CHAPTER 8

Factor Analysis for Binary Data

8.1 Latent trait models

In this chapter, we move to the top right-hand cell of Table 7.1 and discuss methods based on models where the manifest variables are categorical. We start with the case where they are all binary — that is, where they are based on responses of the kind yes/no or right/wrong. Some methods appropriate when there are more than two categories will be given in Chapter 9. The word "trait" in the name of these models is often used because it arises from one of the principal applications for which they were devised, namely the measurement of psychological traits. In this book, they are used in a much broader context and so it seemed appropriate to make this clear in the title of the chapter. Nevertheless, we have also retained the original terminology to keep the link with a very important field of application.

Conceptually there is no difference between the problems treated here and those in the previous chapter on factor analysis. We start with a probability model linking the observed variables to a set of latent variables. We then discuss how to fit the models, judge their goodness-of-fit, interpret their parameters, and so forth. The difference lies in the special problems posed by having to deal with a data matrix consisting of binary items. The basic objectives are the same, namely:

i) to explore the interrelationships between the observed responses

ii) to determine whether the interrelationships can be explained by a small number of latent variables

iii) to assign a score to each individual for each latent variable on the basis of the responses

The binary data matrix

We have already met a data matrix for categorical data in the discussions of cluster analysis, multidimensional scaling and, in passing, correspondence analysis. If the responses are binary, the xs simply record whether the response was positive or negative. A convenient convention, also used in earlier chapters, is to use 1 to indicate a "success" or a positive response, that is "correct" or "yes" as the case may be, and 0 for the "failure" or negative response. This convention has the advantage in the present context that if we sum the responses in any row of the data matrix, we get the total number of positive responses. This is a useful summary measure in its own right and we shall use it in the subsequent analyses. The response coded 1 is sometimes referred to as the *keyed response.* A typical row of the data matrix might then be as follows:

00101110011

The methods about to be described all start from a data matrix consisting of a set of rows like that above, one for each individual or object. However, the restriction to binary data sometimes makes it possible to express the matrix in a more compact and informative way.

Any row of the data matrix is referred to as a *score pattern* or a *response pattern.* If there are p variables there are 2^p possible response patterns. When $p = 3$, for example, they are

000, 001, 010, 011, 100, 101, 110, 111.

If the sample size is much larger than 2^p, many of the response patterns will be repeated. It is, therefore, much more economical to present the matrix as a list of the possible response patterns together with their associated frequencies as follows:

000	175
001	64
010	17
100	12
011	9
101	3
110	33
111	98

The second column records how many times each response pattern occurs in the sample. This grouped form of the data matrix is used whenever the sample size is large. However, when the number of variables p is large, many response patterns may not occur at all, in which case they are omitted from the table to save space.

Latent trait methods were introduced in educational testing where most of their development has taken place; this is now a highly specialised field with a substantial literature of its own. Our emphasis in this chapter will be mainly on their general use as tools for social research in the factor analysis tradition.

An example

To illustrate the various steps in the analysis, we shall use a data set with only four variables extracted from the 1986 British Social Attitudes Survey (McGrath and Waterton, 1986). The data are the responses given by 410 individuals to four out of seven items concerning attitude to abortion. A small proportion of non-response occurs for each item, the proportions being (0.03, 0.03, 0.05, 0.04). In order to avoid the distraction of having to deal with missing values, we have slightly adjusted the data to eliminate missing values.

An analysis that includes all respondents and uses a factor analysis (FA) model for binary items that takes account of missing values was carried out by Knott, Albanese, and Galbraith (1990). The results were not substantially different from those reported here. After eliminating the missing values, we are left with 379 respondents. For each item, respondents were asked if the law should allow abortion under the circumstances presented under each item. The four items used in the analysis are given below:

1. The woman decides on her own that she does not [WomanDecide]
2. The couple agree that they do not wish to have the child [CoupleDecide]
3. The woman is not married and does not wish to marry the man [NotMarried]
4. The couple cannot afford any more children [CannotAfford]

The frequency of each response pattern is given in Table 8.1.

Table 8.1 *Frequencies of response patterns, attitude towards abortion*

Response patterns	Frequency
1111	141
0000	103
0111	44
0011	21
0001	13
1110	12
0010	10
0100	9
0110	7
1011	6
0101	6
1101	3
1100	3
1000	1
1010	0
1001	0
Total	379

We find that the percentage of individuals agreeing that abortion should be legal under circumstances described by the items 1 to 4 are 43.8, 59.4, 63.6, and 61.7%, respectively. If we were doing a factor analysis, we would next compute the correlations between pairs of variables and inspect the result, looking for evidence of positive correlations which suggest that there might be one or more common underlying factors. In the case of binary data, the corresponding things to look at are the pairwise associations between variables. We can do this by constructing 2×2 contingency tables. For example, Table 8.2 cross-tabulates the first two items that show a strong association. A similar analysis for other pairs of variables produces similar results. This suggests that it would

Table 8.2 *Cross-tabulation of items 1 and 2, attitude towards abortion*

	Yes	No
Yes	159	7
No	66	147

be worth asking whether these associations can be attributed to one or more common factors. This is what a latent trait model enables us to do. If we can identify common factors, we may then wish to go on to compute scores for individuals on the latent dimensions.

8.2 Why is the factor analysis model for metrical variables invalid for binary responses?

Since the approach for binary and metrical variables has been so similar up to this point, it is natural to think of treating the binary data as if they were metrical. What is to prevent us from computing the product moment correlations and doing a factor analysis in the usual way? There is no practical bar to doing just that, and one sometimes finds such factor analyses in the research literature. However, such an analysis is inappropriate because it is based on a model which assumes that the observed or manifest variables $(x_1, \ldots, x_p)$ are metrical rather than binary. To see why this is so, we briefly return to the factor analysis model. The model was written as:

$$x_i = \alpha_{i0} + \alpha_{i1} f_1 + \cdots + \alpha_{iq} f_q + e_i \qquad (i = 1, \ldots, p), \tag{8.1}$$

where p denotes the total number of observed items, x_i denotes the ith metrical observed item, $\mathbf{f} = (f_1, \ldots, f_q)$ denotes the vector of latent variables and e_i denotes the residual. We assume that the residual follows a normal distribution with mean 0 and variance σ_i^2, the latent variables are assumed to be independent with standard normal distributions $f_j \sim N(0, 1)$ for all j. Since $\mathbf{f}$ and e_i can take any value and are independent of each other, x_i can also take any value. Therefore, the linear factor model is invalid for categorical variables in general and for binary variables in particular.

We need a different model to relate the latent variables $\mathbf{f}$ to the manifest variables. Two approaches have been adopted to meet this need. The oldest is to try to retain as much as possible of the factor analysis method. This is done by imagining a fictitious variable for each i which is partially revealed to us by x_i. This enables us to retain the factor model for the (unobserved) fictional variable. This method is still widely used and we shall describe it in Section 8.7.

A better approach is to start, as we did in factor analysis, with the idea of a regression model. We want an appropriate model for the regression of each x_i on the latent variables. The usual regression method used for an observable binary response on a set of observable explanatory variables is known as *logistic* regression. It takes its name from the logistic function used in the regression equation.

In order to motivate the choice of this function, we first remind ourselves that the regression of x_i on the latent variables is the expected value of x_i given the fs. Since x_i is binary, the expected value of x_i given the fs is the same as $\Pr(x_i = 1 \mid \mathbf{f}) = \pi_{\mathbf{i}}(\mathbf{f})$ where $\pi_i(\mathbf{f})$ is the conditional probability that binary variable, x_i, equals one given the values of the q latent variables $f_1, \ldots, f_q$. We, therefore, have to specify the form of the probability $\pi_i(\mathbf{f})$ as a function of $f_1, \ldots, f_q$. The function chosen is known as the *link* function.

An identical linear link function would be the simplest giving:

$$\pi_i(\mathbf{f}) = \alpha_{i0} + \alpha_{i1} f_1 + \cdots + \alpha_{iq} f_q \qquad (i = 1, \ldots, p). \tag{8.2}$$

But such a linear relationship between the probability of a correct response and the latent variables has two flaws.

i) The left-hand side of equation (8.2) is a probability that takes values between 0 and 1, and the right-hand side is not restricted in any way and can take any real value.

ii) We might expect that the rate of change in the probability of a correct/positive response will not be the same for the whole range of $\mathbf{f} = (f_1, \ldots, f_q)$. In that case, a curvilinear relationship might be more appropriate.

To take into account both those points, we need to introduce a different link function between the probability and the latent variables. That link should map the range $[0, 1]$ onto the range $(-\infty, +\infty)$. It should also be a monotonic function of each f. Possible links are the logit and the normit. We shall use the logit link mainly because it possesses theoretical and practical advantages (see Section 8.3). The logit model for binary data presented in Section 8.3 is one of the many item response models developed within the Item Response Theory (IRT) approach. IRT developed mainly in connection with educational measurement. Bock and Moustaki (2007) gives an overview of Item Response Theory models. We shall also, in Section 8.7, briefly discuss the use of the normit link (also known as the probit) as an alternative when we consider the underlying variable (UV) approach for analysing binary variables with factor models.

8.3 Factor model for binary data using the Item Response Theory approach

The logistic regression model introduced in Chapter 6, Section 6.12, can be adapted for the factor analysis of binary data, to give the logit model defined as:

$$\text{logit}\pi_i(\mathbf{f}) = \log_e \frac{\pi_i(\mathbf{f})}{1 - \pi_i(\mathbf{f})} = \alpha_{i0} + \sum_{j=1}^{q} \alpha_{ij} f_j. \tag{8.3}$$

By transforming $\pi_i(\mathbf{f})$ using the logit transformation, we have been able to write the model as linear in the latent variables which will greatly facilitate the interpretation. The probability $\pi_i(\mathbf{f})$ denotes the probability of "success" and the ratio $\pi_i(\mathbf{f})/(1-\pi_i(\mathbf{f}))$ is also known as the odds of "success". We can rearrange equation (8.3) to get an expression for $\pi_i(\mathbf{f})$:

$$\pi_i(\mathbf{f}) = \frac{\exp(\alpha_{i0} + \sum_{j=1}^{q} \alpha_{ij} f_j)}{1 + \exp(\alpha_{i0} + \sum_{j=1}^{q} \alpha_{ij} f_j)}. \tag{8.4}$$

It may easily be checked that this expression behaves in the right way, namely that it lies between 0 and 1 and is monotonic in each f.

An important special case is obtained by putting $q = 1$. It is this case with which *item response analysis* is mainly concerned. Thus, we have the unidimensional latent trait model:

$$\pi_i(f_1) = \frac{\exp(\alpha_{i0} + \alpha_{i1} f_1)}{1 + \exp(\alpha_{i0} + \alpha_{i1} f_1)}.$$

The unidimensional latent trait model is also known as the two-parameter model. In the psychometric literature, $\pi_i(f_1)$ is referred to as the item characteristic curve or item response function (IRF). It shows how the probability of a correct response increases with ability, say.

The logit model with one latent variable is plotted on Figure 8.1 for $\alpha_{i0} = 0.5$ and for different positive values of the parameter α_{i1} and on Figure 8.2 for different values of α_{i0} and for $\alpha_{i1} = 0.5$.

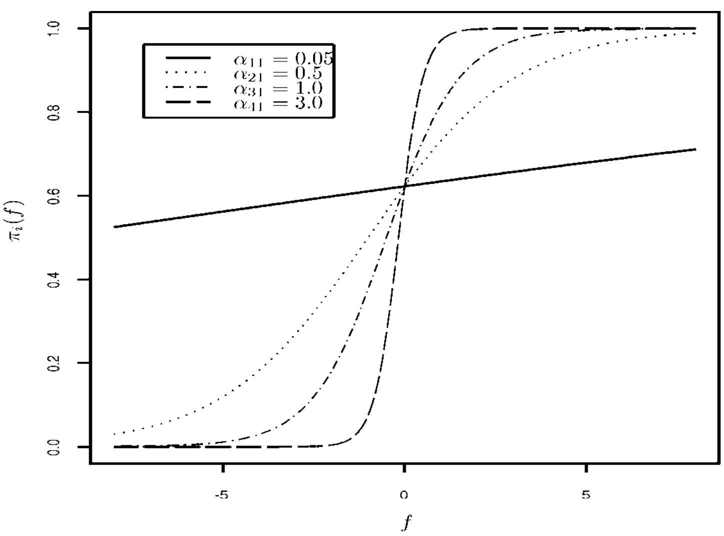

Figure 8.1 *Item characteristic curves for different values of the discrimination coefficient α_{i1} and $\alpha_{i0} = 0.5$*

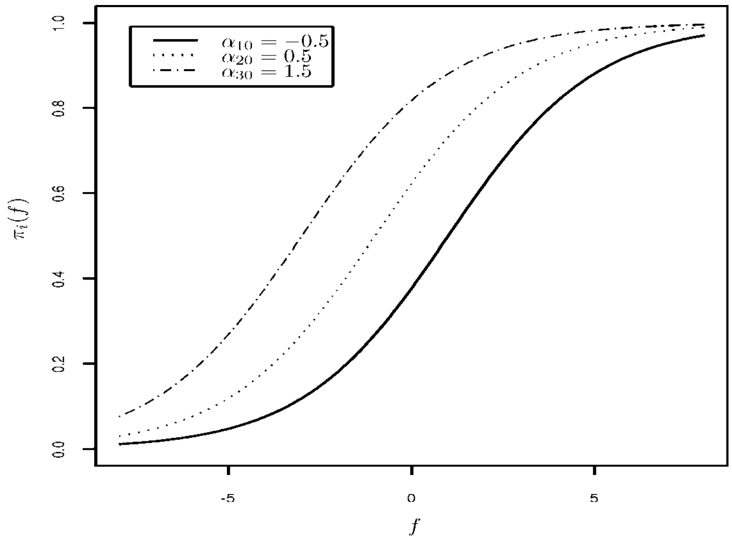

Figure 8.2 *Item characteristic curves for different values of the "difficulty" parameter α_{i0} and $\alpha_{i1} = 0.5$*

It is clear that the parameter α_{i1} determines the steepness of the curve over the middle of the range. This means that a given change in the value of f_1 will produce a larger change in the probability of a positive response when this parameter is large than when it is small. For this reason, it is known in educational testing as the *discrimination* parameter. Increasing the parameter α_{i0} increases the probability for all values of f_1 and so it is referred to as the *difficulty* parameter.

We summarise and complete the specification of the factor model for binary data by listing the assumptions on which it depends as follows:
i) Conditional independence: the responses to the p observed items are independent conditional on the latent variables. In other words, the latent variables (factors) account for all the associations among the observed items. Since the latent variables are unobserved, the assumption of conditional independence can only be tested indirectly by checking whether the model fits the data. A latent variable model is accepted as a good fit when the latent variables account for most of the association among the observed responses.
ii) The link function: $\text{logit}\pi_i(\mathbf{f}) = \alpha_{i0} + \sum_{j=1}^{q} \alpha_{ij} f_j$, where $\Pr(x_i = 1 \mid \mathbf{f}) = \pi_i(\mathbf{f}); (i = 1, \ldots, p)$. A possible alternative would be the normit link, see Section 8.7, which gives very similar results in practice.
iii) The latent variables or factors $f_1, \ldots, f_q$ are independent with standard normal distributions. That is $f_j \sim N(0, 1)$ for $(j = 1, \ldots, q)$. The choice of the normal distribution for the latent variables has rotational advantages as we will see later but other distributions could be used. Fortunately, research

has shown that the form of the distribution of the latent variables does not have much influence on the interpretation of the results of an analysis.

The Rasch model

A special case of the unidimensional model is obtained when all the discrimination parameters are equal ($\alpha_{11} = \alpha_{21} = \cdots = \alpha_{p1}$). This model was first discussed by Rasch (Rasch 1960) and it is usually written as:

$$\Pr(x_i = 1 \mid \alpha_{i0}, \beta_k) = \pi_{ik} = \frac{\exp(\alpha_{i0} + \beta_k)}{1 + \exp(\alpha_{i0} + \beta_k)}.$$

Since the α_{i1} are all equal, $\alpha_{i1} f_k$ has been replaced by β_k, where f_k is the value of f for individual k, ($k = 1, \ldots, n$). This formulation is useful in educational testing where the ability of each individual in the sample is of interest. In other applications the interest is generally in the population from which the sample has been drawn and f is treated as a variable with a probability distribution. The Rasch model is still quite popular in educational testing because of its simplicity and its attractive theoretical properties. In particular:
i) The total score $\sum_{i=1}^{p} x_{ik}$ is sufficient for β_k — that is, it contains all the information in the data about the β_k if the model is true.
ii) The total number of positive/correct responses for item i, $\sum_{k=1}^{n} x_{ik}$ is sufficient for α_{i0}.

Fitting the logit model

Recall that in factor analysis, we fitted the model by choosing the parameter values to make the covariance matrix predicted by the model as close as possible to the observed matrix. For that model, the joint distribution was completely determined by the covariances so, in effect, we were making the observed and predicted distributions as close as possible. We do essentially the same thing when fitting the latent trait model. We choose the parameter values which make the frequency distribution across responses predicted by the model as close as possible to the observed one. As in factor analysis, there are various ways in which this distance can be measured but the one for which software is currently available is based on the likelihood function — the maximum likelihood method.

Interpretation of model parameters

In the latent trait model for each observed item i, we have $q + 1$ parameters to estimate, the intercept α_{i0} and the factor loadings $\alpha_{i1}, \ldots, \alpha_{iq}$. We have already noted that α_{i0} is called the difficulty parameter in educational testing because of its effect on the probability of a positive response. This effect can be seen more clearly if we consider the position when $\mathbf{f} = \mathbf{0}$. Since the fs are assumed to have standard normal distributions, an individual at this point in the latent space may be described as the “median” individual because, on each dimension, half the population lie on either side. In those circumstances we find,

$$\Pr(x_i = 1 \mid \mathbf{f} = 0) = \pi_i(\mathbf{0}) = \frac{\exp(\alpha_{i0})}{1 + \exp(\alpha_{i0})}.$$

This is the probability that the median individual will respond correctly or positively to item i. For example all four curves in Figure 8.1 have the same $\alpha_{i0} = 0.5$ and hence the same value $\pi_i(0) = 0.62$. For practical purposes, $\pi_i(\mathbf{0})$ is more directly interpretable than α_{i0}.

The α_{ij}s ($j = 1, \ldots, q$) are factor loadings, but we have already noted that they are known in educational testing as discrimination parameters. The larger the value of α_{ij}, the greater is the effect of factor j on the probability of a positive response to item i; equivalently, the higher the value of α_{ij} for an item, the greater the difference in the probabilities of getting a correct/positive response between two individuals who are located at some distance apart on the latent dimensions. As a result, it will be easier to discriminate between those two individuals on the evidence of their responses to that item. The factor loadings α_{ij} are not bounded in any way, and for some items they may take very large values, indicating a very steep slope for the item response curve. This phenomenon is known as a "threshold effect", and we shall meet it again in Chapter 10. Large estimates of the discrimination parameters often have large standard errors, which means that their values are poorly determined. The maximum likelihood estimates for the attitude to abortion data are given in Table 8.3 along with their asymptotic (i.e., estimated using large sample theory) standard errors for a one-factor model.

Table 8.3 *Parameter estimates and standard errors in brackets and standardized loadings for the one-factor model, attitude to abortion*

Item	$\hat{\alpha}_{i0}$	s.e.	$\hat{\alpha}_{i1}$	s.e.	st$\hat{\alpha}_{i1}$	$\hat{\pi}_i(0)$
WomanDecide	−0.72	(0.33)	4.15	(0.85)	0.97	0.33
CoupleDecide	1.11	(0.35)	4.50	(0.81)	0.98	0.75
NotMarried	2.18	(0.61)	6.21	(1.54)	0.99	0.90
CannotAfford	1.15	(0.28)	3.49	(0.50)	0.96	0.76

The last column of the table gives the estimated probabilities that the median individual will respond positively to items 1-4. Item 1 stands out from the other items by being much less likely to be answered positively by the median individual. The loadings in the $\hat{\alpha}_{i1}$ column are all positive and very large, suggesting an underlying factor which is common to all items. In this context, one might identify this with a pro/anti-abortion attitude. It should be noted that the standard errors are all fairly large in relation to the differences in the estimates. This should caution us against placing undue weight on small inequalities among the loadings. In the present case, the broad conclusion we have drawn about a common factor seems unlikely to be sensitive to the effects of sampling variation. Taking the loadings at their face value, it appears that "the couple being unmarried" is the best discriminator between a pro- and an anti-abortion attitude and "inability to afford the baby" the worst discriminator.

The column headed $\text{st}\hat{\alpha}_{i1}$ requires some further explanation. In factor analysis, when the correlation matrix is analysed, the factor loading α_{ij} is the correlation between the observed item x_i and the latent variable f_j. This was very convenient as an aid to interpretation. In the latent trait case, the loadings cannot be interpreted as correlation coefficients; indeed, as we have seen, the loadings are not bounded by 0 and 1 as a correlation would be. However, it is possible to transform to standardized loadings that can be interpreted as correlation coefficients in exactly the same way as in factor analysis. This transformation arises naturally out of the alternative way of analysing binary items which we shall consider in Section 8.7. We shall defer consideration of this point to later, but here we merely observe that all the standardized loadings are close to one, indicating a close link between each item and the common factor.

8.4 Goodness-of-fit

The goodness-of-fit of the model can be checked in several different ways.

i) *Global goodness-of-fit test*
One way is to use a standard goodness-of-fit test to compare the observed and expected frequencies across the response patterns. Strictly, we compare observed frequencies and estimates of the expected frequencies under the model being tested — but conventionally, these estimates are referred to as "expected frequencies" when carrying out likelihood ratio or Pearson chi-squared goodness-of-fit tests as below. In fact, since we fit the models by choosing the parameter values so that these distributions are as close as possible, the minimum closeness would be an obvious measure to use for goodness-of-fit. A test based on such a measure is the log-likelihood-ratio test. The log-likelihood-ratio test statistic, G^2, is defined as:

$$G^2 = 2\sum_{r=1}^{2^p} O(r)\log_e \frac{O(r)}{E(r)} \tag{8.5}$$

where r represents a response pattern, and $O(r)$ and $E(r)$ represent the observed and expected frequencies, respectively, of response pattern r. An alternative is to use the Pearson chi-squared goodness-of-fit test statistic, X^2, given by:

$$\mathrm{X}^2 = \sum_{r=1}^{2^p} \frac{(O(r)-E(r))^2}{E(r)}. \tag{8.6}$$

If the model holds, both statistics are distributed approximately as χ^2 with degrees of freedom equal to the number of different response patterns minus the number of independent parameters minus one $(2^p - p(q+1) - 1)$. If the sample size n is much bigger than the total number of distinct responses given by 2^p, then the observed and expected frequencies will be reasonably large and the approximation on which the test is based will be valid. However, when the number of binary variables is large, many response patterns will have expected frequencies which are very small. It is usually recommended that all expected frequencies should be at least five for either test to be valid. If, for example, $p = 20$, there are $2^p = 1048576$ possible response patterns, and even with a sample size of several thousands there will be many expected frequencies which are exceedingly small. In those cases, the chi-squared test and the log-likelihood-ratio test will not follow a chi-squared distribution, and so from the practical point of view these tests cannot be used. The problem can be overcome to some extent by pooling response patterns with expected frequencies less than 5, but that might quickly lead to a situation where no degrees of freedom are left to perform the test. In such cases, we need another approach.

For the attitude to abortion data set in Table 8.6, the five response patterns with small expected frequencies were pooled. The log-likelihood-ratio statistic, is $G^2 = 17.85$ and the chi-squared statistic is $X^2 = 15.09$, both on three degrees of freedom. (This is not the seven degrees of freedom from the formula $2^p - 2p - 1$, because pooling of categories has taken place.)

Both measures indicate a not very good fit (the 1% significance level for chi-squared with three degrees of freedom is 11.35). We could go on to fit a two-factor model, but first it is worth trying to diagnose the reason for the poor fit. The first step is obviously to look for large discrepancies between observed and expected score patterns. These are given in the first two columns of Table 8.6. There are no obviously large deviations except, perhaps, at the two extremes. In a sparse table with many more response patterns it would be much more difficult to judge this, and then other approaches are needed.

ii) *Goodness-of-fit for margins*

Rather than look at the whole set of response patterns, we can look at the two-way margins. That is, we can construct the 2×2 contingency tables obtained by taking the variables two at a time. We have already done this at the beginning of the chapter when we looked at the pairwise associations among variables. The reason for doing that was to bring out the parallel with factor analysis. The two-way tables provided the same sort of information for binary variables as the correlations do for factor analysis. The two-way margins are the cell frequencies in these two-way tables. Comparing the observed and expected two-way margins is therefore analogous to comparing the observed and expected correlations when judging the fit of a factor model. The comparison is made using what we call *chi-squared residuals.* These are the contributions to the chi-squared statistic for the 2×2 table which would arise from the cell. Thus if O is the observed frequency and E the expected frequency, then the residual is $(O-E)^2/E$. Tables 8.4 and 8.5 give the observed and expected frequencies for the two-way and for some of the three-way margins respectively for the attitude to abortion data when the one-factor model is fitted. The last column of the tables gives the chi-squared residual as a measure of the discrepancy between the observed and the predicted frequency. From Table 8.4, we see that 147 respondents responded negatively to items 1 and 2; the model predicted 143.74 responses for that cell giving a residual equal to 0.07. The same calculations are done for all the pairs of items. The residuals computed for each cell are not independent and therefore they cannot be summed to give an overall test distributed as chi-squared. Valid tests are, however, provided in Bartholomew and Leung (2002), Maydeu-Olivares and Joe (2005) and Cai et al. (2006). As a rule of thumb, if we consider the residual in each cell as having a χ^2 distribution with one degree of freedom, then a value of the residual greater than 4 is indicative of poor fit at the 5% significance level. To be able to have a better idea of the discrepancies in the margins, given that the value 4 is only indicative, in the examples later in the chapter, we also report residuals greater than 3. A study of the individual margins provides information about where the model does not fit. For the abortion data, all the residuals are very small. On the evidence from the margins, we have no reason to reject the one-factor model. The overall significant result we obtained from the global goodness-of-fit tests cannot therefore be attributed to the relationships between the pairs and triplets of items.

iii) *Proportion of G^2 explained*

We have remarked at several points in the book that even an incomplete summary of multivariate data can be useful. The same is true of a multivariate model. Even though it may leave something unexplained, it may nevertheless capture some important and interesting features of the data. This is the case with the one-factor model which is serving as our example in this section. This raises the question of whether we can quantify the degree to which a simple model explains the associations between the binary variables. The same general idea proved useful in PCA and FA, where the proportion of variance explained served a similar purpose. Thus, we observed that the proportion of the total variance accounted for by a set of components might be used as a guide to whether the fs were an adequate summary. The same idea can be used here, but we now talk in terms of the proportion of the log-likelihood-ratio statistic for the independence model, which is explained by the model with q factors. The independence model would be appropriate if there were no associations between the binary variables $x_1, \ldots, x_p$. The log-likelihood-ratio statistic, G_0^2, for this model can be regarded as a measure of the associations between the xs. The log-likelihood-ratio statistic, G_q^2, for the model with q latent variables is a measure of the residual associations between the xs which have not been explained by the model.

The percentage of G^2 explained is given by

$$\%G^2 = \frac{G_0^2 - G_q^2}{G_0^2} \times 100$$

Table 8.4 *Chi-squared residuals for the second order margins for the one-factor model, attitude towards abortion*

Response	Item i	Item j	Observed frequency (O)	Expected frequency (E)	$O-E$	$(O-E)^2/E$
(0,0)	2	1	147	143.74	3.26	0.07
	3	1	131	133.17	−2.17	0.04
	3	2	117	119.69	−2.69	0.06
	4	1	129	133.68	−4.68	0.16
	4	2	114	116.09	−2.09	0.04
	4	3	116	111.79	4.21	0.16
(0,1)	2	1	7	11.30	−4.30	1.64
	3	1	7	5.94	1.06	0.19
	3	2	21	19.42	1.58	0.13
	4	1	16	11.99	4.01	1.34
	4	2	31	29.58	1.42	0.07
	4	3	29	33.88	−4.88	0.70
(1,0)	2	1	66	69.89	−3.89	0.22
	3	1	82	80.46	1.54	0.03
	3	2	37	35.35	1.65	0.08
	4	1	84	79.95	4.05	0.21
	4	2	40	38.95	1.05	0.03
	4	3	22	27.32	−5.32	1.04
(1,1)	2	1	159	154.07	4.93	0.16
	3	1	159	159.43	−0.43	0.00
	3	2	204	204.54	−0.54	0.00
	4	1	150	153.38	−3.38	0.07
	4	2	194	194.38	−0.38	0.00
	4	3	212	206.01	5.99	0.17

Table 8.5 *Chi-squared residuals for the third order margins for the one-factor model, response (1,1,1) to items (i, j, k), attitude towards abortion*

Item i	Item j	Item k	Observed frequency (O)	Expected frequency (E)	$O-E$	$(O-E)^2/E$
1	2	3	153	151.18	1.82	0.02
1	2	4	144	145.86	−1.86	0.02
1	3	4	147	150.15	−3.15	0.07
2	3	4	185	185.01	−0.01	0.00

and measures the extent to which the model with q latent variables explains the associations.

For the attitude to abortion data, the percentage of G^2 explained is 96.88%, indicating that the one latent variable model is a much better fit than the independence model or, in other words, there is 96.88% reduction in the log-likelihood-ratio statistic when the one-factor model was fitted.

The above three ways of checking model fit have been discussed in detail in the paper by Bartholomew and Tzamourani (1999).

iv) *Model selection methods*
Another approach, already mentioned in the connection with factor analysis for metrical variables, is based on the use of model selection criteria such as the Akaike information criterion or the Bayesian information criterion (see Sclove 1987).

8.5 Factor scores

Obtaining factor scores for the latent trait model is slightly more complicated than it was for PCA or FA. In PCA, the scores came "ready-made" as linear combinations of the manifest variables. In FA, the position was complicated by the fact that there was no unique value of each f associated with the set of xs. We therefore used a predicted value which turned out to be a linear combination of the xs for which the coefficients were calculated by the standard software. Following the same idea for the latent trait model, we would look for a suitable predictor of each f given the xs. Using regression ideas as before, this would suggest using the conditional mean value or conditional expectation:

$$E(f_j \mid x_1, \ldots, x_p) \qquad (j = 1, \ldots, q). \tag{8.7}$$

Unfortunately, these means are not linear combinations of the xs, although they can easily be computed. However, it turns out that (for the logit link function) they are monotonic functions of what we shall call *component scores* which are given by:

$$X_j = \sum_{i=1}^{p} \alpha_{ij} x_i \qquad (j = 1, \ldots, q). \tag{8.8}$$

In the one-factor case, both the regression function of equation (8.7) and the components give the same ranking to the individuals in the sample. These components are very simply calculated using the estimated weights obtained from fitting the model. For most practical purposes, it makes no difference whether we use the components or the conditional expectations.

For the logit link function, the component score, X_j, includes all the information in the data about the latent variables regardless of the assumption made about the distribution of f_j, whereas the posterior mean $E(f_j \mid x_1, \ldots, x_p)$ itself will vary according to whether we assume the distribution of f_j to be normal or some other distribution. This invariance property is a good reason for preferring the component score.

On the other hand, when a distribution is assumed for the fs it is possible to estimate not only the conditional means, $E(f_j \mid x_1, \ldots, x_p)$, but also the conditional standard deviations, $\sigma(f_j \mid x_1, \ldots, x_p)$ for $(j = 1, \ldots, q)$. The estimated standard deviations should be taken into account in judging the ranking of the response patterns when the conditional means are used.

Table 8.6 gives the estimated conditional means and component scores for all the response patterns for the attitude to abortion data. It also gives the expected frequency for each pattern. The sixth column gives the total score of the response pattern. As we can see, the estimated conditional mean, $\hat{E}(f \mid \mathbf{x})$, and the component score give the same ranking to the individuals. In this particular example, the total score also gives a similar ranking to the individuals, though there are some ties. The reason for this is that all the four items have similar discriminating power. There is also a column headed $\hat{\sigma}(f \mid \mathbf{x})$. This is the estimated conditional standard deviation of the latent variable about its conditional mean. This tends to be larger at the extremes but is fairly constant over the middle range. In all cases it is quite large, indicating that the factor scores are subject to a good deal of uncertainty.

Table 8.6 *Factor scores listed in increasing order, attitude towards abortion*

Observed frequency	Expected frequency	$\hat{E}(f \mid \mathbf{x})$	$\hat{\sigma}(f \mid \mathbf{x})$	Component score (X_1)	Total score	Response pattern
103	100.0	−1.19	0.55	0.00	0	0000
13	16.6	−0.61	0.32	3.49	1	0001
1	1.7	−0.55	0.30	4.15	1	1000
9	9.1	−0.52	0.29	4.50	1	0100
10	12.3	−0.38	0.26	6.21	1	0010
0	1.3	−0.29	0.24	7.64	2	1001
6	7.4	−0.27	0.24	7.99	2	0101
3	1.0	−0.24	0.24	8.65	2	1100
21	14.8	−0.18	0.24	9.70	2	0011
0	2.0	−0.14	0.25	10.37	2	1010
7	12.3	−0.12	0.26	10.71	2	0110
3	1.9	−0.01	0.28	12.14	3	1101
6	6.2	0.14	0.32	13.86	3	1011
44	41.1	0.17	0.32	14.20	3	0111
12	7.2	0.24	0.34	14.87	3	1110
141	143.9	0.95	0.61	18.35	4	1111

8.6 Rotation

As with the factor analysis model, the solution is not unique when we fit more than one latent variable. An orthogonal rotation of the factors coupled with corresponding rotation of the estimated loadings $\hat{\alpha}_{ij}$ leaves the likelihood unchanged. We are therefore free to search for a rotation which is more readily interpretable. The cautionary remarks made in Chapter 7 apply with equal force here. In particular, rotation does not produce a new solution so much as express the original solution in a different way. The main use of rotation is to search for "simple structure". In principle, the same kind of rotations could be used for latent trait models as for factor analysis. However, the uncertainties of estimation increase rapidly with the number of factors. It is doubtful whether there is any value in trying to fit more than two factors with the sample sizes that are commonly available. In any case, we have concentrated in this book on solutions which are capable of being represented in up to two dimensions. Our treatment is therefore consistent with this general approach. For practical purposes, rotation can be carried out in two dimensions graphically, as in Chapter 7.

8.7 Underlying variable approach

In this section we will discuss the alternative approach for constructing and fitting a factor analysis model to binary items. This approach is called the underlying variable (UV) approach. As we explained in Section 8.2, the UV approach is closer in spirit to factor analysis.

In the UV approach, the observed binary variables are assumed to be realisations of fictitious continuous *underlying variables*. Those underlying variables are unobserved but they should not be confused with the latent variables. They might be better described as *incompletely observed variables*, because all we observe is whether or not they exceed some threshold.

For each binary variable x_i, it is assumed that there is an *incompletely observed* continuous variable x_i^* which is normally distributed with mean μ_i and variance σ_i^2.

The connection between x_i and x_i^* is as follows: when the underlying variable x_i^* takes values below a threshold value τ_i, the binary item x_i takes the value 1, otherwise x_i takes the value 0. The parameters τ_i are called threshold parameters. Since no other information is available about x_i^* $(i = 1, \ldots, p)$, its mean and variance are arbitrary and can be set to zero and one respectively without loss of information.

The essence of the method is to treat the x_i^*s as if they had been generated by the classical factor analysis model. That is, we suppose that:

$$x_i^* = \alpha_{i1}^* f_1 + \alpha_{i2}^* f_2 + \cdots + \alpha_{iq}^* f_q + e_i \qquad (i = 1, \ldots, p), \tag{8.9}$$

where the α_{ij}^* are the factor loadings, the f_j are the latent variables, and the e_i are the residuals with zero mean and variance σ_i^2 $(i = 1, \ldots, p; j = 1, \ldots, q)$.

In factor analysis, the x_i^*s are observable variables, whereas here they are underlying, incompletely observed variables.

All that we need to fit a factor model is the matrix of correlations. The correlation can be estimated from each pair of the binary x_is and hence software such as LISREL, Mplus, and EQS can be used to fit the factor model. Correlations estimated in this way are called *tetrachoric* correlations.

There are a number of subtle differences between the fitting of a factor model to tetrachoric correlations and fitting it to product moment correlations. The thresholds are estimated from the univariate marginal distribution of the underlying variable, x_i^*, and the correlations from the bivariate marginal distributions of the x_i^*s for given thresholds. This amounts to saying that the method uses less of the information in the data. The UV approach does make the assumption of conditional independence through the independence of the residual terms, e_i, and it also assumes that the univariate and bivariate distributions of the underlying variables are normal.

The results of carrying out a factor analysis on tetrachoric correlations are very similar to those obtained using the logit latent variable model. This is no accident, because it can be shown that the two types of model are equivalent for binary data. A mathematical proof of this equivalence will be found in Bartholomew and Knott (1999), p.87-88. The logit model and the UV normit model give similar results because the normal and logistic distributions are so similar in shape. There is an exact equivalence between the parameter estimates for the normit UV and the normit IRF model given by:

$$\alpha_{i0} = \frac{\tau_i}{\sigma_i}$$

and

$$\alpha_{ij} = -\frac{\alpha_{ij}^*}{\sigma_i}.$$

The same equivalence holds approximately for the normit UV and the logit IRF models.

Furthermore, we can standardize the factor loadings α_{ij}s to represent correlations between the latent variables f_js and the binary variables x_is.

The standardized αs are given by:

$$\text{st}\alpha_{ij} = \frac{\alpha_{ij}}{\sqrt{\sum_{j=1}^{q} \alpha_{ij}^2 + 1}} = -\alpha_{ij}^*. \tag{8.10}$$

This is the standardization we referred to in Section 8.3 and which was given in Table 8.3.

8.8 Example: sexual attitudes

In order to illustrate the full range of analyses, including the fitting of two factors, we shall take an example extracted from the 1990 British Social Attitudes Survey (Brook, Taylor, and Prior 1991). It concerns contemporary sexual attitudes. The questions addressed to 1077 individuals were as follows.

1. Should divorce be easier?
2. Do you support the law against sexual discrimination?
3. View on pre-marital sex: not at all wrong...always wrong
4. View on extra-marital sex: not at all wrong...always wrong
5. View on sexual relationship between individuals of the same sex: not at all wrong...always wrong
6. Should gays teach in school?
7. Should gays teach in higher education?
8. Should gays hold public positions?
9. Should a female homosexual couple be allowed to adopt children?
10. Should a male homosexual couple be allowed to adopt children?

For those items yielding a binary response (1,2,6,7,8,9,10), a positive response was coded as 1 and a negative response as 0. For items 3, 4, and 5 there were five categories: "always wrong", "mostly wrong", "sometimes wrong", "rarely wrong" and "not at all wrong". Responses "sometimes wrong", "rarely wrong", and "not at all wrong" were coded as 1 and responses "always wrong" and "mostly wrong" as 0. With ten variables, there are $2^{10} = 1024$ possible response patterns. Not all of these occur, but with a sample size of 1077 the data matrix takes up a good deal of space. The full data set is given on the Web site, but the cases with frequencies greater than ten are listed in Table 8.7 in decreasing order of observed frequency as an illustration.

Table 8.8 gives the proportions giving positive and negative responses to each item.

Since we come to the data with no preconceived ideas about what the latent variables might be, we begin by fitting a one-factor model to the ten items. The parameter estimates are listed in Table 8.9. Items 6, 7, and 8 have large discrimination coefficients, $\hat{\alpha}_{i1}$, indicating that the characteristic curves of those items are very steep. From the st$\hat{\alpha}_{i1}$ column, we see that item 1 has the weakest relationship with the latent variable, followed by items 2 and 4. The rest of the items show strong relationships with the latent variable f.

We first investigate the goodness-of-fit of the one-factor model using the methods described in Section 8.4. They all suggest that the one-factor model is not a satisfactory fit to the data. The overall goodness-of-fit measures suggested a very bad fit ($G^2 = 427.39, X^2 = 354.30$ on 32 degrees of freedom). There were also large discrepancies between the observed and expected frequencies for many pairs and triplets of items. Table 8.10 gives all the pairs and the (1,1,1) triplets of items where the chi-squared residuals were greater than 3.

The percentage of G^2 explained is 77.03%, which shows that the model goes a long way in explaining the associations, but taken with the very poor

Table 8.7 *Response frequencies, sexual attitudes data*

Response patterns	Frequency	Response patterns	Frequency
0110000000	117	1110000000	17
0110111100	95	0111000000	15
0100000000	93	0111011100	15
0110011100	90	0010011100	14
0110111111	40	0110111110	14
0010000000	35	0110011110	13
0100011100	32	1110011100	13
0000000000	29	1110111111	12
0110000100	27	0110011000	11
0110001100	21	0110100000	11
0111111100	19	0010000100	11
0100000100	18	0000011100	10
0111111111	18	Other patterns	287

Table 8.8 *Proportions giving positive and negative responses to observed items, sexual attitudes data*

Item	Response 1	Response 0
1	0.13	0.87
2	0.83	0.18
3	0.77	0.23
4	0.13	0.87
5	0.29	0.71
6	0.48	0.53
7	0.55	0.45
8	0.59	0.41
9	0.19	0.81
10	0.11	0.89

fit indicated by the other tests it is clearly desirable to continue by fitting a second latent variable.

The two-factor model is a considerable improvement. The percentage of G^2 explained increased from 77.03 to 86.8%. The log-likelihood-ratio statistic and the chi-squared statistic still indicate a poor fit ($G^2 = 268.50, X^2 = 199.07$, each on 24 degrees of freedom). However, we need to look at the fit on the margins before making a final judgement.

Comparing the results from the one-factor solution given in Table 8.10, we find that the two-factor solution is a great improvement for predicting the

Table 8.9 *Estimated difficulty and discrimination parameters with standard errors in brackets and standardized loadings for the one-factor model, sexual attitudes data*

Items	$\hat{\alpha}_{i0}$	s.e.	$\hat{\alpha}_{i1}$	s.e.	st$\hat{\alpha}_{i1}$	$\hat{\pi}_i(0)$
1	−1.93	(0.09)	0.11	(0.10)	0.11	0.13
2	1.65	(0.10)	0.53	(0.11)	0.47	0.84
3	1.46	(0.10)	1.00	(0.11)	0.71	0.81
4	−2.01	(0.11)	0.60	(0.10)	0.52	0.12
5	−1.29	(0.11)	1.79	(0.16)	0.87	0.22
6	−0.12	(0.45)	10.08	(1.63)	1.00	0.47
7	1.99	(0.84)	10.05	(3.39)	1.00	0.88
8	1.05	(0.17)	3.52	(0.30)	0.96	0.74
9	−2.06	(0.14)	1.64	(0.18)	0.85	0.11
10	−3.72	(0.27)	2.44	(0.25)	0.93	0.02

observed two- and three-way margins. The fit was found to be poor (with residuals greater than 3) for the margins given in Table 8.11.

Although the fit is still somewhat questionable, the large percentage of G^2 explained encourages us to attempt an interpretation of the two-factor model.

Table 8.12 gives the maximum likelihood estimates together with their asymptotic (estimated using large sample theory) standard errors and the standardized parameters for the factor loadings. The last column shows very striking differences in the response of the median individual to the various questions. The last two items on adoption by homosexual couples show virtually no support for the propositions. There are also small probabilities of responding positively to items 1, 4, and 5. The marginal observed proportions given in Table 8.8 give a similar picture but they relate to views in the whole sample rather than to the median individual. As an aid to interpretation, the standardized factor loadings are plotted in Figure 8.3.

We see that items 2, 6, 7, and 8 have high loadings on the first factor and low loadings on the second factor. Items 3, 4 , 9, and 10 have high loadings on the second factor and low on the first factor. Item 5 lies somewhere in between. The interpretation is not entirely clear, but we note that the items in the first group are concerned with public matters whereas items 2 and 6, at least, are concerned with private behaviour. However, the inclusion of items 9 and 10 does not fit with this interpretation. We might hope that the plot of the loadings would suggest a rotation that would help the interpretation. From Figure 8.3, we see that there is no obvious orthogonal rotation that produces a simpler pattern than the one revealed from the original factor solution.

The failure to get two clear-cut factors coupled with the poor fit of the model overall suggests that the analysis should be taken further. The obvious thing would be to try a three-factor model or to re-analyse the data omitting the last two items, which seem to differ in some fundamental way from the

Table 8.10 *Chi-squared residuals greater than 3 for all the second order and (1,1,1) third order margins for the one-factor model, sexual attitudes data*

Response	Items	O	E	$O-E$	$(O-E)^2/E$
(0,0)	5, 3	237	208.20	28.80	3.98
	10, 9	875	814.79	60.23	4.45
(0,1)	3, 1	19	29.08	−10.08	3.49
	10, 9	88	144.77	−56.77	22.26
(1,0)	4, 3	4	22.46	−18.46	15.17
	5, 2	23	37.85	−14.85	5.83
	5, 3	14	36.52	−22.52	13.89
	9, 6	46	25.95	20.05	15.50
	9, 7	36	17.88	18.12	18.35
	9, 8	29	19.87	9.13	4.20
	10, 5	23	34.17	−11.17	3.65
	10, 6	15	4.13	10.87	28.66
	10, 7	12	2.50	9.50	36.16
	10, 8	11	3.73	7.27	14.19
	10, 9	2	54.36	−52.36	50.44
(1,1)	4, 1	29	18.88	10.12	5.42
	9, 6	154	181.92	−27.92	4.29
	9, 7	164	189.98	−25.98	3.55
	10, 9	112	63.09	48.91	37.91
(1,1,1)	1, 2, 6	50	64.53	−14.53	3.27
	1, 3, 4	29	16.10	12.90	10.34
	1, 4, 8	22	14.94	7.06	3.33
	1, 4, 10	8	4.18	3.82	3.49
	1, 5, 10	20	12.21	7.79	4.97
	1, 9, 10	21	9.33	11.67	14.60
	2, 3, 4	122	104.17	17.83	3.05
	2, 6, 9	137	164.02	−27.02	4.45
	2, 7, 9	147	170.74	−23.74	3.30
	2, 9, 10	99	58.25	40.75	28.51
	3, 9, 10	106	60.00	46.00	35.26
	4, 9, 10	33	17.31	15.67	14.21
	5, 9, 10	89	50.37	38.63	29.63
	6, 7, 9	153	180.82	−27.82	4.28
	6, 8, 9	151	176.04	−25.04	3.56
	6, 9, 10	97	62.74	34.26	18.71
	7, 9, 10	100	62.92	37.08	21.85
	8, 9, 10	101	62.55	38.45	23.64

Table 8.11 *Chi-squared residuals greater than 3 for the first, second and (1,1,1) third order margins for the two-factor model, sexual attitudes data*

Response	Items	O	E	$O-E$	$(O-E)^2/E$
(0,0)	7, 6	477	436.9	40.01	3.66
	9, 7	451	413.58	37.42	3.38
	7, 7	487	448.17	38.83	3.37
	8, 7	382	349.33	32.67	3.38
	10, 7	475	436.50	38.49	3.39
(1,0)	4, 3	4	17.65	−13.65	10.55
	5, 2	23	38.38	−15.38	6.16
	5, 3	14	28.06	−14.06	7.04
	10, 3	6	2.62	3.38	4.35
(1,1,1)	1, 3, 4	29	19.51	9.49	4.62

Table 8.12 *Estimated difficulty and discrimination parameters with standard errors in brackets and standardized loadings for the two-factor model, sexual attitudes data*

Items	$\hat{\alpha}_{i0}$	s.e.	$\hat{\alpha}_{i1}$	s.e.	$\hat{\alpha}_{i2}$	s.e.	st$\hat{\alpha}_{i1}$	st$\hat{\alpha}_{i2}$	$\hat{\pi}_i(\mathbf{0})$
1	−2.01	(0.11)	−0.25	(0.14)	0.38	(0.13)	−0.22	0.35	0.12
2	1.67	(0.09)	0.51	(0.12)	0.22	(0.12)	0.44	0.19	0.84
3	1.64	(0.12)	0.40	(0.13)	1.30	(0.16)	0.24	0.77	0.84
4	−2.10	(0.12)	0.11	(0.12)	0.79	(0.14)	0.09	0.62	0.11
5	−1.40	(0.13)	1.12	(0.14)	1.65	(0.17)	0.50	0.74	0.20
6	−0.05	(0.34)	8.12	(1.65)	4.41	(0.88)	0.87	0.48	0.49
7	2.46	(1.46)	10.26	(5.48)	6.22	(2.78)	0.85	0.52	0.92
8	1.06	(0.15)	2.79	(0.26)	1.83	(0.21)	0.80	0.53	0.74
9	−4.14	(0.71)	0.11	(0.23)	4.86	(1.20)	0.02	0.98	0.02
10	−14.82	(202.11)	0.54	(0.77)	10.22	(123.60)	0.05	0.99	0.00

other items. A third possibility, to which we shall return in Chapter 10, is to consider a different kind of model; namely, a latent class model.

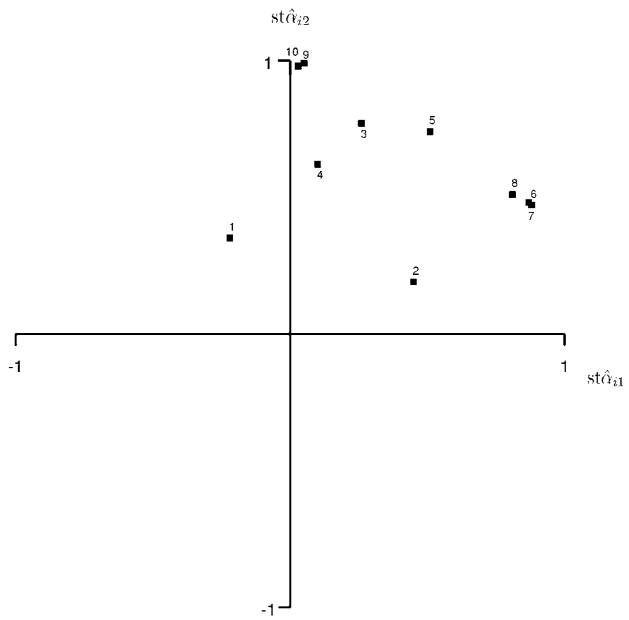

Figure 8.3 *Plots of standardized loadings, sexual attitudes data*

8.9 Additional examples and further work

The Law School Admission Test (LSAT), Section VI

The LSAT example is part of an educational test data set given in Bock and Lieberman (1970). The LSAT is a classical example in educational testing for measuring ability traits. This is a test that was designed to measure a *single* latent ability scale. The test as given in Bock and Lieberman (1970) consisted of five items taken by 1000 individuals. The main interest is whether the attempt to construct items which are indicators solely of this ability has been successful and, if so, what do the parameter estimates tell us about the items. From Table 8.13, you will see that 92% of the students answered item 1 correctly but only 55% answered item 3 correctly. That makes item 3 the most "difficult" among the five items. The full data set is given in Table 8.15. To investigate whether the five items form a *unidimensional* scale, you need to test whether the one-factor model is a good fit to the five items. The overall goodness-of-fit measures show that the one-factor model is a very good fit to the data ($G^2 = 15.30$ and $X^2 = 11.66$ on 13 degrees of freedom). In other words, the associations among the five items can be explained by a single latent variable that in this example is an ability which the test is designed to measure. Since the one-factor model is not rejected by the overall goodness-of-fit test, there is no need to check the fit on the two- and three-way margins. G^2 and X^2 measure how well the model predicts the whole response pattern. The

Table 8.13 *Proportions of positive and negative responses for observed items, LSAT data*

Item	Response 1	Response 0
1	0.92	0.08
2	0.71	0.29
3	0.55	0.45
4	0.76	0.24
5	0.87	0.13

first two columns of Table 8.15 show small discrepancies between the observed frequencies and the expected frequencies under the one-factor model.

Table 8.14 gives the parameter estimates for the one-factor solution. The last column of the table, $\hat{\pi}_i(0)$, gives the probability that the median individual will respond correctly to any of those five items. The five items have different difficulty levels. However the median individual has quite a high chance of getting the items correct indicating that, overall, the items are quite easy.

Table 8.14 *Estimated difficulty and discrimination parameters with standard errors in brackets and standardized loadings for the one-factor model, LSAT data*

item	$\hat{\alpha}_{i0}$	s.e.	$\hat{\alpha}_{i1}$	s.e.	st$\hat{\alpha}_{i1}$	$\hat{\pi}_i(0)$
1	2.77	(0.20)	0.83	(0.25)	0.64	0.94
2	0.99	(0.09)	0.72	(0.19)	0.59	0.73
3	0.25	(0.08)	0.89	(0.23)	0.67	0.56
4	1.28	(0.10)	0.69	(0.19)	0.57	0.78
5	2.05	(0.13)	0.66	(0.20)	0.55	0.89

The factor loadings $\hat{\alpha}_{i1}$ are all positive and of similar magnitude with similar standard errors. The same is true for the standardized loadings st$\hat{\alpha}_{i1}$. That implies that all five items have similar discriminating power and so a similar weight is applied to each response. In that case, the component score that is used to scale individuals on the latent dimension should give results that are close (similar ranking) to the scores obtained when the total score is used (see columns five and six of Table 8.15). This is an example where the Rasch model might be appropriate or you might analyse the five items using a latent class model that is discussed in Chapter 10. You could compare the ranking of the individuals obtained from the latent trait model with the allocation of individuals into two distinct classes.

Table 8.15 *Factor scores in increasing order, LSAT data*

Observed frequency	Expected frequency	$\hat{E}(f \mid \mathbf{x})$	$\hat{\sigma}(f \mid \mathbf{x})$	Component score (X_1)	Total score	Response pattern
3	2.3	−1.90	0.80	0.00	0	00000
6	5.9	−1.47	0.80	0.66	1	00001
2	2.6	−1.45	0.80	0.69	1	00010
1	1.8	−1.43	0.80	0.72	1	01000
10	9.5	−1.37	0.80	0.83	1	10000
1	0.7	−1.32	0.80	0.89	1	00100
11	8.9	−1.03	0.81	1.35	2	00011
8	6.4	−1.01	0.81	1.38	2	01001
29	34.6	−0.94	0.81	1.48	2	10001
14	15.6	−0.92	0.81	1.51	2	10010
1	2.6	−0.90	0.81	1.55	2	00101
16	11.3	−0.90	0.81	1.55	2	11000
3	1.2	−0.88	0.81	1.58	2	00110
3	4.7	−0.79	0.81	1.72	2	10100
16	13.6	−0.55	0.82	2.07	3	01011
81	76.6	−0.48	0.82	2.17	3	10011
56	56.1	−0.46	0.82	2.21	3	11001
4	6.0	−0.44	0.82	2.24	3	00111
21	25.7	−0.44	0.82	2.24	3	11010
3	4.4	−0.42	0.82	2.27	3	01101
2	2.0	−0.40	0.82	2.30	3	01110
28	25.0	−0.35	0.82	2.37	3	10101
15	11.5	−0.33	0.82	2.40	3	10110
11	8.4	−0.30	0.82	2.44	3	11100
173	173.3	0.01	0.83	2.89	4	11011
15	13.9	0.05	0.84	2.96	4	01111
80	83.5	0.13	0.84	3.06	4	10111
61	62.5	0.15	0.84	3.10	4	11101
28	29.1	0.17	0.84	3.13	4	11110
298	296.7	0.65	0.86	3.78	5	11111

Workplace industrial relations data

This example is taken from a section of the 1990 Workplace Industrial Relations Survey (WIRS) dealing with management/worker consultation in firms. A subset of the data is used here that consists of 1005 firms and concerns non-manual workers. The questions asked are given below:

> Please consider the most recent change involving the introduction of new plant, machinery and equipment. Were discussions or consultations of any of the type on this card held either about the introduction of the change or about the way it was to be implemented?

1. Informal discussion with individual workers
2. Meetings with groups of workers
3. Discussions in established joint consultative committee
4. Discussions in specially constituted committee to consider the change
5. Discussions with union representatives at the establishment
6. Discussions with paid union officials from outside

All six items measure the amount of consultation that takes place in firms at different levels of the firm structure. Items 1 to 6 cover a range of informal to formal types of consultation. Those firms which place a high value on consultation might be expected to use all or most consultation practices. The six items are analysed here using the latent trait model. We should mention that the items discussed here were not initially constructed to form a scale as is the case in the LSAT example and in most educational data. Therefore, our analysis is completely exploratory. The full data set is given on the Web site. The proportions giving positive and negative responses to each item are given in Table 8.16. The most common type of consultation among the 1005 firms is the established joint consultative committee. The one-factor model

Table 8.16 *Proportions giving positive and negative responses to observed items, WIRS data*

Item	Response 1	Response 0
1	0.37	0.63
2	0.58	0.42
3	0.28	0.72
4	0.24	0.76
5	0.36	0.64
6	0.15	0.85

gives $G^2 = 269.4$ and $X^2 = 264.2$ on 32 degrees of freedom. Both goodness-of-fit measures indicate that the one-factor model is a poor fit to the data. Table 8.17 gives chi-squared residuals greater than 3 for the second and third-way margins. The largest discrepancies are found between items 1 and 2. As a result, the model fails to explain the associations among the six items, judging by the overall goodness-of-fit measures, and it also fails to explain the pairwise associations.

You should continue the analysis by fitting one more latent variable that might account for the big discrepancies between the observed and expected frequencies. The percentage of G^2 explained increases from 49.35% for the one-factor model to 74.58% for the two-factor model. Clearly, the second latent variable contributes substantially in explaining the associations among the six items. However, the fit of the two-factor model is still poor if we look at the $G^2 = 146.4$ and $X^2 = 131.5$ on 24 degrees of freedom. However, the

Table 8.17 *Chi-squared residuals greater than 3 for the second and (1,1,1) third order margins for the one-factor model, WIRS data*

Response	Items	O	E	$O-E$	$(O-E)^2/E$
(0,0)	2, 1	186	265.75	−79.75	23.93
(0,1)	2, 1	233	153.23	79.77	41.52
(1,0)	2, 1	444	364.25	79.75	17.46
	4, 1	172	145.48	26.52	4.84
	4, 2	61	87.00	−26.00	7.77
(1,1)	2, 1	142	221.77	−79.77	28.69
	4, 1	69	95.65	−26.65	7.43
	4, 2	180	154.13	25.87	4.34
(1,1,1)	1, 2, 3	37	75.79	−38.79	19.85
	1 ,2, 4	23	61.75	−38.75	24.32
	1, 2, 5	53	94.85	−41.85	18.46
	1, 2, 6	26	40.32	−14.32	5.08
	1, 3, 4	30	45.69	−15.69	5.39
	1, 4, 5	35	55.73	−20.73	7.71
	2, 3, 4	93	75.03	17.97	4.31
	2, 4, 5	108	91.39	16.61	3.02

residuals for the two-way margins are all close to zero. The second latent variable accounts for the pairwise associations but the fit is still not satisfactory on the three-way margins. Table 8.18 gives the residuals greater than 3 for the $(1, 1, 1)$ three-way margins. Item 1 appears in all the triplets that show a bad fit. This is the least formal item, which is also vaguely worded and might be interpreted differently by different respondents.

Table 8.18 *Chi-squared residuals greater than 3 for the third order margins for the two-factor model, response (1,1,1) to items (i, j, k), WIRS data*

Item i	Item j	Item k	O	E	$O-E$	$(O-E)^2/E$
1	2	3	37	60.53	−23.53	9.15
1	2	4	23	40.65	−17.65	7.66
1	2	5	53	73.99	−20.99	5.95
1	3	6	31	42.54	−11.54	3.13
1	5	6	36	49.84	−13.84	3.84

Although the model is not good in predicting the three-way margins, it is worth looking at the parameter estimates of the two-factor latent trait model given in Table 8.19. All the loadings ($\hat{\alpha}_{i1}$) of the first factor except that for item 1 (the least formal item) are positive and large indicating a "general" factor relating to amount of consultation which takes place.

Table 8.19 *Estimated difficulty and discrimination parameters with standard errors in brackets and standardized loadings for the two-factor model, WIRS data*

Items	$\hat{\alpha}_{i0}$	s.e.	$\hat{\alpha}_{i1}$	s.e.	$\hat{\alpha}_{i2}$	s.e.	st$\hat{\alpha}_{i1}$	st$\hat{\alpha}_{i2}$	$\hat{\pi}_i(\mathbf{0})$
1	−0.93	(0.31)	−0.97	(0.48)	2.13	(0.96)	−0.38	0.84	0.28
2	0.54	(0.15)	1.51	(0.47)	−0.96	(0.36)	0.74	−0.47	0.63
3	−1.40	(0.14)	1.31	(0.18)	1.11	(0.18)	0.66	0.56	0.20
4	−1.47	(0.11)	1.22	(0.15)	0.12	(0.11)	0.77	0.08	0.19
5	−0.97	(0.14)	1.58	(0.24)	1.24	(0.21)	0.70	0.55	0.27
6	−2.39	(0.20)	1.05	(0.16)	1.06	(0.21)	0.59	0.59	0.08

The analysis may be repeated with item 1 omitted. The items used in the analysis are item 2 to item 6 and those names are used here. The one-factor model gives $G^2 = 50.50$ and $X^2 = 46.29$ on 17 degrees of freedom. The one-factor model is rejected. The fit of the two-way margins is very good except for two pairs, and there is only one large chi-squared residual in the (1,1,1) three-way margins. These residuals are given in Table 8.20 and all include item 2 which is the second least formal item after item 1 (which is omitted from the current analysis). The fit is improved when the two-factor model is fitted giving a $G^2 = 30.16$ and $X^2 = 27.53$ on 13 degrees of freedom. Those statistics still reject the two-factor model. However, the fit on the two-way margins is excellent and the $(1,1,1)$ three-way margins have no residual greater than 0.89. Further analysis of this data set can be found in Bartholomew (1998).

Table 8.20 *Chi-squared residuals greater than 3 for the second and the (1,1,1) third order margins for the one-factor model, WIRS data, item 1 omitted*

Response	Items	O	E	$O-E$	$(O-E)^2/E$
(1,0)	3, 1	61	84.94	−23.94	6.75
(1,1)	4, 2	180	156.28	23.72	3.60
(1,1,1)	(1, 2, 3)	93	77.09	15.91	3.28

The parameter estimates of the one-factor model given in Table 8.21 indicate a clear general factor corresponding to the amount of consultation that takes place. Note that item 2 has the smallest factor loading, while items 3 to 6 have similar factor loadings. It is quite apparent that items 3 to 6 can be considered separately to construct a scale measuring the amount of formal consultation which takes place. Fitting the one-factor model to items 3 to 6 gives $G^2 = 16.6$ and $X^2 = 14.5$ on seven degrees of freedom. All residuals for the two- and three-way margins are smaller than 1.0.

Table 8.21 *Estimated difficulty and discrimination parameters with standard errors in brackets and standardized loadings for the one-factor model with item 1 omitted, WIRS data*

item	$\hat{\alpha}_{i0}$	s.e.	$\hat{\alpha}_{i1}$	s.e.	st$\hat{\alpha}_{i1}$	$\hat{\pi}_i(0)$
2	0.35	(0.07)	0.42	(0.10)	0.39	0.59
3	−1.38	(0.14)	1.69	(0.23)	0.86	0.20
4	−1.40	(0.10)	1.05	(0.14)	0.72	0.20
5	−0.95	(0.13)	1.97	(0.31)	0.89	0.28
6	−2.29	(0.16)	1.34	(0.18)	0.80	0.09

Women's mobility

These data are from the Bangladesh Fertility Survey of 1989 (Huq and Cleland 1990). The rural subsample of 8445 women is analysed here. The questionnaire contains a number of items believed to measure different dimensions of women's status. The particular dimension that we shall focus on here is women's mobility or social freedom. Women were asked whether they could engage in the following activities alone (1=yes, 0=no).

1. Go to any part of the village/town/city
2. Go outside the village/town/city
3. Talk to a man you do not know
4. Go to a cinema/cultural show
5. Go shopping
6. Go to a cooperative/mothers' club/other club
7. Attend a political meeting
8. Go to a health centre/hospital

First, the one-factor model was fitted to the eight items to investigate whether the variables are all indicators of the same type of women's mobility in society. The one-factor model gives a G^2 equal to 364.5 on 39 degrees of freedom indicating a bad fit. Table 8.22 shows the chi-squared residuals greater than 3 for the two-way margins and the (1,1,1) three-way margins of the one-factor model.

The two-factor model is still rejected based on a G^2 equal to 263.41 on 33 degrees of freedom. The percentage of G^2 explained increases only slightly from 94.98% to 96.92%. However, although the contribution of the second factor is small, the fit on the two-way margins and the (1,1,1) three-way margins is generally very good; the margins for which the fit is poor are shown in Table 8.23.

Table 8.22 *Chi-squared residuals greater than 3 for the second and the (1,1,1) third order margins for the one-factor model, women's mobility data*

Response	Items	O	E	$O-E$	$(O-E)^2/E$
(0,1)	3, 2	187	229.19	−42.19	7.76
	6, 2	1986	1899.91	86.09	3.90
	7, 6	532	596.04	−64.04	6.88
	8, 5	194	245.15	−51.15	10.67
	8, 7	108	134.51	−26.51	5.22
(1,0)	2, 1	52	117.29	-65.29	36.35
	5, 1	13	3.02	9.99	32.92
	5, 2	98	77.74	20.25	5.28
	5, 3	20	12.40	7.60	4.66
	5, 4	19	28.75	−9.75	3.31
	6, 2	274	196.34	77.66	30.71
	6, 3	44	32.03	11.97	4.47
	7, 1	6	1.13	4.87	20.97
	7, 2	62	36.82	25.18	17.21
	7, 4	17	8.75	8.25	7.78
	7, 6	41	93.69	−52.69	29.63
	8, 1	28	7.15	20.85	60.83
	8, 3	38	22.74	15.26	10.24
	8, 4	88	67.82	20.18	6.01
	8, 5	340	391.82	−51.82	6.85
(1,1)	6, 2	665	756.15	−91.15	10.99
	7, 6	407	356.45	50.55	7.17
	8, 5	392	348.29	43.71	5.48
(1,1,1)	1, 2, 3	2433	2338.67	94.33	3.80
	1, 2, 6	659	751.02	−92.02	11.27
	1, 5, 8	392	347.45	44.55	5.71
	1, 6, 7	403	355.75	47.25	6.27
	2, 3, 6	653	736.66	−83.66	9.50
	2, 4, 6	637	704.12	−67.12	6.40
	3, 5, 8	389	343.72	45.28	5.96
	3, 6, 7	402	352.32	49.68	7.01
	4, 5, 8	386	341.75	44.25	5.73
	4, 6, 7	396	351.63	44.37	5.60
	5, 6, 7	304	271.48	32.52	3.89
	5, 6, 8	326	279.56	46.44	7.72
	5, 7, 8	276	246.59	29.41	3.51
	6, 7, 8	318	267.09	50.91	9.70

Table 8.23 *Chi-squared residuals greater than 3 for the second and the (1,1,1) third order margins for the two-factor model, women's mobility data*

Response	Items	O	E	$O-E$	$(O-E)^2/E$
(0,1)	8, 5	194	239.58	−45.58	8.67
	8, 7	108	137.09	−29.09	6.17
(1,0)	4, 3	226	253.70	−27.70	3.02
	5, 1	13	7.12	5.88	4.86
	5, 4	19	33.25	−14.25	6.10
	6, 1	15	30.37	−15.37	7.78
	7, 2	62	78.03	−16.03	3.29
	7, 3	8	13.51	−5.51	2.25
	7, 6	41	67.28	−26.28	10.26
	8, 1	28	14.42	13.58	12.78
	8, 2	144	166.51	−22.51	3.04
	8, 4	88	71.84	16.16	3.64
	8, 5	340	388.56	−48.56	6.07
(1,1)	8, 5	392	355.73	36.27	3.70
(1,1,1)	1, 5, 8	392	353.37	38.63	4.22
	2, 5, 8	351	316.27	34.73	3.81
	3, 5, 8	389	348.32	40.68	4.75
	4, 5, 8	386	347.28	38.72	4.32
	5, 7, 8	276	245.75	30.25	3.72
	6, 7, 8	318	287.55	30.45	3.23

The parameter estimates for the two-factor model are given in Table 8.24. The eight items are positively correlated with both factors. However, as we can see from the standardized loadings $\mathrm{st}\hat{\alpha_{i1}}$ and $\mathrm{st}\hat{\alpha_{i2}}$, items 1 to 4 load heavily on the first factor where items 4 to 8 load heavily on the second factor. The loading for item 7 should be interpreted with caution due to its extremely large standard error. The two factors can be interpreted as measuring different dimensions of women's status. Items 5 to 8, and to some extent item 4, indicate a relatively high level of participation in public life; engaging in any of these activities would suggest a high degree of social freedom for a woman in rural Bangladesh. In contrast, items 1 to 3 are less specific indicating a degree of freedom but not necessarily in the public life sphere. The $\hat{\pi}_i(\mathbf{0})$ values show clearly that a woman who is in the middle of both factors has close to zero chances of responding positively to items 5 to 8. You should compare the results obtained here with those obtained in Chapter 10 where a latent class model is fitted.

Table 8.24 *Estimated difficulty and discrimination parameters with standard errors in brackets and standardized factor loadings for the two-factor model, women's mobility data*

Items	$\hat{\alpha}_{i0}$	s.e.	$\hat{\alpha}_{i1}$	s.e.	$\hat{\alpha}_{i2}$	s.e.	st$\hat{\alpha}_{i1}$	st$\hat{\alpha}_{i2}$	$\hat{\pi}_i(\mathbf{0})$
1	2.66	(0.18)	2.46	(0.28)	0.98	(0.17)	0.87	0.34	0.94
2	−1.58	(0.09)	2.48	(0.21)	1.32	(0.15)	0.83	0.44	0.17
3	1.56	(0.05)	1.25	(0.08)	0.86	(0.10)	0.69	0.47	0.83
4	−1.17	(0.06)	1.97	(0.16)	2.26	(0.17)	0.62	0.72	0.24
5	−6.58	(0.30)	1.98	(0.23)	3.57	(0.22)	0.47	0.85	0.00
6	−5.11	(0.27)	1.32	(0.23)	3.60	(0.24)	0.33	0.91	0.01
7	−17.24	(94.82)	2.20	(0.43)	10.01	(58.02)	0.21	0.97	0.00
8	−4.94	(0.17)	1.51	(0.17)	2.80	(0.15)	0.45	0.84	0.01

8.10 Software

The software GENLAT (Moustaki 2001) for estimating the logit model is available on the Web site associated with the book. An important feature of the software is that it also produces estimated asymptotic standard errors for the estimates. These are based on asymptotic theory (large samples) and are only approximations but they often serve to add a note of caution to the interpretation. The program provides the goodness-of-fit measures and scaling methods discussed in this chapter. Software GLLAMM (Rabe-Hesketh, Pickles, and Skrondal 2004), MULTILOG (Thissen, Chen, and Bock 1991) and PARSCALE (Muraki and Bock 1997) can also be used to fit factor analysis model for binary data using the IRF approach. The UV approach is implemented in commercial software such as Amos (Arbuckle 2006) LISREL (Jöreskog and Sörbom 1999), Mplus (Muthén and Muthén 2007), and EQS (Bentler 1996).

8.11 Further reading

Bartholomew, D. J. and Knott, M. (1999). *Latent Variable Models and Factor Analysis* (2nd ed.). London: Arnold.

Bollen, K. A. (1989). *Structural Equations with Latent Variables.* New York: Wiley and Sons.

Fischer, G. H. and Molenaar, I. W. (Eds.) (1995). *Rasch Models: Foundations, Recent Developments, and Applications.* New York: Springer-Verlag.

Hambleton, R. K., Swaminathan, H. and Rogers, H. J. (1991). *Fundamentals of Item Response Theory.* Newbury Park, California: Sage Publications.

Heinen, T. (1996). *Latent Class and Discrete Latent Trait Models.* Thousand Oaks: Sage Publications.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized Latent Variable Modelling: Multilevel, Longitudinal, and Structural Equation Models.* Boca Raton, FL: Chapman and Hall/CRC.

van der Linden, W. J. and Hambleton, R. (Eds.) (1997). *Handbook of Modern Item Response Theory.* New York: Springer-Verlag.