CHAPTER 3

FACTOR ANALYSIS

Exploratory

In social science research it is often necessary to ascertain how a measured collection of factors have some common single variable affect. These are typically known as latent variables. A latent variable is inferred from observed variables in the data set.

One of the main ways to obtain rich data is by using a Likert scale. Interestingly the Likert scale is named after Rensis Likert a social psychologist who developed the 5-point Likert scale in 1932 to measure people's attitudes in behavioral research as part of his PhD. Data obtained through Likert scales are discrete, categorical and ordinal.

The number of responses for Likert scale questions can vary. The Likert scale is considered to be an ordinal scale as the person is asked to consider their answer to a question from one of the ordered categories. The distance between each response in the item for a Likert scale is not considered to be always equal. Each of the Likert questions is called an item. With a group of items measuring a construct(s) that cannot be directly measured. A five point Likert scale could be for example 5=strongly agree, 4=agree, 3=undecided, 2=disagree, 1=strongly disagree.

Factor analysis, both exploratory and confirmatory, allows you to identify groups or clusters of variables and understand the structure of these variables in questionnaires including Likert scales. Before we explore the detailed structures of such questionnaires using factor analysis let's consider a basic technique first.

A numerical addition of Likert scores in the questionnaire gives the researcher a rank order. This will help to give us an understanding of the data in a similar way to which we started exploring contingency tables in the last chapter.

When performing this calculation a researcher would use the numerical coded values to create the total. With positively worded items being scored directly and negatively worded items recorded in a reverse scale. To see what this means let's take the case of 5=strongly agree, 4=agree, 3=undecided, 2=disagree, 1=strongly disagree as shown in Table 3.1.

To illustrate how this would be performed we will use three questions from the happiness data study. In future chapters the whole of this study will be explored using a range of statistical techniques.

Table 3.1 Simplistic Likert scale scoring

	Strongly disagree (1)	*Disagree (2)*	*Neutral (3)*	*Agree (4)*	*Strongly agree (5)*
a) I feel that life is very rewarding.	☐	☐	☐	x	☐
b) I feel able to take anything on.	x	☐	☐	☐	☐
c) I feel that I am not especially in control of my life (R).	☐	☐	☐	x	☐

Note that the first two questions in Table 3.1 are positively worded and the third is negatively worded. If a person responded (a) 4; (b) 1; (c) 4 then their total score would be 4+1+2 = 7. Note that question (c) has been reversed scored from 4 to 2. When analysing Likert scale questions it can be difficult to interpret what the statistics are actually inferring about the responses. It is often advised that the median measure should be used for Likert scale data as opposed to calculating a mean value or as we have done here summing the total item scores.

Simplistic numerical addition of Likert scale questionnaires as we have just described does NOT reveal the underlying structure or themes within the data. This type of analysis would give you a rank order of the responses to this part of the questionnaire.

This chapter around factor analysis will enable you as a researcher to explore Likert scale data from a more sophisticated statistical standpoint and hence being able to unlock a greater wealth of information.

EXPLORATORY FACTOR ANALYSIS

Exploratory factor analysis (EFA) is the name given to a range of statistical techniques to evaluate the dimensionality of items in a questionnaire. The item responses could range from ‘strongly disagree’ to ‘strongly agree’. Exploratory factor analysis explores and uncovers the smallest number of underlying constructs (called latent structures) in a questionnaire. Exploratory factor analysis is an ‘exploratory’ tool as no priori restrictions are put on the relationship found between the observed measures and the resulting latent variables. With confirmatory factor analysis the number of factors and their structure is specified in advance (see Chapter 5). This is the key difference between exploratory factor analysis and confirmatory factor analysis (CFA). Exploratory factor analysis is sometimes used as a precursor to confirmatory factor analysis. The estimates from exploratory factor analysis can be ‘confirmed’ in confirmatory factor analysis through detailed statistical evaluation (Dixon et al., 2016).

Factor extraction methods

Factor analysis is used to uncover the underlying constructs and identify associated items. The constructs can be thought of as themes contained in the questionnaire (Conway and Huffcutt, 2003). In this section we will explore how two types of exploratory factor analysis – principal factor (principal axis factoring in SPSS) and maximum likelihood – are used to identify latent factors created by a set of measured items.

Maximum likelihood (ML) is a method to find latent factors, which have the highest likelihood value of giving the best overall fit to the data. Its main advantage is that it allows for a detailed statistical evaluation of the factor solution. However maximum likelihood estimation requires the assumption of multivariate normality. Care needs to be taken when using maximum likelihood as it can produce improper solutions. An improper solution is when the factor model does not converge to give a final set of estimates or it produces an out of range estimate with indicators greater than 1.0.

Principal factor analysis (PF) has the advantage of having no distribution assumptions. If your data is markedly non-normal then principal factor analysis might be a preferred option. However, unlike a maximum likelihood estimation, principal factor analysis does not provide detailed statistical fit indices that are important if you are considering in the future performing confirmatory factor analysis.

DISCOVERING LATENT FACTORS

When carrying out factor analysis certain elements need to be considered in order to decide whether factor analysis is suited to your data. Being able to interpret the results given in the tables generated through statistical packages allows you to be confident that the analysis you are performing are providing meaningful results. The sections that follow take you through that interpretation to make and inform the decisions required.

Communality

There are a number of checks than can be made to see if variables should be included in the analysis. The first looks to see if variables share enough in common with each other. The total variance for individual items can be viewed as having two parts. One aspect being the amount of variance it shares with other items. This is called the common variance and the other aspect is its own unique variance. The proportion of common variance in an item is called communality.

To explore communality we will use the data that were developed by taking items from the Oxford happiness questionnaire (Argyle and Hills, 2002). Table 3.2 shows the ten items used in this data set taken from the original 29. Table 3.3 shows the five hundred responses provided to the questions around happiness. Looking at the figures it would seem that those responding to the questions are relatively happy, but not particularly healthy or in control of

Table 3.2 Happiness scale

	Strongly disagree (1)	*Disagree (2)*	*Neutral (3)*	*Agree (4)*	*Strongly agree (5)*
I am very happy.	☐	☐	☐	☐	☐
I laugh a lot.	☐	☐	☐	☐	☐
Life is good.	☐	☐	☐	☐	☐
I feel I have a great deal of energy.	☐	☐	☐	☐	☐
I am always committed and involved.	☐	☐	☐	☐	☐
I feel that life is very rewarding.	☐	☐	☐	☐	☐
I feel able to take anything on.	☐	☐	☐	☐	☐
I feel that I am not especially in control of my life.	☐	☐	☐	☐	☐
I don't feel particularly healthy.	☐	☐	☐	☐	☐
I don't have particularly happy memories.	☐	☐	☐	☐	☐

their own lives. This is not the whole story, so this is why we use factor analysis to gain a deeper understanding about what our data tells us.

In order to consider communality, initial and extraction values can be explored. These communality values are given in the range from zero to one. A value of zero means that a variable shares none of its variance with any of the other variables and a value of one means that all of the variance that is associated with the variable is common.

When performing exploratory factor analysis the communalities output is generated and provides information about the shared and proportional variance between items. The communalities table, Table 3.4, has two columns labelled 'initial' and 'extraction'. The initial communalities values indicate the amount of variance that is shared by the item and the others in the list. The extraction communalities values give the proportion of the variance that is explained by the extracted factors.

First check the values of the initial communalities to see if any of the items has a value that is much lower than the others. In this data set none of the values vary greatly. Since there is no great difference in the item scores there is no reason to suggest that any of the items should be removed from the analysis. After making these checks on the initial values you should next look at the column labelled 'extraction'. If you look for example at the item

Table 3.3 Responses to the happiness study

	Strongly disagree (1)	*Disagree (2)*	*Neutral (3)*	*Agree (4)*	*Strongly agree (5)*
I am very happy.	40	35	40	77	308
I laugh a lot.	29	37	62	88	284
Life is good.	32	31	38	93	306
I feel I have a great deal of energy.	40	39	61	80	280
I am always committed and involved.	79	65	74	85	197
I feel that life is very rewarding.	166	96	78	75	85
I feel able to take anything on.	81	82	116	107	114
I feel that I am not especially in control of my life.	38	54	107	118	183
I don't feel particularly healthy.	50	55	72	120	203
I don't have particularly happy memories.	126	96	120	81	77

Table 3.4 Communalities

Communalities

	Initial	*Extraction*
I am very happy.	.349	.518
I laugh a lot.	.315	.436
Life is good.	.279	.372
I feel I have a great deal of energy.	.268	.348
I am always committed and involved.	.146	.330
I feel that life is very rewarding.	.075	.172
I feel able to take anything on.	.071	.117
I feel that I am not especially in control of my life.	.130	.300
I don't feel particularly healthy.	.124	.210
I don't have particularly happy memories.	.089	.195

Extraction Method: Principal Axis Factoring.

'I am very happy', this has an extraction communality of 0.518. This means that 51.8% of the variance associated with this item is a shared (common) variance. All the extraction values are larger than the initial values. This then implies that there is more variance explained by each of the individual items.

Table 3.5 KMO and Bartlett's test

KMO and Bartlett's Test		
Kaiser-Meyer-Olkin Measure of Sampling Adequacy		.759
Bartlett's Test of Sphericity	Approx. Chi-square	629.219
	df	45
	Sig.	.000

Tests suitability of factor analysis

Kaiser-Meyer-Olkin (KMO) is generally considered to be the best measure of sampling adequacy for carrying out factor analysis. Output tables can be generated in statistical packages providing you with the KMO measure of sampling adequacy. A value closer to 1 implies that the data set is most appropriate to be analysed using factor analysis giving factors that are robust and useful. A KMO value of 0.5 is considered to be mediocre (and the minimum to carry out factor analysis), values between 0.7 and 0.8 as good, between 0.8 and 0.9 very good and between 0.9 and 1.0 excellent (Kaiser, 1970; Hutcheson and Sofroniou, 1999).

Table 3.5 provides a KMO of 0.759 for the happiness data set, which would imply a good suitability of the data for structure detection. The Bartlett Test of Sphericity is usually given when performing factor analysis to verify whether items correlate. If the test produces a significant result ($p<0.05$) then factor analysis is appropriate because relationships between variables have been detected. The assumption of this test is that the data are normally distributed and if this is not true for your data then the Bartlett results should be used with caution.

Looking at the communality, KMO and Bartlett results factor analysis would be an appropriate statistical technique to use with the happiness data.

How many factors to include

It is not always obvious how many latent variables (factors) should be retained. It can vary due to the particular constructs expected in the research or the level of required simplicity required from the data reduction. As with our data it would not be sensible to reduce the ten items into nine latent variables as this would clearly not achieve our goal of using factor analysis to simplify the problem. Experience plays a part here and you will learn with practice to decide on what is the optimal number of latent variables in different situations. Yet saying this there are guidelines to follow and criteria you can use to help you make an informed decision.

The statistical software gives the table in Table 3.6 showing how much variation is accounted for by each factor. These have the technical name of eigenvalues and are calculated as the sum of all the squared factor loadings

Table 3.6 Total variance explained

	Initial eigenvalues		
Factor	*Total*	*% of variance*	*Cumulative %*
1	2.608	26.076	26.076
2	1.429	14.290	40.366
3	1.075	10.750	51.116
4	.941	9.411	60.527
5	.814	8.136	68.662
6	.755	7.551	76.213
7	.689	6.891	83.104
8	.636	6.355	89.460
9	.563	5.630	95.089
10	.491	4.911	100.000

for each factor. Eigen in German means 'own' or 'belonging to'. In terms of our context with eigenvalues and eigenvectors this relates to the value and the vector that belong to that particular matrix. If you would like more information concerning factor analysis, eigenvalues and eigenvectors then please see Appendix 3. As with all the information in the appendices this is not necessary to understand and perform the statistical techniques.

As we have ten items in our Happiness questionnaire there are ten eigenvalues in Table 3.6. The eigenvalues for each factor give the variance explained by it. Therefore factor 1 in Table 3.6 explains 26.076% of the total variance. In the table it is clear that the first few eigenvalues explain the large proportion of the variance, with smaller and smaller percentages for subsequent factors.

The default setting in the software for SPSS and Stata is to retain eigenvalues greater than one as those smaller than one explain less variance this is called the Kaiser criterion. It is also good if total percentage of variance explained by the factors with eigenvalues greater than one is larger than 50%. You can see in our particular example this is 51.116%

However as suggested in the introduction to this section the researcher needs to be aware that a factor with an eigenvalue of 1.075 (factor 3 in Table 3.6) is very similar to an eigenvalue of 0.941 (factor 4 in Table 3.6). When selecting a cut-off value use your research questionnaire rationale and decisions around why you are undertaking factor analysis to check that the correct number of latent factors are taken into consideration. Yet as a rule of thumb the Kaiser criterion is a sensible starting point (Kaiser, 1970; Jolliffe, 1986; Stevens, 1992).

To support you in making your decision in the number of factors you retain, it is also useful to look at the scree plot. The scree plot shown in

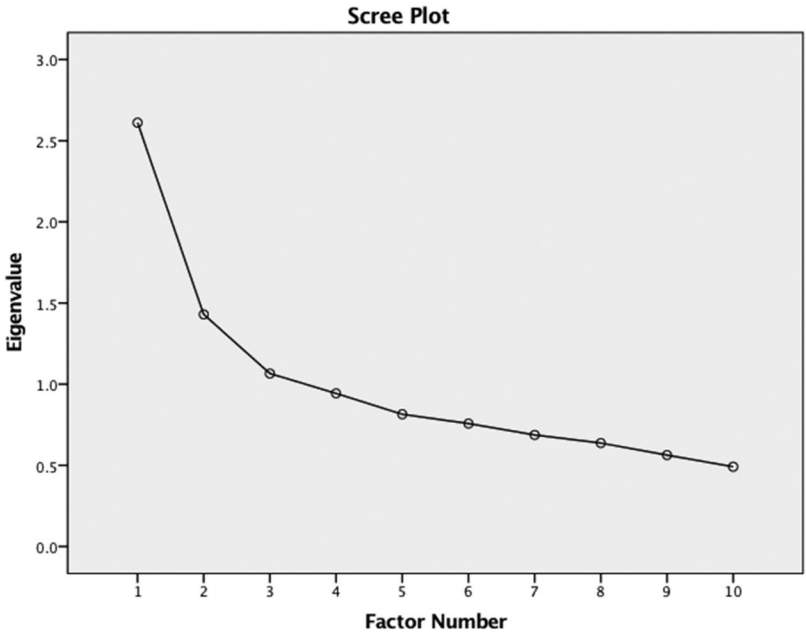

Figure 3.1 Scree plot

Figure 3.1 shows the relative sizes of the eigenvalues for our happiness data. With the scree diagram you are looking to see where the graph loses its steepness. What we mean by steepness in this context is if you look at the left hand side of the graph the line is clearly descending quickly, but at some point around eigenvalue three and four there is a flatting out and a clear change in direction after eigenvalue three.

We will assume that in our particular example we are going to move forward with our three latent variables, as the tests detail above suggest that these offer the best possible solution.

The next section will consider the output that factor analysis gives in relation to the latent factor groupings of the items themes.

Rotation methods

Rotation of factors in factor analysis is an important mathematical procedure in order to help to produce a solution that clarifies its interpretation. The system is rotated until there is a maximized solution of the sum of the variances of the squared loadings. This mathematical procedure called rotation preserves the relationship between individual variables. With our happiness study an oblique rotation technique is more appropriate as we are looking for correlations within factors. The oblique rotation technique called Promax should be used as it allows for the inter-woven correlated structures between items.

Table 3.7 is called the pattern matrix giving the factor loading for each of the items. These values can be thought of equivalent to the correlations between items and factors, with factors between 1 and -1. In Table 3.7 'I am very happy' has a correlation with factor 1 with a loading of 0.744. As with Pearson correlation the square of these coefficients obtains a measure of the importance of a particular variable to a factor. Hence in this case $0.744^2 = 0.554$ and so we can say as with r^2 values that 55.4% of the variance is explained by this latent factor and 44.6% ($1–0.744^2$) unexplained. In other words the first latent factor accounts for 55.4% of the variance of the item 'I am very happy'. Looking at the next item 'I laugh a lot' in the same first factor, then by squaring this value you can obtain the level of variance explained as 42.25% ($0.650^2 = 0.4225$) by this latent factor.

The happiness data shown in Table 3.7 has three latent factors which we could offer having titles as follows:

- 'happy' (factor 1) with four items; 'I am very happy'; 'I laugh a lot'; 'Life is good'; 'I feel I have a great deal of energy';
- 'positive' (factor 2) with three items; 'I am always committed and involved'; 'I feel that life is very rewarding'; 'I feel able to take anything on';
- 'unhappy' (factor 3) with three items; 'I am not particularly optimistic about the future'; 'I don't have particularly happy memories'; 'I don't feel particularly healthy'.

Table 3.7 Pattern matrix

	Factor		
	1	*2*	*3*
I am very happy.	.744		
I laugh a lot.	.650		
Life is good.	.576		
I feel I have a great deal of energy.	.545		
I am always committed and involved.	.317	.473	
I feel that life is very rewarding.		.414	
I feel able to take anything on.		.301	
I feel that I am not especially in control of my life.			.503
I don't feel particularly healthy.			.357
I don't have particularly happy memories.			.334

SPSS output table with extraction method: principal axis factoring. Rotation method: Promax with Kaiser normalization. Only loadings of magnitude above 0.30 are shown

Stata command: factor <variables>, ipf and postestimation of rotate, promax oblique kaiser factors(3) blanks(0.27) gives a similar three factor output.

Exploratory factor analysis produces a solution with the best simple structure maximizing factor loadings close to one and minimizing those close to zero. Loadings greater than or equal to 0.3 are said to be salient, relating meaningfully to a primary or secondary factor (Brown, 2006). According to Guadagnoli and Velicer (1988) a factor with ten item loadings greater than 0.4 is stable for a sample size greater than 150, with Field (2000) suggesting that retained factors should have at least three items with a loading greater than 0.4.

Before we move to look at how we can use exploratory factor analysis as a technique in data reduction let us consider the other factor extraction method that we discussed at the start of this chapter – maximum likelihood (ML).

Maximum likelihood

Research shows that maximum likelihood is robust for small kurtosis values (Chou and Bentler, 1995). As discussed in Chapter 1 skewness and kurtosis values of zero are said to be perfectly normal. Deviations of less than ±1 from zero are considered very good. Values lying outside of this range between ±1 and ±2 considered acceptable (Field, 2000; Trochim and Donnelly, 2006; Muijs, 2010; Gravetter and Wallnau, 2014). Maximum likelihood has been shown to be a very well behaved estimator for non-normal data so long as there are no extreme outliers (Curran et al., 1996). Using these conditions on normality we can check our happiness data. Table 3.8 shows descriptive statistics including the

Table 3.8 Skewness and Kurtosis statistics

	Mean	*Std. Deviation*	*Skewness*	*Kurtosis*
I am very happy.	4.16	1.294	–1.378	.562
I laugh a lot.	4.12	1.224	–1.242	.391
Life is good.	4.22	1.211	–1.509	1.123
I feel I have a great deal of energy.	4.04	1.310	–1.147	.021
I am always committed and involved.	3.51	1.500	–.493	–1.231
I feel that life is very rewarding.	2.63	1.490	.348	–1.321
I feel able to take anything on.	3.18	1.382	–.179	–1.187
I feel that I am not especially in control of my life.	3.71	1.270	–.652	–.650
I don't feel particularly healthy.	3.74	1.352	–.770	–.671
I don't have particularly happy memories.	2.77	1.390	.189	–1.192

values for skewness and *excess* kurtosis, which mainly fall into the very good and acceptable regions.

This analysis shows that we can have confidence to apply the maximum likelihood extraction when performing exploratory factor analysis. Table 3.9 shows the pattern matrix. Notice that the extraction produces the same three latent factors as found previously using principal axis factoring. The principal axis factoring factors from the first factor analysis are shown in brackets. As well as producing the same latent factors, it can be seen that maximum likelihood extraction method individual item scores are very similar to principal axis factoring. The relative error being small for all items (1-2%), with only the one item 'I feel that I am not especially in control of my life' having a larger relative error of 7%

Table 3.9 Comparison of extraction methods

Pattern Matrix

	Factor		
	1	*2*	*3*
I am very happy.	0.753 (0.744)		
I laugh a lot.	0.646 (0.650)		
Life is good.	0.585 (0.576)		
I feel I have a great deal of energy.	0.545 (0.545)		
I feel that I am not especially in control of my life.		0.536 (0.503)	
I don't feel particularly healthy.		0.363 (0.357)	
I don't have particularly happy memories.		0.318 (0.334)	
I am always committed and involved.	0.303 (0.317)		0.466 (0.473)
I feel that life is very rewarding.			0.433 (0.414)
I feel able to take anything on.			0.295 (0.301)

SPSS output in the table with extraction method maximum likelihood. In brackets extraction method is principal axis factoring. Rotation method: Promax with Kaiser normalization.

Stata command: factor <variables>, ml factors(3) blanks(0.29) and postestimation command: rotate, promax oblique kaiser factors(3) detail blanks(0.29) gives a similar three factor output.

FACTOR ANALYSIS FOR DATA REDUCTION

Factor analysis can also be used as a data reduction technique. This next section will look at an example of how data reduction can be used with socio-economic factors. This data reduction into a smaller number of more manageable factors allows for greater levels of interpretation.

Principal components analysis (PCA)

The factor extraction technique that is used for data reduction is called principal components analysis. Both exploratory factor analysis and principal components analysis methods are similar as they are both used to examine correlations and covariance's between sets of items. The two methods are often confused due to this similarity.

Principal components analysis accounts for the variance in the observed measures rather than explain the correlations amongst them as with maximum likelihood and principal factor analysis. Principal components analysis reduces a larger set of measured items into a smaller number of composite variables. These composite variables can then be used in subsequent analysis, without any consideration of prior constructs of the underlying structure. As with maximum likelihood and principal factor analysis, principal components analysis takes a purely exploratory approach.

A difference with principal components analysis is that it requires orthogonal rotation technique such as Varimax as the items under investigation are assumed to be unrelated to each other. If this is not the case with your data reduction, and the items are related, then you should as we have seen earlier in the chapter use an oblique rotation such as Promax.

In the following study we will illustrate the use of principal components analysis for data reduction of socio-economic variables. The sample is of 500 households from poor informal settlements in the Global South. The questionnaire in Table 3.10 shows the possible family possessions of these households.

Table 3.10 Socio-economic data

	Yes	*No*
The family have a generator	☐	☐
The family own a car or jeep	☐	☐
The family have a computer	☐	☐
The family have a gas stove	☐	☐
The family own land	☐	☐
The family have electricity	☐	☐
The family have a television	☐	☐
The family own a radio	☐	☐
The family own a cell/mobile	☐	☐

Table 3.11 Frequencies for socio-economic data

	Yes	*No*
The family have a generator	64	436
The family own a car or jeep	169	331
The family have a computer	152	348
The family have a gas stove	146	354
The family own land	159	341
The family have electricity	439	61
The family have a television	433	67
The family own a radio	427	73
The family own a cell/mobile	471	29

Table 3.12 Eigenvalues: socio-economic data

	Initial eigenvalues		
Factor	*Total*	*% of variance*	*Cumulative %*
1	2.281	25.340	25.340
2	1.366	15.175	40.515
3	1.081	12.012	52.527
4	.980	10.887	63.414
5	.824	9.154	72.567
6	.734	8.154	80.721
7	.693	7.696	88.417
8	.643	7.140	95.557
9	.400	4.443	100.000

The questionnaire contains the following responses shown in Table 3.11. It can be seen from this table that there is a clear split with these poor households. A number of the households have electrical devices such as mobile phone, television and radio, with less respondents saying that they have a computer, car, own land or have an electrical generator.

When principal components analysis data reduction is performed it suggests that there could be a three factor solution. This can be seen in Table 3.12 as three of the eigenvalues are greater than one, having values of 2.281, 1.366 and 1.081.

This three factor solution offers the following common themes:

- Factor one is the most affluent households possessing cars, generator, gas stove and computers.

- Factor 2 being a fairly common factor with television and electricity. As we have seen in the descriptive statistics that 433 have a television and 439 have electricity out of 500.
- Factor 3, is mobile phone, radio and land. We have seen from the descriptive data that most families have a mobile phone (471 of the 500 households).

This data reduction suggests a scale of the most affluent households to the least being factor 1 to factor 3 respectively.

It can be seen that one of the eigenvalues is close to 1.0 with a value of 1.081 (factor 3 in Table 3.12). In this situation it is always worth checking to see if a lower solution gives a better set of real life factors. Running principal components analysis specifying a fixed number of factors as two yields the results shown in Table 3.14.

Looking at these two latent factors it is possible to define differences between latent factors 1 and 2. This two factor solution seems to generate more common themes than the initial three factor solution obtained by simply allowing the Kaiser criterion of all eigenvalues greater than 1.0 to be factors. Factor 1 suggests a more affluent household who can afford extras in a Global South context such as computer, car and generator. A generator supply for electricity in most Global South countries tends to imply a wealthier household owing to publically supplied electricity having erratic provision, with daily cases of power cuts.

Table 3.13 Three factor rotated PCA solution

Rotated Component Matrix

	Component		
	1	*2*	*3*
The family own a car or jeep	.673		
The family have a generator	.671		
The family have a computer	.655		
The family have a gas stove	.627		
The family have a television		.865	
The family have electricity		.826	
The family own a cell/mobile			.793
The family own a radio			.539
The family own land	.359		.425

SPSS extraction method is principal component analysis. Rotation Method: Varimax with Kaiser Normalization.

Stata command principle component factor: factor <variables>, pcf and postestimation command: rotate, kaiser factors(3)

Table 3.14 Two factor rotated PCA solution

Rotated Component Matrix

	Component	
	1	*2*
The family have a generator	.680	
The family own a car or jeep	.672	
The family have a computer	.654	
The family have a gas stove	.564	
The family own land	.436	
The family have electricity		.855
The family have a television		.818
The family own a radio		.366
The family own a cell/mobile		.362

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization. In Stata use principal-component factor. Stata command principle component factor: factor <variables>, pcf and postestimation command: rotate, kaiser factors(2)

Clearly we can argue the pros and cons of both of these outputs. The decision around which of these you would choose to use for your research as the wealth factors would be based on your understanding of the data and/or your literature review relating to the context you are working in.

CALCULATING AND USING LATENT FACTORS IN FUTURE ANALYSIS

A linear regression example

In later chapters we look at how factor analysis can be used to explore links between variables. To illustrate this with an example we will look to see how the happiness latent factors are affected by socio-economic backgrounds using the data we have created in the examples in this chapter.

In the happiness data one of the latent factors was having a positive attitude to life. The three items in this latent factor are 'I am always committed and involved', 'I feel that life is very rewarding' and 'I feel able to take anything on'. As you can see these suggest a very positive outlook to life. Using this as the dependent variable we can ask the following research question:

Does your socio-economic background affect your commitment attitudes to life?

Table 3.15 Linear regression for positive happiness

Coefficients

Model	*Unstandardized Coefficients*		*Standardized Coefficients*	*t*	*Sig.*
	B	*Std. Error*	*Beta*		
(Constant)	1.988E-5	.030		.001	.999
Wealth 1 richer	.071	.030	.106	2.379	.018
Wealth 2 poorer	–.005	.030	–.008	–.183	.855

Dependent Variable: Positive happiness

Table 3.15, a linear regression table (regression will be covered in greater detail later in the book), suggests due to the positive significance B-value of 0.071 and associated p-value of 0.018 that if you come from a more affluent household then you have a greater likelihood of having a positive attitude to life.

Calculating the appropriate latent factor score

Having discovered the factor scores for the latent variables it is possible to interpret these results relative to the different individuals who have answered the questionnaire. To demonstrate how this can be done we shall use the three factor scores relating to a 'positive attitude to life' from the happiness questionnaire in the early part of this chapter.

We can write down an equation for this latent factor where the numbers are factor loadings given in Table 3.16 and can be thought of as weightings for the various variables:

Positive = 0.473 x Committed + 0.414 x Rewarding + 0.301 x Anything

From this we can calculate the latent factor scores for each person. If for example a person had given answers to these three items as 2, 4, and 3 in the questionnaire this would result in a factor score of 3.505 for that particular individual. The calculation to obtain this value is:

Positive = 0.473 x (2) + 0.414 x (4) + 0.301 x (3)
= 0.946 + 1.656 + 0.903 = 3.505

Table 3.16 Factor scores: positive attitude to life

Item	*Variable name*	*Factor loading*
I am always committed and involved.	Committed	0.473
I feel that life is very rewarding.	Rewarding	0.414
I feel able to take anything on.	Anything	0.301

This value can be computed by SPSS and Stata, for all the participants in the survey giving the latent variable that can be used in further analysis.

The way we calculate the factor score of 3.505 is called a coarse factor score and created by a non-refined method. This illustrates how to derive a simplistic weighted value of the raw score.

Statistical packages offer the user a range of alternative more sophisticated methods to evaluate the latent factors with greater internal consistency. These are called refined methods. There are three main refined methods used in factor score extraction, Regression, Bartlett and Anderson-Rubin. All three have advantages and the method of choice depends on how the extracted factors are going to be used.

The Regression method (Thurstone, 1935) modifies the factor loadings to compensate for the initial correlations between the variables. In most cases the regression method is the most frequently used method to estimate refined factor scores. With this method the factor scores have a mean of zero. The resulting factor scores correlate not only with the items in the latent factor but also with the items in the other latent factors. These factor scores are standardised to reflect a Z-score metric with the values ranging from -3 to +3 (Brown, 2006).

The Bartlett method, in contrast to the regression method produces factor scores that are only correlated with items within that latent factor. One advantage of Barlett factor scores is that they are most likely to give 'true' factor scores as they are produced using maximum likelihood estimates (Hershberger, 2005).

The third method used to extract factor scores is the Anderson-Rubin. With this method the factor scores are uncorrelated and are standardised to have a mean of zero and standard deviation of one. Then these are often standardised using a T-score which is a shifted Z-score scaled to have a mean of 50 and standard deviation of 10, using a data transformation such as trunc(50 + 10*Fac1_1 + 0.5). If you wish to create factor scores that are uncorrelated for data reduction then the use of Anderson-Rubin method is often preferred (Harman, 1976; Grice, 2001; Tabachnick and Fidell, 2001; Carifio and Perla, 2008; Di Stefano et al., 2009)

MISSING VALUES

Factors score values will only be calculated for items that contain no missing values. If there are any missing item values no factor scores will be produced pertaining to that person's questionnaire category. It is possible to avoid these missing data issues by using one of the many imputation techniques. In an ideal world data imputation would always be avoided but for a multitude of reasons this is not always the case. In Chapter 1 we gave more detail on the various options that are available to deal with missing values.

HOW TO REPORT FACTOR ANALYSIS

As we have seen in this chapter it is important to report the Kaiser-Meyer-Olkin measure, the Bartlett Test of Sphericity, a Scree plot and the percentage of variance explained by the eigenvalue factors included. The report should also include factor weightings in the relevant matrix with the extraction and rotation method used. When using factor score the extraction method should be given, i.e. Regression, Bartlett or Anderson-Rubin method.

Calculating factor analysis with Stata and SPSS

For SPSS, select *Analyze – Dimension Reduction – Factor* to open a dialog box that allows variables and options to be selected. First select the variables to be used in factor analysis and drop these into the variable box. Then on the right hand side of this window select *Descriptives* and tick KMO. Next select the *Extraction* window to obtain options for factor method – Principal components, Maximum likelihood, or Principal axis factoring. In this window you can also select Scree plot and the number of factors to extract. The third dialog box is *Rotation* to select the appropriate rotation, either orthogonal Varimax or oblique Promax. The fourth window *Scores* allows you to calculate and save latent score factors by ticking the *Save as variables* box and then selecting the appropriate method. The final window *Options* gives display format with options *Sorted by size* and *Suppress small coefficients.* It is usual to set the *Absolute value below:* as 0.3

For Stata, select *Statistics – multivariate analysis – Factor and principal component analysis* to open the dialog box. In the menu (Model) select the variables (items) that are required for your particular factor analysis. In 'model2' identify the method to be used principal factor, maximum likelihood or principle-component factor. To select the appropriate rotation, either orthogonal Varimax or oblique Promax then select *Statistics – multivariate analysis – Factor and principal component – Postestimation – rotate loadings*. To generate a Scree plot use the same commands as this, changing only the final option to 'scree plot of eigenvalues', *Statistics – multivariate analysis – Factor and principal component – Postestimation – Scree plot of eigenvalues*. In the '*Statistics – multivariate analysis – postestimation reports and statistics'* section you can select the estat tools such as KMO.

To calculate and save latent score factors type in the command line 'predict' and your variable names for these factor scores. The default setting is the Regression Method. Hence if you have two factors you would type *predict var1 var2* or if you wished to use the Bartlett method then you would type *predict var1 var2, bartlett*.

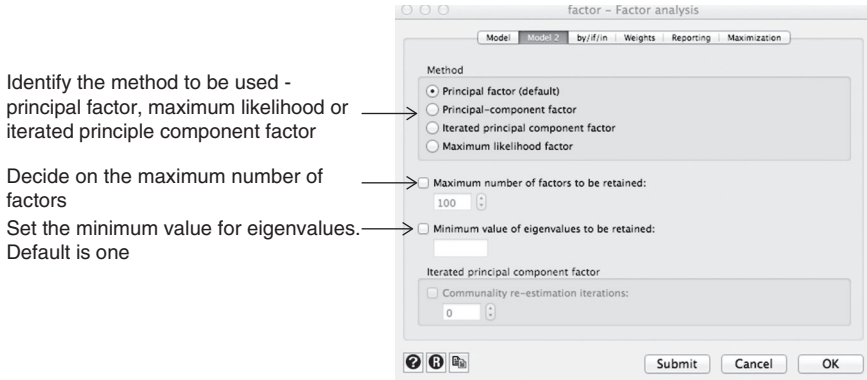

Figure 3.2 Stata factor analysis: select model

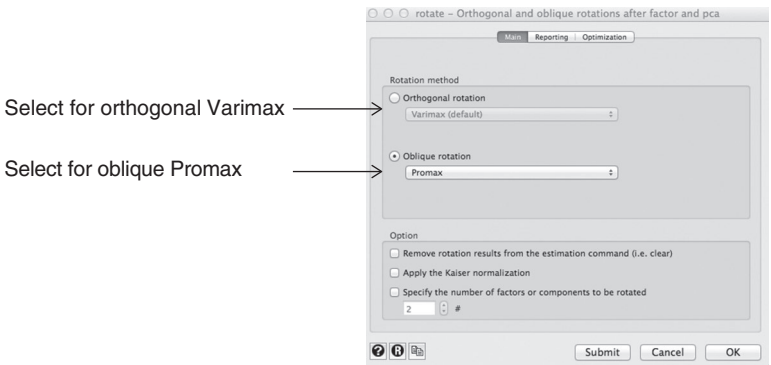

Figure 3.3 Stata Varimax or Promax rotation

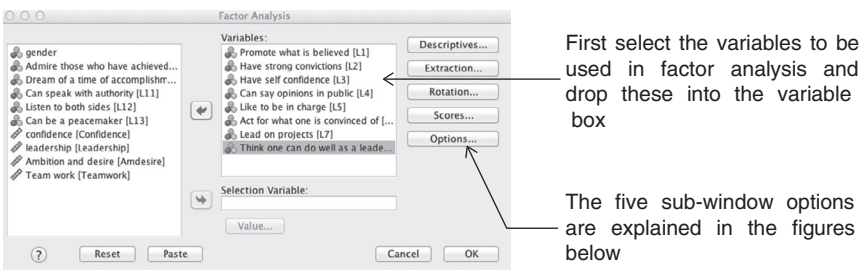

Figure 3.4 SPSS factor analysis

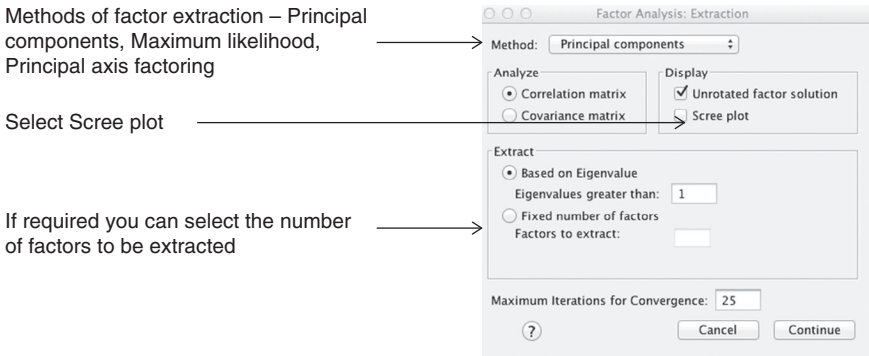

Figure 3.5 SPSS extraction window

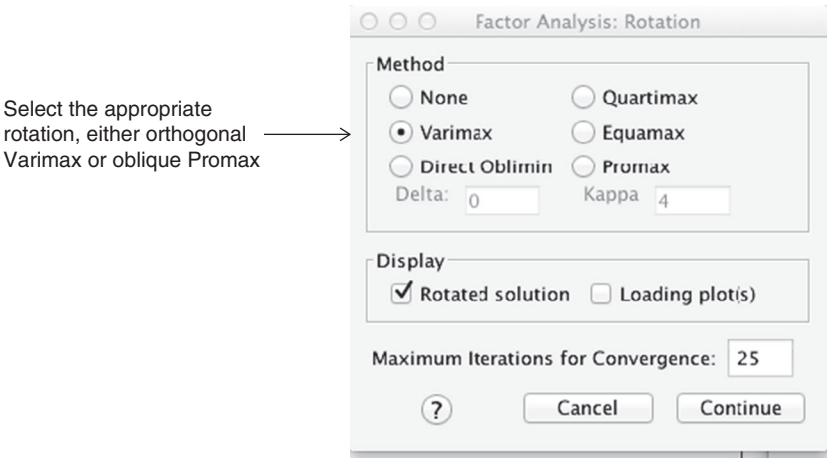

Figure 3.6 SPSS rotation window

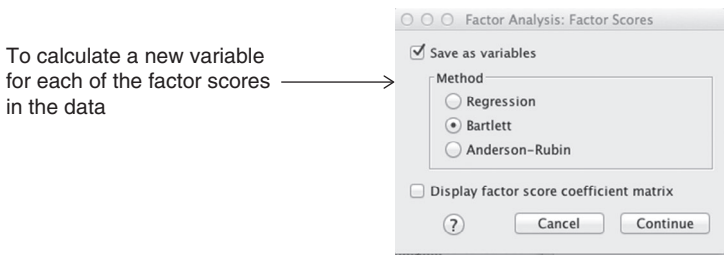

Figure 3.7 SPSS factor scores

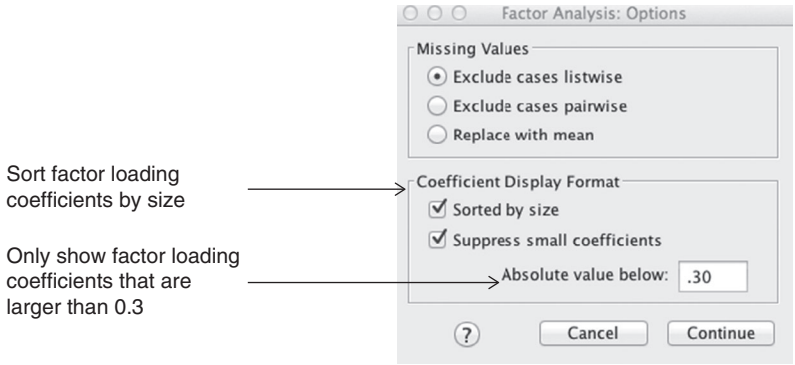

Figure 3.8 SPSS options window