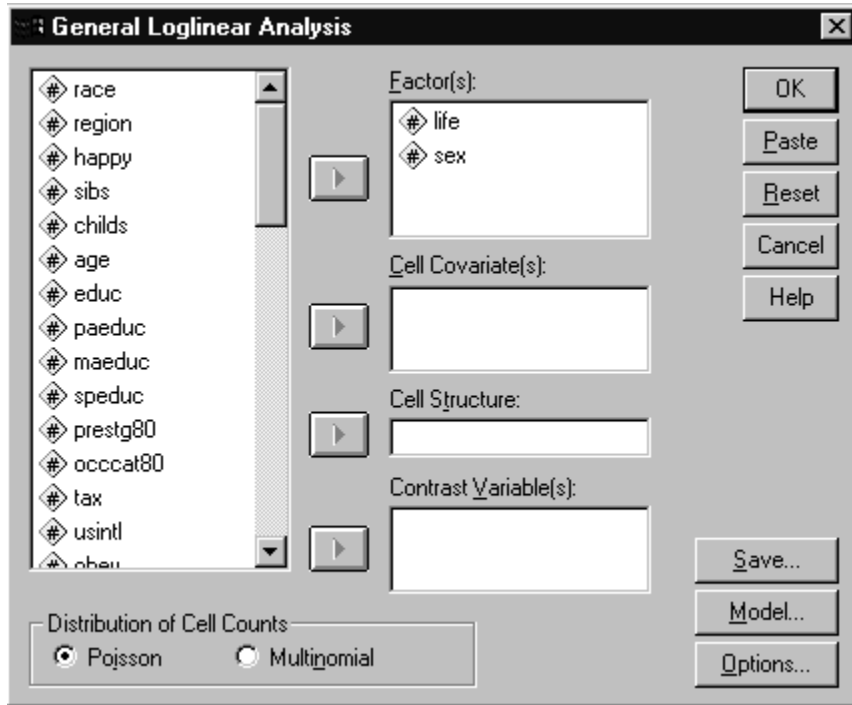
$$\begin{aligned} \log p(\text{SEX}=i, \text{LIFE}=j) &= \log p(\text{SEX}=i) + \log p(\text{LIFE}=j) \\ &= a_i + b_j \end{aligned}$$

We have seen how to deal with such models using Factors in General Linear Models. Thus, on a log scale the model is linear and is often referred to as a **log-linear** model. In this form the parameters are the logs of the probabilities so are more difficult to interpret immediately.

Thus, we can see that this is an example of a simple non-linear model with a particular Loss function derived from the Multinomial distribution and thus fits into our general description of a Statistical Model.

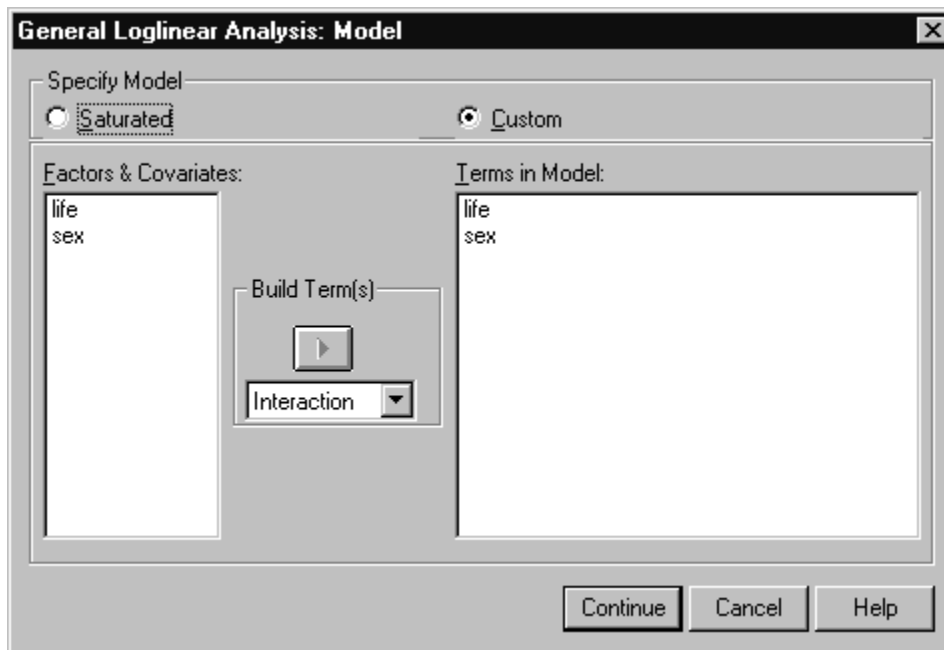
These models are fitted through the module:

Analyze  
LogLinear  
General



Select the categorical variables as **Factors**.

Open the **Model** dialog box:



The default is **Saturated**, switch to **Custom** and build a model in the usual way.

Also the default output includes tables and plots of residuals, which are not usually needed until a final model has been selected. Switch these off in the **Options** dialog. If parameter estimates are required switch this on in the **Options** dialog.

### General Loglinear

#### Table Information

Factor	Value	Observed		Expected	
		Count	%	Count	%
LIFE	Exciting				
SEX	Male	213.00	( 21.73)	188.21	(19.21)
SEX	Female	221.00	(22.55)	245.79	(25.08)
LIFE	Routine				
SEX	Male	200.00	( 20.41)	219.01	( 22.35)
SEX	Female	305.00	( 31.12)	285.99	( 29.18)
LIFE	Dull				
SEX	Male	12.00	(1.22)	17.78	(1.81)
SEX	Female	29.00	(2.96)	23.22	(2.37)

#### Goodness-of-fit Statistics

	Chi-Square	DF	Sig.
Likelihood Ratio	12.1095	2	.0023
Pearson	11.9941	2	.0025

#### Correspondence Between Parameters and Terms of the Design

Parameter	Aliased	Term
1		Constant
2		[LIFE = 1]
3		[LIFE = 2]
4	x	[LIFE = 3]
5		[SEX = 1]
6	x	[SEX = 2]

Note: 'x' indicates an aliased (or a redundant) parameter. These parameters are set to zero.

## Parameter Estimates

Parameter	Estimate	SE	Z-value	Asymptotic 95% CI	
				Lower	Upper
1	3.1450	.1586	19.83	2.83	3.46
2	2.3595	.1633	14.45	2.04	2.68
3	2.5110	.1623	15.47	2.19	2.83
4	.0000	.	.	.	.
5	-.2669	.0645	-4.14	-.39	-.14
6	.0000	.	.	.	.

## Multi-way Tables

With this formulation we can generalise it to model multi-way contingency tables. The simplest model would assume all Factors are independent of each other leading to more additive terms in the log-linear model. We can then build more complex models by adding interaction terms between pairs of Factors allowing for the non-independence of these Factors.

Examples:

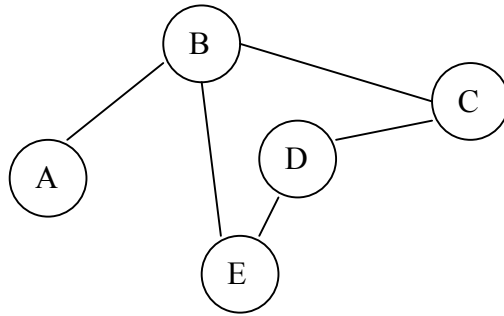
<b>A + B + C</b>	A, B and C are mutually independent
<b>A + B*C</b>	A is independent of B and C, B is dependent on C
<b>A*B + B*C</b>	A is dependent on B, and C is dependent on B, but there is no direct dependence of A on C

This last example is very important and can be expressed as “A and C are *conditionally* independent given B”. The practical importance of this is that if we wish to predict A we need only know B, C is irrelevant. However, if we considered just the two-way table  $A \times C$  we may well get a statistically significant association due to the mutual dependence of A and C on B. This is analogous to *spurious correlation* in Normal regression models. Thus it is important when we have many factors that are apparently inter-related that we separate the direct associations from the indirect (spurious) associations. This we can only do by considering models for the multi-way tables.

Even more complex interactions between triples of Factors can be added, allowing for more complex inter-relationships. Interpretation of parameter values becomes increasingly complicated. For example, the two-way interaction parameters are equivalent to **log odds-ratios**, which can be transformed back to odds and odds-ratios.

While quantitative interpretation of interactions may be difficult, qualitative insight may be gained by simply testing for the presence of significant interactions. The absence of an interaction will mean a simplification in the

inter-relationships. For complex models a **Graphical** representation can be very helpful.



The represents the model:  $A + B + C + D + E + A*B + B*C + B*E + C*D + D*E$

The rule for drawing this graphical representation is simple: The circles represent the factors and we connect factors if they have an interaction. The interpretation of the graph is also simple: By considering “routes” from one node to another along the connecting lines we can derive **conditional independence’s**.

For example, consider all routes from all nodes to A. We see that they must go through node B. This can be interpreted as: A is **conditionally independent** of C, D and E **given** B. The practical interpretation of this is that, to predict A we need only know the value of B, all other information is superfluous.

Searching for the simplest model that adequately represents the data can be laborious. For example, with 5 factors there are 32 possible models to test. Stepwise procedures can be used to step through each interaction in turn removing those that are not significant.

In SPSS we can use a stepwise model selection procedure through

**Analyze**  
**Loglinear**  
**Model Selection...**

In this procedure we can only select **Factors** (note you will have to provide the range of factor levels for each factor).

The only procedure is **Backward Selection**. The default starting point is the saturated model, use the **Model** dialog to change this.

By default the maximum number of steps is set to 10, you may need to increase this for complex models.

## Example 1

The datafile **vote.sav** contains data on the voting intentions (only Labour vs Conservative recorded) of a sample of people and their **CLASS**, **SEX** and **AGE** (grouped). However, this data is already in aggregate form, that is the original data has been cross-tabulated and only the cell frequencies (freq) retained. Below is a portion of the data:

	class	age	sex	vote	freq
1	1	1	1	1	4
2	1	1	1	2	0
3	1	1	2	1	10
4	1	1	2	2	0
5	1	2	1	1	27
6	1	2	1	2	8
7	1	2	2	1	26
8	1	2	2	2	9
9	1	3	1	1	27
10	1	3	1	2	4
11	1	3	2	1	25
12	1	3	2	2	9
13	1	4	1	1	17
14	1	4	1	2	12
15	1	4	2	1	28
16	1	4	2	2	9

In order to analyze such data correctly we must declare **FREQ** as a **Weight Cases** variable, as follows:

**Data → Weight Cases**  
**Weight Cases by** select **FREQ**

Now we proceed with Log-linear analysis

**Analyze → Loglinear → General...**

**Factors = class age sex vote**  
**Model = age class sex vote**

Goodness-of-fit Statistics

	Chi-Square	DF	Sig.
Likelihood Ratio	234.2176	51	3.E-25
Pearson	222.1450	51	3.E-23

Thus these factors are clearly not independent.  
If we now include a **VOTE** by **CLASS** interaction:

**Factors = class age sex vote**  
**Model = age class sex vote vote\*class**



Goodness-of-fit Statistics

	Chi-Square	DF	Sig.
Likelihood Ratio	82.8230	49	.0018
Pearson	75.0415	49	.0098

Then the Likelihood Ratio Chi-Square has reduced from 234.22 to 82.82, a difference of 152.4 on 2 degrees of freedom. This is highly significant (though we do not get this information from SPSS) and indicates that the VOTE\*CLASS interaction is significant. However, we cannot rely on this without testing all the other possible interactions.

We will use Model Selection to search for the simplest relationship.

**Analyze → Loglinear → Model Selection...**

**Factors = class(1 3) age(1 5) sex(1 2) vote(1 2)**  
**Model = saturated**

\*\*\*\*\* HIERARCHICAL LOG LINEAR \*\*\*

Tests that K-way and higher order effects are zero.

K	DF	L.R. Chisq	Prob	Pearson Chisq	Prob	Iteration
4	8	8.086	.4251	8.186	.4155	4
3	30	39.493	.1151	36.038	.2069	4
2	51	234.218	.0000	222.145	.0000	2
1	59	882.822	.0000	985.052	.0000	0

-----  
 Tests that K-way effects are zero.

K	DF	L.R. Chisq	Prob	Pearson Chisq	Prob	Iteration
1	8	648.604	.0000	762.907	.0000	0
2	21	194.725	.0000	186.107	.0000	0
3	22	31.407	.0881	27.852	.1806	0
4	8	8.086	.4251	8.186	.4155	0

Backward Elimination (p = .050) for DESIGN 1 with generating class

CLASS\*AGE\*SEX\*VOTE

Likelihood ratio chi square = .00000 DF = 0 P = 1.000

-----

If Deleted Simple Effect is	DF	L.R. Chisq Change	Prob	Iter
CLASS*AGE*SEX*VOTE	8	8.086	.4251	4

Step 1

The best model has generating class

CLASS\*AGE\*SEX  
CLASS\*AGE\*VOTE  
CLASS\*SEX\*VOTE  
AGE\*SEX\*VOTE

Likelihood ratio chi square = 8.08611 DF = 8 P = .425

-----

If Deleted Simple Effect is	DF	L.R. Chisq Change	Prob	Iter
CLASS*AGE*SEX	8	3.158	.9240	3
CLASS*AGE*VOTE	8	17.195	.0281	4
CLASS*SEX*VOTE	2	2.056	.3577	4
AGE*SEX*VOTE	4	6.153	.1880	3

Step 2

The best model has generating class

CLASS\*AGE\*VOTE  
CLASS\*SEX\*VOTE  
AGE\*SEX\*VOTE

Likelihood ratio chi square = 11.24431 DF = 16 P = .794

-----

If Deleted Simple Effect is	DF	L.R. Chisq Change	Prob	Iter
CLASS*AGE*VOTE	8	17.958	.0215	3
CLASS*SEX*VOTE	2	1.550	.4608	3
AGE*SEX*VOTE	4	8.004	.0914	3

Step 3

The best model has generating class

CLASS\*AGE\*VOTE  
AGE\*SEX\*VOTE  
CLASS\*SEX

Likelihood ratio chi square = 12.79409 DF = 18 P = .804

.... more steps

Step 6

The best model has generating class

CLASS\*AGE\*VOTE  
SEX\*VOTE

Likelihood ratio chi square = 22.91803 DF = 28 P = .737

-----

If Deleted Simple Effect is	DF	L.R. Chisq Change	Prob	Iter
CLASS*AGE*VOTE	8	18.281	.0192	4
SEX*VOTE	1	10.890	.0010	2

The final model has generating class

CLASS\*AGE\*VOTE  
SEX\*VOTE

Goodness-of-fit test statistics

Likelihood ratio chi square =	22.91803	DF = 28	P = .737
Pearson chi square =	22.18494	DF = 28	P = .773

Thus non-significant interaction terms are removed one at a time until all those left are significant. For this data the final model is:

**VOTE\*SEX + VOTE\*CLASS\*AGE**

Thus the relation of VOTE to CLASS and AGE requires a three-way interaction, indicating that the relation of voting preference to AGE is not the same in each CLASS. However SEX only enters as a two-way interaction with VOTE, indicating that gender differences in voting preference are the same for all AGE groups and CLASS groups.

## Interpretation of Parameters

Having used **Model Selection** to decide on a suitable model we must refit using **General** to obtain details such as parameter estimates.

Here is the edited output from the analysis of the VOTE data.

### GENERAL LOGLINEAR ANALYSIS

Design: Constant + AGE + CLASS + SEX + VOTE + SEX\*VOTE + CLASS\*AGE\*VOTE

Parameter	Aliased	Term
1		Constant
12		[VOTE = 1]
13	x	[VOTE = 2]
14		[SEX = 1]*[VOTE = 1]
15	x	[SEX = 1]*[VOTE = 2]
16	x	[SEX = 2]*[VOTE = 1]
17	x	[SEX = 2]*[VOTE = 2]

(+more lines)

Chi-Square	DF	Sig.		
Likelihood Ratio		22.9198	28	.7370

Parameter	Estimate	SE	Z-value
1	3.4387	.1295	26.56
12	-.5566	.2220	-2.51
14	-.3732	.1133	-3.29

(+more parameters)

As this is a **log-linear** model the main effect parameters are differences in **log-probabilities**, which are related to **odds**. Consider AGE group 5 ( <26 ) and CLASS 3 ( Working ). Then for **females** we compute the odds of Conservative versus Labour as:

$$\text{odds Cons:Lab} = \exp(-0.5566) : 1 = 0.573 : 1$$

which can be converted to probabilities:  $(0.573/1.573, 1/1.573) = (0.364, 0.636)$

For **males** we have:

$$\text{odds Cons:Lab} = \exp(-0.5566 - 0.3732) : 1 = 0.395 : 1$$

which are equivalent to probabilities (0.283,0.717)

The **ratio** of the odds for males to females =  $0.395/0.573 = 0.689 = \exp(-0.3732)$

Thus from the interaction parameter (-0.3732) we can get a direct interpretation in terms of an **odds ratio = 0.689**. Since SEX only occurs in this model in the interaction **VOTE\*SEX** this odds ratio is constant for all ages and classes. Thus the odds for voting Conservative are about 31% less for males of all ages and class.

The **qualitative** interpretation is straightforward. The interaction parameter for **male** and **Conservative** is **negative** indicating that the probability of voting Conservative is lower for males. A useful rule-of-thumb is that the maximum change in probability (occurring with probabilities around 0.5) is the interaction parameter divided by 4, i.e. about -0.1 in our example. This approximate result is accurate for values of the parameter between  $\pm 1$ .

## Continuous Covariates

Interactions involving continuous covariates can be built and included as terms in a log-linear model. The covariate is specified as a **Cell Covariate** and can then be used in the **Model** specification.

The interaction is then modelled as a “trend” with respect to this covariate. Such models are closely related to logistic regression models. In fact, for a binary dependent variable the two ways of specifying the model are exactly equivalent.

## General Loglinear

Covariates: XAGE (copy of AGE values)

Model: AGE + CLASS + SEX + VOTE + CLASS\*VOTE + SEX\*VOTE + VOTE\*XAGE

Correspondence Between Parameters and Terms of the Design

Parameter	Aliased	Term
14		[CLASS = 1]*[VOTE = 1]
20		[SEX = 1]*[VOTE = 1]
24		[VOTE = 1]*XAGE

Goodness-of-fit Statistics

	Chi-Square	DF	Sig.
Likelihood Ratio	51.9665	47	.2866

Parameter	Estimate	SE	Z-value	Asymptotic 95% CI	
				Lower	Upper
14	1.6497	.1674	9.85	1.32	1.98
16	1.2893	.1510	8.54	.99	1.59
20	-.3732	.1133	-3.29	-.60	-.15
24	-.0156	.0035	-4.42	-.02	-8.690E-03

### Logistic Regression (Dependent = VOTE)

Parameter Estimates

VOTE		B	Std. Error	Wald	df	Sig.	Exp(B)
Conservative	Intercept	.394	.202	3.807	1	.051	
	[CLASS=1]	1.659	.170	95.832	1	.000	5.256
	[CLASS=2]	1.309	.153	72.922	1	.000	3.702
	[CLASS=3]	0 <sup>a</sup>	.	.	0	.	.
	[SEX=1]	-.411	.122	11.406	1	.001	.663
	[SEX=2]	0 <sup>a</sup>	.	.	0	.	.
	XAGE	-.016	.004	17.965	1	.000	.984

a. This parameter is set to zero because it is redundant.

## SPSS commands for Log-Linear Models

### Loglinear General

Fits general loglinear models. Continuous covariates included as **Cell Covariates**

Build model through **Model** dialog box, specifying interactions in usual way. Parameter estimates requested in **Options**, store predicted values through **Save**.

### Loglinear Model Selection

Allows backward stepwise procedures to select model. Only Factors allowed. Starting model defined in **Model** dialog box. Can specify model as **All n-way interactions**.

Parameter estimates are not available.

### **Loglinear Logit**

Fits Logistic models via the equivalent log-linear model.  
Need to take care in specifying the model to get expected results.  
Continuous covariates specified as **Cell Covariates**.  
Saved predicted values are expected cell counts not predicted probabilities.

### **Regression Multinomial Logistic**

Fits generalised Logistic models for multi-category data.  
Factors and Covariates specified as in General Linear Models and terms built in the same way through **Model**.  
Predicted probabilities can be displayed through **Statistics** but cannot be saved.

### **Regression Binary Logistic**

If one of the categorical variables is Binary and is taken as the **dependent** variable gives results equivalent to specific log-linear model.  
**Model** specification more cumbersome, Factors have to be included then declared as **Categorical**.  
Predicted probabilities can be **Saved**.

## **Practical Session 7: Log-Linear Models**

Using the STATLAB data (statlaba.sav)

1. Examine the interactions (associations) between MTE, MTO, FTE and MTO.
  - (i) Firstly by specifying models and comparing them.
  - (ii) Secondly, using the model selection procedure to find the “best” model.

Using the BSAS data (bsas91.sav)

2. The interaction between PRSOCCL and SRSOCCL is a social mobility effect.
  - Does this effect differ between males and females?
3. There is a strong association between HEDQUAL and PRSOCCL. Does this vary with RSEX?