LATENT VARIABLE MODELS

FOR BINARY RESPONSE DATA

Maria Teresinha Albanese

A Thesis submitted for a degree of

Doctor of Philosophy

The London School of Economics and

Political Science

September 1990

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

ACKNOWLEDGEMENTS

I wish to express my deep gratitude to my supervisor, Dr. Martin Knott. His constant guidance and encouragement has been invaluable.

I am also very grateful to my supervisor, Prof. David Bartholomew, for many helpful discussions and extremely valuable comments.

My deep thanks are due to my family and many friends for their love and understanding.

I gratefully acknowledge financial support from the Brazilian Government through the Universidade Federal do Rio Grande do Sul and the Conselho Nacional de Desenvolvimento Cientifico e Tecnologico.

ABSTRACT

Most of the results in this thesis are obtained for the logit/probit model for binary response data given by Bartholomew (1980), which is sometimes called the two-parameter logistic model. In most the cases the results also hold for other common binary response models.

By profiling and an approximation, we investigate the behaviour of the likelihood function, to see if it is suitable for ML estimation. Particular attention is given to the shape of the likelihood around the maximum point in order to see whether the information matrix will give a good guide to the variability of the estimates.

The adequacy of the asymptotic variance-covariance matrix is investigated through jackknife and bootstrap techniques.

We obtain the marginal ML estimators for the Rasch model and compare them with those obtained from conditional ML estimation. We also test the fit of the Rasch model against a logit/probit model with a likelihood ratio test, and investigate the behaviour of the likelihood function for the Rasch model and its bootstrap estimates together with approximate methods.

For both fixed and decreasing sample size, we investigate the stability of the discrimination parameter estimates $\hat{\alpha}_{i,}$, when the number of items is reduced.

We study the conditions which give rise to large discrimination parameter estimates. This leads to a method for the generation of a (p+1)th item with any fixed $\hat{\alpha}_{p+1,1}$ and $\hat{\alpha}_{p+1,0}$.

In practice it is importante to measure the latent variable and this is usually done by using the posterior mean or the component scores. We give some theoretical and applied results for the relation between the linearity of the plot of the posterior mean latent variable values, the component scores and the normality of those posterior distributions.

SUMMARY

Chapter 3- ADEQUACY of the ASYMPTOTIC VARIANCE-COVARIANCE MATRIX using BOOTSTRAP and JACKKNIFE TECHNIQUES

Chapter 4- RASCH MODEL

LIST of TABLES

Table 3.3- Comparison between the bootstrap, original ML(in brackets) and the jackknife parameter estimates $\hat{\alpha}_{i,0}$ to the Attitudes towards the U.S.Army.

Table 3.4- Comparison between the bootstrap, original ML(in brackets) and the jackknife parameter estimates $\hat{\alpha}^*_{i,0}$ to the Attitudes towards the U.S.Army.

Table 3.5- Correlations between the original ML parameter estimates based on the observed 2nd derivative matrix (under the diagonal) and on the information matrix (above the diagonal) to the Attitudes towards the U.S.Army.

Table 3.6- Bootstrap (under the diagonal) and jackknife (above the diagonal) estimates of the correlations between the parameter estimates of the Attitudes towards the U.S.Army.

Table 3.7- Comparison between the bootstrap, original ML(in brackets) and the jackknife parameter estimates $\hat{\alpha}_{i,1}$ to the ART on white women.

Table 3.8- Comparison between the bootstrap, original ML(in brackets) and the jackknife parameter estimates $\hat{\alpha}_{i,0}$ to the ART on white women.

Table 3.9- Comparison between the bootstrap, original ML(in brackets) and the jackknife parameter estimates $\hat{\alpha}^*_{i,0}$ to the ART on white women.

Table 3.10- Correlations between the original ML parameter estimates based on the observed 2nd derivative matrix (under the diagonal) and on the information matrix (above the diagonal) to the ART on white women.

Table 3.11- Bootstrap (under the diagonal) and jackknife (above the diagonal) estimates of the correlations between the parameter estimates of the ART on white women.

Table 3.12- Score distribution and results obtained by fitting a logit/probit model for the Stouffer and Toby data.

Table 3.13- Comparison between the bootstrap, original ML(in brackets) and the jackknife parameter estimates $\hat{\alpha}_{i,1}$ for the Stouffer and Toby data.

Table 3.14- Comparison between the bootstrap, original ML(in brackets) and the jackknife parameter estimates $\hat{\alpha}_{i,0}$ for the Stouffer and Toby data.

Table 3.15- Comparison between the bootstrap, original ML(in brackets) and the jackknife parameter estimates $\hat{\alpha}^*_{i,0}$ for the Stouffer and Toby data.

Table 3.16- Correlations between the original ML parameter estimates based on the observed 2nd derivative matrix (under the diagonal) and on the information matrix (above the diagonal) for the Stouffer and Toby data.

Table 3.17- Bootstrap (under the diagonal) and jackknife (above the diagonal) estimates of the correlations between the parameter estimates of the Stouffer and Toby data.

Table 3.18- Comparison between the bootstrap, original ML(in brackets) and the jackknife parameter estimates $\hat{\alpha}_{i,1}$ for the Lombard and Doering data.

Table 3.19- Comparison between the bootstrap, original ML(in brackets) and the jackknife parameter estimates $\hat{\alpha}_{i,0}$ for the Lombard and Doering data.

Table 3.20- Comparison between the bootstrap, original ML(in brackets) and the jackknife parameter estimates $\hat{\alpha}^*_{i,0}$ for the Lombard and Doering data.

Table 3.21- Correlations between the original ML parameter estimates based on the observed 2nd derivative matrix (under the diagonal) and on the information matrix (above the diagonal) for the Lombard and Doering data.

Table 3.22- Bootstrap (under the diagonal) and jackknife (above the diagonal) estimates of the correlations between the parameter estimates of the Lombard and Doering data.

Table 3.23- Comparison between the bootstrap, original ML(in brackets) and the jackknife parameter estimates $\hat{\alpha}_{i,1}$ to the ART on black women.

Table 3.24- Comparison between the bootstrap, original ML(in brackets) and the jackknife parameter estimates $\hat{\alpha}_{i,0}$ to the ART on black women.

Table 3.25- Comparison between the bootstrap, original ML(in brackets) and the jackknife parameter estimates $\hat{\alpha}^*_{i,0}$ to the ARt on black women.

Table 3.26- Correlations between the original ML parameter estimates based on the observed 2nd derivative matrix (under the diagonal) and on the information matrix (above the diagonal) to the ART on black women.

Table 3.27- Bootstrap (under the diagonal) and jackknife (above the diagonal) estimates of the correlations between the parameter estimates of the ART on black women.

Table 3.28- Comparison between the bootstrap, original ML (in brackets) and the normal bootstrap parameter estimates $\hat{\alpha}_{i,1}$ to the Attitudes towards the U.S.Army.

Table 3.29- Comparison between the bootstrap, original ML (in brackets) and the normal bootstrap parameter estimates $\hat{\alpha}_{i,1}$ to the ART on white women.

Table 3.30- Comparison between the bootstrap, original ML (in brackets) and the normal bootstrap parameter estimates $\hat{\alpha}_{i,1}$ for the Stouffer and Toby data.

Table 3.31- Comparison between the bootstrap, original ML (in brackets) and the normal bootstrap parameter estimates $\hat{\alpha}_{i,1}$ for the Lombard and Doering data.

Table 3.32- Comparison between the bootstrap, original ML (in brackets) and the normal bootstrap parameter estimates $\hat{\alpha}_{i,1}$ to the ART on black women.

Table 4.1- Parameter estimates and asymptotic standard deviations from fitting the Rasch model to the Attitudes towards the U.S.Army.

Table 4.2- Parameter estimates and asymptotic standard deviations from fitting the Rasch model to the ART on white women.

Table 4.3- Parameter estimates and asymptotic standard deviations from fitting the Rasch model to the Stouffer and Toby data.

Table 4.4- Parameter estimates and asymptotic standard deviations from fitting the Rasch model to the Lombard and Doering data.

Table 4.5- Parameter estimates and asymptotic standard deviations from fitting the Rasch model to the ART on black women.

Table 4.6- Parameter estimates and asymptotic standard deviations (in brackets) from fitting two- and one-parameter logistic (a logit/probit and the Rasch) models to each of the 30 normal bootstrap samples.

Table 4.7- Goodness-of-fit results from fitting a logit/probit and the Rasch model to each one of the 30 normal bootstrap samples obtained from ART on black women.

Table 4.8- Frequency distribution of the parameter estimates $\hat{\pi}_i$, i=1,...,4, from fitting a logit/probit model and the Rasch model to each one of the 30 normal bootstrap samples obtained from the ART on black women.

Table 4.9- Discrimination parameter estimates $\hat{\alpha}_{i,1}$ from MLE, approximate methods 1 and 2, respectively, for the 30 normal bootstrap samples obtained from the ART on black women.

Table 4.10- Difficulty parameter estimates from fitting the Rasch model to the LSAT, section 6.

Table 4.11- Difficulty parameter estimates from fitting the Rasch model to the LSAT, section 7.

Table 4.12- Comparison between Rasch and logit/probit models in terms of the LR statistic, when fitting subtests of Test 11A for England.

Table 4.13- Comparison between Rasch and logit/probit models in terms of the LR statistic, when fitting subtests of Test 11A for Wales.

Table 4.12- Comparison between Rasch and logit/probit models in terms of the LR statistic, when fitting subtests of Test 11A for Ireland.

Table 5.1- Parameter estimates $\hat{\alpha}_{i,1}$ and asymptotic standard deviations (in brackets) from fitting a logit/probit model to Test 11A.

Table 5.2- Parameter estimates $\hat{\alpha}_{i,1}$ and asymptotic standard deviations (in brackets) from fitting a logit/probit model to Test 11A.

Table 5.3- Parameter estimates $\hat{\alpha}_{i,1}$ and asymptotic standard deviations (in brackets) from fitting a logit/probit model to Test 12.

Table 5.4- Parameter estimates $\hat{\alpha}_{i,1}$ and asymptotic standard deviations (in brackets) from fitting a logit/probit model to Test 12.

Table 5.5- Parameter estimates $\hat{\alpha}_{i,1}$ and asymptotic standard deviations (in brackets) from fitting a logit/probit model to Test 11A- Ireland.

Table 5.6- Parameter estimates $\hat{\alpha}_{i,1}$ and asymptotic standard deviations (in brackets) from fitting a logit/probit model to Test 11A- Ireland.

Table 5.7- Correlations between the component scores from fitting a logit/probit to different subsets of items from Test 11A- Ireland.

Table 5.8- Parameter estimates $\hat{\alpha}_{i,1}$ and asymptotic standard deviations (in brackets) from fitting a logit/probit model to Test 11A- Wales.

Table 5.9- Parameter estimates $\hat{\alpha}_{i,1}$ and asymptotic standard deviations (in brackets) from fitting a logit/probit model to Test 11A- Wales.

Table 5.10- Correlations between the component scores from fitting a logit/probit to different subsets of items from Test 11A- Wales.

Table 5.11- Parameter estimates $\hat{\alpha}_{i,1}$ and asymptotic standard deviations (in brackets) from fitting a logit/probit model to Test 11A- England.

Table 5.12- Parameter estimates $\hat{\alpha}_{i,1}$ and asymptotic standard deviations (in brackets) from fitting a logit/probit model to Test 11A- England.

Table 5.13- Correlations between the component scores from fitting a logit/probit to different subsets of items from Test 11A- England.

Table 5.14- Frequency distribution of the parameter estimates $\hat{\alpha}_{i,1}$ from fitting a logit/probit model to Test 8, according to the location of the school.

Table 5.15- Parameter estimates and asymptotic standard deviations (in brackets) from fitting a logit/probit model to Reading Ability Tests (NFER data).

Table 6.1- Distribution of the score patterns and ratios of the conditional frequencies.

Table 6.2- Distribution of the score patterns and ratios of the conditional frequencies.

Table 6.3- Distribution of the score patterns and ratios of the conditional frequencies.

Table 6.4- Distribution of the score patterns and ratios of the conditional frequencies.

Table 6.5- Distribution of the score patterns and ratios of the conditional frequencies for the normal bootstrap sample 5 from the ART on black women.

Table 6.6- Distribution of the score patterns and ratios of the conditional frequencies for the Macready and Dayton's data.

Table 6.7- Distribution of the score patterns and ratios of the conditional frequencies for the ART on white women.

Table 6.8- Frequency distribution of N_{ts}, the expected number of individuals with score pattern x_s (s=1,...,4) at z_t (t=1,...,16).

Table 6.9- Frequency distribution of the data generated by chopping the distribution of N_{ts} at t' equal to 10 for every score pattern x_s.

Table 6.10- Frequency distribution of the score patterns of the Lombard and Doering data (Table 2.5) after deleting item 2.

Table 6.11- Distribution of N_{ts}, the expected number of individuals with score pattern x_s, s=1,...,8 at z_t, t=1,...,16.

Table 6.12- Sum of N_{ts}, $t \geq t'$, for every score pattern x_s, s=1,...,8.

Table 6.13- Frequencies of the data set generated by chopping the posterior distribution of z_t given x_s at t' for every score pattern x_s.

Table 6.14- Parameter estimates and asymptotic standard deviations (in brackets) from fitting a logit/probit model to the data presented in Table 6.12.

Table 6.15- Comparison between the pattern of $\hat{\alpha}_{i,1}$ and the predictability of an item from the others.

Table 7.1- Estimates of E(Y|x), E(Z|x) and the component score $\sum \alpha_{i,1} x_i$ when fitting a logit/probit model to the LSAT VI.

Table 7.2- Estimates of E(Y|x), E(Z|x) and the component score $\sum \alpha_{i,1} x_i$ when fitting a logit/probit model to Test 12.

Table 7.3- Estimates of E(Y|x), E(Z|x) and the component score $\sum \alpha_{i,1} x_i$ when fitting a logit/probit model to Test 13.

Table 7.4- Frequency distribution of the number of positive responses given to the six items with $\hat{\alpha}_{i,1} \geq 3.0$ for some intervals of E(Z|x).

LIST of FIGURES

Figure 2.15- Maximum loglikelihood value over $\hat{\alpha}_{i,0}$ for each $\hat{\alpha}^*_{i,1}$ fixed to the ART on white women, using methods A and B.

Figure 2.16- Loglikelihood values as a function of $\hat{\alpha}^*_{1,1}$ and $\hat{\alpha}^*_{1,0}$, using profile method to the ART on black women.

Figure 2.17- Loglikelihood values as a function of $\hat{\alpha}^*_{1,1}$ and $\hat{\alpha}^*_{1,0}$, using approximate method A to the ART on black women.

Figure 2.18- Maximum loglikelihood value over $\hat{\alpha}_{i,1}$ for each $\hat{\alpha}^*_{i,0}$ fixed to the ART on black women, using methods A and B.

Figure 2.19- Maximum loglikelihood value over $\hat{\alpha}_{i,0}$ for each $\hat{\alpha}^*_{i,1}$ fixed to the ART on black women, using methods A and B.

Figure 2.20- Loglikelihood values as a function of $\hat{\alpha}^*_{2,1}$ and $\hat{\alpha}^*_{2,0}$, using profile method for the Lombard and Doering data.

Figure 2.21- Loglikelihood values as a function of $\hat{\alpha}^*_{2,1}$ and $\hat{\alpha}^*_{2,0}$, using approximate method A for the Lombard and Doering data.

Figure 2.22- Maximum loglikelihood value over $\hat{\alpha}_{2,1}$ for each $\hat{\alpha}^*_{2,0}$ fixed for the Lombard and Doering data, using methods A and B.

Figure 2.23- Maximum loglikelihood value over $\hat{\alpha}_{2,0}$ for each $\hat{\alpha}^*_{2,1}$ fixed for the Lombard and Doering data, using methods A and B.

Figure 3.1- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}_{1,1}$ to the Attitudes towards the U.S.Army (original ML $\hat{\alpha}_{1,1} = 1.64$, bootstrap $\hat{\alpha}_{1,1} = 1.68$ and $R^2 = 97.9\%$).

Figure 3.2- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}_{1,0}$ to the Attitudes towards the U.S.Army (original ML $\hat{\alpha}_{1,0} = 0.85$, bootstrap $\hat{\alpha}_{1,0} = 0.88$ and $R^2 = 93.9\%$).

Figure 3.3- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}^*_{1,0}$ to the Attitudes towards the U.S.Army (original ML $\hat{\alpha}^*_{1,0} = 0.44$, bootstrap $\hat{\alpha}^*_{1,0} = 0.45$ and $R^2 = 99.4\%$).

Figure 3.4- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}_{1,1}$ to the ART on white women (original ML $\hat{\alpha}_{1,1} = 1.04$, bootstrap $\hat{\alpha}_{1,1} = 1.14$ and $R^2 = 88.5\%$).

Figure 3.5- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}_{1,0}$ to the ART on white women (original ML and bootstrap $\hat{\alpha}_{1,0}$ equal to 0.59 and R^2 = 99.2%).

Figure 3.6- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}^*_{1,0}$ to the ART on white women (original ML and bootstrap $\hat{\alpha}^*_{1,0}$ equal to 0.41 and R^2 = 99.2%).

Figure 3.7- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}_{1,1}$ for the Stouffer and Toby data (original ML $\hat{\alpha}_{1,1}$ = 1.15, bootstrap $\hat{\alpha}_{1,1}$ = 1.19 and R^2 = 93.5%).

Figure 3.8- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}_{1,0}$ for the Stouffer and Toby data (original ML $\hat{\alpha}_{1,0}$ = 1.66, bootstrap $\hat{\alpha}_{1,0}$ = 1.70 and R^2 = 94.3%).

Figure 3.9- Normal probability plotting of the bootstrap parameter estimate $\hat{\pi}_1$ for the Stouffer and Toby data (original ML and bootstrap $\hat{\pi}_1$ = 0.84 and R^2 = 99.2%).

Figure 3.10- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}^*_{1,0}$ for the Stouffer and Toby data (original ML $\hat{\alpha}^*_{1,0}$ = 1.09, bootstrap $\hat{\alpha}^*_{1,0}$ = 1.10 and R^2 = 98.0%).

Figure 3.11- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}_{4,1}$ for the Stouffer and Toby data (original ML $\hat{\alpha}_{4,1}$ = 2.10, bootstrap $\hat{\alpha}_{4,1}$ = 2.72 and R^2 = 35.3%).

Figure 3.12- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}_{4,0}$ for the Stouffer and Toby data (original ML $\hat{\alpha}_{4,0}$ = -1.33, bootstrap $\hat{\alpha}_{4,0}$ = -1.53 and R^2 = 55.0%).

Figure 3.13- Normal probability plotting of the bootstrap parameter estimate $\hat{\pi}_4$ for the Stouffer and Toby data (original ML $\hat{\pi}_4$ = 0.21, bootstrap $\hat{\pi}_4$ = 0.18 and R^2 = 94.4%).

Figure 3.14- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}^*_{4,0}$ for the Stouffer and Toby data (original ML $\hat{\alpha}^*_{4,0}$ = -0.57, bootstrap $\hat{\alpha}^*_{4,0}$ = -0.56 and R^2 = 99.0%).

Figure 3.15- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}_{2,1}$ for the Lombard and Doering data (original ML $\hat{\alpha}_{2,1}$ = 3.40, bootstrap $\hat{\alpha}_{2,1}$ = 4.14 and R^2 = 67.4%).

Figure 3.16- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}_{2,0}$ for the Lombard and Doering data (original ML $\hat{\alpha}_{2,0} = 0.60$, bootstrap $\hat{\alpha}_{2,0} = 0.74$ and $R^2 = 66.4\%$).

Figure 3.17- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}^*_{2,0}$ for the Lombard and Doering data (original ML and bootstrap $\hat{\alpha}^*_{2,0} = 0.17$ and $R^2 = 98.8\%$).

Figure 3.18- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}_{4,1}$ for the Lombard and Doering data (original ML $\hat{\alpha}_{4,1} = 0.77$, bootstrap $\hat{\alpha}_{4,1} = 0.82$ and $R^2 = 99.2\%$).

Figure 3.19- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}_{4,0}$ for the Lombard and Doering data (original ML and bootstrap $\hat{\alpha}_{4,0}$ equal to -2.75 and $R^2 = 99.4\%$).

Figure 3.20- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}^*_{4,0}$ for the Lombard and Doering data (original ML $\hat{\alpha}^*_{4,0} = -2.18$, bootstrap $\hat{\alpha}^*_{4,0} = -2.12$ and $R^2 = 98.1\%$).

Figure 3.21- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}_{1,1}$ to the ART on black women (original ML $\hat{\alpha}_{1,1} = 14.39$, bootstrap $\hat{\alpha}_{1,1} = 6.79$ and $R^2 = 83.8\%$).

Figure 3.22- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}_{1,0}$ to the ART on black women (original ML $\hat{\alpha}_{1,0} = 0.24$, bootstrap $\hat{\alpha}_{1,0} = 0.01$ and $R^2 = 93.2\%$).

Figure 3.23- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}^*_{1,0}$ to the ART on black women (original ML $\hat{\alpha}^*_{1,0} = 0.02$, bootstrap $\hat{\alpha}^*_{1,0} = 0.01$ and $R^2 = 97.1\%$).

Figure 3.24- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}_{4,1}$ to the ART on black women (original ML $\hat{\alpha}_{4,1} = 0.19$, bootstrap $\hat{\alpha}_{4,1} = 0.14$ and $R^2 = 44.0\%$).

Figure 3.25- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}_{4,0}$ to the ART on black women (original ML $\hat{\alpha}_{4,0} = -1.08$, bootstrap $\hat{\alpha}_{4,0} = -1.56$ and $R^2 = 36.7\%$).

Figure 3.26- Normal probability plotting of the bootstrap parameter estimate $\hat{\pi}_4$ to the ART on black women (original ML and bootstrap $\hat{\pi}_4$ equal to 0.23 and $R^2 = 77.0\%$).

Figure 3.27- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}^*_{4,0}$ to the ART on black women (original ML $\hat{\alpha}^*_{4,0}$ = -1.06, bootstrap $\hat{\alpha}^*_{4,0}$ = -1.04 and R^2 = 99.0%).

Figure 4.1- Loglikelihood values as a function of the parameter estimates $\hat{\alpha}_1$ and $\hat{\alpha}_{1,0}$ from fitting the Rasch model to the ART on black women.

Figure 4.2- Maximum likelihood values over $\hat{\alpha}_{1,0}$ for a fixed $\hat{\alpha}_1$ from fitting the Rasch model to the ART on black women.

Figure 4.3- Comparison between $\hat{\pi}_1$'s from fitting a logit/probit and the Rasch model to 30 normal bootstrap samples from the ART on black women.

Figure 4.4- Comparison between $\hat{\pi}_2$'s from fitting a logit/probit and the Rasch model to 30 normal bootstrap samples from the ART on black women.

Figure 4.5- Comparison between $\hat{\pi}_3$'s from fitting a logit/probit and the Rasch model to 30 normal bootstrap samples from the ART on black women.

Figure 4.6- Comparison between $\hat{\pi}_4$'s from fitting a logit/probit and the Rasch model to 30 normal bootstrap samples from the ART on black women.

Figure 7.1- Relation between $E(Z|\mathbf{x})$ and $\sum \alpha_{i,1}x_i$ when fitting a logit/probit model to the Law School Admission Test, section 6.

Figure 7.2- Posterior densities $h(z|\mathbf{x})$ when fitting a logit/probit model to the LSAT VI, for the score patterns '00000', '01000', '00101', '01101', '10111' and '11111'.

Figure 7.3- Relation between $E(Z|\mathbf{x})$ and $\sum \alpha_{i,1}x_i$ when fitting a logit/probit model to Test 11A (Ireland, items 1 to 6, 8 to 10).

Figure 7.4- Relation between $E(Z|\mathbf{x})$ and $\sum \alpha_{i,1}x_i$ when fitting a logit/probit model to Test 11A (Ireland, items 1 to 10).

Figure 7.5- Relation between $E(Z|\mathbf{x})$ and $\sum \alpha_{i,1}x_i$ when fitting a logit/probit model to Test 12.

Figure 7.6- Relation between $E(Z|\mathbf{x})$ and $\sum \alpha_{i,1}x_i$, assuming $\alpha_{15,1}$ and $\alpha_{16,1}$ equal to infinity, when fitting a logit/probit model to Test 12.

Figure 7.7- Relation between $E(Z|x)$ and $\sum \alpha_{i,1} x_i$ when fitting a logit/probit model to Test 13.

Figure 7.8- Posterior densities $h(z|x)$ for the first ten different score patterns of Test 12, for which $-2.26 \leqslant E(Z|x) \leqslant -1.67$).

Figure 7.9- Posterior densities $h(z|x)$ for some score patterns of Test 12, for which $-0.81 \leqslant E(Z|x) \leqslant -0.66$).

Figure 7.10- Posterior densities $h(z|x)$ for the last ten different score patterns of Test 12, for which $1.39 \leqslant E(Z|x) \leqslant 1.90$).

Figure 7.11- Representative collection of posterior densities $h(z|x)$ for the observed score patterns of Test 12.

Figure 7.12- Representative collection of posterior densities $h(z|x)$ for the observed score patterns of Test 13.

Chapter 1

LATENT TRAIT MODELS FOR BINARY RESPONSES

1- *Introduction*

In this thesis variables will be either categorical or metrical. Categorical variables are measured in a nominal or ordinal scale while metrical variables assume values in an interval or ratio scale (discrete or continuous).

Whether categorical or metrical a variable will be either manifest (directly observable) or latent (not directly observable, and generally called a factor in factor analysis).

Bartholomew (1987) has classified latent variable models according to the type of latent and manifest variables: factor analysis (both variables are metrical), latent trait analysis and factor analysis of categorical data (latent variables are metrical and manifest variables are categorical), latent structure analysis (latent variables are categorical) divided into latent profile analysis (metrical manifest variables) and latent class analysis (categorical manifest variables). Bartholomew discusses these models from a new point of view, starting from a general model(1.5) that allows these techniques and some ones to emerge as special cases.

When discussing factor analysis for categorical data and latent trait analysis, two approaches are considered for the construction of the models: the ***Underlying Variable*** (UV) approach used in the factor analysis tradition, where the categorical manifest variables are supposedly produced by underlying continuous variables, and the *Response Function* (RF) approach with its origin in the theory of educational testing and developed further in Bartholomew (1980,1981). It starts with a response function giving the probability of a positive response for an individual with variable value y.

He shows that the two models can be equivalent for binary variables, but not for polytomous variables. He also points out that when they are equivalent the choice between them depends on taste.

In this thesis we shall use the Response Function approach to latent trait analysis, for which follows a review of the literature.

2- *Definition of the Model*

2.1- Notation and Assumptions

We shall consider the case when the manifest variables are binary and the latent variables are metrical. This situation happens, for example, in a survey where individuals are asked to answer questions by yes or no, agree or disagree, or in educational testing where the students may answer an item in a test right or wrong. Usually the two possible outcomes are coded as 1 (positive), otherwise 0(negative).

Thus if the test has p items and is answered by n individuals then the data matrix will be an (n x p) array of zeros and ones. We shall

refer to any row of the data matrix as a **response or score pattern**, which is the set of responses of a given individual. Therefore there are 2P different possible response patterns, which number increases quickly with p so that some patterns will probably not appear in the sample. For practical purposes in the samples we shall list only those response patterns which occur at least once.

Notation

Let $X=(X_1,X_2,\ldots,X_n)$ be a vector of p manifest variables, where X_i is equal to 1 or 0 for all i, and $Y=(Y_1,Y_2,\ldots,Y_p)'$ a vector of latent variables. Then the joint distribution of the X's is given by

$$f(x) = \int_{R_y} h(y)\, g(x|y)\, dy \qquad (1.1)$$

where

R_y the range space of y,

h(.) is the prior density of y,

g(.|y) is the conditional density of x given y.

Our main interest is what we can know about Y after X has been observed. This comes from the conditional density

$$h(y|x) = h(y)\, g(x|y) \,/\, f(x),$$

which depends on our knowledge about h, g and f.

Obviously f(x) is the only density function about which inferences can be directly made, and therefore all the information we can get

about g and h comes from knowledge of f. It follows that they are not uniquely determined (Bartholomew,1980). As we cannot obtain a complete specification of $h(y|x)$, we need to make some restrictions on the class of functions to be considered.

The assumption of *conditional independence*

$$g(x|y) = \prod_{i=1}^{p} g_i(x_i|y) \qquad (1.2)$$

is usually considered necessary for effective theoretical work with response functions. It is almost the definition of the concept of underlying factor in factor analysis. For it means that the association between X's is wholly explained by their dependence on the Y's. Consequently, if Y is held fixed there will be no correlation between X's. This assumption cannot be tested empirically, since it is part at the definition of Y. We will come back to this point later on.

Conditional independence for g means that the set of latent variable is complete, i.e, Y is sufficient to explain the dependence between the X's.

As the X's are binary,

$$g_i(x_i|y) = [\pi_i(y)]^{x_i} [1-\pi_i(y)]^{1-x_i} \quad i=1,\ldots,p \qquad (1.3)$$

where $\pi_i(y)=P[X_i=1|y]$ is called ***response function*** by Bartholomew (1980). In educational testing, where most of the models have been developed for a one-dimensional latent variable representing an ability of some kind, $\pi_i(y)$ is called ***item characteristic curve*** (ICC) or *item response function* (IRF).

Consequently from (1.2) and (1.3), the joint density function of the x's (1.1) can be written as

$$f(x) = \int_{R_y} h(y) \prod_{i=1}^{p} \left\{ [\pi_i(y)]^{x_i} [1-\pi_i(y)]^{1-x_i} \right\} dy. \qquad (1.4)$$

2.2- Response Function

Many suggestions about the shape of response functions and prior distribution of the latent variables have been made over the years. These have given rise to different models.

We shall present some of these models, starting from a general model (1.5) and deriving them as special cases.

The choice of a suitable response function was discussed by Bartholomew (1980), where he gave a set of properties that a family of response functions is desirable to possess. For instance, he says that the response function should be monotonic nondecreasing in the latent space, a property also implied by the normal ogive and logistic models, as we will see later. This implies, for example, that the probability of a correct response increases with ability (educational testing). Bartholomew also proposed a class of linear models with response functions satisfying:

$$G^{-1}(\pi_i(y)) = \alpha_{i,0} + \sum_{j=1}^{q} \alpha_{i,j} H^{-1}(y_j), \quad i=1,\ldots,p \qquad (1.5)$$

where

$\pi_i(y)$ is the response function,

y_j (j=1,2,...,q) are independently and uniformly distributed on (0,1) and functions G and H are distribution functions of random variables symmetrically distributed about zero.

In practice he limits the choice of G^{-1} and H^{-1} to the commonly used functions, the logit {logit(v)=log[v/(1-v)]} and the probit [probit(v)=Φ^{-1}(v), where Φ is the standard normal distribution]. Considering these functions, the following models can be derived:

Logit when both G^{-1} and H^{-1} are logit functions

Probit when both G^{-1} and H^{-1} are probit functions and

Logit/Probit when G^{-1} is a logit and H^{-1} is a probit function.

The logistic and normal distributions are very similar in shape, and the choice between them is without practical importance so that from one model we can obtain the approximately corresponding estimates of the parameters for the other since

$$\text{logit}(v) \cong \pi/\sqrt{3}\ \Phi^{-1}(v)$$

Hence the logit/probit model

$$\text{logit}[\pi_i(y)] = \alpha_{i,0} + \sum_{j=1}^{q} \alpha_{i,j}\ \Phi^{-1}(y_j)$$

is approximately the same as the logit model, i.e,

$$\text{logit}[\pi_i(y)] = \alpha_{i,0} + \sum_{j=1}^{q} (\sqrt{3}/\pi)\ \alpha_{i,j}\ \Phi^{-1}(y_j).$$

Thus if we fit the logit/logit model we would expect the slope parameter $\alpha_{i,j}$ to be $\sqrt{3}/\pi$ times what we would have obtained from the logit/probit model.

Similarly, the probit model

$$\Phi^{-1}[\pi_i(y)] = \alpha_{i,0} + \sum_{j=1}^{q} \alpha_{i,j}\, \Phi^{-1}(y_j)$$

is approximately the same as the logit/probit model

$$\text{logit}[\pi_i(y)] = \pi/\sqrt{3}\, \left[\, \alpha_{i,0} + \sum_{j=1}^{q} \alpha_{i,j}\, \Phi^{-1}(y_j)\right]$$

In this case both parameters $\{\alpha_{i,0}\}$ and $\{\alpha_{i,j}\}$ have to be multiplied by $\pi/\sqrt{3}$.

By transformation of $H^{-1}(y) = z$, Bartholomew (1987, Chapter 5) has proved that logit/probit model can be written in terms of normally distributed variables, as

$$\text{logit}[\pi_i(z)] = \alpha_{i,0} + \sum_{j=1}^{q} \alpha_{i,j}\, z_j \qquad i=1,\ldots,p \qquad (1.6)$$

Several response functions presented in the literature are special cases of the general model (1.5).

We shall give the logistic and normal ogive models as they are usually presented in the literature, following by showing how they can be written as special cases of model (1.5).

2.2.1- Normal Ogive or Probit model

Lawley (1943) introduced a response function called the normal ogive model (Lord and Novick, 1968, p365) specified by

$$\pi_i(y) = \Phi(a_i\ (y - b_i)) \qquad\qquad i=1,\ldots,p$$

where

Φ is the cumulative distribution function of the standard normal distribution,

y, for q=1, is the latent ability parameter normally distributed with mean μ and variance σ^2, which characterizes the individuals and

a_i and b_i are parameters characterizing the item, called discriminating power and difficulty of item i.

Furthermore, it is assumed that $a_i \geq 0$, which means that $\pi_i(y)$ is a nondecreasing function of y.

The normal ogive model is equivalent to the probit model

$$\pi_i(y) = \Phi(\alpha_{i,0} + \alpha_{i,1}\ z),$$

where

$$b_i = -\alpha_{i,0} / \alpha_{i,1} \quad \text{and} \quad a_i = \alpha_{i,1}.$$

As the location and scale of the parameter estimates depend on the mean and variance of the distribution of the latent variable (or ability parameter), the equivalence between the parameter estimates is done by scaling. Bock and Aitkin (1981), for example, suggest that

$$\sum_{i=1}^{p} b_i = \sum_{i=1}^{p} \frac{-\alpha_{i,0}}{\alpha_{i,1}} = 0$$

and

$$\prod_{i=1}^{p} a_i = \prod_{i=1}^{p} \alpha_{i,1} = 1.$$

This model is the basis for numerous developments in psychometric theory, see for example, Lord (1952), Lord and Novick (1968), Bock and Lieberman (1970), Samejima (1974). Bock and Aitkin (1981) give also an extension for more than one latent variable for binary response.

2.2.2- Logistic or Logit/Probit Model

Two-parameter logistic model

Birnbaum (Lord and Novick (1968, Chapter 17) gave the two-parameter logistic model determined by assuming that the response function has the form of a logistic cumulative distribution function

$$\pi_i(y) = \frac{\exp[-da_i(y - b_i)]}{1 + \exp[-da_i(y - b_i)]}$$

where

y, is a latent ability parameter normally distributed with mean μ and variance σ^2,

a_i and b_i have the same meaning as in the normal ogive model and

d is a number that serves, at our convenience, as a unit scaling factor, with a value 1.7 corresponding to the maximum agreement between normal and logistic distributions.

The equivalence between the two-parameter logistic and the logit/probit model, i.e,

$$\pi_i(y) = \frac{\exp(\alpha_{i,0} + \alpha_{i,1} z)}{1 + \exp(\alpha_{i,0} + \alpha_{i,1} z)}$$

may be seen by taking

$$d = -1, \quad b_i = -\alpha_{i,0}/\alpha_{i,1} \quad \text{and} \quad a_i = \alpha_{i,1}.$$

As for the normal ogive model, $a_i = \alpha_{i,1}$ and the possible different mean and variance for the normal distribution of Y is corrected by scaling.

Lord and Novick (1968b, Chapter 17) estimate the parameters a_i and b_i assuming that Y is N(0,1).

Rasch model

A random effect form of the model due to Rasch (1960), is a simplified form of the two-parameter logistic model with

$$\pi_i(y) = \frac{\exp(y - b_i)}{1 + \exp(y - b_i)} .$$

Here all the item discriminating powers are equal to 1, i.e, a_i=1, i=1,...,p. Thus $\pi_i(y)$ depends only on the distance between the latent value y and the item difficulty b_i and as the value of b_i increases fewer individuals will be likely to answer correctly or positively item i.

Therefore the Rasch model is equivalent to the logit/probit model when $\alpha_{i,1}=1$ and $b_i=-\alpha_{i,0}$ for all manifest variables or items i. The equivalence between the parameter estimates is obtained by scaling if a standard normal distribution is not assumed for Y.

The main advantage of this simplification is the existence of a sufficient statistic for Y, the total number of positive responses of an individual (or the total raw score of the data matrix).

There are many papers on the Rasch model and its extensions, among them Andersen (1970,1972,1973b), Gustafsson (1980a,1980b), Fischer (1981), Molenaar (1983), Thissen (1982) and many others to which we will give references later on.

Three-parameter logistic model

If in addition it is assumed that if an individual does not know the answer he will guess, and with probability c_i will guess positively then according to Lord and Novick (1968b, Chapter 17), the response function for the three-parameter logistic model is given by

$$\pi_i(y) = c_i + \frac{(1-c_i)\exp[da_i(y-b_i)]}{1+\exp[da_i(y-b_i)]}$$

Then the two-parameter logistic model is a particular case of this model, when $c_i=0$ for all i.

The three-parameter logistic model cannot be written in the general form (1.5), since that does not have guessing parameters.

This model has been applied by, for example, Lord (1968a,1983a), Hullin, Lissak and Drasgow (1982), Lord and Wingersky (1985), Thissen and Wainer (1982).

Since the normal ogive is equivalent to the probit model and the two-parameter logistic is equivalent to the logit/probit model, we shall use both names to refer to the same model, although we shall generally use the notation following the general model(1.5) and consider Y as a latent variable.

2.2.3- Properties of the Response Function

Let us consider a logit/probit model, though the same approach is valid also for the logit and probit models.

The two most important properties which response functions produce are:

(1) The choice of which the two possible outcomes is to be regarded as positive is totally arbitrary. If the positive answer has probability $\pi_i(z)$ then the negative has probability $1-\pi_i(z)$ which are given by

$$\pi_i(z) = \frac{\exp\left(\alpha_{i,0} + \sum_{j=1}^{q} \alpha_{i,j}\, z_j\right)}{1 + \exp\left(\alpha_{i,0} + \sum_{j=1}^{q} \alpha_{i,j}\, z_j\right)}$$

$$= \left\{ 1 + \exp\left(-\alpha_{i,0} - \sum_{j=1}^{q} \alpha_{i,j}\, z_j\right)\right\}^{-1}$$

and

$$1-\pi_i(z) = \left\{ 1 + \exp\left[\alpha_{i,0} + \sum_{j=1}^{q} \alpha_{i,j}\, z_j \right] \right\}^{-1}$$

This means that increasing any **z**, increases the probability of a positive response and decreases, as expected, the probability of a negative response by the same amount. Thus, when **q=1**, it is possible to obtain all $\alpha_{i,1}$'s positive or zero by suitable choice of which outcome is to be considered as positive.

(2) The direction in which most latent variables are measured is arbitrary. Changing the direction of measurement involves replacing z_j by $-z_j$ in (1.6). This is equivalent to changing the sign of the corresponding $\alpha_{i,j}$ without changing the model.

2.3- Interpretation of the Parameters

The parameters of the logit/probit model may be interpreted in several ways.

The coefficient $\alpha_{i,0}$ is the value of logit $\pi_i(z)$ at **z=0** and thus π_i is the probability of a positive response from a median individual. In the context of educational testing, $\alpha_{i,0}$ or π_i would be called the item difficulty.

The coefficients $\alpha_{i,j}$ may be interpreted in three related ways.

First, as a measure of the extent to which Z_j discriminates between individuals. For two individuals a given distance apart on the Z_j-scale, the bigger the absolute value of $\alpha_{i,j}$ the greater the difference in their probabilities of given a positive response to

item i and thus easier to discriminate between them in relation to item i. Therefore $\alpha_{i,j}$ is a parameter that indicates the value of an item in the sense of the amount of information that the item provides about Z_j. In educational testing, this is the interpretation usually adopted, and $\alpha_{i,j}$ is called item discriminating power.

A second interpretation of the $\alpha_{i,j}$ is by analogy with linear factor analysis or principal components, where the $\alpha_{i,j}$'s are equivalent to the loadings. They are the weights of the x_i's in the determination of the component scores X_j's, i.e, $X_j = \sum_{i=1}^{p} \alpha_{i,j} x_i$.

Finally the $\alpha_{i,j}$ are related to correspondence analysis, where they are equivalent to the category scores. This is done by attributing the value of $\alpha_{i,j}$ to a positive response on manifest variable j and zero to a negative response. Then for each latent variable Z_j the data matrix constituted by 1 and 0 is replaced by $\alpha_{i,j}$ and 0. The individual score is thus the sum of the category scores for that latent variable Z_j.

For the general model (1.5), considering Y either as a parameter or a variable has given rise to different procedures when looking for more information about Y, after the model has been fitted.

In educational testing, where Y is usually treated as a parameter, some work has been done in estimating the parameters of the latent distribution function; see for example, Lord (1983b), Andersen and Madsen (1977), Samanthanan and Blumenthal (1978) and Mislevy (1984).

Bartholomew (1984), treating Y as a latent variable has deal with the situation by scaling the latent variable, i.e, locating the

individuals in the Y-space on the basis of their observed values of X.

Since in this thesis we are treating Y as a latent variable, we look at the scaling, instead of the estimation of the parameters of the latent distribution.

2.4- Scaling a Latent Variable

According to Bartholomew (1984) scaling a latent variable should be done via the posterior distribution of y given **x**, and he suggested the mean $E(Y|\mathbf{x})$. Since the prior distribution of Y is uniform on (0,1), this measure may be interpreted as the expected proportion of the population lying below an individual with that value of **x**. The practical advantage is that when q=1, $E(Y|\mathbf{x})$ is approximately a linear function of the quantity $X = \sum_{i=1}^{p} \alpha_{i,1} x_i$ if all $\alpha_{i,1}$'s are small for G^{-1}=logit, regardless of the form of H. However, if the G^{-1} is the probit function this relation does not work though the similarity of the logit and probit models should ensure that the linear form is still a good approximation.

Bartholomew (1984) shows that an approximation can be obtained doing

$$E(Y|\mathbf{x}) \cong (1 + X)/(2 + A)$$

where

$$X = \sum_{i=1}^{p} \alpha_{i,1}\, x_i \quad \text{and} \quad A = \sum_{i=1}^{p} \alpha_{i,1}.$$

This result is exact if π_i=0.5 and $\alpha_{i,1}$=1 for all i.

He also points out that even when the approximation is not good, then E(Y|x) still provides the correct ranking of individuals. Obviously, X and E(Y|x) are almost equivalent for purposes of scaling, since both give the same ranking on the latent scale. This result does depend on the choice of a uniform prior density for Y.

If the $\alpha_{i,1}$'s are very similar then the ranking determined by $X = \sum_{i=1}^{p} \alpha_{i,1} x_i$ and $\sum_{i=1}^{p} x_i$ are likely to be the same whichever latent models (logit/probit, logit or probit) we are using.When this situation happens the convergence of the algorithm for estimation of the parameters (section 3.2) is obtained quicker than when at least one of the estimates differs from the other.

The definition of X implies that, we may interpret the $\alpha_{i,1}$ as item discriminating power, and thus the item with larger $\alpha_{i,1}$ will carry more weight in the determination of X.

We shall come back to the scaling of a latent variable in Chapter 7, in which we present some new results.

3- *Maximum Likelihood Estimation*

3.1- Estimation Procedures: joint, conditional and marginal likelihood functions

In the literature we have found that the parameter of latent models for binary data are estimated essentially through 3 different procedures: joint maximum likelihood (ML), conditional ML and marginal ML.

As we have already pointed out when describing different shapes of response functions for the general latent model(1.5), Y is usually defined in the literature as a person parameter rather than a latent variable, as used in the context of this thesis. However we shall refer to Y as a person or ability parameter, if necessary, when reporting research using that approach.

Joint Maximum Likelihood

A joint maximum likelihood estimation was proposed by Birnbaum (1968) for the two- and three-parameter logistic model, and for the Rasch model by Wright and Panchapakesan (1969), among others. In this approach, person abilities and item parameters (discrimination and difficulty) are estimated simultaneously so that the procedure is not conditioned on the ability parameters.

The joint ML estimation of the person and item parameters is not generally possible because the number of parameters increases with the sample size and thus standard limit theorems do not apply. Several researchers, including Wood, Wingerkly and Lord (1976) have avoided

this problem by assuming that respondents who have the same score pattern, or same number of positive responses or who have been assigned provisionally to homogeneous ability groups, have the same ability. On this assumption, the number of parameters is finite and standard asymptotic theorems apply.

The assumption that abilities are fixed in size, when in fact they are not identifiable and have a distribution in the population of persons, is difficult to justify statistically. A better approach to estimation in the presence of a random nuisance parameter (person ability) is that of integrating over the parameter distribution and estimating the item parameters by maximum likelihood in the marginal distribution, which is done when using marginal ML procedure.

Lord (1983a) derives asymptotic formulas for the statistical bias in the joint ML estimation of the parameters for the three-parameter logistic model. The derivation deals with a single fixed manifest variable and assumes the single latent variable as a known parameter.

In order to investigate the characteristics of the asymptotic biases, Lord used simulated data having parameter values roughly equal to the estimates yielded by 2995 respondents on a Verbal Scholastic Aptitude Test of length 90. The results showed that, in general, if the parameter estimates had large standard deviations the biases were also large. However, the magnitude of the bias of an estimator was typically about 0.1 of and seldom greater than 0.2 of its standard deviation. Lord concluded that because the standard deviations are inversely proportional to the sample size, when the later is large the numerical value of the biases are probably negligible.

Van den Wollenberg, Wierda and Jansen (1988) have shown through simulation studies that the joint ML estimation procedure for the Rasch model gives rise to biased estimators. This bias cannot be removed by a correction factor (p-1)/p (where p is the number of items). The bias is dependent not only on the number of items, but also on the distribution of the item parameters, which makes correction for bias practically impossible. They concluded that the joint ML method is not a good alternative to the conditional ML method, at least when small number of items are involved. However when the number of items becomes large, the bias becomes relatively small and a correction is no longer needed. In that case the joint ML could be used as a fast alternative to the conditional ML estimation procedure.

Baker (1988) reviews the ML estimation procedures for the one-, two- and three-logistic models.

Conditional Maximum Likelihood (CML)

Rasch (1960) showed that under his probabilistic model the 'item totals' (number of positive responses given by every person) and the 'row scores' (number of positive responses given to every item) are sufficient statistics for the person and difficulty parameters. Using Rasch results as a starting point, Andersen (1970, 1972, 1973a) developed a conditional ML procedure to estimate the difficulty parameters that did not involve the latent individual parameters. The difficulty parameter estimates are obtained from the likelihood function conditioned upon the item total scores.

Wright and Douglas (1977) have shown that the conditional ML estimation is inaccurate when a test has more than 10 or 15 items due to round-off-error. They proposed a simplified alternative procedure for conditional estimation, which is limited to 20 or 30 items due to the same precision problem, especially in the presence of extreme difficulty parameter estimates.

In order to compare the joint ML and conditional ML for tests with more than 20 items, Wright and Douglas carried out a simulation study based on 15 replications of 500 individuals each for tests of size 20 and 40. They assumed that the ability was normally distributed with mean 0, 1 and 2, and the difficulty parameters were generated from a normal distribution with mean zero. The comparison between both procedures was done in terms of MAX DIFF (maximum difference between a generated difficulty parameter and the mean over the 15 replications of its estimates), RMS (root mean square of these differences over items) and the MEAN ABS (mean of the absolute value of these differences over items). They found out that in terms of RMS and MEAN ABS both estimation procedures, conditional ML and joint ML, give approximately the same results, while the MAX DIFF's tend to increase for both algorithms, but strongly for conditional, when the mean of the sample shifts away from zero (equal 1 or 2). This later result was found to be due to the increasing discrepancy between item and sample characteristics, which made estimation difficult for the conditional ML because of accumulated round-off-error.

Marginal Maximum Likelihood (MML)

Thissen (1982) developed marginal ML procedures for the Rasch model making use of the fact that all response patterns which have the same number of positive responses have proportional likelihoods for the single latent variable. Unlike the conditional solution (CML), this estimation procedure is not conditional on the sufficient statistic for the person parameter and requires specification of the prior distribution for the person parameters.

The formulation of the model explicitly includes the item discriminating power common to all items and it is assumed that the latent ability is distributed as N(0,1).

Two algorithms have been described by Thissen (1982) for MML estimation:

(a) A gradient solution, following Bock and Lieberman (1970), where the parameters are estimated by maximum likelihood and

(b) An alternative solution, following the algorithm described by Bock and Aitkin (1981), uses Gauss-Hermite quadrature points for the N(0,1) prior distribution for latent ability (person parameters).

They also show that the MML procedure is similar to a combination of CML of the item parameters with estimation of the mean and variance of the population distribution as described by Andersen and Madsen (1977). The mean of the item difficulty parameters is essentially equivalent to Andersen and Madsen's population mean and the estimated discrimination parameter is the same as the standard deviation of the population distribution (normal) for conventionally

standardized CML estimates. For this procedure the population distribution is not required to be normal, but must have finite mean and variance.

Tsutakawa (1984) derived a MML procedure employing the two-parameter logistic model. His method differs from Bock and Aitkin's method in the manner in which the prior distribution of the latent variable is handled, but it is equivalent for the special case of a discrete empirical prior. He analysed a 50-item arthritis knowledge test administered to 162 individuals, using both the joint and the marginal ML procedures. After appropriate scaling to take metric differences into account, the values of the discrimination and difficulty parameters yielded by the two methods were very similar.

Tsutakawa also used simulated data to evaluate the parameter recovery capability of the two procedures. This investigation involved two hundred simulated respondents having a unit normal distribution and a 50-item test with representative values of the item parameters. The estimated item parameters were plotted against tha underlying parameter values. The plots showed a close agreement between the two methods as well as a general 45^0 line relating the estimates and the parameters. The scatter of the item discriminating power about the line was much greater than of the item difficulty.

It follows a description of the ML procedure used in this thesis to estimate the parameters of the general model(1.5).

3.2- Marginal Maximum Likelihood Estimation

For any model of the family (1.5) the joint probability function of $x_1,\ldots,x_p$ is

$$f(x) = \int_{Ry} \prod_{i=1}^{p} [\pi_i(y)]^{x_i} [1-\pi_i(y)]^{1-x_i} h(y)dy \qquad (1.7)$$

If x_s is the observed response vector for the s^{th} sample member then the loglikelihood is

$$L = \sum_{s=1}^{n} \log f(x_s) \qquad (1.8)$$

Bock and Lieberman (1970) maximised this function with the normal ogive model for the response function and for one latent variable distributed as N(0,1), i.e, using the probit model. The likelihood equations were solved iteratively by a Newton-Raphson method and Gauss-Hermite quadrature was employed to perform the necessary integrations. Due to the heavy numerical integration the method was considered to be limited to one latent variable and not more than 10 manifest variables.

Bock and Aitkin (1981) by a simple transformation of the Bock and Lieberman (1970) likelihood equations, found a computational solution so that the method could be applied for more than one latent variable and a large number of manifest variables. This reformulation is related to the E-M algorithm for maximum likelihood estimation as discussed by Dempster, Laird and Rubin (1977).

We shall give the main results of this method as described by Bartholomew (1987, Chapter 6).

3.2.1- An E-M Algorithm

We shall consider the logit/probit model for one latent variable expressed by Z as defined in (1.6), i.e,

$$\text{logit}(\pi_i(z)) = \alpha_{i,0} + \alpha_{i,1}\, z$$

Each iteration of this E-M algorithm involves two steps called the expectation step (E) and the maximization step (M) and the method starts with arbitrary values for the parameters.

E-step: Using the current values for $\{\alpha_{i0}\}$ and $\{\alpha_{i1}\}$, predict z_s for $s=1,2,\ldots,n$, through

$$E(Z_s|x_s) = \int_{-\infty}^{\infty} z_s \prod_{i=1}^{p} [\pi_i(z_s)]^{x_i} [1-\pi_i(z_s)]^{1-x_i} h(z_s)\, dz_s \;/\; f(x)$$

The value of $E(Z_s|x_s)$ has to be found by numerical integration.

M-step: Treating the expected values $E(Z_s|x_s)$, $s=1,2,\ldots,n$, as if they were true values z_s, estimate the parameters $\{\alpha_{i0}\}$ and $\{\alpha_{i1}\}$ by maximum likelihood, as follows:

Let the conditional loglikelihood defined by

$$L = \sum_{s=1}^{n} \sum_{i=1}^{p} \left\{ x_{is} [\log\pi_i(z_s)] + (1-x_{is}) [\log(1-\pi_i(z_s)] \right\}$$

$$- \sum_{s=1}^{n} \sum_{i=1}^{p} \left\{ x_{is} \text{ logit}[\pi_i(z_s)] + \log(1-\pi_i(z_s)) \right\}, \qquad (1.9)$$

where logit $\pi_i(z_s) = \alpha_{i,0} + \alpha_{i,1} z_s$.

Then the partial derivatives with respect to the parameters $\{\alpha_{i,0}\}$ and $\{\alpha_{i,1}\}$ are

$$\frac{\partial L}{\partial \alpha_{i,0}} = \sum_{s=1}^{n} [x_{is} - \pi_i(z_s)]$$

$$\frac{\partial L}{\partial \alpha_{i,1}} = \sum_{s=1}^{n} z_s [x_{is} - \pi_i(z_s)] \qquad \text{for } i=1,2,\ldots,p. \qquad (1.10)$$

Thus estimating equations are obtained setting (1.10) equal to zero and for each variable i there is a pair of non-linear equations which can be solved for $\alpha_{i,0}$ and $\alpha_{i,1}$. Methods of solving these equations are reviewed by McFadden (1982).

Having completed the M-step, the E-step is done again, and the cycle is repeated until the estimates become stable, according to some criterion.

Bock and Aitkin (1981) reported that the convergence of the E-M algorithm is only geometric and slows up as the solution point is approached. They suggested using the acceleration technique of Ramsey (1975) to speed convergence.

The convergence properties of the E-M algorithm has been studied analytically by Wu (1983). He showed that if the likelihood function

is unimodal and certain differentiability conditions are satisfied, any E-M sequence converges to the unique ML estimates of the parameters.

More than one latent variable

If there is more than one latent variable, the term $\alpha_{i,1}\, z_s$ in (1.9) is replaced by $\sum_{j=1}^{q} \alpha_{i,j}\, z_{j,s}$; q equations replace the second member of (1.10)– one for each $\{\alpha_{i,j}\}$ – and $\pi_i(z_s)$ becomes $\pi_i(\mathbf{z})$. In this case, the determination of $\hat{\alpha}_{i,0}$ and $\hat{\alpha}_{i,j}$, for $j=1,2,\ldots,q$, involves the solution of q+1 simultaneous non-linear equations for each i.

In order to get unique solutions, when q>1, we must impose some constraints. One possibility is to fix the values of enough α's to ensure a unique solution. For example, it is sufficient to fix $\alpha_{i,1}=0$ for some i, when q=2.

3.2.2- A Variation of the E-M Algorithm

A variation of the E-M algorithm was proposed by Bock and Aitkin (1981) also for the probit model. Bartholomew (1987, Chapter 6) discusses the same variation from a rather different perspective setting G^{-1} in (1.5) as the logit function. It follows the main results for one latent variable.

Even though the latent variable Z is distributed as N(0,1), it is proposed as an approximation that Z assumes values $z_1, z_2, \ldots, z_k$ with probabilities $h(z_1), h(z_2), \ldots, h(z_k)$ chosen so that the joint probability function

$$f(x_s) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} g(x_s|z)\, h(z)\, dz \qquad s=1,2,\ldots,n$$

can be approximated with high accuracy by Gauss-Hermite quadrature, i.e,

$$f(x_s) = \sum_{t=1}^{k} g(x_s|z_t)\, h(z_t) \qquad s=1,2,\ldots,n \qquad (1.11)$$

where z_t is a tabled quadrature point (node) and $h(z_t)$ is the corresponding weight (see Straud and Sechrest,1966).

The quadrature weights, $h(z_t)$, are approximately the normalized, i.e., $\sum_{t=1}^{k} h(z_t)=1$, values of the probability density of a N(0,1) random variable at the points z_t, which are choosen to best approximate the marginal probability function $f(x_s)$. This approximation becomes more accurate as the number of quadrature points increases.

From the maximization of

$$L = \sum_{s=1}^{n} \log f(x_s)$$

we obtain, for v=0,1

$$\frac{\partial L}{\partial \alpha_{i,v}} = \sum_{t=1}^{k} \frac{\partial \pi_i(z_t)}{\partial \alpha_{i,v}} \frac{[R_{i,t} - N_t \pi_i(z_t)]}{\pi_i(z_t) [1-\pi_i(z_t)]} \qquad (1.12)$$

where

$$R_{it} = \sum_{s=1}^{n} x_{is} h(z_t|x_s) \qquad (1.13)$$

$$N_t = \sum_{s=1}^{n} h(z_t|x_s) \qquad (1.14)$$

and $h(z_t|x_s)$ is the posterior probability of Z_t given x_s.

Before defining an E-M algorithm in this approach, it is useful look at the meaning of N_t and R_{it}. As the quantity $h(z_t|x)$ is the probability that an individual with response vector x is located at z_t, N is the expected number of individuals at z_t. By analogy, R_{it} is the expected number of positive responses to item i among those individuals at z_t.

Consequently if we know the allocation of each individual on the Z-scale then N_t is the number of individuals at z_t and R_{it} is the observed frequency of positive response at z_t.

The estimation of the parameters is performed by choosing any starting values for $\{\alpha_{i,0}\}$ and $\{\alpha_{i,1}\}$ followed by repeated applications of (1.12), (1.13) and (1.14) over the set of items, using an E-M algorithm defined as

E-step: Calculate the values of R_{it} and N_t using equations (1.13) and (1.14).

M-step: Obtain improved estimates of the $(\alpha_{i,0})$ and $(\alpha_{i,1})$ solving equation (1.12), using the values of R_{it} and N_t from the E-step.

The E-M cycles are continued until convergence is obtained. In this case the number of values that the latent variable assumes is fixed and the set of values constitues the distribution of Z.

If we use the logit/probit model for $\pi_i(z)$ then

$$\pi_i(z) = [\ 1 + \exp(-\alpha_{i,0} - \alpha_{i,1}\ z)\ ]^{-1}$$

and

$$\frac{\partial\ \pi_i(z)}{\partial\ \alpha_{i,v}} = z^v\ \pi_i(z)\ [\ 1-\pi_i(z)\ ] \qquad (1.15)$$

for v=0,1.

Substituting (1.12) and (1.15) the equations become

$$\sum_{t=1}^{k} z^{v-1}\ [\ x_{is} - \pi_i(z_t)\ ]\ h(z_t|x_s) = 0 \qquad (1.16)$$

for v=0,1 and i=1,2,...,p, which may be compared with (1.10).

Even though we have presented these methods for a response function, in which G^{-1} in (1.5) was the logit function and the prior distribution of the single latent variable was approximated using Gauss-Hermite quadrature points, it may be applied for any response function and any discrete prior distribution.

If some other prior distribution of the latent variable is assumed, other points may be chosen and a normalized density point t substituted for $h(z_t)$ in (1.11). For example, if a rectangular prior is assumed then k points may be set at equal intervals over an appropriate range and the quadrature weight set $(h(z_t))$ equal to 1/k.

Bock and Aitkin (1981) have considered besides a prior standard normal distribution, a rectangular and an empirical distribution for the single latent variable and taken k=10 (see their paper for more details). Working through the data of Section 6 and 7 for the Law School Aptitude Test (LSAT) presented in Bock and Lieberman (1970), they have obtained practically the same estimates of the parameters from these three different prior distributions. They also suggested that adequate solutions could be obtained with even smaller k, for example, k=3,5 or 7 and this would make it feasible to generalize the method to several latent variables. On the other hand, investigations made by Shea (1984) show that at least k=20 may be necessary to obtain reasonable accuracy and this puts much greater demands on computing resources.

The general model (1.5) as defined by Bartholomew (1980) involves an arbitrary assumption about the form of the prior distribution of the latent variable. Although the form of the analysis does not depend on this assumption, as shown by Bartholomew (1984), it does affect the estimation of the parameters. Therefore it is important to know whether the values of the estimates are sensitive to the choice of the prior distribution.

Bartholomew (1988) answered this question mainly through numerical evidence that the choice of the prior has negligible effect on the

expected first- and second-marginal proportions. He concludes that the estimates are not sensitive to the choice of the prior based upon some results reported in Bartholomew (1980,1987) that the ML estimates which depend on margins of all order are usually very close to those depending only on the first-and-second order margins.

4-On the Existence and Uniqueness of the ML Estimates in a Rasch Model

If we intend to apply the Rasch model to a set of data, it may seem worthwhile to check first whether the parameters can be estimated.

The necessary and sufficient (n.s.) condition for the existence and uniqueness of the joint and conditional ML estimates for the Rasch model has been studied by Fischer (1981), Haberman (1977) and Andersen (1980), among others.

Fischer (1981)'s paper deals with both ML estimation procedures in a dichotomous Rasch model for complete and incomplete (omitted responses) data matrix. The basic condition is essentially the same, but we shall just present the main results for the joint ML estimation for a complete data matrix.

Let x_{ij} be equal to 1 if individual j has answered item i positively, and x_{ij} equal to 0, otherwise for i=1,...,p. Then

$t_j = \sum_{i=1}^{p} x_{ij}$ is the total number of positive responses given by individual j for j=1,...,n, and

$s_i = \sum_{j=1}^{n} x_{ij}$ is the total number of positive responses of item i for i=1,...,p.

Fischer (1981) defines a data matrix A as well-conditioned iff in every partition of the items into two nonempty subsets I_1 and I_2, some individual has answered positively to some item in the first set I_1 and answered negatively to some item in the second set I_2. Otherwise, A is called ill-conditioned.

Let I_1 and I_2 be some partition of the items. Then the subjects can be partitioned into the following three mutually exclusive subsets, any of which may be empty: S_1 consists of all subjects who solved all the items in I_2; S_2 consists of all subjects who solved none of the items in I_1, except for those subjects who already belong to S_1; S_3 consists of all subjects not belonging to S_1 or to S_2.

He shows that if A is well-conditioned, for every partition of the items into two non-empty subsets I_1 and I_2, the set S_3 is non-empty, i.e., after appropriate permutation of rows, the data matrix A attains the following structure:

$$A = \left[\begin{array}{c|c} A_1 & A_2 \\ \hline A_3 & A_4 \\ \hline A_5 & A_6 \end{array}\right] = \begin{array}{c} \begin{array}{cc} I_1 \quad & \quad I_2 \end{array} \\ \left[\begin{array}{c|c} & 1\ldots1 \\ A_1 & \ldots\ldots \\ & 1\ldots1 \\ \hline 0\ldots0 & \\ \ldots\ldots & A_4 \\ 0\ldots0 & \\ \hline A_5 & A_6 \end{array}\right] \end{array} \begin{array}{l} \left.\begin{array}{c} \\ \\ \\ \end{array}\right\} S_1 \\ \left.\begin{array}{c} \\ \\ \\ \end{array}\right\} S_2 \\ \left.\begin{array}{c} \\ \end{array}\right\} S_3 \end{array}$$

That is, there is at least one row in S_3 with one or more 1's in A_5 and one or more 0's in A_6, and S_1 and/or S_2 may be empty. Consequently, if S_3 is empty, A is ill-conditioned.

Finally, the ML estimates of the Rasch model are finite and unique iff the data matrix A is well-conditioned and $0<t_j<p$, for $j=1,\ldots,n$.

It is a consequence of A being well-conditioned that $0<s_i<n$, $i=1,\ldots,p$, that is, we neglect all items that have got all responses positive or negative, and the same is valid for the individuals.

In practice, for determining whether a complete data matrix A is well-conditioned or not, all we have to do is to order the item statistics s_i, $s_1 \leqslant s_2 \leqslant \ldots \leqslant s_p$, and to check whether the following equality is <u>not</u> fulfilled for any of the every index value p', $1 \leqslant p' \leqslant p-1$,

$$\sum_{t=p-p'}^{p} t\, n_t = \sum_{i=1}^{p'} s_i + (p-p') \sum_{t=p-p'}^{p} n_t$$

where n_t is the number of individuals with $t_j=t$.

If this equality is fulfilled by some index p', the data matrix A is ill-conditioned and thus the parameter of the Rasch model cannot be estimated.

5- *Breakdown of the Estimation Procedure*

There are some configurations of data, analogous to those called Heywood cases in factor analysis (underlying variable model), in which the 'true' ML estimate is infinite and the iterative system proceeds in that direction indefinitely. In this situation, after some number of cycles most of the discrimination parameter estimates and the

likelihood remain roughly constant while one or a few parameters increase indefinitely. The difference in goodness-of-fit with such a high discrimination parameter estimate (3.0 or bigger) is negligible. In practice in these cases the value of the estimate is a function of the stopping rule of the iterative procedure.

A Heywood case, on the other hand, is the occurence of a negative or zero estimate of the error variance Ψ for one or more variables. In the underlying model representation, $\alpha_{i,j} = -\lambda_{ij} / (\Psi_i)^{\frac{1}{2}}$, where λ_{ij} is the factor loading, so that a diverging discrimination parameter $\alpha_{i,j}$ in the response function model is equivalent to a Ψ_i (error variance) approaching zero.

In summary, according to Anderson and Gerbing (1984), Boomsma (1985) and Fachel (1986), the occurrence of Heywood cases increases as

(1)- the sample size decreases;

(2)- the number of indicators per factor and consequently the number of variables decreases, although Fachel has observed small variation between 5 and 100 variables;

(3)- the population values of the error variance are close to zero;

(4)- the factor loading are not uniform, for example, when only one factor loading increases up to 0.90 while the others remain equal to 0.5.

Van Driel (1978) identifies 3 causes for Heywood cases:

- sampling fluctuations combined with true values of the error variance close to zero;
- there does not exist any factor analysis model that fits the data;

- indefiniteness of the model (i.e.,too many true factor loadings are zero).

Bartholomew (1987) affirms, from his experience with binary estimation procedures, that the circumstances under which a slope parameter become larger and larger in the response function model are when

- the sample size is small, a few hundred or less,
- the number of variables is small and
- the discrimination parameters are very unequal,

which are equivalent to those leading to (1), (2) and (4) given above for the Heywood cases in factor analysis.

6- *Simulation Studies: comparison between the Rasch and the Logit/Probit Models*

Dinero and Haertel (1977) investigate the impact of variation in discrimination parameters on the correlation between parameter values (difficulties as well as abilities). From a Monte Carlo study, responses of 75 individuals to 30 items were simulated under a two-parameter logistic (logit/probit) model, and then fitted with the Rasch model.

The degree of fit was examined as a function of the variance of the item discriminations (0.05, 0.10, 0.15, 0.20 and 0.25) within distributions of different forms (uniform, normal and positively skewed), all with mean equal to 1.0.

For each distribution there was only a slight increase in the lack of fit as the variances increased. The poorest overall fit was when

the discrimination parameters were uniformly distributed.

They also investigate the impact of variation in item discriminations on the correlation between difficulty parameters and estimates. Again, the uniform distribution yielded very low correlations (-0.20) for the difficulty parameter estimates, while for either the normal or skewed distributions, there was no evidence that the variance of the distribution has any affect on the accuracy of the difficulty estimates.

A major part of this study was replicated by Van de Vijver (1986), but he found that the shape of the discrimination parameter distribution does not influence the robustness of the Rasch estimates dramatically, even for the uniform distribution.

Another simulation study was done by Van de Vijver (1986) in order to investigate the robustness of the Rasch model against violations of the homogeneity of discrimination parameters. The study was carried out simulating discrimination parameters that assumed values between 0.0 and 2.0, for sets of 10 to 50 items and sample size of 25 to 500 individuals.

He observed that the correlation between difficulty parameters and corresponding estimates increases with the sample size, while the correlation between person parameters and estimates increases with test length. Furthermore, correlations were not very sensitive to the heterogeneity of the discrimination parameters, but a decrease could be observed between the difficulty parameters and estimates when discrimination parameters assumed values equal to 2.0 (the most extreme situation). The same relation is valid for bias and RMSE (root mean squared error). They concluded that even in small samples and for short tests, heterogeneity of the discrimination parameter hardly

affects the accuracy of the Rasch estimates.

Hulin, Lissak and Drasgow (1982) investigate the accuracy of simultaneous estimation of item and person parameters from simulated two-parameter logistic model, samples of 200 to 2000 individuals and tests of 15, 30 and 60 items. The ability values were drawn from N(0,1) distribution, the discrimination and difficulty parameters from U(0.3;1.4) and U(-3,3), respectively..

The accuracy of the item parameters (difficulty and discrimination) was measured by the RMSE (root mean squared error) between recovered and actual response function, while the accuracy of ability was measured by both correlation and RMSE. This measure of the accuracy of the recovered response function corresponds to the mean squared error of prediction in multiple regression.

The main result was that for a fixed test length, sample size has a small influence on the accuracy of the ability parameter estimates, while the effect of decreasing the number of items is pronounced. Correlations between difficulty parameters and estimates are all high ($\geq$0.94) and stable, and larger than those between discrimination parameters and estimates. Correlations are less well behaved than the RMSE's, since they do not display the effects of test length within a constant sample size, for example. RMSE show a substantial increase in estimation accuracy as the test length increased from 15 to 60 items. Finally, they found tradeoffs between test length and sample size, since doubling the number of items and halving sample size resulted in comparable response function average RMSE's, at least for tests of 30 and 60 items and sample sizes of 500, 1000 and 2000.

7- *Goodness-of-fit*

If the sample size (n) is large compared with 2^p (number of possible response patterns) a chi-square or log-likelihood goodness-of-fit test can be carried out on the observed and expected frequencies of the response patterns. Often, there are many small expected frequencies so that pooling becomes necessary. Since the number of degrees of freedom in the unpooled case is $2^p-p(q+1)-1$, then situations may occur where there will be no degrees of freedom to judge the goodness of fit.

When a formal test cannot be carried out and p is not too large , the goodness of fit of the model may be judged by comparing the observed and expected frequencies of the response patterns. An additional check maybe done by comparing the observed and the fitted values of the one-and-two way marginal frequencies.

There are other checks which can be made on the data before or after fitting a model. For example, Bartholomew (1980) showed that if a one-latent variable model applies then it must be possible to label the categories so that, in the population, all the cross-product ratios exceed one. A systematic approach to the question of whether the data are consistent with an unidimensional model has been developed by Holland (1981) and extended by Rosenbaum (1984).

According to Rosenbaum, theorem 1, if a latent variable model is unidimensional for P[X=x] with nondecreasing response function then X is conditional associated, i.e, for all nondecreasing functions g(.) and f(.), all functions h(.) and all partitions and rearrangements of

X into two nonoverlapping groups of items, (S,T),

$$\mathrm{Cov}(g(S), f(S) \mid h(T)) \geq 0$$

where Cov(.,.|.) denotes conditional population covariance.

In particular, if we take $S=(X_i,X_j)$, and T equal to the remaining p-2 items with $h(T) = \sum_{k \neq i,j} X_k$ then a unidimensional latent variable model for $P[X = x]$ with nondecreasing response function implies that

$$\mathrm{Cov}((X_i, X_j) \mid \sum_{k \neq i,j} X_k = t) \geq 0$$

for all pairs of manifest variables and all values of t in the population.

Equivalently, such a model implies that there is a population cross-product ratio of at least equal 1 in every 2x2 subtable of the (p-1) layer of the 2x2x(p-1) population contingency table recording $X_i \times X_j \times \sum_{k \neq i,j} X_k$.

Tatsuoka (1984) describes the use of caution indices to identify individuals with unusual response patterns relative to a given model.

Due to the special properties of the Rasch model several goodness-of-fit test have been developed, of these, the conditional likelihood ratio test, introduced by Andersen (1973b), is perhaps the best known. The test is based on a comparison between item difficulties estimated from different subsamples formed according to the number of positive responses on the test and overall estimates

obtained from the whole sample. If the Rasch model fits the data well then consistent difficulty estimates should be obtained for any subdivision of the sample into two or more groups. It was shown that, when the sample is large, the test statistic has approximately a χ^2 distributed random variable with (p-1)(g-1) degrees of freedom, where p is the number of items and g is the number of subsamples considered.

The problem of fit of the Rasch model has been further discussed by Gustafsson (1980b), Van den Wollenberg (1982), Molenaar (1983), and Kelderman (1984), among others.

Gustafsson (1980b) presented a test for the hypothesis that two disjoint groups of items measure the same construct. This provides a test of unidimensionality when items are grouped a priori. Van den Wollenberg (1982) also developed test statistics for lack of equality of discrimination parameters and unidimensionality.

These type of tests are global measures of how all the items in a test fit the Rasch model. To assess goodness-of-fit to a given item response function, Gustafsson (1980b) suggested using graphical procedures. On the other hand, Molenaar (1983) has provided procedures for a more detailed analysis under the Rasch model, which also involves information about the goodness-of-fit for a given item response function.

McKinley and Mills (1985) conducted an extensive investigation of goodness-of-fit indices for a given item response fuction. They compared four such indices, those developed by Bock (1972), Yen (1981), Wright and Mead (1978), and the Likelihood Ratio (LR) statistic. The first three of these employ the standard chi-square

goodness-of-fit formula and vary only with respect to the number of groups and the definition of the latent level used to compute the expected proportion of positive responses.

Nine tests with length 75 and sample sizes of 500, 1000 and 2000 individuals were used to generate the simulated data under each of the one-, two-,and three-parameter models. In addition, the normally distributed samples had means of -1, 0 and 1 on the latent scale. When the data generated by the two and three parameter models were analysed under a one-parameter model, the results indicated a consistent lack of fit. As was the case with Yen's (1981) study, analising three-parameter data using the two parameter model worked quite well.

McKinley and Mill (1985) concluded that the LR index appeared to yield the fewest erroneous rejections of the hypothesis of fit, while the Bock index yielded fewer erroneous conclusions of fit. However, the differences were slight. They also applied the four procedures to an additional 9 tests having an underlying multidimensional latent structure. In all cases, the analysis yielded a high proportion of misfits. Thus, the underlying assumption of unidimensionality appears to be critical to obtaining good fit between the ICC and the observed data.

8- *Sampling Variation of the Maximum Likelihood Estimators*

There has been little discussion about the magnitudes of the standard deviations of estimated parameters for commonly used latent variable models. This may be due to the fact that no simple closed formulae exist for the standard deviations as a function of the sample size and the parameters. The usual way to estimate the standard deviations for maximum likelihood estimates of the parameters is to compute the asymptotic variance-covariance matrix, using the elements of the inverse of the information matrix evaluated at the solution point. Thus if we have a set of parameters β then

$$\left[D(\hat{\beta}) \right] = E \left[\frac{-\partial^2 L}{\partial\beta_i \, \partial\beta_j} \right]^{-1} \Bigg|_{\beta=\hat{\beta}} \qquad (1.17)$$

where

$$\frac{\partial^2 L}{\partial\beta_i \, \partial\beta_j} = \left[\sum_{s=1}^{n} \frac{1}{f_s} \frac{\partial^2 f_s}{\partial\beta_i \, \partial\beta_j} - \frac{1}{f_s^2} \frac{\partial f_s}{\partial\beta_i} \frac{\partial f_s}{\partial\beta_j} \right] \Bigg|_{\beta=\hat{\beta}}$$

and $f_s \equiv f(x_s)$. On taking the expectation, the first term vanishes leaving

$$\left[D(\hat{\beta}) \right]^{-1} = n \, E \left[\frac{1}{f_s^2} \frac{\partial f_s}{\partial\beta_i} \frac{\partial f_s}{\partial\beta_j} \right] \Bigg|_{\beta=\hat{\beta}} \qquad (1.18)$$

In our case X is a response pattern taking 2^P different values and the expectation in (1.18) is thus

$$\sum_{s=1}^{n} \frac{1}{f^2(x_s)} \frac{\partial f(x_s)}{\partial\beta_i} \frac{\partial f(x_s)}{\partial\beta_j} \qquad (1.19)$$

If p is small it is feasible to evaluate this sum for all i and j and then to invert the resulting matrix. However if p is large some probabilities will become very small so that the computation of $1/f(x_s)$ will cause overflow on most of computers. In this case an approximation may be used, replacing the expectation of the information matrix by its observed value. This requires the computation of (1.17) and the inversion of the resulting matrix. Since the first term has expectation zero a further approximation may be obtained from

$$D^*(\hat{\beta}) = \left[\sum_{s=1}^{n} \frac{1}{f^2(x_s)} \frac{\partial f(x_s)}{\partial \beta_i} \frac{\partial f(x_s)}{\partial \beta_j} \right]^{-1} \qquad (1.20)$$

The number of distinct terms in the sum of (1.20) will usually be less than n since more than one individual may have the same response pattern.

Louis (1982) developed a technique for computing the observed information matrix when the E-M algorithm is used to find the maximum likelihood estimates in incomplete data problems. It requires computation of the complete-data gradient and second derivative matrix which can be implemented quite simply in the E-M iterations. This procedure can be applied to obtain the asymptotic variance-covariance matrix in latent variable models, since they involve observable (manifest) variables and not directly observable (latent) variables which corresponds to a case of incomplete data, as defined by Dempster, Laird and Rubin (1977).

Thissen and Wainer (1982) investigated the asymptotic standard errors of the item parameters for the one-, two-,and three-parameter models under the assumption that the latent value of the respondents were known and normally distributed with zero mean and unit variance. Tables of the minimum asymptotic standard errors were reported for combinations of parameter values under the three models.

An interesting set of results was given by the two-parameter and the three-parameter model with c=0 (guessing parameter). Even though the numerical values of the difficulty and the discrimination parameters would be the same, the information matrices are not. The three-parameter matrix still has a row and column corresponding to the guessing parameter. When one item was easy and had low discriminating power, the standard errors under the two-parameter model were roughly 0.09 of those reported for the three-parameter model. Clearly, the two and the three-parameter model with c=0 are not the same with respect to the standard errors of the item parameter estimates. The asymptotic standard errors for the item difficulty under the Rasch model were consistently smaller than those obtained for the other two models. In particular, the increase in standard error with departure of item difficulty from zero was much less pronounced.

Based upon the results, Thissen and Wainer (1982) suggested that when working with logistic response models we should try to fit the simplest model first, and only if it is found to be inadequate move to the more complex ones.

Lord and Wingersky (1983) have developed a method for computing the asymptotic variance-covariance matrix for the three-parameter logistic model, considering the unidimensional latent variable as a person parameter. The derivation assumes that both item and person

parameters are unknown. They demonstrate that the size of the error variances are affected strongly by the restrictions introduced in order to fix the latent scale. One disadvantage of this method is that the information matrix to be inverted is very large.

On the other hand, Gruijter (1985) has shown that the method can be simplified for the Rasch model when we are only interested in the item parameters. This is done under a suitable restriction on the difficulty parameters, as for example, assuming that the mean b is equal to zero,i.e, $b_p = -\sum_{i=1}^{p-1} b_i$, or setting $b_p=0$. Although the variance-covariance matrix for the item parameters can be obtained without difficulty, it depends on the restrictions. He points out that the first restriction seems to be preferable to the others due to its simplicity and relative accuracy of the mean.

9- *Adequacy of the Asymptotic Variance-Covariance Matrix*

When interpreting the asymptotic variance-covariance matrix of the parameter estimates it is assumed that the model is appropriate for the data. Since this assumption may be false in practice, or the sample size is not large enough for the number of parameters which have been estimated or even the standard asymptotic theory does not apply, the standard deviation and covariances obtained asymptotically will probably represent lower limits for the actual ones, and they must be analysed carefully.

Another way to look at the variance-covariance matrix that may give some idea about how the asymptotic theory is working, is through jackknife and bootstrap techniques.

9.1- Jackknife

Jackknifing is a statistical technique first proposed by Quenouille (1956), which is used for reducing bias in the estimation of parameters and for estimating the variance-covariance matrix of the estimates. Miller (1974) gives an review of the subject.

In the basic jackknife the observations are randomly divided into g groups of size h each.

Let $X_1, X_2, \ldots, X_p$ be a sample of independent and identically distributed(iid) random variables and $\hat{\beta}$ be an estimator of the parameter vector β based on the sample size n, where n=gh.

Let $\hat{\beta}_{-i}$ be the corresponding estimator based on the sample of size (g-1)h, where the i^{th} group of size h has been deleted.

Then jackknife pseudovalues are defined by

$$\tilde{\beta}_i = g\,\hat{\beta} - (g-1)\,\hat{\beta}_{-i}$$

for i=1,2,...,g.

The jackknife estimates $\tilde{\beta}$ and its estimated variance-covariance matrix are obtained from the g pseudovalues by treating them as independently identically distributed observations from a multivariate normal distribution (Tukey,1958). These estimates are given by

$$\tilde{\beta} = \frac{\sum \tilde{\beta}_i}{g} \qquad (1.21)$$

$$\tilde{\beta} = g\hat{\beta} - \frac{g-1}{g} \sum \hat{\beta}_{-i}$$

and

$$\sum(\tilde{\beta}) = \frac{\sum (\tilde{\beta}_i - \tilde{\beta})(\tilde{\beta}_i - \tilde{\beta})^t}{g(g-1)} \qquad (1.22)$$

Since it often happens that $\tilde{\beta}$ and $\hat{\beta}$ are asymptotically equivalent, $\sum(\tilde{\beta})$ is sometimes used to estimate the variance-covariance matrix of $\hat{\beta}$.

The jackknife estimate of bias is the difference between the parameter estimate $\hat{\beta}$ and $\tilde{\beta}$ multiplied by the correction factor n/(n-1), i.e.,

$$\text{bias} = \frac{n}{n-1}(\hat{\beta} - \tilde{\beta}) \qquad (1.23)$$

In most of the applications the number of groups, g, is equal to n, i.e, each observation corresponds to one group and the $\hat{\beta}_{-i}$ is obtained from the sample, deleting the i^{th} observation, i.e, h=1.

The jackknife technique has been applied in many areas, including factor analysis. Pennell (1972) demonstrated how the method can be used to find confidence intervals for the factor loadings, while Clarkson (1979) discussed the results of simulation studies using jackknife techniques and proposed modifications.

Clarkson's studies do not include the jackknife samples which provide Heywood cases. He argue that in these cases the jackknife

estimates of the factor loadings are not representative of the 'usual' jackknife results because they are too large in absolute value.

Jorgensen (1987) gave a modification of the jackknife method for estimating the dispersion of the parameter estimates that are obtained as limits of iterative processes. He also gave examples to show how the method can be applied to the E-M algorithm and to iteratively reweighted least-squares.

9.2- Bootstrap

The bootstrap is a general resampling procedure introduced by Efron (1979) to estimate the distribution of statistics based on independent observations. It can be carried out non-parametrically and parametrically, depending on the distribution from which the bootstrap samples are drawn.

We shall first present the non-parametric or empirical bootstrap method.

Suppose $X_1, X_2, \ldots, X_p$ are independent and identically distributed(iid) random variables from a population with unknown distribution function F, and suppose the goal is to make inferences about the parameter vector β of the population.

Let $\hat{\beta}(x_1, x_2, \ldots, x_p)$ be an estimator of β based on the sample size n and let $\hat{F}$ be the empirical distribution, that is, the distribution function that assign mass 1/n to each X_i.

The bootstrap approximates the sampling distribution of β under F by the sampling distribution of $\hat{\beta}$ under $\hat{F}$. This procedure is carried out using Monte Carlo method as follow:

(1) Construct $\hat{F}$

(2) Draw a bootstrap sample, $X^*_1, X^*_2, \ldots, X^*_p$ iid with cdf $\hat{F}$ and calculate

$$\hat{\beta}^* = \hat{\beta}(X^*_1, X^*_2, \ldots, X^*_p)$$

(3) Independently do B times the step 2 (for some large B), obtaining $\hat{\beta}^*_b$, b=1,2,...,B. The distribution function of $\hat{\beta}$ is approximated by

$$\hat{F}_B(y) = \#(\hat{\beta}^*_b \leq y) / B.$$

The bootstrap estimate of β based on the B replications is the mean of the $\hat{\beta}^*_b$ estimates, i.e,

$$\hat{\beta}^*_\cdot = \sum \hat{\beta}^*_b / B \qquad (1.24)$$

and the bootstrap variance-covariance matrix estimate of β based on the B replications is the variance-covariance matrix of the $\hat{\beta}^*_b$ estimates, i.e,

$$\Sigma_B = (B-1)^{-1} \sum (\hat{\beta}^*_b - \hat{\beta}^*_\cdot) (\hat{\beta}^*_b - \hat{\beta}^*_\cdot)^t \qquad (1.25)$$

As the number of replications $B \to \infty$, $\hat{\beta}^*_\cdot$ will approach the bootstrap estimate of β and Σ_B the corresponding bootstrap estimate of the variance-covariance matrix Σ.

For example, if $X_1, X_2, \ldots, X_n$ are drawn from a normal distribution with mean μ and standard error σ. Then

$$\bar{X}^*_{\cdot} = \sum \bar{X}^*_b / B \quad \text{and}$$

$$\hat{\sigma} = \{ (B-1)^{-1} \sum (\bar{X}^*_b - \bar{X}^*_{\cdot})^2 \}^{\frac{1}{2}}$$

can be used to estimate μ and σ.

The bootstrap estimate of bias is the difference between the parameter estimate $\hat{\beta}$ and the bootstrap estimate $\hat{\beta}^*_{\cdot}$, that is,

$$\text{bias} = \hat{\beta} - \hat{\beta}^*_{\cdot} \tag{1.26}$$

The basic result of the bootstrap theory is that the empirical distributions of the parameter estimates obtained by this method are asymptotically the same as the sampling distribution of those parameters in sampling from the population from which the original sample was drawn.

There is nothing which says that the bootstrap must be carried out non-parametrically. If we have reason to believe that the true distribution F is Normal, for example, then we can estimate F by its parametric ML estimate $\hat{F}$. The bootstrap samples at step (1) of the algorithm could them be drawn from $\hat{F}_{normal}$ instead of $\hat{F}$ (empirical distribution) and steps (2) and (3) carried out as before.

Efron (1979) also suggests that Taylor series expansion method can be used to obtain the approximate mean and variance of the bootstrap distribution of $\hat{\beta}^*_{\cdot}$, and he shows that it turns out to be the same as Jaeckel's infinitesimal jackknife (Miller,1974), which differ only in detail from the standard jackknife described before.

Efron and Tibshirani (1986) discuss the number of replications B necessary to give reasonable results when we are estimating the standard deviation of one parameter. They set out the following approximation

$$CV(\hat{\sigma}_B) = \{ CV(\hat{\sigma})^2 + [(E(\hat{\delta}) + 2)/4B] \}^{\frac{1}{2}}$$

where $CV(\hat{\sigma})$ is the limiting coefficient of variation of σ as $B \to \infty$, $\hat{\delta}$ is the kurtosis of the bootstrap distribution of $\hat{\beta}^*$, given the observed data $\mathbf{x}=(x_1,x_2,\ldots,x_n)$, and $E(\hat{\delta})$ its expected value average over $\mathbf{x}$. For typical situations, $CV(\hat{\sigma})$ lies between 0.10 and 0.30.

From this approximation and assuming that $E(\hat{\delta})=0$, they point out that for values of $CV(\hat{\sigma}) \geq 0.10$, there is little improvement when B is bigger than 100. In fact B as small as 25 gives reasonable results. However the situation is quite different for setting bootstrap confidence intervals.

Efron (1984) discusses different kinds of confidence intervals using the bootstrap and he shows that it is necessary to have at least 1000 samples to compute the BC (bias corrected percentile interval) as defined in the same paper, and BC_α intervals while for the simplest method, percentile interval, 250 replications can give useful results.

The percentile interval is obtained by taking $\beta \in (\hat{F}^{-1}(\alpha), \hat{F}^{-1}(1-\alpha))$ as an approximate $1-2\alpha$ central interval for β.

Confidence intervals are a fundamentally more ambitious measure of statistical accuracy than standard errors, so it is not surprising that they require more computational effort.

Chatterjee (1984) gives an application of the non-parametric bootstrap method to the problem of estimating the variability of the estimates of factor loadings. The number of bootstrap samples was settled empirically; it appeared that 300 gave reasonable stability. Combining the bootstrap with graphical techniques he examines the variability of the estimator of the factor loadings. He points out that bootstrap may very well reveal when the asymptotic results are poor approximations.

Grönroos (1985) applies bootstrap methods to confirmatory factor analysis of a LISREL submodel (Jöreskog and Sörbom, 1984) to estimate factor loadings and their standard deviations.

His simulation studies involve artificial data with sample size 100, 150 and 300 and initially 300 replications. However the number of bootstrap samples become smaller, since he deletes from the analysis all those which provide the occurrence of Heywood cases.

Comparing asymptotic theory with bootstrap and Normal bootstrap results, he points out that the difference between the two bootstrap methods is very small, but it is larger, even though not essential significant, when compared with the asymptotic results.

Beran and Srivastava (1985) use bootstrap test and confidence regions for functions of the population covariance matrix, for example, eigenvalues and eigenvectors, which have the desired asymptotic levels if model restrictions, such as multiple eigenvalues, are taken into account in designing the bootstrap algorithm.

Efron and Tibshirani (1986) give a review of bootstrap methods for estimating standard errors and confidence intervals. The bootstrap is also extended to other measures of statistical accuracy such as bias and prediction error, and to complicated data structures such as time series, censored data, and regression models.

Bootstrap confidence intervals have been discussed with new improvements by Efron (1987) and their applications to problems in a wide range of situations is given by Diciccio and Tibshirani (1987).

Chapter 2

BEHAVIOUR of the LIKELIHOOD FUNCTION

1- *Comparison between the Profile and an Approximate Method*

Since the parameters of the latent variable models under investigation are usually estimated by the method of maximum likelihood (ML), it is very important to check if the behaviour of the likelihood function is suitable for the method.

We are interested in checking on whether the likelihood has a smooth unimodal shape, or whether it has multiple relative maxima. The shape of the likelihood around the maximum point will show whether the information matrix will give a good guide to the variability of the estimates. It is a counter-indication to the use of maximum likelihood estimates if there is a flat plateau, or a ridge moving off to infinity.

A badly behaved likelihood function suggests either that a reparametrization is necessary, or that the model is a poor fit for the data, or that the inference is particularly difficult.

How can we investigate the behaviour of the likelihood function?

Let us consider the latent variable model for fitting binary responses given by (1.5) in the case of a single latent variable. Thus the likelihood is a function of $\alpha_{i,0}$ and $\alpha_{i,1}$, i=1,2,...,p. A profile likelihood can be obtained for $\alpha_{i,0}$ and $\alpha_{i,1}$ by maximising the

likelihood over the remaining variables j, j=1,2,...,p and j≠i. We repeat this procedure to get the profile likelihood at a representative set of values of $(\alpha_{i,0},\alpha_{i,1})$.

We usually choose to look at the profile likelihood for those parameters for which the likelihood seems to be less satisfactory. One guide to possible poor behaviour is the size of the ML estimate $\hat{\alpha}_{i,1}$. A value of $\hat{\alpha}_{i,1}$ greater than 3.0 may be a sign of a badly behaved likelihood function.

Obtaining the behaviour of the likelihood function using the profile method, described above, takes much computer time, since if we evaluate it for eighty $(\hat{\alpha}_{i,0},\hat{\alpha}_{i,1})$ points we have to maximise the likelihood function that number of times.

Clearly it would be useful to have a quicker method that gives the same information as the profile likelihood.

A simple alternative is to replace the maximisation procedure by some approximation. We have tried using the original marginal ML estimates for $\alpha_{j,0}$ and $\alpha_{j,1}$ for j≠i instead of maximising again for each new choice of values for $\alpha_{i,0}$ and $\alpha_{i,1}$.

We shall call the latter approach method A, the profile likelihood method B . Put

$L_A(\alpha_{i,0},\alpha_{i,1})$ = loglikelihood value obtained by fixing the remaining parameter at these ML values $\hat{\alpha}_{j,0}$ and $\hat{\alpha}_{j,1}$, j=1,2,...,p, j≠i.

$L_B(\alpha_{i,0},\alpha_{i,1})$ = loglikelihood value obtained by maximising over $\alpha_{j,0}$ and $\alpha_{j,1}$, j=1,2,...,p, j≠i.

We apply and compare both methods by contouring the values for L_A, L_B as a function of $\alpha_{i,0}$ and $\alpha_{i,1}$, as defined above, using the subroutine library GINO-SURF. This is done using 3 sets of data that have been analysed in Bartholomew (1987, Chapter 9), using a single latent variable logit/probit model and marginal maximum likelihood estimation procedure. The computer program used for fitting the model was FACONE written by Dr. Brian Shea at London School of Economics, using 48 quadrature points and considering that the convergence of the estimation procedure is obtained when the maximum gradient of the parameter estimates is smaller or equal 0.001. The asymptotic standard deviations of the parameter estimates are obtained by inverting the observed second derivative matrix at the ML solution point. We shall take Bartholomew's results of the fitting model as a starting point for our analysis.

1.1- Arithmetic Reasoning Test on White Women

The frequency distribution of the response patterns for the first and second examples are samples of the Arithmetic Reasoning Test (ART) from the American Youth on the Armed Services Vocational Aptitude Battery, given by Mislevy (1985). The individuals were classified by sex and colour, but the results given here relate to white and black women.

Table 2.1- Score distribution and results obtained by fitting a logit/probit model to the Arithmetic Reasoning Test on white women.

Response pattern	Observed frequency	Expected frequency	Total score	Component score
0000	20	26.79	0	0.00
0010	14	9.83	1	1.00
1000	23	18.43	1	1.04
0100	20	15.78	1	1.24
0001	8	4.86	1	1.44
1010	9	11.24	2	2.04
0110	11	10.55	2	2.24
1100	18	20.21	2	2.28
0011	2	3.57	2	2.44
1001	8	6.86	2	2.48
0101	5	6.70	2	2.68
1110	20	21.87	3	3.28
1011	6	8.16	3	3.48
0111	7	8.74	3	3.68
1101	15	17.18	3	3.72
1111	42	37.23	4	4.72
Total	228	228.00	-	-

$\chi^2 = 8.39$ on 6 degrees of freedom ($p = 0.21$)

Thus it is reasonable to infer that the data are consistent with a single latent variable indicating the arithmetic reasoning ability. The scaling given by the component is consistent with that of the total score because the $\hat{\alpha}_{i,1}$'s are very similar as we can see in Table 2.2.

Table 2.2- Parameter estimates and asymptotic standard deviations from fitting a logit/probit model to the Arithmetic Reasoning Test on white women.

Item i	$\hat{\alpha}_{i,1}$	$SE(\hat{\alpha}_{i,1})$	$\hat{\alpha}_{i,0}$	$SE(\hat{\alpha}_{i,0})$	$\hat{\pi}_i$
1	1.04	0.32	0.59	0.17	0.64
2	1.24	0.39	0.56	0.17	0.64
3	1.00	0.30	-0.06	0.16	0.48
4	1.44	0.45	-0.51	0.21	0.38

The parameter estimates show that the items are neither very easy nor too difficult with approximately equal discriminating power.

We apply below methods A and B to discover the behaviour of the likelihood for the data in Table 2.1 and parameter estimates in Table 2.2.

Let us choose the first item as our item i. Since all the slope parameters are approximately the same, we would expect to get the same behaviour by choosing any other item.

Figures 2.1 and 2.2 have been obtained from 183 pairs $(\hat{\alpha}_{1,0},\hat{\alpha}_{1,1})$, where $\hat{\alpha}_{1,0} \in (-3.50,3.50)$ and $\hat{\alpha}_{1,1} \in (0.10,12.00)$.

According to Table 2.2, the ML estimates for item 1 are $\hat{\alpha}_{1,1}$=1.04 and $\hat{\alpha}_{1,0}$=0.59. However Figure 2.1 suggests that the value of the likelihood does not change much along a whole straight line of values for $\hat{\alpha}_{1,1}$ and $\hat{\alpha}_{1,0}$. Close inspection of the input data shows that there is a slight decrease but not enough to show up in the contouring. Figure 2.2 shows a result much closer to Figure 2.1 than one might expect, though the peak is slightly better defined. Comparing both

graphs this is the only difference between them and it is due to the fact that in method A the likelihood decrease faster than in method B.

The most striking aspect of both figures is the long ridge in the picture going off in a vaguely North Easterly direction. This suggests that there is very little information in the data to choose between $(\alpha_{1,0},\alpha_{1,1})$ values along that ridge, and casts doubt on the validity of the ML estimates for $(\alpha_{1,0},\alpha_{1,1})$.

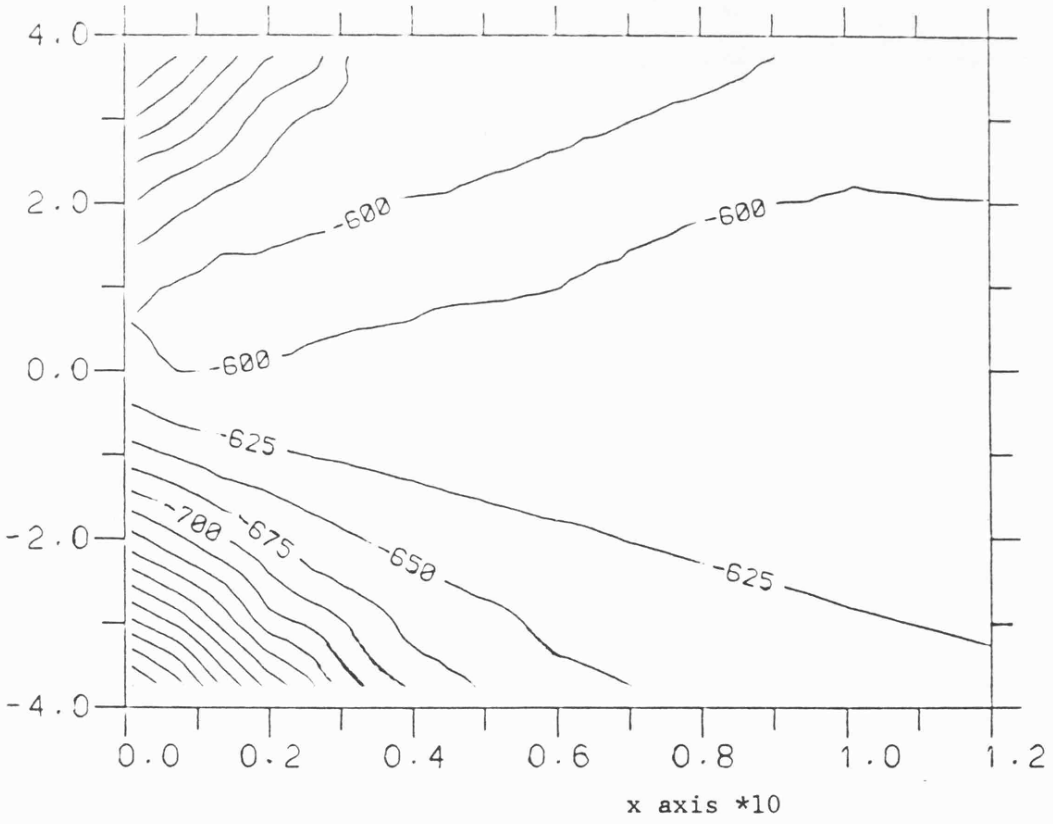


Figure 2.1- Loglikelihood values as a function of $\hat{\alpha}_{1,1}$ and $\hat{\alpha}_{1,0}$, using method B (profile) to the ART on white women.

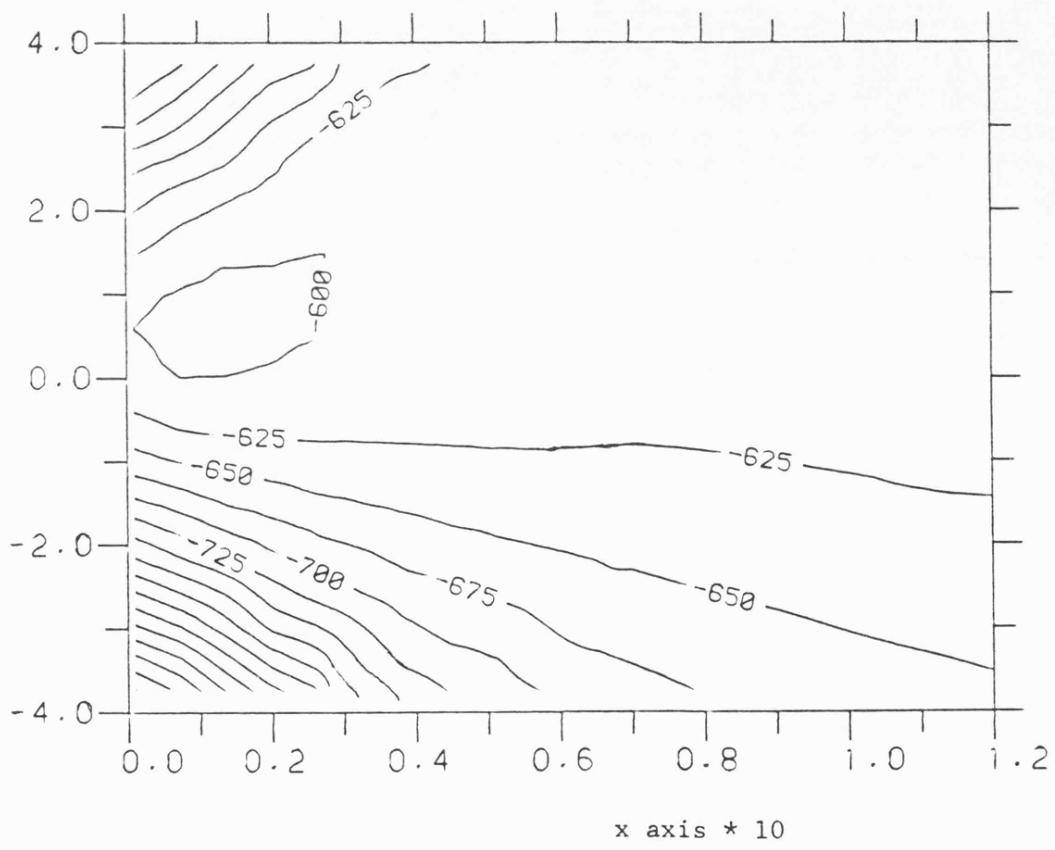


Figure 2.2- Loglikelihood values as a function of $\hat{\alpha}_{1,,1}$ and $\hat{\alpha}_{1,,0}$, using approximate method A to the ART on white women.

1.2- Arithmetic Reasoning Test on Black Women

As a second example, we analyse the results of the Arithmetic Reasoning Test on black women.

Table 2.3- Score distribution and results obtained by fitting a logit/probit model to the Arithmetic Reasoning Test on black women.

Response pattern	Observed frequency	Expected frequency	Total score	Component score
0000	29	28.39	0	0.00
0001	8	8.19	1	0.19
0010	7	7.99	1	0.37
0100	14	14.95	1	0.38
0011	3	2.36	2	0.56
0101	5	4.42	2	0.57
0110	6	4.41	2	0.75
0111	0	1.33	3	0.94
1000	14	17.74	1	14.39
1001	10	6.88	2	14.58
1010	11	8.90	2	14.76
1100	19	16.77	2	14.77
1011	2	3.54	3	14.95
1101	5	6.66	3	14.96
1110	8	8.84	3	15.14
1111	4	3.62	4	15.33
Total	145	145.00	-	-

$\chi^2 = 6.42$ on 3 degrees of freedom ($p = 0.10$)

As for the test on white women (Table 2.1) we can also infer that the logit/probit model with one latent variable fits reasonably well.

Note that Table 2.4 below shows significant differences between the slope parameter estimates($\hat{\alpha}_{i,1}$, i=1,...,4).

Table 2.4- Parameter estimates and asymptotic standard deviations from fitting a logit/probit model to the Arithmetic Reasoning Test on black women.

Item i	$\hat{\alpha}_{i,1}$	$SE(\hat{\alpha}_{i,1})$	$\hat{\alpha}_{i,0}$	$SE(\hat{\alpha}_{i,0})$	$\hat{\pi}_i$
1	14.39	67.78	0.25	4.63	0.56
2	0.38	0.22	-0.33	0.16	0.42
3	0.37	0.24	-0.96	0.20	0.28
4	0.19	0.24	-1.08	0.21	0.25

The results show that item 1, due its large discriminating power, divides the sample into two totally separate groups, those answering the item positively and those who do not. On the other hand, its standard deviation is too large to be trusted. Even for the other $\hat{\alpha}_{1,1}$ the standard deviations may be considered so large that little information is present about them.

Due to the very large slope parameter estimate of item 1 and its strikingly wild standard deviation, it is an obvious choice to look at the behaviour of the likelihood function for 185 pairs $(\hat{\alpha}_{1,0},\hat{\alpha}_{1,1})$.

Since both methods give exactly the same picture, we present just one (Figure 2.3). There is only a tiny difference between the 185 loglikelihood values from methods A and B, for $\hat{\alpha}_{1,1}$ bigger than 3.0 and any $\hat{\alpha}_{0,1}$.

Figure 2.3 shows that the likelihood function assumes practically the same values for all $\hat{\alpha}_{1,1}$, and as $\hat{\alpha}_{1,1}$ increases the best values for $\hat{\alpha}_{1,0}$ cover all its interval of variation. Although the subroutine used to draw the graph does not show small differences, analysing the input data we can confirm that the likelihood continues to increase indefinitely, indicating that the actual value for $\hat{\alpha}_{1,1}$ is infinity, which is not sensible.

This is one example where the loglikelihood does not behave appropriately for ML method of estimation.

The broad ridge going from West to East strongly suggests that $\alpha_{1,0}$ is not a meaningful parameter for values of $\alpha_{1,1}$ giving the highest likelihood, since every value of $\hat{\alpha}_{1,0}$ larger than -1.0 will provide the same maximum for loglikelihood function.

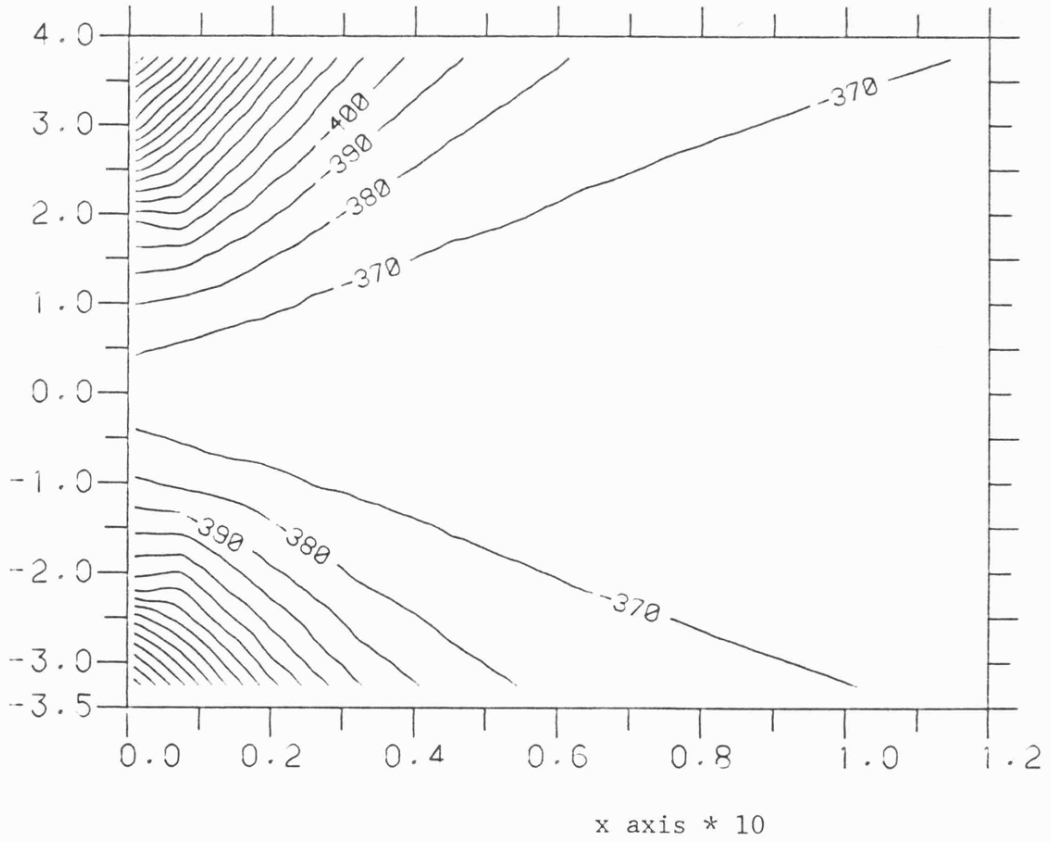


Figure 2.3- Loglikelihood values as a function of $\hat{\alpha}_{1,1}$ and $\hat{\alpha}_{1,0}$, using methods A or B to the ART on black women.

1.3- Cancer Knowledge

The data in Table 2.5 comes from a study on knowledge about cancer by Lombard and Doering (1947). Questions were asked about whether or not the following were sources of general information:

(1)radio (2)newspaper (3)solid reading (4)lectures

Table 2.5- Score distribution and results obtained by fitting a logit/probit model to the Lombard and Doering's data.

Response pattern	Observed frequency	Expected frequency	Total score	Component score
0000	477	467.37	0	0.00
1000	63	70.80	1	0.72
0001	12	16.62	1	0.77
0010	150	155.93	1	1.34
1001	7	3.10	2	1.49
1010	32	33.30	2	2.06
0011	11	7.98	2	2.11
1011	4	2.02	3	2.83
0100	231	240.52	1	3.40
1100	94	82.16	2	4.12
0101	13	20.29	2	4.16
0110	378	362.29	2	4.74
1101	12	8.51	3	4.89
1110	169	181.61	3	5.46
0111	45	46.04	3	5.51
1111	31	30.49	4	6.23
Total	1729	1729.00	-	-

χ^2 = 11.68 with 6 degrees of freedom ($0.05<p<0.10$)

Table 2.5 shows that these data are fitted reasonably well by a logit/probit model with one single latent variable as a measure of how well-informed a person is.

The scaling of the sample is not exactly the same when using the total and the component scores. This is due to the large value assumed by $\hat{\alpha}_{2,1}$ as showed in Table 2.6.

Table 2.6- Parameter estimates and asymptotic standard deviations from fitting a logit/probit model to the Lombard and Doering data.

Item i	$\hat{\alpha}_{i,1}$	$SE(\hat{\alpha}_{i,1})$	$\hat{\alpha}_{i,0}$	$SE(\hat{\alpha}_{i,0})$	$\hat{\pi}_i$
1	0.72	0.09	-1.29	0.06	0.22
2	3.40	1.14	0.60	0.17	0.64
3	1.34	0.17	-0.14	0.08	0.46
4	0.77	0.14	-2.70	0.18	0.06

The large value for the discriminating power of item 2 indicates that the newspaper has the largest effect on getting information about cancer. Its standard deviation, however, is relatively large. The difficulty parameter estimates range from 'popular source of information' (item 4) to 'not very popular' (item 2).

To carry out the analysis of the behaviour of the likelihood, we used 138 values for $(\hat{\alpha}_{2,0},\hat{\alpha}_{2,1})$, since $\hat{\alpha}_{2,1}$ is very large compared with the other parameter estimates $\hat{\alpha}_{i,1}$.

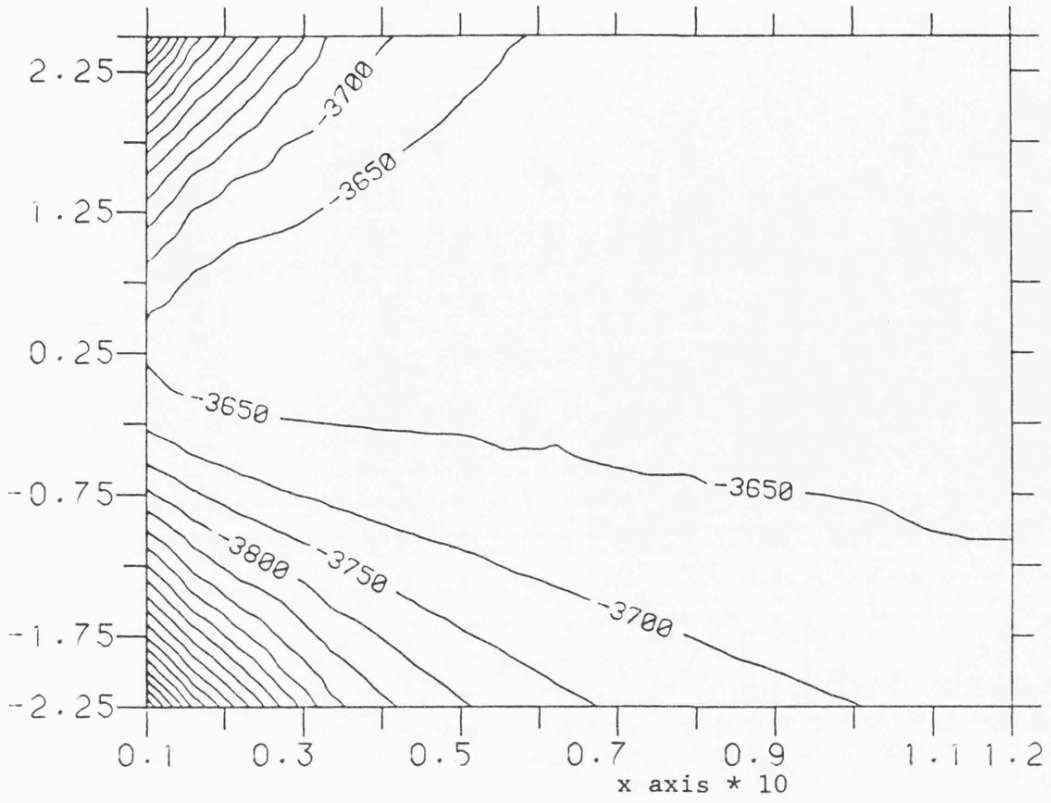


Figure 2.4- Loglikelihood values as a function of $\hat{\alpha}_{2,1}$ and $\hat{\alpha}_{2,0}$, using method B (profile) for the Lombard and Doering data.

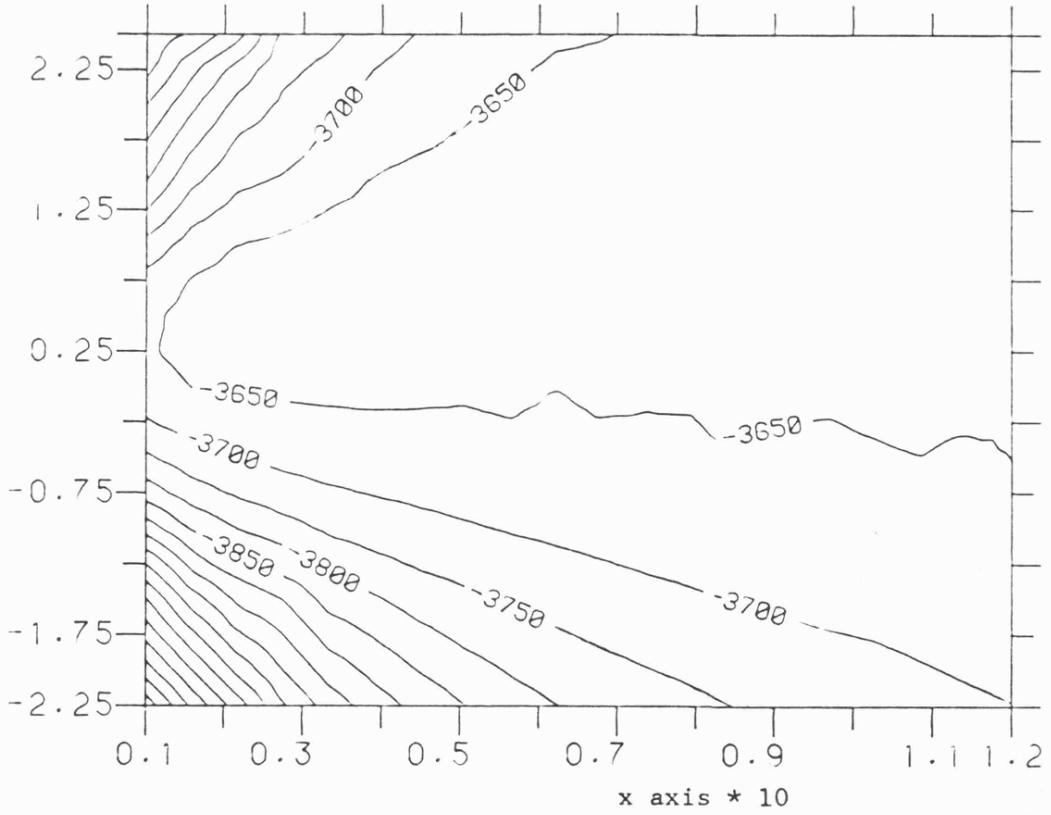


Figure 2.5- Loglikelihood values as a function of $\hat{\alpha}_{2,1}$ and $\hat{\alpha}_{2,0}$, using approximate method A for the Lombard and Doering data.

According to Table 2.6 the likelihood function assumes its maximum value when $\hat{\alpha}_{i,1}$=3.40 and $\hat{\alpha}_{i,0}$=0.60 for item 2. Both Figures 2.5 and 2.6 show that $\hat{\alpha}_{2,1}$ could be equal to any number bigger than 1.0 and the range of $\hat{\alpha}_{2,0}$ increases as $\hat{\alpha}_{2,1}$ also increases. As when analysing Figures 2.1 and 2.2, this happens because the likelihood values change very little for $\hat{\alpha}_{2,1}$ bigger than 3.40.

In this case too, methods A and B give the same information about the shape of the likelihood function, which does not seem suitable for the ML method.

Conclusion

We have compared 3 sets of data for which a logit/probit model with one latent variable seemed to fit reasonably well.

The results suggest that when one of the $\hat{\alpha}_{i,1}$ is large this probably indicates bad behaviour of the likelihood.

It is difficult to say exactly how large each $\hat{\alpha}_{i,1}$ can be before the ridge in the likelihood appears and the second observed derivatives or the information matrix are not good guides to the variability of this estimates.

There is strong evidence that we can use the approximate method A instead of the profile likelihood, since they give the same information about the behaviour of the likelihood function.

2- *Another Look at the Likelihood Function*

Working with the contoured likelihood is not always easy, since a lot of points are required and it is hard to see small changes in the likelihood values. It is useful to plot the shape of the likelihood function along the ridge that is evident in Figures 2.1 to 2.6. This corresponds to maximising the previously obtained loglikelihood values over $\alpha_{i,0}$. Using the data points $(\hat{\alpha}_{i,1},\hat{\alpha}_{i,0})$ from Figures 2.1 to 2.5, results are in the plotting points of Tables 2.7 to 2.9 and the likelihood functions in Figures 2.6 to 2.8.

Table 2.7-Maximum loglikelihood value over $\hat{\alpha}_{1,0}$, fixing $\hat{\alpha}_{1,1}$ to the ART on white women.

$\hat{\alpha}_{1,1}$	L_A	L_B
0.0	-601.37	-601.14
1.0	-592.14	-592.12
2.0	-594.59	-594.22
3.0	-598.28	-596.62
4.0	-601.03	-597.87
5.0	-602.99	-598.51
6.0	-604.17	-598.85
7.0	-604.96	-599.06
8.0	-605.50	-599.19
9.0	-605.85	-599.28
10.0	-605.99	-599.33
11.0	-606.13	-599.37

Table 2.8-Maximum loglikelihood value over $\hat{\alpha}_{1,0}$, fixing $\hat{\alpha}_{1,1}$ to the ART on black women.

$\hat{\alpha}_{1,1}$	L_A	L_B
0.0	-368.08	-367.48
1.0	-365.66	-365.33
2.0	-365.01	-364.90
3.0	-364.83	-364.78
4.0	-364.77	-364.74
5.0	-364.74	-364.72
6.0	-364.72	-364.71
7.0	-364.71	-364.70
8.0	-364.71	-364.69
9.0	-364.70	-364.69
10.0	-364.70	-364.69
11.0	-364.70	-364.69

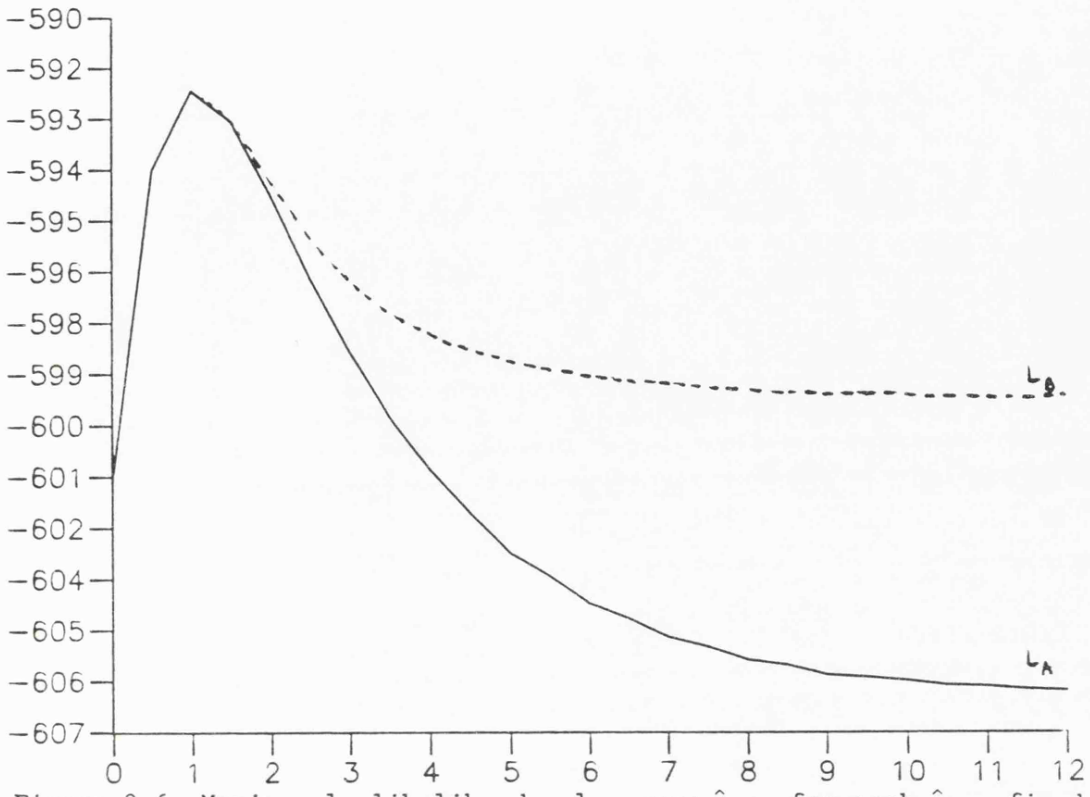


Figure 2.6- Maximum loglikelihood value over $\hat{\alpha}_{1,0}$ for each $\hat{\alpha}_{1,1}$ fixed, for the ART on white women presented in Table 2.7.

Table 2.9- Maximum loglikelihood value over $\hat{\alpha}_{2,0}$, fixing $\hat{\alpha}_{2,1}$ to the Lombard and Doering data.

$\hat{\alpha}_{2,1}$	L_A	L_B	$\hat{\alpha}_{2,1}$	L_A	L_B
0.1	-3758.59	-3755.13	6.0	-3624.05	-3622.90
1.0	-3656.02	-3645.49	7.0	-3624.71	-3623.17
2.0	-3637.84	-3625.53	8.0	-3625.10	-3623.24
3.0	-3622.71	-3622.47	9.0	-3625.29	-3623.27
4.0	-3622.68	-3622.52	10.0	-3625.47	-3623.31
5.0	-3623.54	-3622.80	11.0	-3625.62	-3623.39

Figure 2.6 shows that both methods give approximately the same loglikelihood values for $\hat{\alpha}_{1,1}$ smaller than 2, increasing up to $\hat{\alpha}_{1,1}$=1.04 (ML estimate) and decreasing faster when using method A than the profile likelihood. This result agrees with our analysis of Figures 2.1 and 2.2.

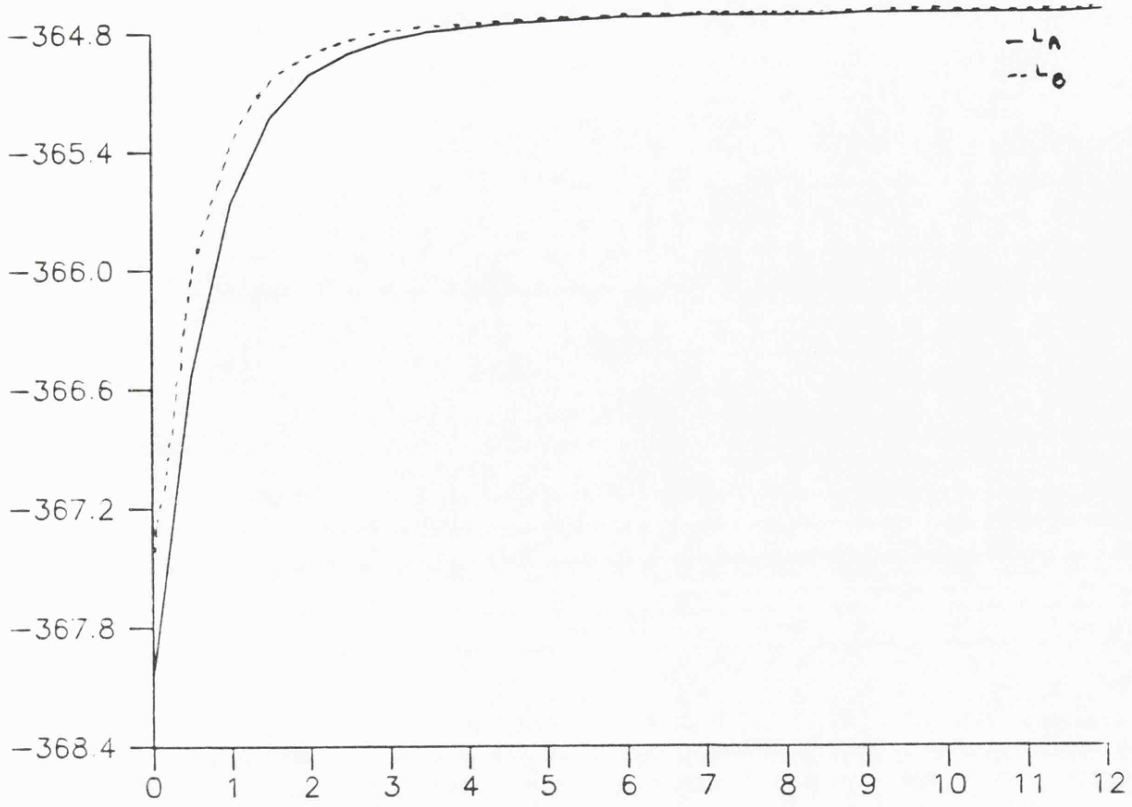


Figure 2.7- Maximum loglikelihood value over $\hat{\alpha}_{1,0}$ for each $\hat{\alpha}_{1,1}$ fixed, to the ART on black women, presented in Table 2.8.

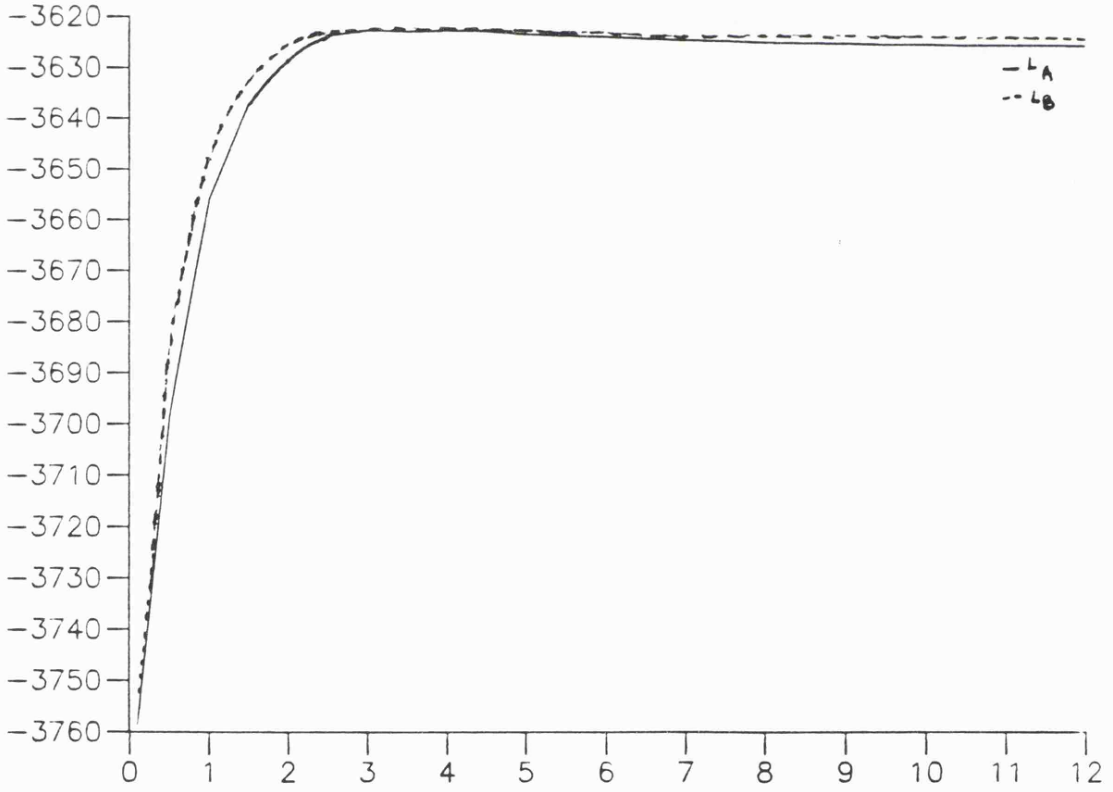


Figure 2.8- Maximum loglikelihood value over $\hat{\alpha}_{2,0}$ for each $\hat{\alpha}_{2,1}$ fixed, to the Lombard and Doering data, presented in Table 2.9.

As in the three dimensional graph, Figures 2.7 and 2.8 confirm that both methods give roughly the same information about the behaviour of the likelihood. Inspection of the data in Table 2.8 shows that the likelihood continues increasing, while in Table 2.9 the likelihood assumes a maximum value, but after that decreases so slightly that the change is insignificant when plotting the data.

Plotting the results for all items

Since approximate method A followed by a simple plot is easy to apply, we shall look at the shape of the likelihood for all items, instead of only one, for the ART on white and black women, and the Lombard and Doering data.

Table 2.10- Maximum loglikelihood value, $L_A(i)$, over $\hat{\alpha}_{i,0}$, fixing $\hat{\alpha}_{i,1}$, i=1,2..,4, to the ART on white women, using approximate method A.

$\hat{\alpha}_{i,1}$	$L_A(1)$	$L_A(2)$	$L_A(3)$	$L_A(4)$
0.0	-601.37	-603.57	-601.63	-606.09
1.0	-592.14	-592.22	-592.05	-592.74
2.0	-595.06	-593.84	-595.18	-592.68
3.0	-598.28	-595.98	-599.43	-595.06
4.0	-601.22	-598.18	-602.58	-596.31
5.0	-602.88	-599.46	-604.71	-597.47
6.0	-604.17	-600.42	-606.19	-598.33
7.0	-604.90	-600.99	-607.26	-598.71
8.0	-605.50	-601.37	-608.07	-599.13
9.0	-605.77	-601.66	-608.71	-599.27
10.0	-606.10	-601.79	-609.22	-599.44
11.0	-606.13	-601.93	-609.63	-599.51

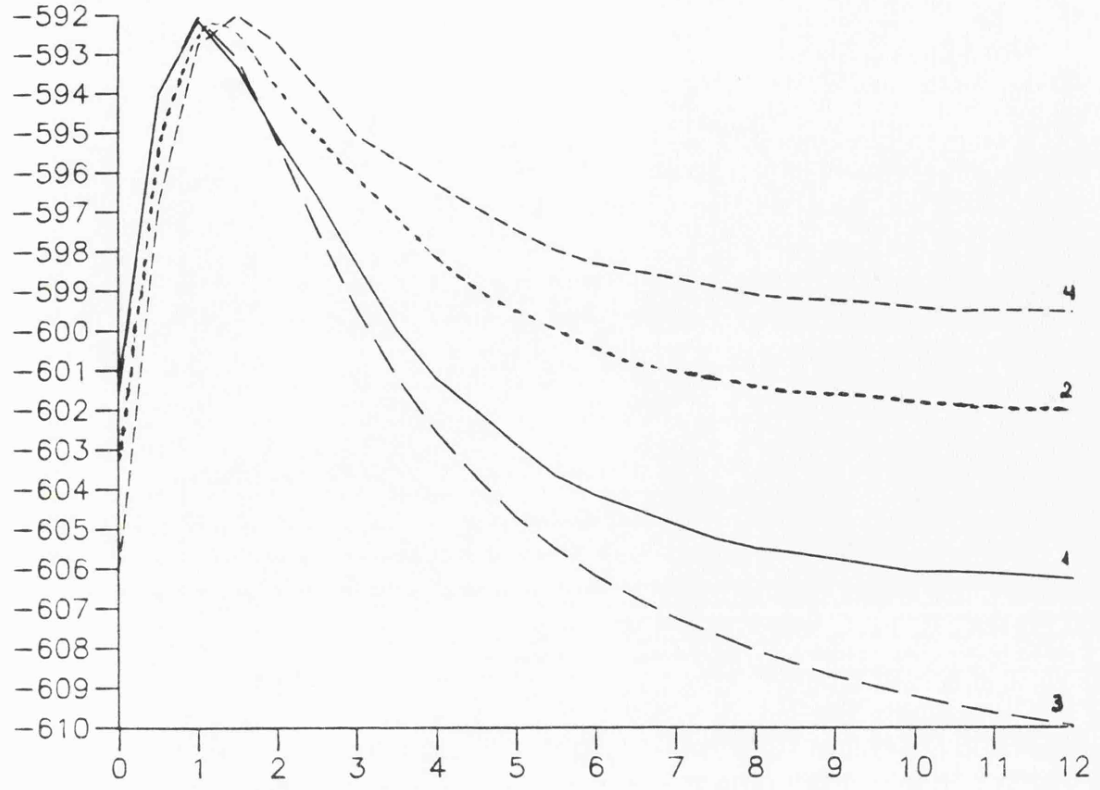


Figure 2.9- Maximum loglikelihood value over $\hat{\alpha}_{i,0}$ for each $\hat{\alpha}_{i,1}$ fixed, i=1,...,4, to the ART on white women, presented in Table 2.10, using approximate method A.

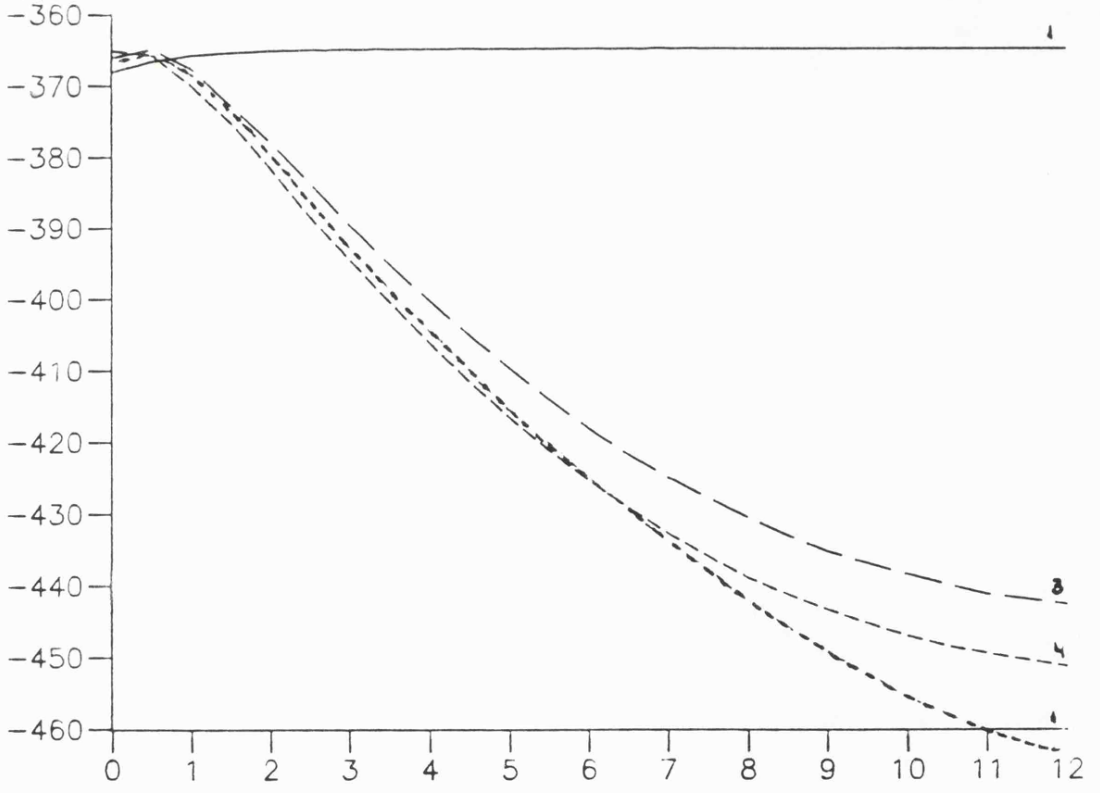


Figure 2.10- Maximum loglikelihood value over $\hat{\alpha}_{i,0}$ for each $\hat{\alpha}_{i,1}$ fixed, i=1,...,4, to the ART on black women, presented in Table 2.11, using approximate method A.

Table 2.11- Maximum loglikelihood values, $L_A(i)$, over $\hat{\alpha}_{i,0}$ fixing $\hat{\alpha}_{i,1}$, i=1,2,..,4, to the ART on black women, using approximate method A.

$\hat{\alpha}_{i,1}$	$L_A(1)$	$L_A(2)$	$L_A(3)$	$L_A(4)$
0.0	-368.08	-366.79	-366.01	-365.07
1.0	-365.66	-368.07	-367.59	-369.97
2.0	-365.01	-379.56	-377.88	-381.54
3.0	-364.83	-392.35	-389.50	-394.39
4.0	-364.77	-404.22	-400.20	-406.07
5.0	-364.74	-415.00	-409.66	-416.58
6.0	-364.72	-424.78	-418.04	-425.29
7.0	-364.71	-433.57	-424.92	-432.68
8.0	-364.71	-441.96	-430.51	-438.91
9.0	-364.70	-449.16	-435.21	-443.32
10.0	-364.70	-455.29	-438.39	-446.98
11.0	-364.70	-460.03	-441.16	-449.43

Table 2.12- Maximum loglikelihood values, $L_A(i)$, over $\hat{\alpha}_{i,0}$ fixing $\hat{\alpha}_{i,1}$, i=1,2,..,4, to the Lombard and Doering data, using the approximate method A.

$\hat{\alpha}_{i,1}$	$L_A(1)$	$L_A(2)$	$L_A(3)$	$L_A(4)$
0.0	-3666.35	-3790.44	-3755.44	-3640.81
1.0	-3626.13	-3660.79	-3630.09	-3624.84
2.0	-3680.93	-3627.58	-3635.98	-3650.53
3.0	-3739.91	-3622.71	-3666.38	-3682.70
4.0	-3785.70	-3623.91	-3693.62	-3710.37
5.0	-3818.87	-3623.95	-3714.12	-3730.04
6.0	-3842.86	-3624.05	-3728.73	-3743.71
7.0	-3859.40	-3624.97	-3740.07	-3752.89
8.0	-3871.61	-3625.30	-3749.21	-3759.02
9.0	-3881.08	-3625.29	-3756.80	-3763.18
10.0	-3886.42	-3625.97	-3762.16	-3765.29
11.0	-3890.69	-3625.62	-3766.36	-3766.72

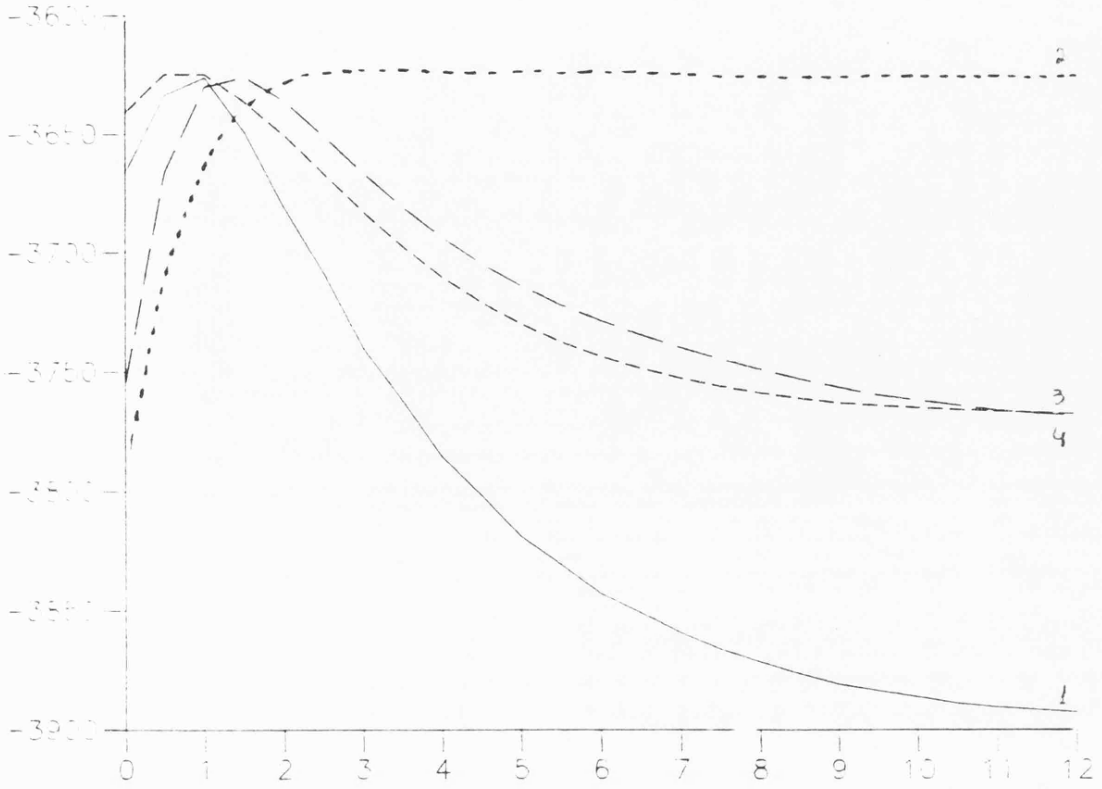


Figure 2.11- Maximum loglikelihood value over $\hat{\alpha}_{i,0}$ for each $\hat{\alpha}_{i,1}$ fixed, i=1,...,4, for the Lombard and Doering data, presented in Table 2.12, using approximate method A.

Figure 2.9 shows that whichever item we choose, all items are well-behaved. However it is interesting to point out that the order of the curves is inversely related to size of the $\hat{\alpha}_{i,1}$,i=1,...,4, since here they all have the same coefficient of variation (0.31).

As we can see in Figure 2.10, the bad behaviour of the likelihood is indicated by item 1, with very large $\hat{\alpha}_{1,1}$ and its large standard deviation. Item 2 and 3 present very similar $\hat{\alpha}_{i,1}$ (0.38 and 0.37), with coefficient of variation 0.58 and 0.65, respectively, but the latter loglikelihood decreases slowly. The value of $\hat{\alpha}_{4,1}$ is half the size of item 2 or 3, but with a large coefficient of variation (1.26).

Working through the values of $\hat{\alpha}_{i,1}$ and Figure 2.11 we see that $\hat{\alpha}_{3,1}$ is bigger than $\hat{\alpha}_{4,1}$ (1.34 and 0.77), but the former has a smaller coefficient of variation, and both items give approximately the same likelihood shape. Item 1 has the smallest $\hat{\alpha}_{1,1}$ and the smallest coefficient of variation (0.12) and the biggest likelihood function decrease, while item 2 has a large value for $\hat{\alpha}_{i,1}$ and large coefficient of variation (0.34) and effectively its likelihood function never decreases.

Conclusions

These results suggest that there is strong evidence that we can look at the behaviour of the likelihood function by the approximate method A, using a graph like those in Figures 2.6 to 2.8. However, we should remember that the likelihood values from this method are equal or smaller than the real values and small decreases in the likelihood function should actually be still smaller.

Finally we can conclude that large discriminating power values ($\hat{\alpha}_{i,1}$) and large standard deviation point to bad likelihood behaviour. The results also indicate that for the same test the shapes of the approximate profile likelihoods obtained for different items i are related to the size of $\hat{\alpha}_{i,1}$ and its coefficient of variation.

3- *Reparametrization*

The investigation of the behaviour of the likelihood function that has been carried out suggests that, at least for the ART on black women (Table 2.3) and the Lombard and Doering data (Table 2.5) a reparametrization is necessary.

We have worked through many reparametrizations, as for example,

$$\hat{\alpha}^{*}_{i,1} = \hat{\alpha}_{i,1} / (1 + \exp(\hat{\alpha}_{i,1}))$$

$$\hat{\alpha}^{*}_{i,1} = 1 / \hat{\alpha}_{i,1}$$

$$\hat{\alpha}^{*}_{i,1} = \hat{\alpha}_{i,1} / (1 + \hat{\alpha}^{2}_{i,1})^{\frac{1}{2}}$$

$$\hat{\alpha}^{*}_{i,0} = \hat{\alpha}_{i,0} / (1 + \hat{\alpha}^{2}_{i,1})^{\frac{1}{2}}$$

$$\hat{\alpha}^{*}_{i,0} = - \hat{\alpha}_{i,0} / \hat{\alpha}_{i,1}$$

where i=1 for the ART on black women data and i=2 for the Lombard and Doering data.

We shall present the results just for the reparametrizations that gave useful results, in the sense that it showed better behaviour of the likelihood function, that is, for

$$\hat{\alpha}_{i,0} = \hat{\alpha}^{*}_{i,0} / (1 + \hat{\alpha}^{2}_{i,1})^{\frac{1}{2}} \quad \text{and} \quad \hat{\alpha}^{*}_{i,1} = \hat{\alpha}_{i,1} / (1 + \hat{\alpha}^{2}_{i,1})^{\frac{1}{2}}$$

using the profile and the approximate methods (B and A, respectively).

3.1- Arithmetic Reasoning Test on White Women

The data related to this example are presented in Table 2.1.

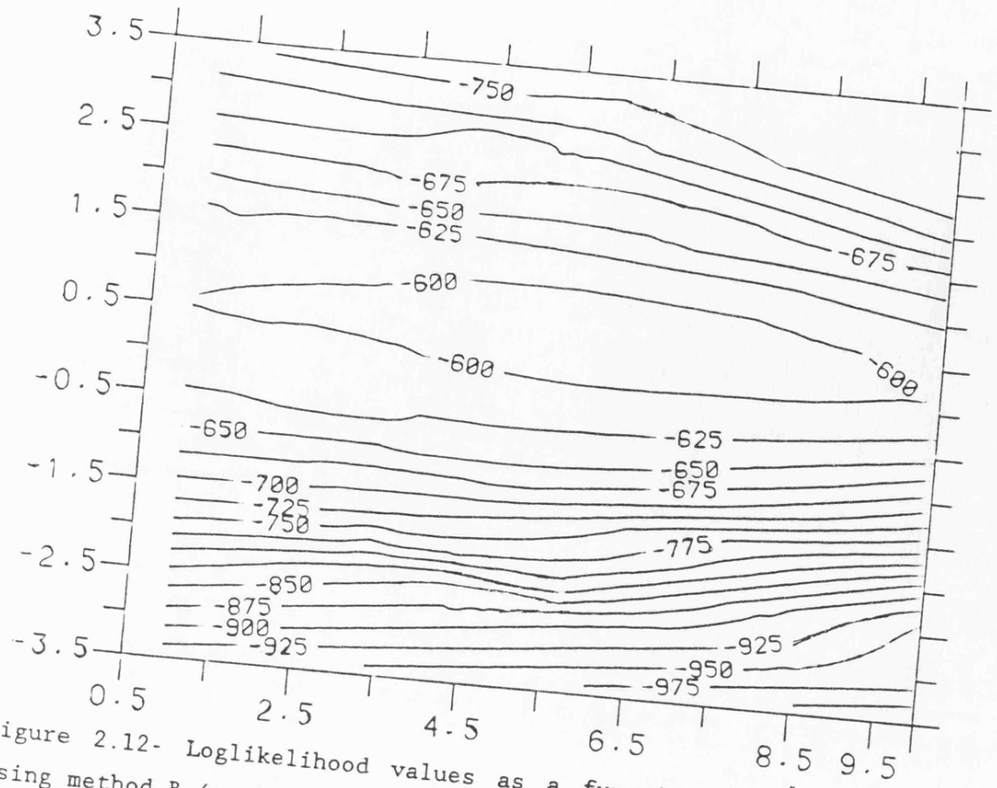


Figure 2.12- Loglikelihood values as a function of $\hat{\alpha}^*_{1,1}$ and $\hat{\alpha}^*_{1,0}$ using method B (profile) to the ART on white women.

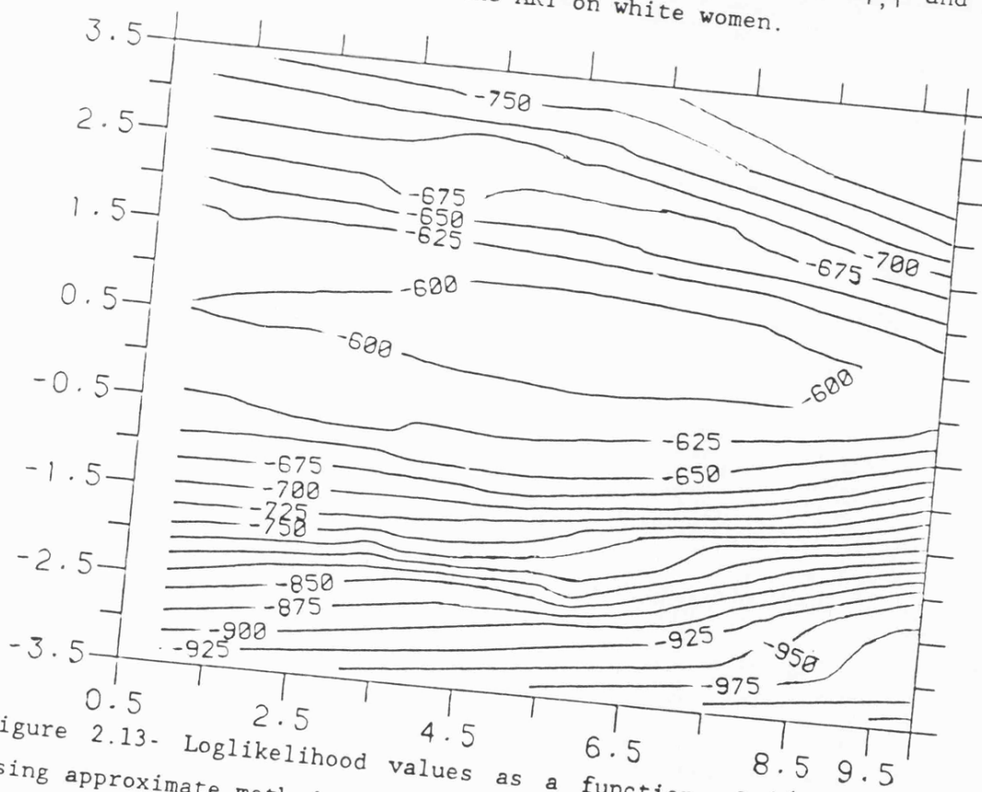


Figure 2.13- Loglikelihood values as a function of $\hat{\alpha}^*_{1,1}$ and $\hat{\alpha}^*_{1,0}$ using approximate method A to the ART on white women.

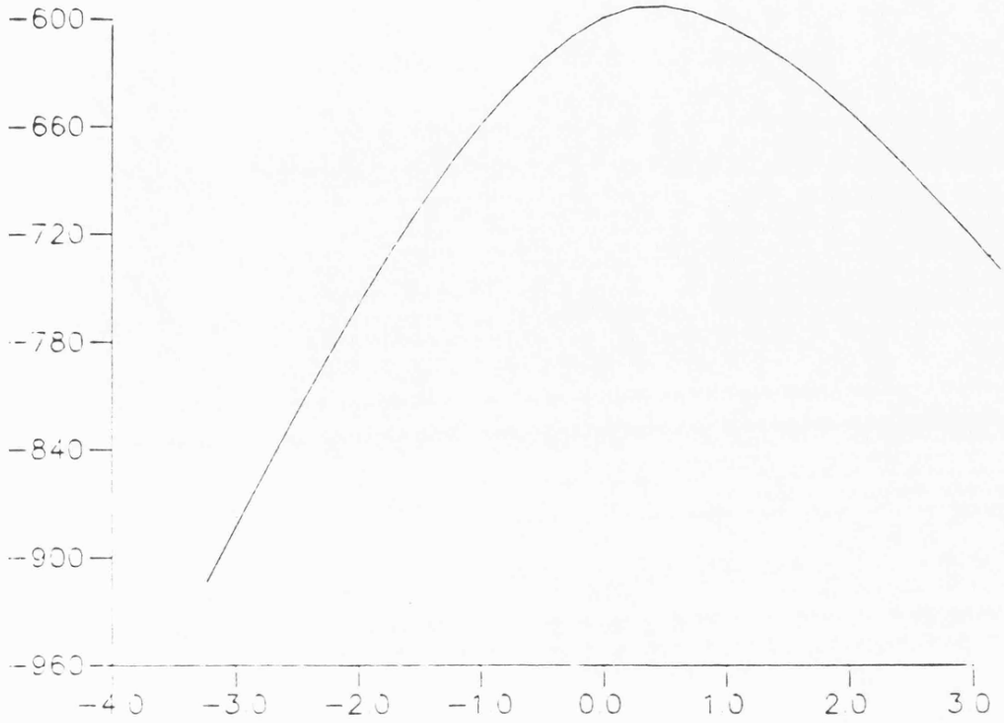


Figure 2.14- Maximum loglikelihood value over $\hat{\alpha}_{1,,1}$ for each $\hat{\alpha}^{*}_{1,,0}$ fixed to the ART on white women, using methods A and B.

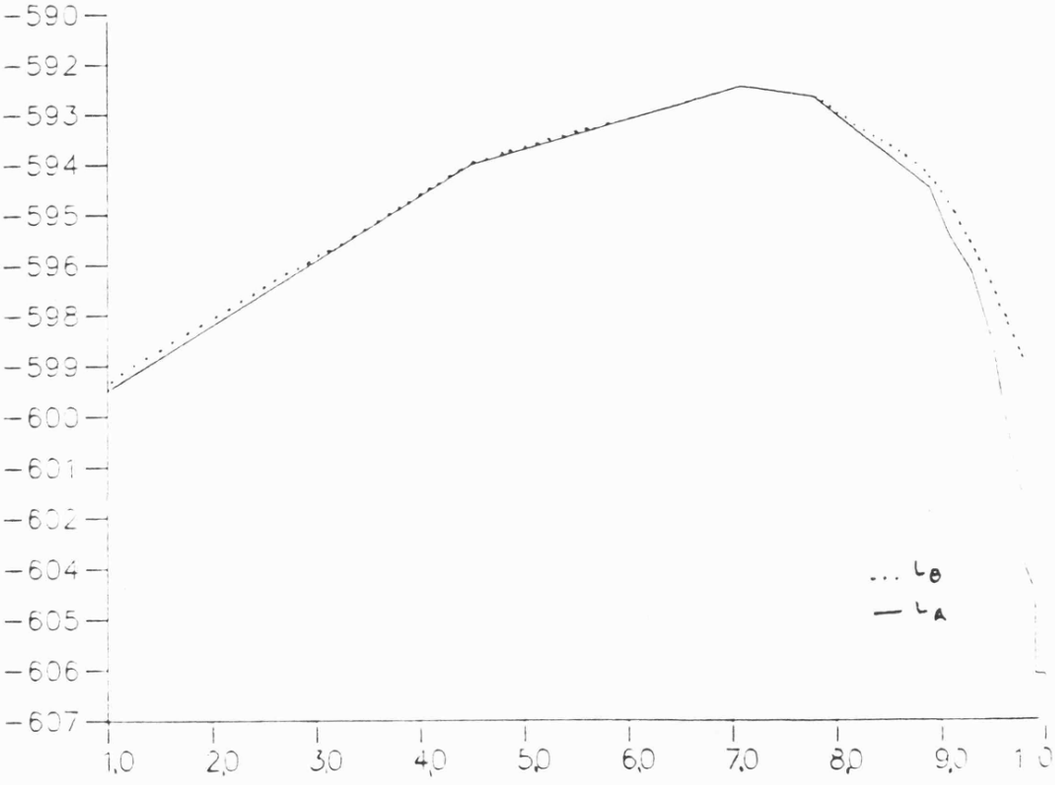


Figure 2.15- Maximum loglikelihood value over $\hat{\alpha}_{1,,0}$ for each $\hat{\alpha}^{*}_{1,,1}$ fixed to the ART on white women, using methods A and B.

Both Figures 2.12 and 2.13 show the same shape of the likelihood function and their parallel and almost horizontal lines indicate that the values of the loglikelihood almost do not change for a fixed $\hat{\alpha}^*_{1,0}$ over all range of $\hat{\alpha}^*_{1,1}$. There is a peak inside the ellipse, although the contouring does not show the small differences in the loglikelihood values. We can see it in Figure 2.15, where we have the maximum loglikelihood values over $\hat{\alpha}^*_{1,0}$ for each $\hat{\alpha}^*_{1,1}$ fixed.

That only one line represents the behaviour of the likelihood function in Figure 2.14 is due to the fact that methods A and B give the same results for all values assumed by $\hat{\alpha}^*_{1,0}$.

From Figures 2.14 and 2.15 we can see that the loglikelihood function behaves well in both reparametrizations and that the maximum loglikelihood values for $\hat{\alpha}^*_{1,1}$ range in a larger interval than for $\hat{\alpha}^*_{1,0}$, since maximum loglikelihood $\hat{\alpha}^*_{1,1} \in (-916.11;-592.27)$ while the maximum loglikelihood $\hat{\alpha}^*_{1,0} \in (-606.23;-592.14)$.

Comparing Figures 2.1, 2.2 and 2.6 with 2.12 to 2.15 we can conclude that the the reparametrization $\hat{\alpha}^*_{1,0}$ and $\hat{\alpha}^*_{1,1}$ give a likelihood function with better behaviour than $\hat{\alpha}_{1,0}$ and $\hat{\alpha}_{1,1}$.

3.2- Arithmetic Reasoning Test on Black Women

The following graphs refer to the data in Table 2.3.

As in the first example, the Figures 2.16 and 2.17, 2.18 and 2.19 shows the same shape for the likelihood function, whether using profile (method B) or approximate method A.

Figures 2.16 and 2.17 show that the parallel lines are becoming horizontal as the loglikelihood function approximates to the maximum value, where we can see a broad bridge going from West to East and $\hat{\alpha}^*_{1,1}$ assuming all values while $\hat{\alpha}^*_{1,0}$ ranges from -0.35 to 0.35.

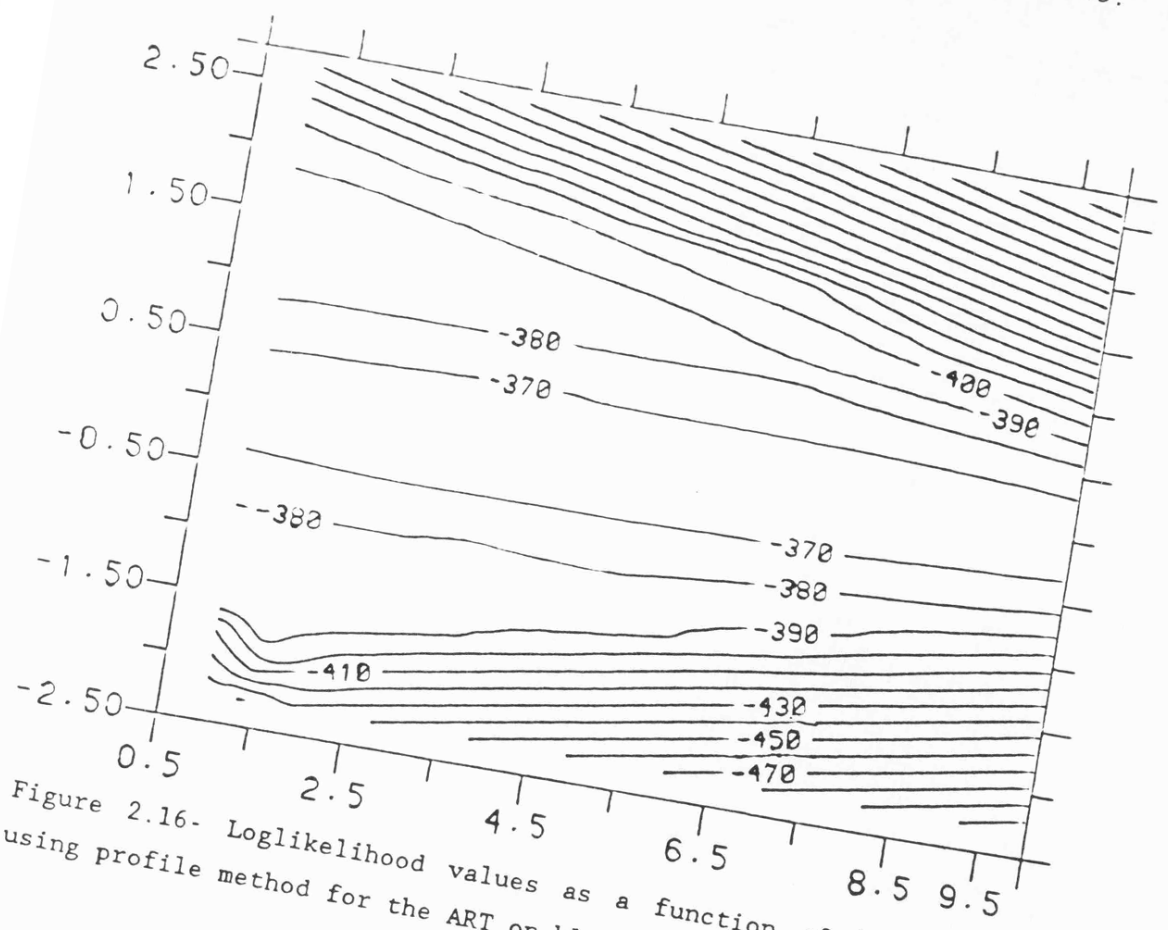


Figure 2.16- Loglikelihood values as a function of $\hat{\alpha}^*_{1,1}$ and $\hat{\alpha}^*_{1,0}$ using profile method for the ART on black women.

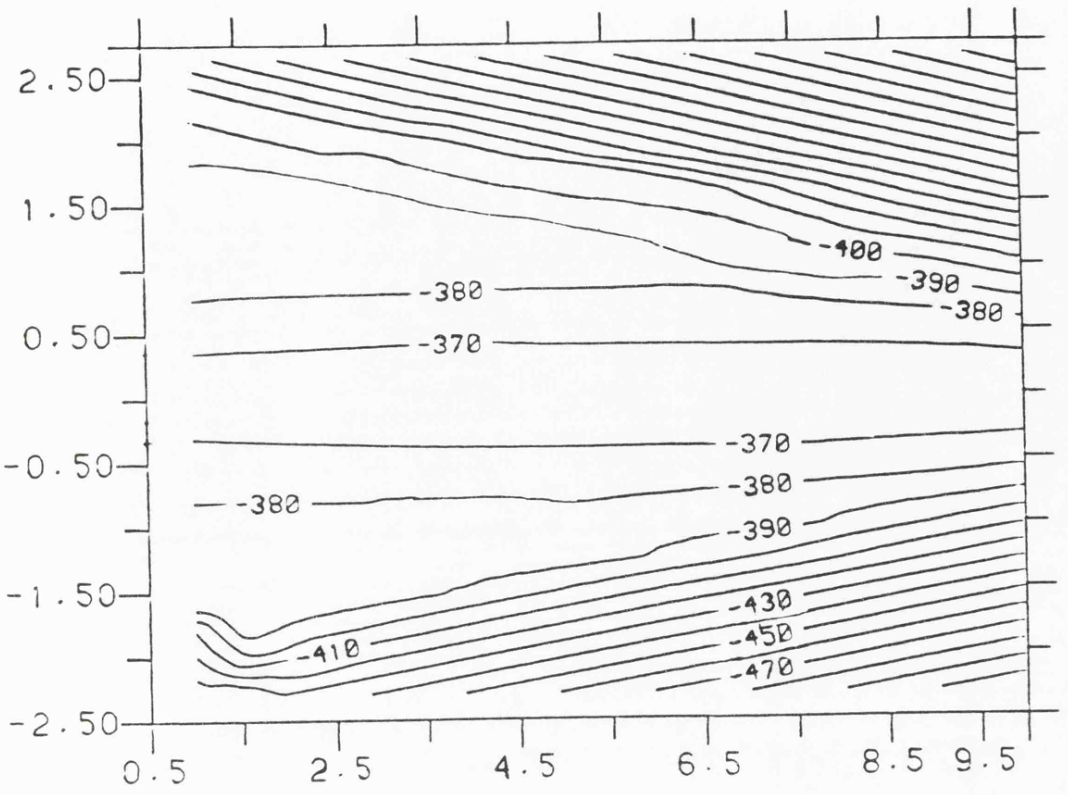


Figure 2.17- Loglikelihood values as a function of $\hat{\alpha}^*_{1,1}$ and $\hat{\alpha}^*_{1,0}$, using approximate method A for the ART on black women.

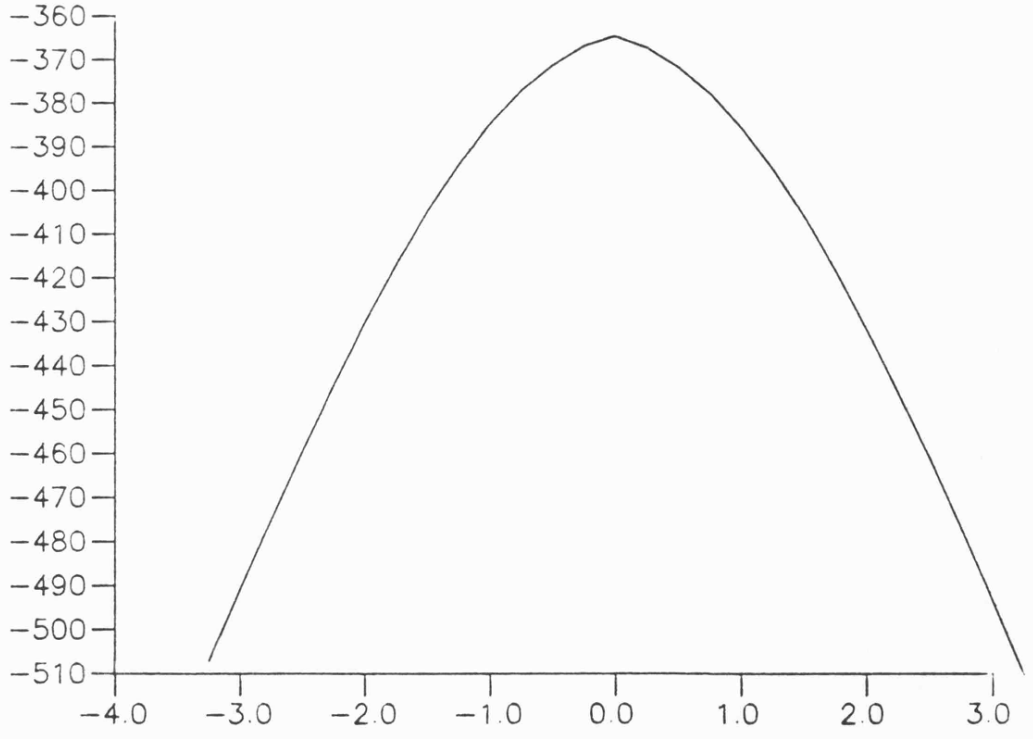


Figure 2.18- Maximum loglikelihood value over $\hat{\alpha}_{1,1}$ for each $\hat{\alpha}^*_{1,0}$ fixed for the ART on black women, using methods A and B.

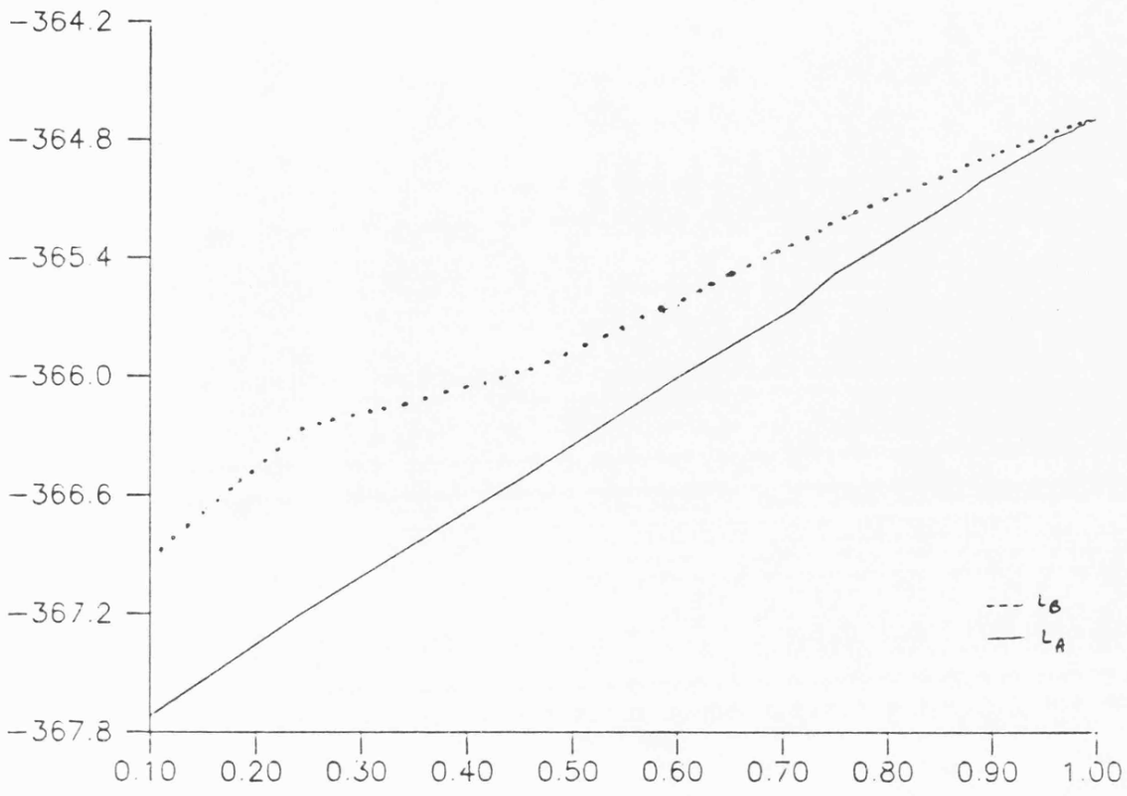


Figure 2.19- Maximum loglikelihood value over $\hat{\alpha}_{1,0}$ for each $\hat{\alpha}^*_{1,1}$ fixed for the ART on black women, using methods A and B.

The apparent increased likelihood function shown in Figure 2.19 is,actually, almost constant since it assumes values in a small interval (in the profile method from -367.48 to -364.68 and in the approximate method from -368.08 to -364.69), corresponding to an increase of 0.9%. Thus the reparametrization $\hat{\alpha}^*_{1,1}$ provides a likelihood function that is monotone increasing.

On the other hand, Figure 2.18 indicates that the reparametrization

$\hat{\alpha}^*_{1,0} = \hat{\alpha}_{1,0}/(1+\hat{\alpha}^2_{1,1})^{\frac{1}{2}}$ works very well, since the likelihood function is unimodal, assuming values from -510.35 to -364.68 in both methods (profile and approximate).

3.3- Cancer Knowledge

This example corresponds to the Lombard and Doering data(Table 2.5).

The small difference between methods A and B (Figures 2.20 and 2.21) is because the loglikelihood function for $\hat{\alpha}_{2,1} < 1.1$ in the profile method is bigger than in the approximate method.

The behaviour of the likelihood function after reparametrization in these example is very similar to the former one.

Although Figure 2.23 seems to show an increased loglikelihood function for $\hat{\alpha}^*_{2,1}$, it is almost constant, since it ranges from -3758.59 to -3755.13 which represents a small increase of 3.8%. Therefore the reparametrization $\hat{\alpha}^*_{2,1}$ provides a likelihood function that is monotone increasing.

As in the preceding example, the only useful reparametrization is given by $\hat{\alpha}^*_{2,0}$, as we can see in Figure 2.22 an unimodal likelihood function that assumes values between -5813.38 and -3625.14.

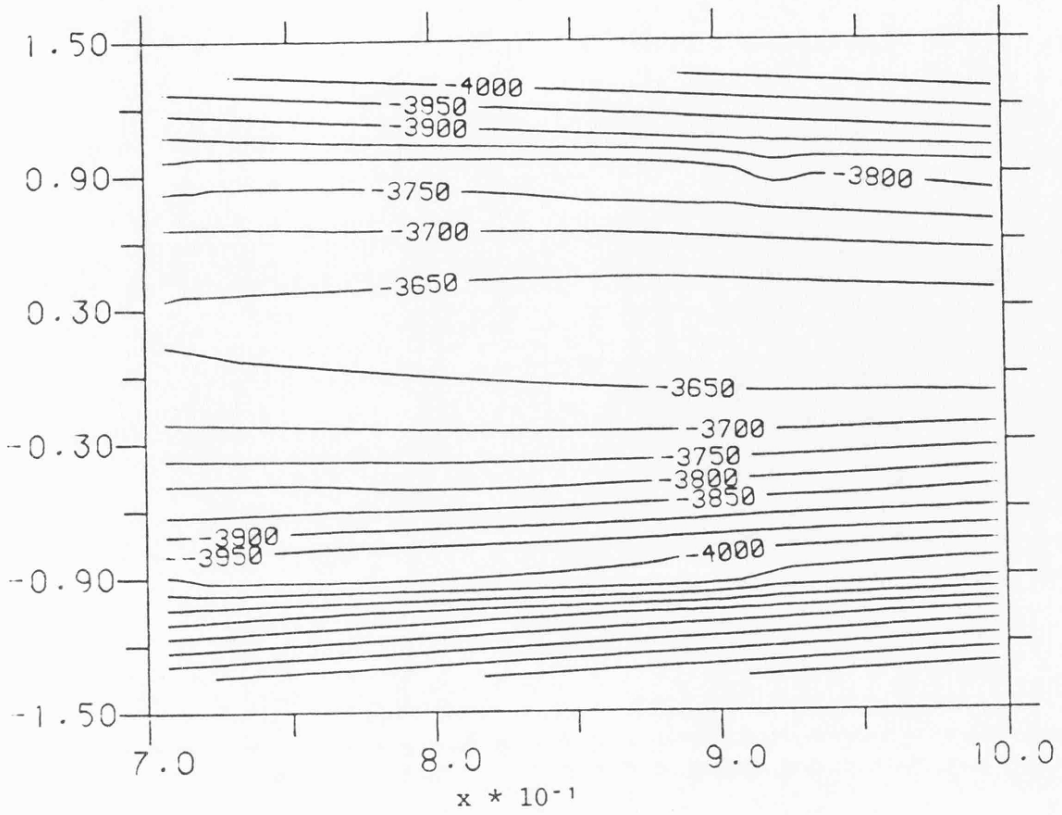


Figure 2.20- Loglikelihood values as a function of $\hat{\alpha}^{*}_{2,1}$ and $\hat{\alpha}^{*}_{2,0}$, using profile method for the Lombard and Doering data.

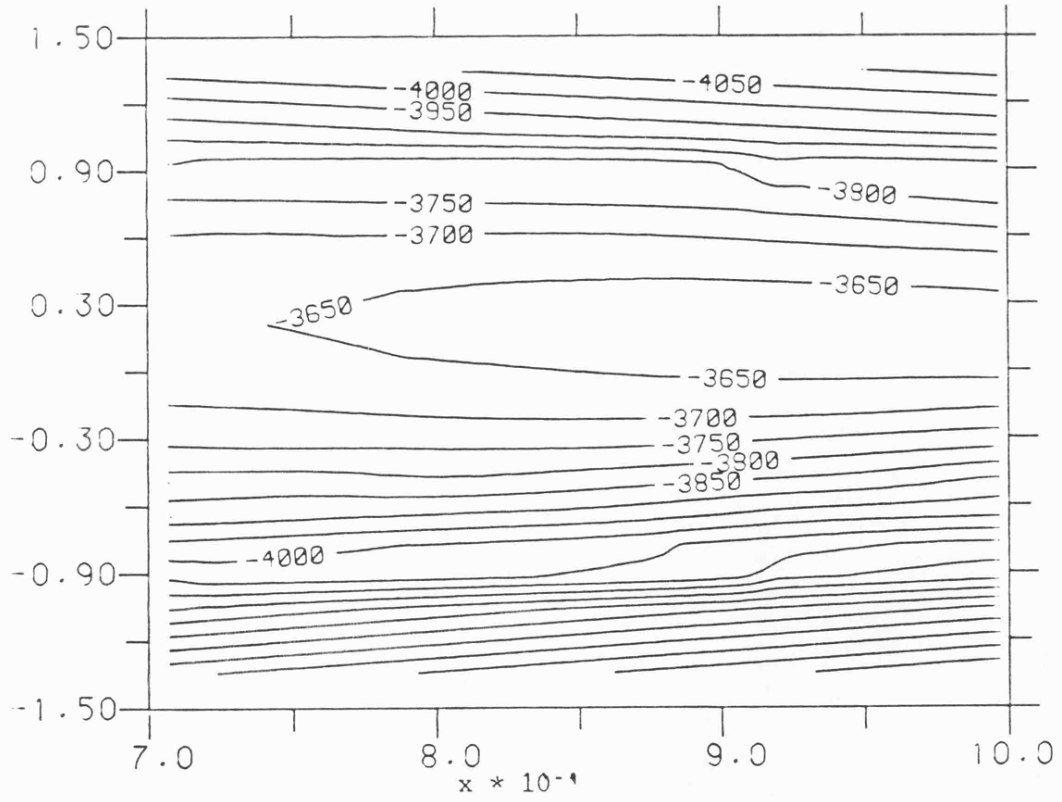


Figure 2.21- Loglikelihood values as a function of $\hat{\alpha}^{*}_{2,1}$ and $\hat{\alpha}^{*}_{2,0}$, using approximate method A for the Lombard and Doering data.

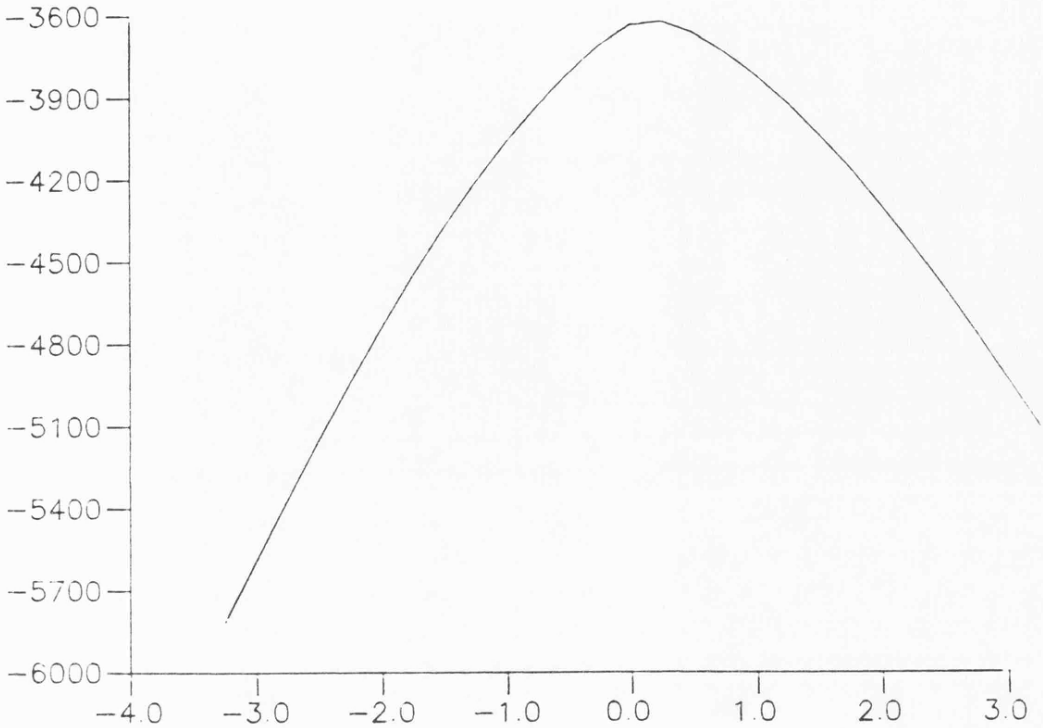


Figure 2.22- Maximum loglikelihood value over $\hat{\alpha}_{2,1}$ for each $\hat{\alpha}^{*}_{2,0}$ fixed to the Lombard and Doering data, using methods A and B.

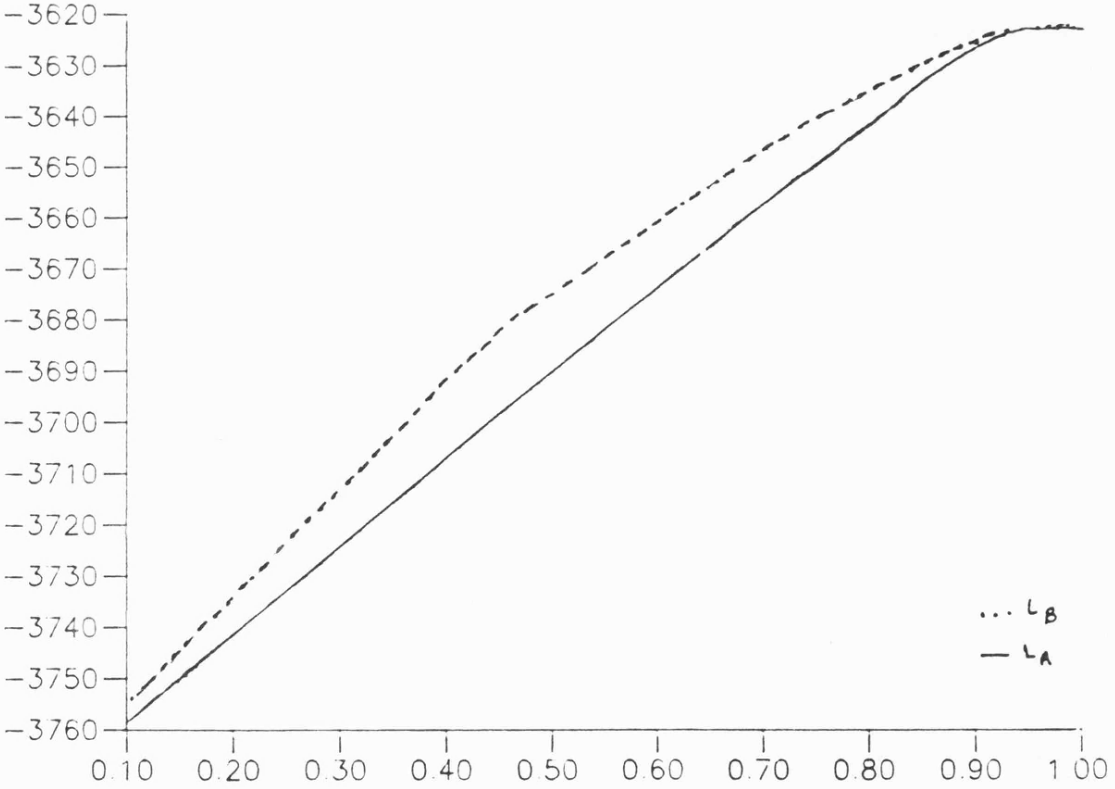


Figure 2.23- Maximum loglikelihood value over $\hat{\alpha}_{2,0}$ for each $\hat{\alpha}^{*}_{2,1}$ fixed to the Lombard and Doering data, using methods A and B.

Conclusion

The search of better behaviour for the likelihood function through reparametrization indicated that the only one that works well for the 3 different sets of data is given by

$$\hat{\alpha}^{*}_{i,0} = \hat{\alpha}_{i,0} / (1 + \hat{\alpha}^{2}_{i,1})^{\frac{1}{2}}$$

where i=1 for the ART on white and black women data and i=2 for the Lombard and Doering data.

Interpretation of $\alpha^{*}_{i,0}$

The reparametrization $\alpha^{*}_{i,0} = \hat{\alpha}_{i,0} / (1 + \hat{\alpha}^{2}_{i,1})^{\frac{1}{2}}$ corresponds to the probit of the expected value of $\Phi(\alpha_{i,0}+\alpha_{i,1}z)$, the response function of a probit model, i.e.,

$$\alpha^{*}_{i,0} = \Phi^{-1} \{ E (\Phi(\alpha_{i,0} + \alpha_{i,1}z)) \}.$$

For convenience, let us consider

$$\alpha_{i,0} = a \quad \text{and} \quad \alpha_{i,1} = b$$

Then

$$E(\Phi(a+bz)) = \int_{-\infty}^{\infty} \Phi(a+bz) \, (2\pi)^{-\frac{1}{2}} \exp(-\tfrac{1}{2} z^{2}) \, dz$$

If we take

$$bz = u \quad \text{and} \quad bdz=du$$

Then

Then

$$E(\Phi(a+bz)) = \int_{-\infty}^{\infty} \Phi(a+u)\ (2\pi)^{-\frac{1}{2}}\ b^{-1}\ \exp(-\tfrac{1}{2}\ u^2\ b^{-2})du$$

$$= \int_{-\infty}^{\infty} P(Z-u \leqslant a)\ (\text{density for } W \sim N(0,b^2) \text{ at } u)du$$

$$= P(Z+W \leqslant a), \quad Z+W \sim N(0,1+b^2)$$

and therefore

$$E(\Phi(a+bz)) = \Phi(\ a/(1+b^2)^{\frac{1}{2}}\ )$$

or

$$a/(1+b^2)^{\frac{1}{2}} = \Phi^{-1}\{E(\Phi(a+bz))\}.$$

Chapter 3

ADEQUACY of the ASYMPTOTIC VARIANCE-COVARIANCE MATRIX using BOOTSTRAP and JACKKNIFE TECHNIQUES

1- *Introduction*

The aim of this chapter is to investigate the adequacy of the asymptotic variance-covariance matrix in latent trait analysis for binary data through the jackknife and bootstrap techniques described in Chapter 1, sections 9.1 and 9.2. This investigation will be carried out using 5 examples, the three sets of data that we have worked in Chapter 2 and two, which will be introduced in this chapter, for the logit/probit model. These examples have a good range of different patterns of parameter estimates and sample sizes.

We shall compare the bootstrap, jackknife and the original ML parameter estimates for $\hat{\alpha}_{i,1}$, $\hat{\alpha}_{i,0}$ and $\hat{\alpha}^{*}_{i,0} = \hat{\alpha}_{i,0} / (1 + \hat{\alpha}^{2}_{i,1})^{\frac{1}{2}}$, $i=1,\ldots,4$, their variability and how close is the bootstrap distribution of these parameter estimates and their jackknife pseudovalues to a Normal distribution.

We shall look also at the jackknife and bootstrap bias of the parameter estimates $\hat{\alpha}_{i,1}$, $i=1,\ldots,4$, as defined in Chapter 1, sections 9.1 and 9.2.

We shall look at standard deviations and correlations between the parameter estimates, rather than covariances.

The asymptotic covariance matrix will be obtained in two different ways: from both the observed second derivative matrix and the information matrix, and they will be also compared.

The standard deviation of the ML parameter estimate $\hat{\alpha}^*_{i,0}$ is obtained from the following approximation:

$$\mathrm{Var}(\alpha^*_{i,0}) \approx \left[\frac{\partial \alpha^*_{i,0}}{\partial \alpha_{i,0}} \right]^2 \mathrm{Var}(\alpha_{i,0}) + 2 \frac{\partial \alpha^*_{i,0}}{\partial \alpha_{i,0}} \frac{\partial \alpha^*_{i,0}}{\partial \alpha_{i,1}} \mathrm{Cov}(\alpha_{i,0},\alpha_{i,1}) +$$

$$+ \left[\frac{\partial \alpha^*_{i,0}}{\partial \alpha_{i,1}} \right]^2 \mathrm{Var}(\alpha_{i,1}).$$

Therefore

$$\mathrm{Var}(\alpha^*_{i,0}) \approx \frac{1}{1 + \alpha^2_{i,0}} \mathrm{Var}(\alpha_{i,0}) - 2 \frac{\alpha_{i,0}\,\alpha_{i,1}}{(1 + \alpha^2_{i,1})^2} \mathrm{Cov}(\alpha_{i,0},\alpha_{i,1}) +$$

$$+ \frac{(\alpha_{i,0}\,\alpha_{i,1})^2}{(1 + \alpha^2_{i,1})^3} \mathrm{Var}(\alpha_{i,1}),$$

when $\alpha_{i,1}$ and $\alpha_{i,0}$ are replaced by their ML parameter estimates $\hat{\alpha}_{i,1}$ and $\hat{\alpha}_{i,0}$, i=1,...,p.

As pointed out in Chapter 1, the empirical distributions of these bootstrap parameter estimates are asymptotically the same as the sampling distribution of those parameters in sampling from the population from which the original sample was drawn. On the other

hand, so far we have not found in the literature any reference about the shape of the distribution of the pseudovalues. For this reason we expect that the asymptotic theory will work better for the bootstrap than the jackknife results.

The investigation of the normality of the bootstrap distribution of the parameter estimates and the jackknife distribution of the pseudovalues will be done by Normal probability plotting and looking at R^2, the proportion of variance explained by the fitted straight line.

As we show later that, at least in this set of 5 examples, the fitting of the jackknife pseudovalues by a Normal distribution is not associated with the degree of similarity between the jackknife and the original ML parameter estimates, so we only present Normal probability plots of some bootstrap distributions. Maybe this apparent non-association is due to the small number of different jackknife pseudovalues (16), and a larger number of variable would provide satisfactory results.

The results will take into account all bootstrap and jackknife samples, even those when fitted by a logit/probit model provide very large estimates for $\hat{\alpha}_{i,1}$.

The decision about what number of bootstrap samples to take in order to check the adequacy of the asymptotic variance-covariance matrix, was based on the bootstrap parameter estimates $\hat{\alpha}_{i,1}$ and $\hat{\alpha}_{i,0}$ and their standard deviations, obtained from 50, 100 and 200 replications.

In every example we have compared the bootstrap parameter estimates obtained from 50 and 100 replications. The results show stability. To
check further on stability we have increased the bootstrap sample size to 200 for the Arithmetic Reasoning Test (ART) on white women data, and Stouffer and Toby data. Since the doubling of the number of replication still shows the same stability observed when comparing 50 and 100 replications, we decided to give results based on 100 bootstrap samples in all examples.

In the application of the jackknife technique to the 5 set of data, we consider the case where we delete only one observation each time. Although the number of jackknife estimates is equal to the number of observations, a score pattern with frequency n provides n equal jackknife samples. Therefore the number of different jackknife parameter estimates for each parameter is equal to the number of different score patterns in the sample.

If the jackknife gives the same information as the bootstrap about the applicability of the asymptotic theory to estimating of the variability of the parameter estimates then it is more practical to use the jackknife, since it is quicker .

In the following we shall compare the original ML parameter estimates with the bootstrap and jackknife results for the five sets of data referred before.

2- *Attitudes towards the U.S.Army*

The data in Table 3.1 was presented by Stouffer, Guttman, Suchman, Lazarfeld, Star and Clauser (1950, p.21-22), where the four items were intended to measure attitudes towards the U.S.Army held by 1000 noncomissioned officers in 1945. The questions asked were:

(i) how well is the Army run
(ii) whether you will return to civilian life with a favourable attitude towards the Army
(iii) whether you have got a square deal in the Army and
(iv) whether the Army has tried its best to look out for the welfare of enlisted men.

Table 3.1-Score distribution and results obtained by fitting a logit/probit model to the Attitudes towards the U.S.Army.

Response pattern	Observed frequency	Expected frequency	Total score	Component score
0000	229	227.39	0	0.00
0100	52	52.57	1	1.12
0010	25	27.91	1	1.41
0001	16	17.78	1	1.60
1000	199	194.97	1	1.64
0110	16	13.00	2	2.53
0101	8	9.03	2	2.72
1100	96	100.99	2	2.76
0011	10	5.84	2	3.02
1010	60	65.60	2	3.06
1001	45	47.40	2	3.25
0111	3	5.57	3	4.13
1110	69	63.72	3	4.18
1101	55	50.03	3	4.36
1011	42	39.24	3	4.66
1111	75	78.95	4	5.78
Total	1000	1000.00	-	-

χ^2= 7.39 with 7 degrees of freedom ($p \approx 0.40$).

It is reasonable to infer from the low χ^2 value that the data are consistent with a single latent variable measuring attitudes towards the U.S.Army. The scaling given by the component scores is consistent with that of the total scores because the $\hat{\alpha}_{i,1}$'s are very similar as we can see in Table 3.2.

Table 3.2- Comparison between the bootstrap, original ML(in brackets) and the jackknife parameter estimates $\hat{\alpha}_{i,1}$ to the Attitudes towards the U.S.Army.

i	$\hat{\alpha}_{i,1}$	$SD(\hat{\alpha}_{i,1})$	$CV(\hat{\alpha}_{i,1})$	$R^2(\hat{\alpha}_{i,1})$
1	1.68 (1.64) 2.14	.25 (.24) .22	.15 (.15) .10	97.9 77.0
2	1.13 (1.12) 1.08	.15 (.14) .14	.13 (.13) .13	98.0 82.1
3	1.45 (1.41) 1.50	.20 (.19) .19	.14 (.13) .13	99.0 81.5
4	1.63 (1.60) 2.24	.20 (.22) .22	.12 (.14) .10	97.9 77.5

Table 3.3- Comparison between the bootstrap, original ML(in brackets) and the jackknife parameter estimates $\hat{\alpha}_{i,0}$ to the Attitudes towards the U.S.Army.

i	$\hat{\alpha}_{i,0}$	$SD(\hat{\alpha}_{i,0})$	$CV(\hat{\alpha}_{i,0})$	$R^2(\hat{\alpha}_{i,0})$
1	0.88 (0.85) 1.53	.11 (.09) .12	.12 (.10) .08	93.9 95.6
2	-0.65 (-0.66) -0.90	.08 (.09) .08	.12 (.14) .09	98.5 94.6
3	-1.18 (-1.15) -0.82	.12 (.11) .11	.10 (.10) .13	97.1 84.9
4	-1.57 (-1.58) -1.81	.14 (.14) .14	.09 (.09) .08	98.0 78.2

Table 3.4- Comparison between the bootstrap, original ML(in brackets) and the jackknife parameter estimates $\hat{\alpha}^{*}_{i,0}$ to the Attitudes towards the U.S.Army.

i	$\hat{\alpha}^{*}_{i,0}$	$SD(\hat{\alpha}^{*}_{i,0})$	$CV(\hat{\alpha}^{*}_{i,0})$	$R^2(\hat{\alpha}^{*}_{i,0})$
1	0.45 (0.44) 0.69	.04 (.04) .05	.09 (.09) .07	99.4 83.7
2	-0.43 (-0.44) -0.60	.05 (.05) .05	.12 (.11) .08	99.1 82.2
3	-0.67 (-0.67) -0.44	.06 (.05) .06	.09 (.09) .14	98.6 85.3
4	-0.82 (-0.84) -0.71	.06 (.06) .06	.07 (.07) .08	99.4 80.2

Tables 3.2 to 3.4 show an excellent agreement between all the bootstrap results and original ML parameter estimates. This is, perhaps, to be expected since all the bootstrap distributions of the parameter estimates are approximated very well by a normal distribution. The asymptotic theory works well in this example, where the sample size is 1000 and the ML parameter estimates $\hat{\alpha}_{i,1}$ are nearly equal.

The jackknife parameter estimates $\hat{\alpha}_{i,1}$, for i=2,3, and their standard deviations are very similar to the corresponding original ML, while for items 1 and 4, they are slightly bigger with smaller coefficients of variation.

The relation between the jackknife and the original ML $\hat{\alpha}^{*}_{i,0}$ has the same pattern as $\hat{\alpha}_{i,0}$, showing similar results, but not as close as that given by the bootstrap.

Bootstrap biases of $\alpha_{i,1}$ (equation 1.26), for i=1,...,4, are equal to 0.04, 0.01, 0.04 and 0.03, while the jackknife biases (equation 1.23), are equal to 0.50, -0.04, 0.09 and 0.64,

respectively. Thus, boostrap has provide estimates $\hat{\alpha}_{i,1}$ with equal or less bias than the jackknife.

The bootstrap distribution of the parameter estimates are as well or better fitted by a normal distribution then the corresponding jackknife pseudovalues.

Figures 3.1 to 3.3 present the bootstrap distribution of the parameter estimates $\hat{\alpha}_{1,1}$, $\hat{\alpha}_{1,0}$ and $\hat{\alpha}^{*}_{1,0}$ and their fit by a normal distribution.

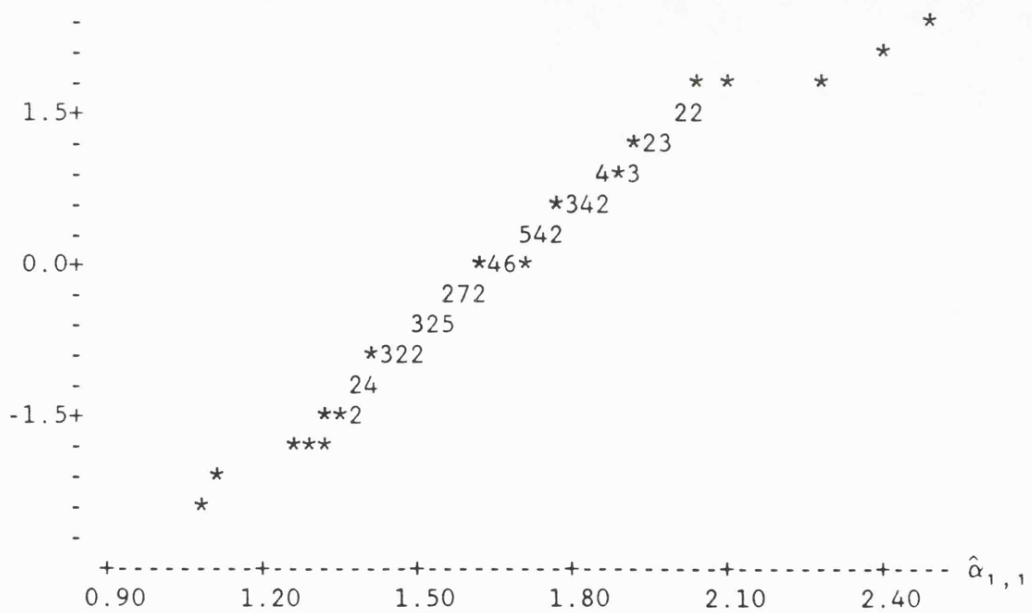


Figure 3.1- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}_{1,1}$ to the Attitudes towards the U.S.Army (original ML $\hat{\alpha}_{1,1}$ = 1.64, bootstrap $\hat{\alpha}_{1,1}$ = 1.68 and R^2 = 97.9%).

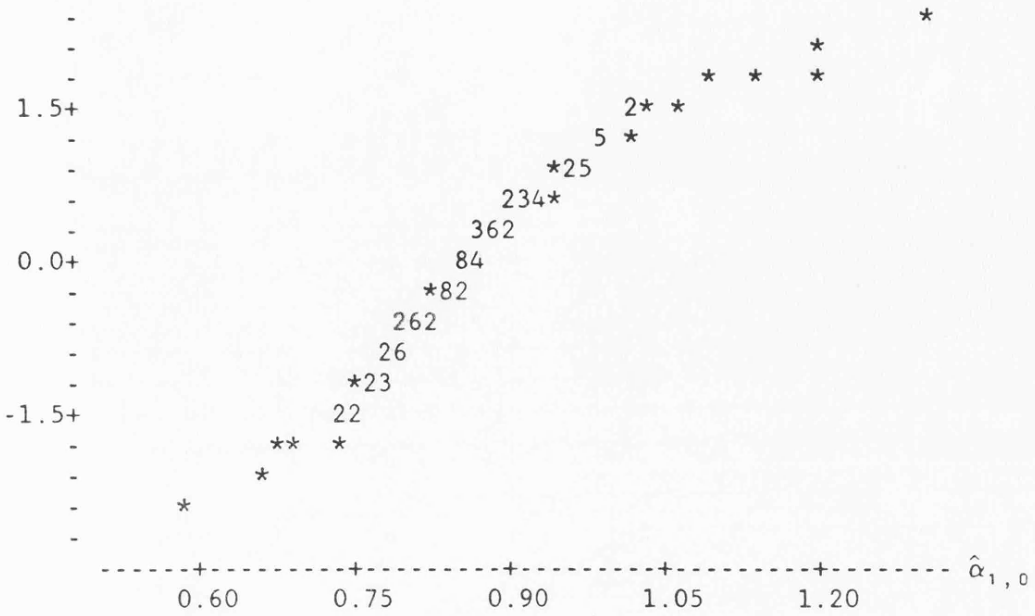


Figure 3.2- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}_{1,0}$ to the Attitudes towards the U.S.Army (original ML $\hat{\alpha}_{1,0} = 0.85$, bootstrap $\hat{\alpha}_{1,0} = 0.88$ and $R^2 = 93.9\%$).

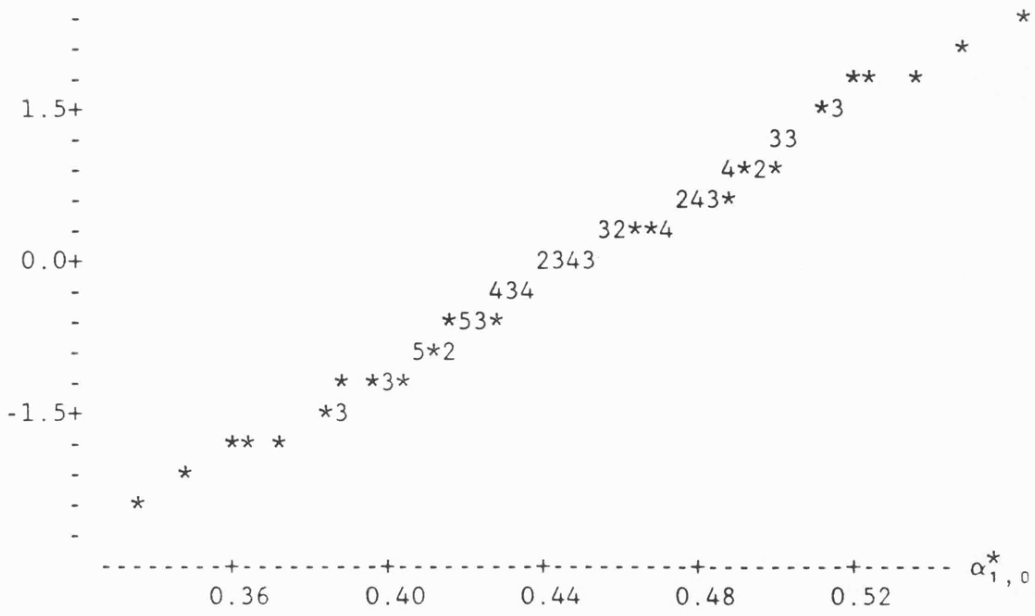


Figure 3.3- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}^*_{1,0}$ to the Attitudes towards the U.S.Army (original ML $\hat{\alpha}^*_{1,0} = 0.44$, bootstrap $\hat{\alpha}^*_{1,0} = 0.45$ and $R^2 = 99.4\%$).

The correlation matrix of the original ML, bootstrap, the jackknife parameter estimates are displayed in Tables 3.5 and 3.6.

Table 3.5- Correlations between the original ML parameter estimates based on the observed 2nd derivative matrix (under the diagonal) and on the information matrix (above the diagonal) to the Attitudes towards the U.S.Army.

	$\hat{\alpha}_{1,1}$	$\hat{\alpha}_{2,1}$	$\hat{\alpha}_{3,1}$	$\hat{\alpha}_{4,1}$	$\hat{\alpha}_{1,0}$	$\hat{\alpha}_{2,0}$	$\hat{\alpha}_{3,0}$	$\hat{\alpha}_{4,0}$
$\hat{\alpha}_{1,1}$		-0.11	-0.13	-0.14	0.54	0.05	0.10	0.13
$\hat{\alpha}_{2,1}$	-0.10		-0.04	-0.06	-0.08	-0.33	0.04	0.07
$\hat{\alpha}_{3,1}$	-0.18	-0.01		-0.13	-0.10	0.02	-0.58	0.11
$\hat{\alpha}_{4,1}$	-0.12	-0.10	-0.11		-0.10	0.03	0.09	-0.72
$\hat{\alpha}_{1,0}$	0.54	-0.08	-0.12	-0.09		0.22	0.24	0.24
$\hat{\alpha}_{2,0}$	0.05	-0.33	0.01	0.04	0.22		0.15	0.13
$\hat{\alpha}_{3,0}$	0.13	0.02	-0.59	0.08	0.26	0.16		0.07
$\hat{\alpha}_{4,0}$	0.12	0.09	0.10	-0.72	0.23	0.12	0.08	

From Table 3.5 we can see that both methods give approximately the same asymptotic correlations between the original parameter estimates $\hat{\alpha}_{i,1}$ and $\hat{\alpha}_{i,0}$. The highest correlations (-0.72 to 0.54) are between $\hat{\alpha}_{i,1}$ and $\hat{\alpha}_{i,0}$ for all items. The remaining correlations are smaller than 0.26 (in absolute value), showing weak or no correlation between these parameter estimates.

Table 3.6- Bootstrap (under the diagonal) and jackknife (above the diagonal) estimates of correlations between the parameter estimates of the Attitudes towards the U.S.Army.

	$\hat{\alpha}_{1,1}$	$\hat{\alpha}_{2,1}$	$\hat{\alpha}_{3,1}$	$\hat{\alpha}_{4,1}$	$\hat{\alpha}_{1,0}$	$\hat{\alpha}_{2,0}$	$\hat{\alpha}_{3,0}$	$\hat{\alpha}_{4,0}$
$\hat{\alpha}_{1,1}$		-0.06	-0.17	-0.09	0.60	0.02	0.14	0.10
$\hat{\alpha}_{2,1}$	0.10		-0.03	-0.14	-0.05	-0.36	0.03	0.10
$\hat{\alpha}_{3,1}$	-0.39	-0.23		-0.12	-0.16	0.00	-0.59	0.07
$\hat{\alpha}_{4,1}$	0.01	-0.23	0.04		-0.12	0.03	0.06	-0.73
$\hat{\alpha}_{1,0}$	0.67	0.07	-0.24	0.07		0.20	0.22	0.34
$\hat{\alpha}_{2,0}$	-0.01	-0.27	-0.03	0.06	0.26		0.11	0.14
$\hat{\alpha}_{3,0}$	0.41	0.05	-0.62	-0.03	0.35	0.12		0.09
$\hat{\alpha}_{4,0}$	0.14	0.19	-0.13	-0.66	0.23	0.14	0.30	

We shall compare the correlations in Table 3.6 to those in Table 3.5 obtained from the observed second derivatives, though the same results are valid if we use the information matrix instead.

The jackknife estimates of the correlations are nearly equal to the asymptotic correlations between the original ML parameter estimates $\hat{\alpha}_{i,1}$ and $\hat{\alpha}_{j,0}$, for i and j equal to 1,...,4.

The largest differences between bootstrap estimates of the correlations and the asymptotic of the original ML are the correlations between $\hat{\alpha}_{3,1}$ and $\hat{\alpha}_{2,1}$ (-0.23 compared to -0.001); $\hat{\alpha}_{3,1}$ and $\hat{\alpha}_{4,0}$ (-0.13 compared to 0.10); $\hat{\alpha}_{3,0}$ and $\hat{\alpha}_{1,1}$ (0.41 compared to 0.13);$\hat{\alpha}_{3,0}$ and $\hat{\alpha}_{4,0}$ (0.30 compared to 0.08). Whether these correlations are different of zero is difficult to say.

Tables 3.2 and 3.3 seem not to provide any straight reason for the biggest differences being associated to item 3.

The agreement between the bootstrap estimates of the correlations with the corresponding asymptotic ML correlations are not as good as that showed by the jackknife results. Actually the biggest differences between bootstrap and the original ML , shown above, are between the same parameter estimates as showed by the comparison of the former with jackknife estimates of the correlations.

3- *Arithmetic Reasoning Test on White Women*

The full set of data for the Arithmetic Reasoning Test (ART) on white women is described in Table 2.1, followed by an extensive analysis for the logit/probit model in Chapter 2.

Table 3.7- Comparison between the bootstrap, original ML(in brackets) and the jackknife parameter estimates $\hat{\alpha}_{i,\cdot}$ for the ART on white women.

i	$\hat{\alpha}_{i,\cdot}$	$SD(\hat{\alpha}_{i,\cdot})$	$CV(\hat{\alpha}_{i,\cdot})$	$R^2(\hat{\alpha}_{i,\cdot})$
1	1.14 (1.04) 1.37	.42 (.32) .36	.37 (.31) .26	88.5 89.4
2	1.26 (1.24) 1.14	.40 (.39) .40	.32 (.31) .35	94.2 80.9
3	1.04 (1.00) 0.93	.34 (.30) .32	.33 (.30) .34	96.9 78.7
4	1.51 (1.44) 1.41	.56 (.45) .50	.37 (.31) .35	94.2 73.0

Table 3.8- Comparison between the bootstrap, original ML(in brackets) and the jackknife parameter estimates $\hat{\alpha}_{i,0}$ for the ART on white women.

i	$\hat{\alpha}_{i,0}$	$SD(\hat{\alpha}_{i,0})$	$CV(\hat{\alpha}_{i,0})$	$R^2(\hat{\alpha}_{i,0})$
1	0.59 (0.59) 0.59	.17 (.18) .17	.29 (.29) .29	99.0 85.7
2	0.56 (0.56) 0.56	.19 (.19) .18	.34 (.30) .32	97.1 87.8
3	-0.08(-0.08)-0.06	.16 (.16) .16	.00 (2.00)2.67	98.4 68.7
4	-0.53(-0.51)-0.51	.23 (.21) .20	.43 (.41) .39	99.2 86.1

Table 3.9- Comparison between the bootstrap, original ML(in brackets) and the jackknife parameter estimates $\hat{\alpha}^*_{i,0}$ for the ART on white women.

i	$\hat{\alpha}^*_{i,0}$	$SD(\hat{\alpha}^*_{i,0})$	$CV(\hat{\alpha}^*_{i,0})$	$R^2(\hat{\alpha}^*_{i,0})$
1	0.41 (0.41) 0.32	.12 (.12) .12	.29 (.29) .38	99.2 82.8
2	0.35 (0.35) 0.35	.11 (.10) .11	.31 (.28) .38	98.4 79.4
3	-0.06 (-0.06)-0.04	.10 (.10) .11	1.67 (1.67)2.75	98.3 72.2
4	-0.29 (-0.29)-0.28	.11 (.11) .10	.38 (.38) .36	99.4 77.2

The bootstrap $CV(\hat{\alpha}_{i,1})$ of items 1 and 4 are slightly bigger than the original ML (0.37 compared with 0.31). In item 1 this difference is probably due to the two extreme values of $\hat{\alpha}_{1,1}$, as we can see in Figure 3.4. If we take them out then the fit of a normal distribution will be even better and the similarity between the bootstrap and original ML results will increase. This result illustrates that large values ($\hat{\alpha}_{1,1}$=3.19) can happen even when the original ML parameter estimate $\hat{\alpha}_{1,1}$ is small ($\hat{\alpha}_{1,1}$=1.04). However two bootstrap samples give 2 extreme values for $\hat{\alpha}_{1,1}$, the same does not happen when estimating $\hat{\alpha}_{1,0}$ and $\hat{\alpha}^*_{1,0}$, as we can see from Figures 3.5 and 3.6.

Tables 3.8 and 3.9 display a very good agreement between all the bootstrap parameter estimates $\hat{\alpha}_{i,0}$ and $\hat{\alpha}^*_{i,0}$ and they are fitted well by a normal distribution, giving strong evidence that the original ML estimates and their standard deviations can probably be trusted.

Therefore there is very good agreement between the bootstrap and original ML parameter estimates and their corresponding standard deviations, though in the former example is better for $\hat{\alpha}_{i,1}$, probably because the sample size is bigger (1000), since the $\hat{\alpha}_{i,1}$'s are also very similar to each other.

From Table 3.7 we can see that jackknife gives very close results to the original ML, except for item 1, for which the jackknife $CV(\hat{\alpha}_{1,1})$ is slightly smaller, 0.26 compared to 0.31, and $\hat{\alpha}_{1,1}$ is bigger than the corresponding the original ML and bootstrap estimates (1.37 compared to 1.04 and 1.14).

The jackknife parameter estimates $\hat{\alpha}_{i,0}$ and $\hat{\alpha}^*_{i,0}$ are nearly equal to the corresponding original ML estimates, except for the coefficient of variation of $\hat{\alpha}_{3,0}$, which is bigger (2.67 compared to 2.0), $\hat{\alpha}^*_{1,0}$ is smaller (0.32 compared to 0.41) with larger coefficient of variation (0.38 compared to 0.29) and the coefficient of variation of $\hat{\alpha}^*_{3,0}$is larger (2.75 compared to 1.67).

The boostrap estimates of the bias of $\alpha_{i,1}$, i=1,...,4, are equal to 0.10, 0.02, 0.04 and 0.07, while the jackknife estimates are 0.33, -0.10, -0.07 and -0.03, respectively. Therefore, both methods are nearly equally biased, except for $\alpha_{1,1}$, for which the bootstrap bias is slightly smaller.

As in the preceding example, bootstrap parameter estimates are equal or closer to the original ML estimates than the corresponding jackknife ones.

While the bootstrap $R^2 \geq 88.55\%$, R^2 for the jackknife pseudovalues varies between 68.7% and 89.4%, indicating that bootstrap distribution of $\hat{\alpha}_{i,1}$ and $\hat{\alpha}_{i,0}$ is fitted better by a normal distribution than the jackknife pseudovalues.

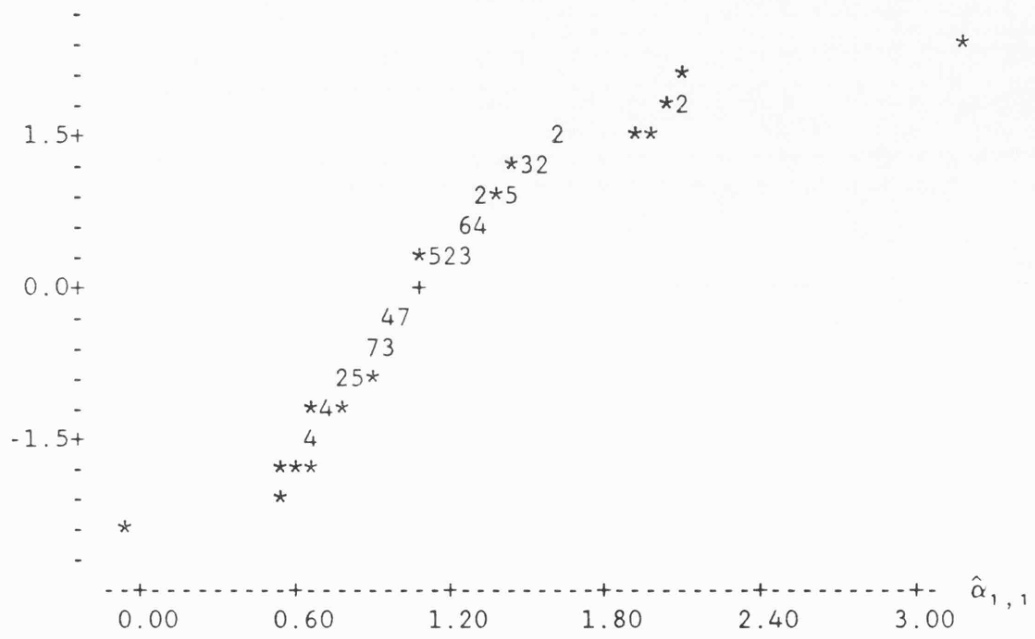


Figure 3.4- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}_{1,1}$ to the ART on white women (original ML $\hat{\alpha}_{1,1} = 1.04$, bootstrap $\hat{\alpha}_{1,1} = 1.14$ and $R^2 = 88.5\%$).

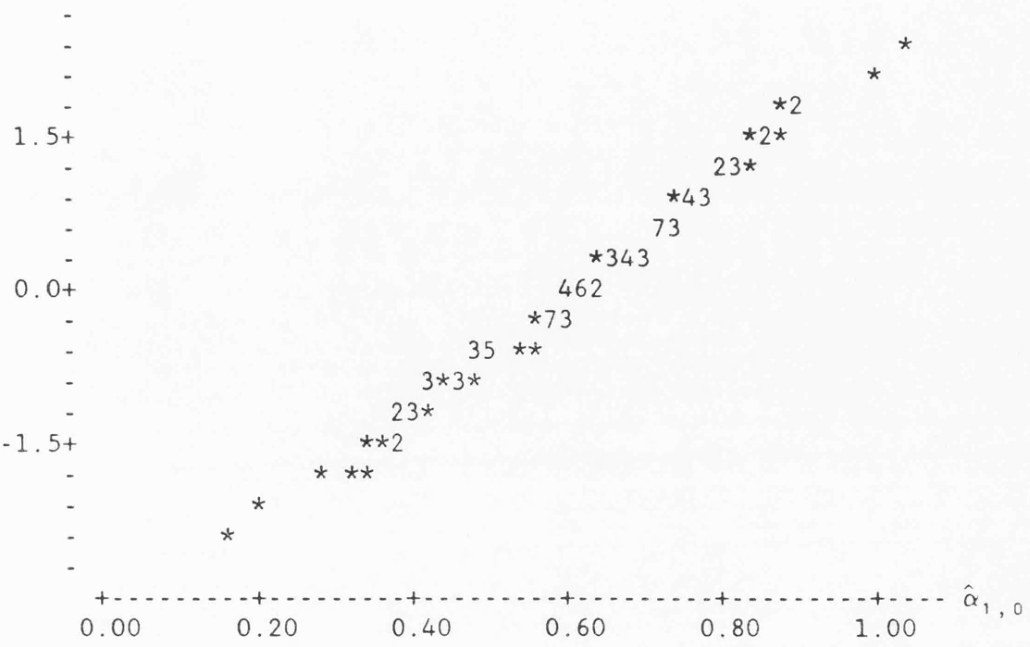


Figure 3.5- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}_{1,0}$ to the ART on white women (original ML and bootstrap $\hat{\alpha}_{1,0}$ equal to 0.59 and R^2 = 99.2%).

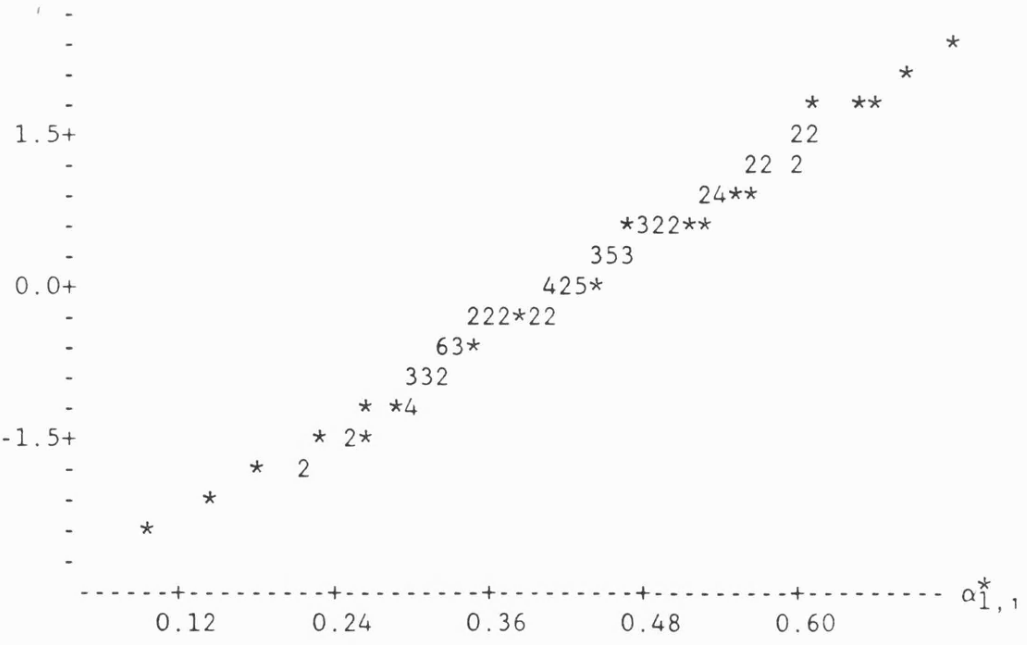


Figure 3.6- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}^*_{1,0}$ to the ART on white women (original ML and bootstrap $\hat{\alpha}^*_{1,0}$ equal to 0.41 and R^2 = 99.2%).

In the following two tables we present the correlation matrix of the original ML, bootstrap and the jackknife parameter estimates.

Table 3.10- Correlations between the original ML parameter estimates based on the observed 2nd derivative matrix (under the diagonal) and on the information matrix (above the diagonal) for the ART on white women.

	$\hat{\alpha}_{1,1}$	$\hat{\alpha}_{2,1}$	$\hat{\alpha}_{3,1}$	$\hat{\alpha}_{4,1}$	$\hat{\alpha}_{1,0}$	$\hat{\alpha}_{2,0}$	$\hat{\alpha}_{3,0}$	$\hat{\alpha}_{4,0}$
$\hat{\alpha}_{1,1}$		-0.08	-0.04	-0.13	0.32	-0.04	0.00	0.06
$\hat{\alpha}_{2,1}$	-0.18		-0.07	-0.20	-0.03	0.35	0.01	0.09
$\hat{\alpha}_{3,1}$	-0.13	0.07		-0.14	-0.02	-0.03	-0.03	0.06
$\hat{\alpha}_{4,1}$	0.02	-0.28	-0.21		-0.05	-0.08	0.00	-0.36
$\hat{\alpha}_{1,0}$	0.34	-0.07	0.05	0.01		0.16	0.17	0.21
$\hat{\alpha}_{2,0}$	-0.07	0.38	0.02	-0.12	0.15		0.19	0.25
$\hat{\alpha}_{3,0}$	0.01	0.00	-0.02	0.00	0.17	0.18		0.20
$\hat{\alpha}_{4,0}$	0.01	0.11	0.09	-0.34	0.20	0.26	0.20	

Both methods give approximately the same correlations with the maximum difference equal to 0.15, related to the correlation between $\hat{\alpha}_{1,1}$ and $\hat{\alpha}_{4,1}$ (0.02, based on the observed second derivatives compared to -0.13 obtained from the information matrix).

Table 3.11- Bootstrap (under the diagonal) and jackknife (above the diagonal) estimates of correlations between the parameter estimates of the ART on white women.

	$\hat{\alpha}_{1,1}$	$\hat{\alpha}_{2,1}$	$\hat{\alpha}_{3,1}$	$\hat{\alpha}_{4,1}$	$\hat{\alpha}_{1,0}$	$\hat{\alpha}_{2,0}$	$\hat{\alpha}_{3,0}$	$\hat{\alpha}_{4,0}$
$\hat{\alpha}_{1,1}$		-0.22	-0.20	0.13	0.22	-0.10	0.02	-0.04
$\hat{\alpha}_{2,1}$	-0.28		0.19	-0.38	-0.02	0.36	0.07	0.23
$\hat{\alpha}_{3,1}$	-0.33	0.06		-0.33	-0.01	0.13	-0.08	0.22
$\hat{\alpha}_{4,1}$	0.08	-0.23	-0.33		0.07	-0.08	0.08	-0.46
$\hat{\alpha}_{1,0}$	0.21	-0.14	0.08	0.04		0.17	0.17	0.20
$\hat{\alpha}_{2,0}$	-0.03	0.40	0.04	-0.07	0.07		0.24	0.25
$\hat{\alpha}_{3,0}$	0.13	0.09	-0.07	-0.02	0.23	0.18		0.14
$\hat{\alpha}_{4,0}$	-0.10	0.15	0.26	-0.47	0.22	0.07	0.28	

We shall compare the bootstrap and jackknife correlations with the asymptotic ones obtained from the observed second derivative matrix, although the same conclusions are valid when comparing with the information matrix.

The largest differences between the bootstrap estimates of the correlations and the original ML correlations between $\hat{\alpha}_{i,1}$ and $\hat{\alpha}_{j,0}$ are equal to 0.20 between $\hat{\alpha}_{1,1}$ and $\hat{\alpha}_{3,1}$ (-0.33 compared to -0.13); 0.19 between $\hat{\alpha}_{2,0}$ and $\hat{\alpha}_{4,0}$ (0.07 compared to 0.26) and 0.17 between $\hat{\alpha}_{3,1}$ and $\hat{\alpha}_{4,0}$ (0.26 compared to 0.09).

The magnitude of the discrepancies between the jackknife estimates of the correlations and the original ML correlations are up to 0.13, and it is between $\hat{\alpha}_{3,1}$ and $\hat{\alpha}_{4,0}$.

In this example the results show that the jackknife estimates of the correlations between $\hat{\alpha}_{i,1}$ and $\hat{\alpha}_{j,0}$, i,j=1,..,4, are equal or closer to the asymptotic ML correlations than the bootstrap estimates.

4- *Attitudes towards Situations of Conflict*

Stouffer and Toby (1951) report the answers of 216 respondents in 4 situations of conflict. For each situation the respondents can react either by a universalistic attitude (negative response) or particularistic attitude (positive response), which results are given in Table 3.12.

Table 3.12- Score distribution and results obtained by fitting a logit/probit model for the Stouffer and Toby data.

Response pattern	Observed frequency	Expected frequency	Total score	Component score
0000	20	22.49	0	0.00
1000	38	38.40	1	1.15
0010	9	6.92	1	1.35
0100	6	5.62	1	1.58
0001	2	1.17	1	2.10
1010	24	22.93	2	2.50
1100	25	20.67	2	2.73
0110	4	4.21	2	2.93
1001	7	5.39	2	3.25
0011	2	1.14	2	3.45
0101	1	1.11	2	3.68
1110	23	27.38	3	4.08
1011	6	9.18	3	4.60
1101	6	9.86	3	4.83
0111	1	2.34	3	5.03
1111	42	37.19	4	6.18
Total	216	216.00	-	-

χ^2=5.85 with 3 degrees of freedom ($0.20<p<0.10$).

Table 3.12 shows that these data are fitted well by a logit/probit model with one single latent variable as a measure of the attitude of a person when under different situations of conflict. The scaling given by the total and the component scores is the same though the ML $\hat{\alpha}_{i,,1}$ varies from 1.15 to 2.10, as we can see in Table 3.13.

Andersen and Madsen (1977), using conditional maximum likelihood estimation found that the Rasch model fits these data very well, since

the item difficulties estimated from different subsamples formed according to the number of positive responses on the test are very similar to the overall estimates obtained from the whole sample of items. However when investigating the latent distribution, taking into account these values for the parameter estimates, they found a lack of fit by a latent normal distribution.

There is no contradiction between a conditional fit of the Rasch model and the logit/probit model, since the Rasch model is fitted without any assumption about the distribution of the latent variable, and as we pointed out in Chapter 1, section 3.2.2, the parameter estimates are little affected by the choice of the prior distribution. What may also happen is that the $\hat{\alpha}_{i,1}$'s in the logit/probit model are not statistically different from each other.

Table 3.13- Comparison between the bootstrap, original ML(in brackets) and the jackknife parameter estimates $\hat{\alpha}_{i,1}$ for the Stouffer and Toby data.

i	$\hat{\alpha}_{i,1}$	$SD(\hat{\alpha}_{i,1})$	$CV(\hat{\alpha}_{i,1})$	$R^2(\hat{\alpha}_{i,1})$
1	1.19 (1.15) 1.33	.38 (.36) .28	.32 (.31) .21	93.5 90.8
2	1.82 (1.58) 1.44	.83 (.44) .42	.46 (.28) .29	66.4 75.5
3	1.44 (1.35) 1.24	.46 (.36) .34	.32 (.27) .27	93.0 77.0
4	2.72 (2.10) 2.18	2.99 (.66) .66	1.10 (.31) .30	35.3 63.5

The similarity between the bootstrap and the original ML results related to $\hat{\alpha}_{i,1}$, i=1 and 3, can be considered very good, since the parameter estimates and their standard deviations are nearly equal. This similarity could improve even more for item 1 if we delete the bootstrap sample that have provided an extreme value equal to 3.0, as

we can see in Figure 3.7. However there are significant differences for the remaining items (i=2 and 4), especially for item 4, where R^2 is equal to 35.30%. Looking at the bootstrap distribution of $\hat{\alpha}_{4,1}$ (Figure 3.11), we can see that this is due to a sample with $\hat{\alpha}_{4,1}$=28.12. On deleting this sample it is found that the only changes on the bootstrap results in Tables 3.13 to 3.15 are related to item 4,that is,

$\hat{\alpha}_{4,1}$ = 2.46 $SD(\hat{\alpha}_{4,1})$= 1.41 $CV(\hat{\alpha}_{4,1})$= 0.57 and R^2 = 71.1%

$\hat{\alpha}_{4,0}$ =-1.47 $SD(\hat{\alpha}_{4,0})$= 0.38 $CV(\hat{a}_{4,0})$= 0.26 and R^2 = 68.4%.

The improvement is not significant since we still have a large difference between the bootstrap and the ML parameter estimates $\hat{\alpha}_{i,1}$, i=2 and 4 (1.82 compared to 1.58 and 2.46 compared to 2.10, respectively). Their R^2 are equal to 66.4% and 71.1%, respectively, which do not indicate a good approximation of the bootstrap distribution of parameter estimates by a normal distribution, probably responsible for the values 0.89 and 1.14 for the ratios of bootstrap standard deviations to the corresponding asymptotic standard deviations.

On the other hand, a joint analysis of Figures 3.11 and 3.12 suggests that the bootstrap distribution $\hat{\alpha}_{2,1}$ and $\hat{\alpha}_{2,0}$ that is either very skewed or a mixture of a normal distribution and some infinite values for $\alpha_{2,1}$, which are estimated to be only large due to inaccurate computing.

We can get a better idea about the bootstrap distribution of the parameter estimates for items 1 and 4 by looking at Figures 3.7 and 3.13.

There is good agreement between jackknife and original ML parameter estimates $\alpha_{i,1}$ for items 1 and 3. As bootstrap is very close to the original ML parameter estimates for these items, both methods give practically the same information about them. However for items 2 and 4, jackknife results are closer to the original ML than the bootstrap results. This is not desirable, especially because bootstrap is warning that the standard deviations probably are bigger than those given by the asymptotic theory.

Bootstrap biases of $\alpha_{i,1}$, i=1,...,4, are equal to 0.04, 0.24, 0.11 and 0.62, while the jackknife estimates are 0.18, -0.14, -0.11 and 0.08, respectively. The largest biases are yielded by bootstrap and related to $\alpha_{4,1}$, followed by $\alpha_{2,1}$. hence, in terms of bias of $\alpha_{i,1}$, jackknife provides estimates equal or less biased than bootstrap.

Table 3.14- Comparison between the bootstrap, original ML(in brackets) and the jackknife parameter estimates $\hat{\alpha}_{i,0}$ for the Stouffer and Toby data.

i	$\hat{\alpha}_{i,0}$	$SD(\hat{\alpha}_{i,0})$	$CV(\hat{\alpha}_{i,0})$	$R^2(\hat{\alpha}_{i,0})$
1	1.70 (1.66) 1.81	.28 (.22) .20	.16(.13) .11	94.3 72.6
2	-0.01 (0.01) 0.01	.25 (.24) .19	25.00(24.00)14.32	92.5 71.9
3	0.08 (0.08) 0.08	.19 (.20) .18	2.38(2.50) 2.12	98.0 74.2
4	-1.53 (-1.33)-1.41	.90 (.36) .35	.59(.27) 0.24	55.0 70.7

Table 3.15- Comparison between the bootstrap, original ML(in brackets) and the jackknife parameter estimates $\hat{\alpha}^*_{i,0}$ for the Stouffer and Toby data.

i	$\hat{\alpha}^*_{i,0}$	$SD(\hat{\alpha}^*_{i,0})$	$CV(\hat{\alpha}^*_{i,0})$	$R^2(\hat{\alpha}^*_{i,0})$
1	1.10 (1.09) 1.04	.16 (.14) .15	.14 (.13) .14	98.0 72.3
2	0.002 (0.005)0.07	.10 (.13) .10	50.00(26.00) 1.43	98.0 71.1
3	-0.05 (0.05) 0.05	.10 (.12) .10	2.00 (.40) 2.00	97.8 68.9
4	-0.56 (-0.57)-0.55	.11 (.10) .11	.20 (.18) .20	99.0 80.0

According to the bootstrap results, the asymptotic theory can probably be applied to the estimation of the variability of all parameter estimates $\hat{\alpha}_{i,0}$ and $\hat{\alpha}^*_{i,0}$, except for $\hat{\alpha}_{4,0}$, which is not well fitted by a normal distribution, since R^2=55.0% (Figure 3.12). In contrast the bootstrap distribution of the parameter estimate $\hat{\pi}_4$ can be approximated by a normal distribution since the R^2 is equal to 94.4% (Figure 3.13). It seems that the fit of a normal distribution to the bootstrap distribution of $\hat{\pi}_i$ is usually as good as for $\hat{\alpha}_{i,0}$ but much better when $\hat{\alpha}_{i,0}$ is not well approximated by a normal distribution.

The bootstrap and the original ML coefficient of variation of $\hat{\alpha}^*_{2,0}$ are very large, 50.00 and 26.00, respectively, due to the large $CV(\hat{\alpha}_{2,0})$ in both cases.

The bootstrap distribution for $\hat{\alpha}_{1,1}$, $\hat{\alpha}_{1,0}$ and $\hat{\pi}_1$ given by Figures 3.7 to 3.9 show a good fit to a normal distribution ($R^2 \geq 93.5\%$) and great similarity in the display of the points. The biggest values for the estimates in these graphs come from the same sample, but it is not the one that has $\hat{\alpha}_{4,1}$ =28.12.

Bootstrap results suggest that the application of the asymptotic theory to determine the variance matrix may not be adequate when both the sample size is small and at least one of the $\hat{\alpha}_{i,1}$ is bigger than 2.0 and almost double size of the smallest (1.15).

Tables 3.14 and 3.15 show that the jackknife parameter estimates $\hat{\alpha}_{i,0}$ and $\hat{\alpha}^{*}_{i,0}$ and their coefficients of variation are very similar to the original ML, except for items 2 and 3, for which they are smaller, indicating that the jackknife standard deviations are probably underestimating the true ones.

As in the preceding examples, the pattern of R^2 of the jackknife pseudovalues seems not to be associated to the degree of similarity between its estimates and the original ML as showed by the bootstrap results.

Jackknife parameter estimates are equal or closer to the original ML than the bootstrap ones. Furthermore, while some jackknife estimates of the standard deviation are even smaller than the asymptotic ones, bootstrap results are warning that the true standard deviations are probably bigger.

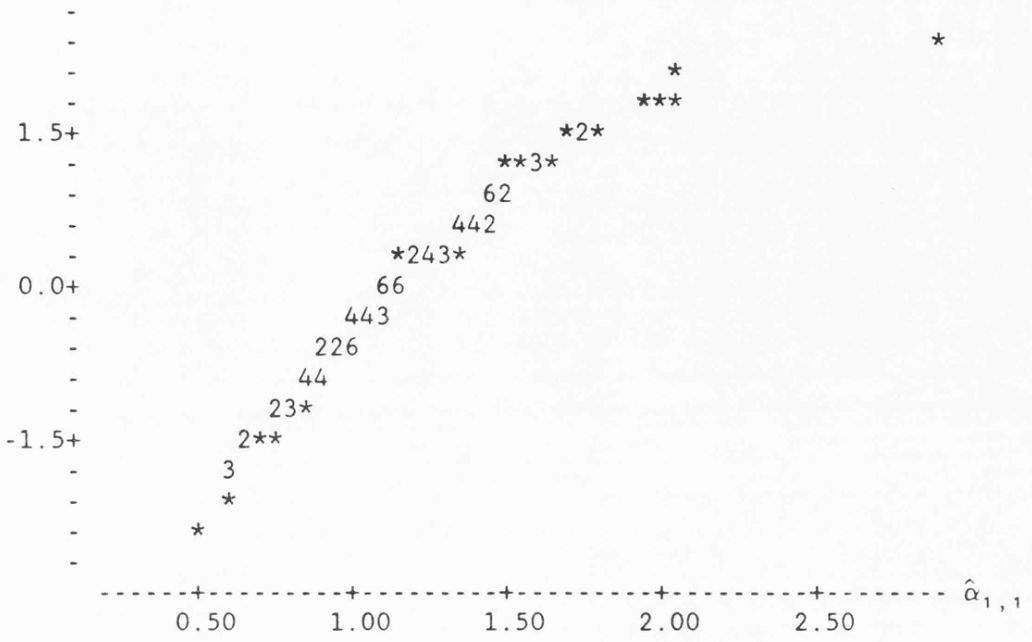


Figure 3.7- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}_{1,1}$ for the Stouffer and Toby data (original ML $\hat{\alpha}_{1,1} = 1.15$, bootstrap $\hat{\alpha}_{1,1} = 1.19$ and $R^2 = 93.5\%$).

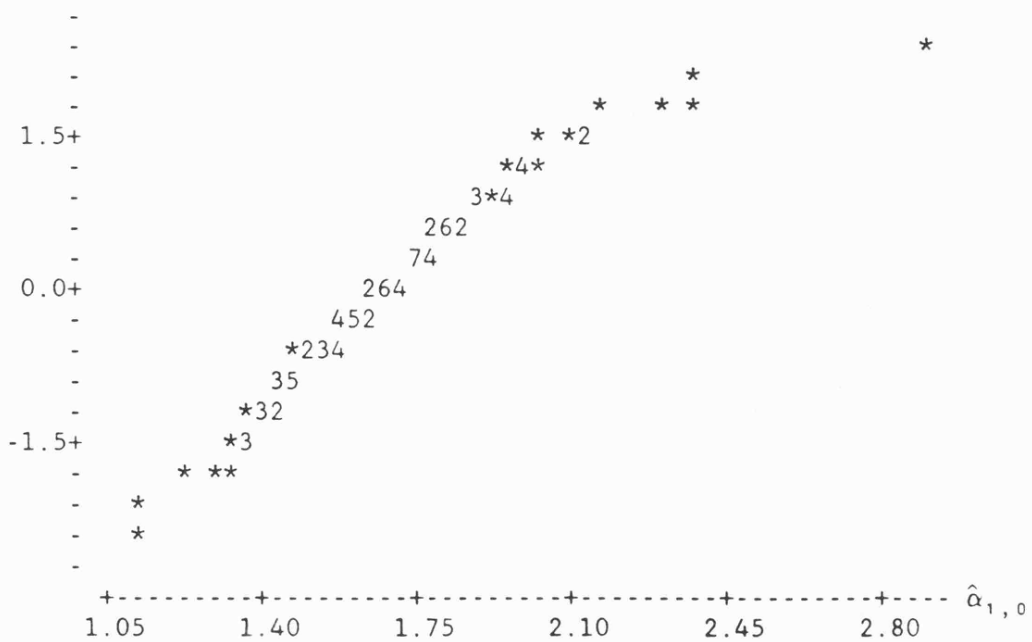


Figure 3.8- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}_{1,0}$ for the Stouffer and Toby data (original ML $\hat{\alpha}_{1,0} = 1.66$, bootstrap $\hat{\alpha}_{1,0} = 1.70$ and $R^2 = 94.3\%$).

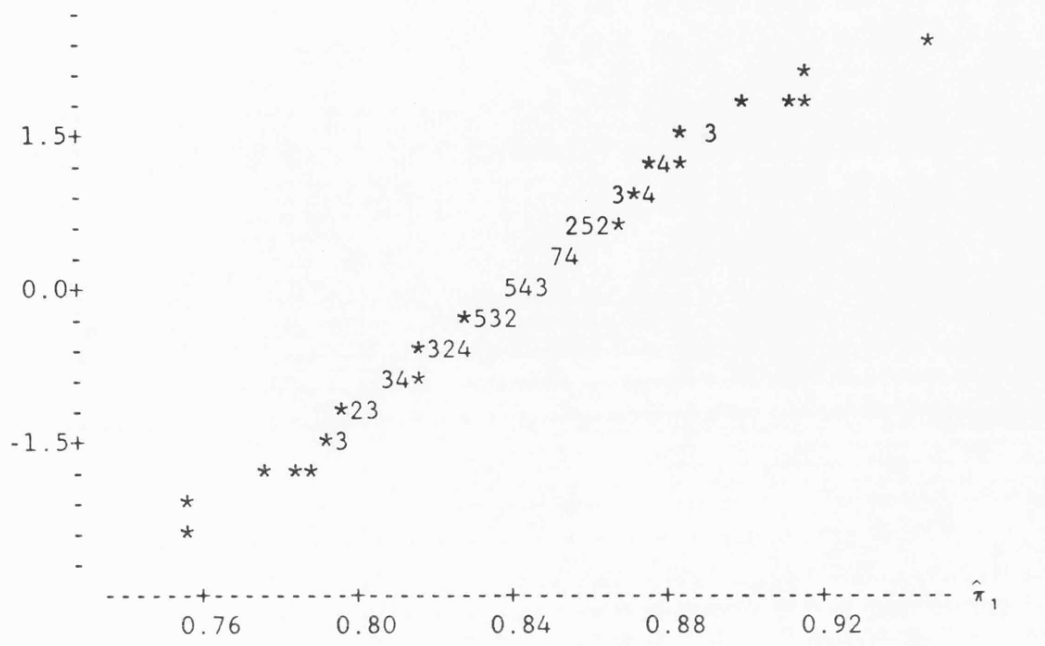


Figure 3.9- Normal probability plotting of the bootstrap parameter estimate $\hat{\pi}_1$ for the Stouffer and Toby data (original ML and bootstrap $\hat{\pi}_1 = 0.84$ and $R^2 = 99.2\%$).

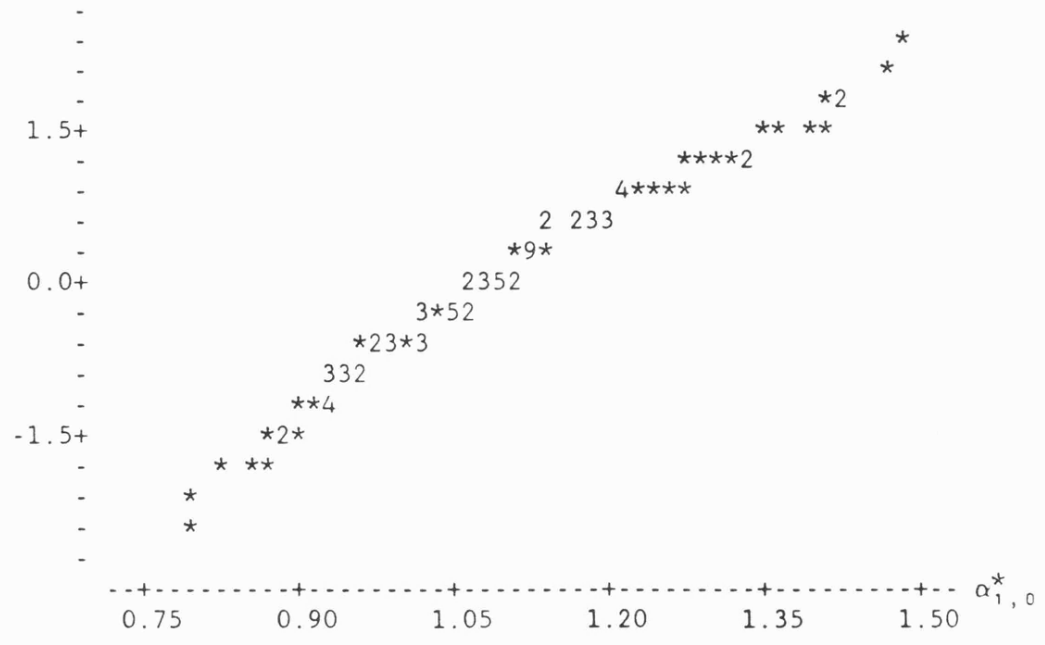


Figure 3.10- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}^*_{1,0}$ for the Stouffer and Toby data (original ML $\hat{\alpha}^*_{1,0} = 1.09$, boostrap $\hat{\alpha}^*_{1,0} = 1.10$ and $R^2 = 98.0\%$).

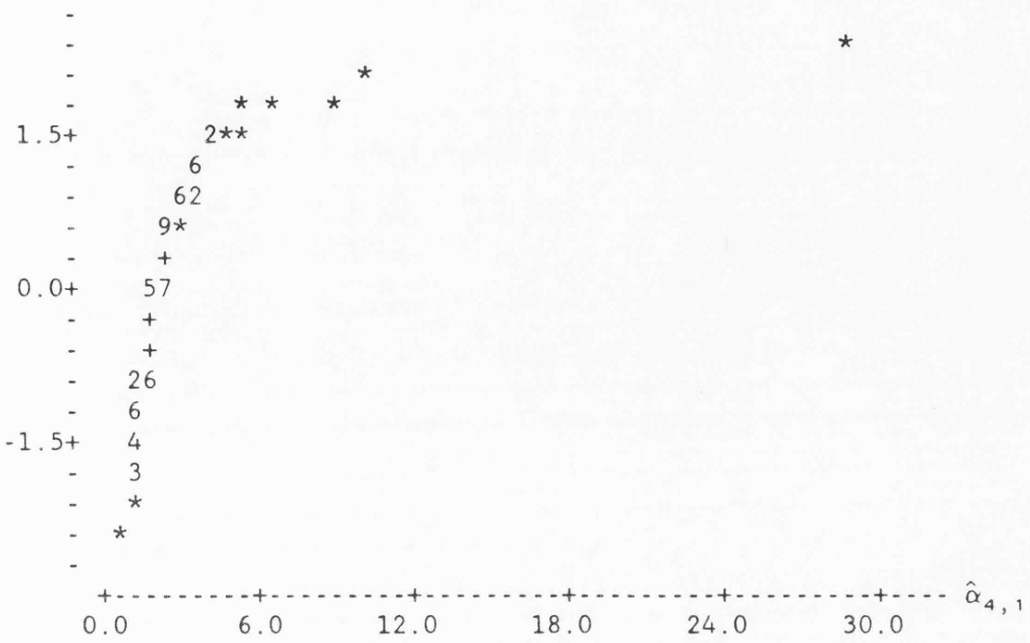


Figure 3.11- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}_{4,1}$ for the Stouffer and Toby data (original ML $\hat{\alpha}_{4,1}$ = 2.10, bootstrap $\hat{\alpha}_{4,1}$ = 2.72 and R^2 = 35.3%).

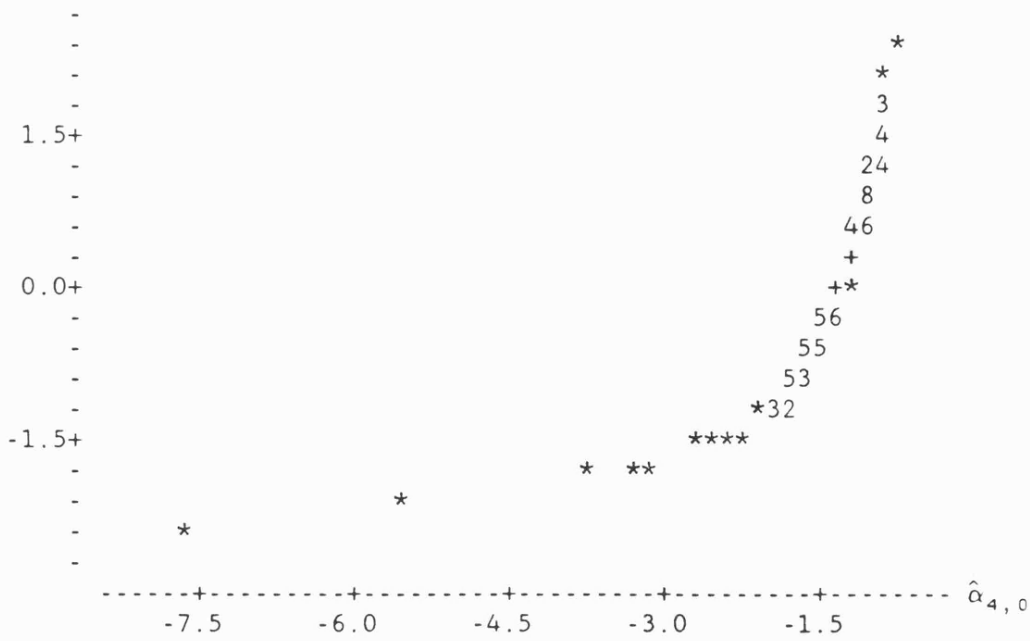


Figure 3.12- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}_{4,0}$ for the Stouffer and Toby data (original ML $\hat{\alpha}_{4,0}$ = -1.33, bootstrap $\hat{\alpha}_{4,0}$ = -1.53 and R^2 = 55.0%).

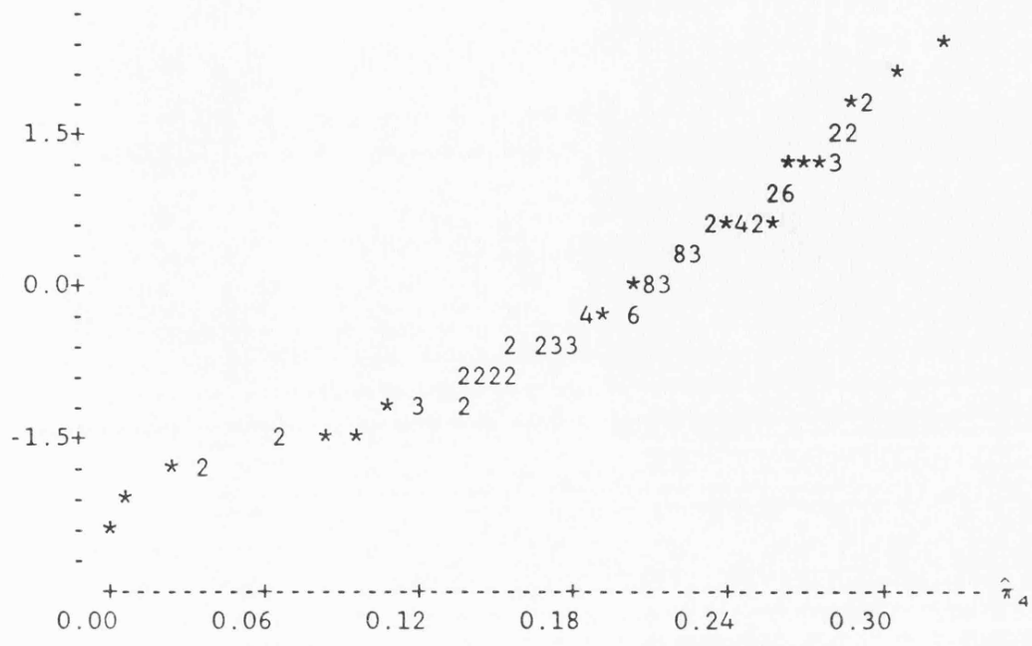


Figure 3.13- Normal probability plotting of the boostrap parameter estimate $\hat{\pi}_4$ for the Stouffer and Toby data (original ML $\hat{\pi}_4 = 0.21$, bootstrap $\hat{\pi}_4 = 0.18$ and $R^2 = 94.4\%$).

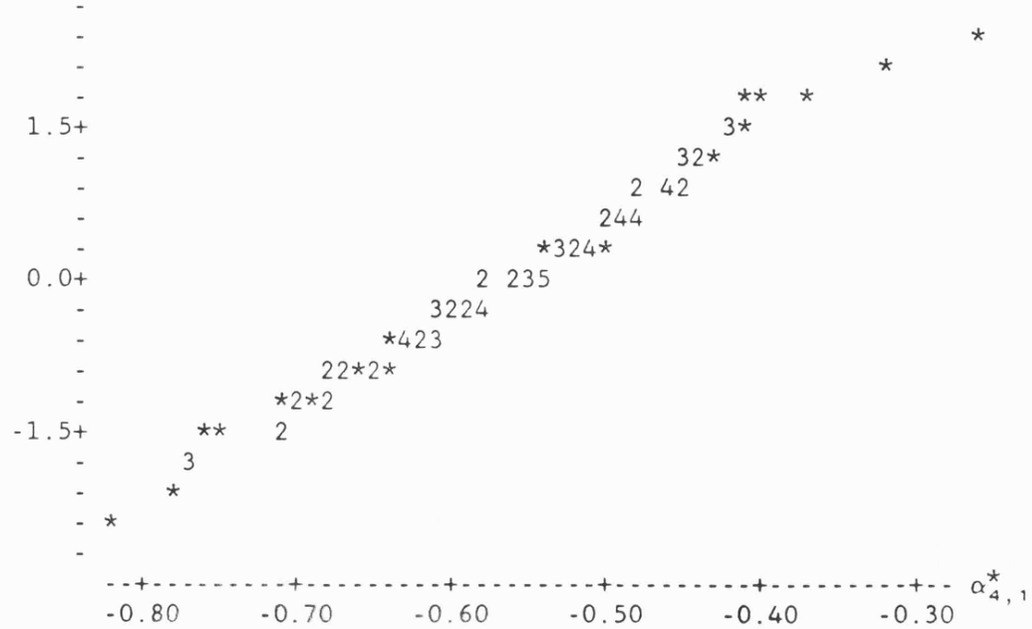


Figure 3.14- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}^*_{4,0}$ for the Stouffer and Toby data (original ML $\hat{\alpha}^*_{4,0} = -0.57$, bootstrap $\hat{\alpha}^*_{4,0} = -0.56$ and $R^2 = 99.0\%$).

The analysis of the adequacy of the asymptotic covariance matrix will be complemented with a comparison between the correlation matrices of the original ML, the bootstrap and the jackknife parameter estimates $\hat{\alpha}_{i,1}$ and $\hat{\alpha}_{i,0}$, for i=1,...,4, based on Tables 3.16 and 3.17.

Table 3.16- Correlations between the original ML parameter estimates based on the observed 2nd derivative matrix (under the diagonal) and on the information matrix (above the diagonal) for the Stouffer and Toby data.

	$\hat{\alpha}_{1,1}$	$\hat{\alpha}_{2,1}$	$\hat{\alpha}_{3,1}$	$\hat{\alpha}_{4,1}$	$\hat{\alpha}_{1,0}$	$\hat{\alpha}_{2,0}$	$\hat{\alpha}_{3,0}$	$\hat{\alpha}_{4,0}$
$\hat{\alpha}_{1,1}$		-0.05	-0.03	-0.05	0.64	0.01	0.01	0.06
$\hat{\alpha}_{2,1}$	0.05		-0.06	-0.26	-0.05	0.01	0.00	0.22
$\hat{\alpha}_{3,1}$	-0.10	-0.13		-0.16	-0.04	0.00	0.05	0.15
$\hat{\alpha}_{4,1}$	-0.11	-0.29	-0.06		-0.05	-0.01	-0.02	-0.74
$\hat{\alpha}_{1,0}$	0.66	0.02	-0.09	-0.09		0.17	0.15	0.16
$\hat{\alpha}_{2,0}$	0.02	0.04	0.00	-0.02	0.17		0.28	0.24
$\hat{\alpha}_{3,0}$	0.00	-0.01	0.07	-0.02	0.15	0.28		0.22
$\hat{\alpha}_{4,0}$	0.11	0.24	0.07	-0.71	0.19	0.25	0.22	

The correlations based on the observed second derivative are nearly equal to those obtained from the information matrix, since the maximum difference is 0.10, the asymptotic correlation between $\hat{\alpha}_{1,1}$ and $\hat{\alpha}_{2,1}$ (0.05 compared to -0.05).

The strongest correlations are between $\alpha_{1,1}$ and $\alpha_{1,0}$ (0.66 and 0.64) and between $\alpha_{4,1}$ and $\alpha_{4,0}$ (-0.71 and -0.74), while the remaining parameter are not or weakly correlated.

Table 3.17- Bootstrap (under the diagonal) and jackknife (above the diagonal) estimates of correlations between the parameter estimates for the Stouffer and Toby data.

	$\hat{\alpha}_{1,1}$	$\hat{\alpha}_{2,1}$	$\hat{\alpha}_{3,1}$	$\hat{\alpha}_{4,1}$	$\hat{\alpha}_{1,0}$	$\hat{\alpha}_{2,0}$	$\hat{\alpha}_{3,0}$	$\hat{\alpha}_{4,0}$
$\hat{\alpha}_{1,1}$		0.22	-0.11	-0.10	0.49	0.11	0.09	0.16
$\hat{\alpha}_{2,1}$	0.12		-0.13	-0.34	0.10	-0.04	0.08	0.32
$\hat{\alpha}_{3,1}$	-0.29	-0.22		0.01	-0.06	0.10	0.01	0.08
$\hat{\alpha}_{4,1}$	-0.16	-0.20	0.07		-0.03	0.07	0.06	-0.80
$\hat{\alpha}_{1,0}$	0.66	0.06	-0.24	-0.03		0.22	0.15	0.13
$\hat{\alpha}_{2,0}$	0.09	-0.24	0.20	0.17	0.29		0.27	0.14
$\hat{\alpha}_{3,0}$	0.01	-0.16	0.09	0.22	0.22	0.27		0.16
$\hat{\alpha}_{4,0}$	0.16	0.22	0.01	-0.92	0.12	-0.14	-0.19	

Comparing Tables 3.15 and 3.16, we can see that the largest differences between the bootstrap estimates of the correlations and the asymptotic ML correlations based on the observed second derivative matrix are equal to 0.41 between $\hat{\alpha}_{4,0}$ and $\hat{\alpha}_{3,0}$ (-0.19 compared to 0.22) and 0.39, between $\hat{\alpha}_{4,0}$ and $\hat{\alpha}_{2,0}$ (-0.14 compared to 0.25). They are followed by a difference of 0.24 for the correlation between $\hat{\alpha}_{3,0}$ and $\hat{\alpha}_{4,1}$ (0.22 compared to -0.02) and other 4 differences around 0.20, while the remaining are smaller than |0.10|. There are no significant differences between these results, for which the correlations are based on the observed second derivatives, and those based on the information matrix.

The maximum difference between the jackknife estimates of the correlations and the asymptotic correlations based on the second observed derivative matrix is 0.17 between $\hat{\alpha}_{1,1}$ and $\hat{\alpha}_{1,0}$ (0.49 compared to 0.66) and also between $\hat{\alpha}_{1,1}$ and $\hat{\alpha}_{2,1}$ (0.22 compared to

0.05). If we use the results from the information matrix instead all the differences remain nearly the same, except for the correlation between $\hat{\alpha}_{1,1}$ and $\hat{\alpha}_{2,1}$ that increases to 0.27.

Therefore the jackknife estimates of the correlations between $\hat{\alpha}_{i,1}$ and $\hat{\alpha}_{j,0}$, for i,j=1,..,4, are equal or closer to the asymptotics than the bootstrap estimates.

5- *Cancer Knowledge*

The data from a study on knowledge about cancer were given by Lombard and Doering (1947). They are displayed in Table 2.5, Chapter 2.

Table 3.18- Comparison between the bootstrap, original ML(in brackets) and the jackknife parameter estimates $\hat{\alpha}_{i,1}$ for the Lombard and Doering data.

i	$\hat{\alpha}_{i,1}$	$SD(\hat{\alpha}_{i,1})$	$CV(\hat{\alpha}_{i,1})$	$R^2(\hat{\alpha}_{i,1})$
1	0.73 (0.72) 0.72	.09(.09) .09	.12 (.12) .12	98.3 87.5
2	4.14 (3.40) 3.01	2.71(1.14)1.19	.65 (.34) .40	67.4 73.9
3	1.39 (1.34) 1.31	.19(.17) .17	.14 (.13) .13	95.4 72.5
4	0.82 (0.77) 0.80	.14 (.14) .15	.17 (.18) .19	99.2 65.7

Table 3.19- Comparison between the bootstrap, original ML(in brackets) and the jackknife parameter estimates $\hat{\alpha}_{i,0}$ for the Lombard and Doering data.

i	$\hat{\alpha}_{i,0}$	$SD(\hat{\alpha}_{i,0})$	$CV(\hat{\alpha}_{i,0})$	$R^2(\hat{\alpha}_{i,0})$
1	-1.29 (-1.29) -1.28	.07 (.06) .07	.05 (.05) .05	99.4 79.8
2	0.74 (0.60) 0.55	.50 (.17) .20	.68 (.28) .36	66.4 81.3
3	-0.14 (-0.14) -0.13	.07 (.08) .06	.50 (.57) .46	99.2 79.8
4	-2.75 (-2.75) -2.74	.12 (.18) .12	.04 (.06) .04	99.4 48.2

From Tables 3.18 and 3.19 we can see that, except for item 2, there is very good agreement between all the bootstrap and the original ML results and the bootstrap distributions of the parameter estimates $\alpha_{i,1}$ and $\alpha_{i,0}$ are fitted very well by a Normal distribution ($R^2 \geq 95.4\%$).

For item 2, the differences between the two results are significant and the fittings by a normal distribution are not good, since R^2 equals 67.4% and 66.4% (see Figures 3.15 and 3.16). Actually, these two figures suggest that either the bootstrap distributions of $\hat{\alpha}_{2,1}$ and $\hat{\alpha}_{2,0}$ are fitted by two distributions, one normal and another with $\alpha_{2,1}$ equal to infinity) or they are fitted by only one normal distribution extremely skewed. The bootstrap parameter estimates are larger than the original ML so that the bootstrap $CV(\hat{\alpha}_{2,1})$ and $CV(\hat{\alpha}_{2,0})$ are 91% and 143% larger than the corresponding original ML ones.

There is very good agreement between all jackknife estimates and their standard deviations and the corresponding original ML ones,

except for $\alpha_{2,1}$, for which the jackknife estimate is even smaller (3.01 compared to 3.40).

For item 2, jackknife estimates are closer to the original ML than to the bootstrap (4.14), indicating that jackknife tends to be closer to ML estimates than to bootstrap when one of the $\hat{\alpha}_{i,1}$ is large compared with the other.

Bootstrap and jackknife estimates of $\alpha_{i,1}$, for i=1,3,4, are approximately unbiased, since the boostrap biases are 0.01, -0.05 and -0.05 and the jackknife biases are 0.00, -0.03 and 0.03, respectively. The jackknife bias of $\alpha_{2,1}$ (-0.39) is smaller than the corresponding bootstrap one (0.74).

The fit by a normal distribution of the pseudovalues does not give information, as in bootstrap, about the relation between the jackknife results and original ML ones.

Table 3.20- Comparison between the bootstrap, original ML(in brackets) and the jackknife parameter estimates $\hat{\alpha}^*_{i,0}$ for the Lombard and Doering data.

i	$\hat{\alpha}^*_{i,0}$	$SD(\hat{\alpha}^*_{i,0})$	$CV(\hat{\alpha}^*_{i,0})$	$R^2(\hat{\alpha}^*_{i,0})$
1	-1.04 (-1.04) -1.03	.06 (.06) .05	.06 (.06) .05	98.7 77.6
2	0.17 (0.17) 0.18	.03 (.03) .03	.18 (.19) .17	98.8 79.8
3	-0.08 (-0.08) -0.09	.04 (.05) .04	.50 (.62) .44	99.5 70.7
4	-2.12 (-2.18) -2.13	.11 (.12) .11	.05 (.05) .05	98.1 56.7

Table 3.20 shows that the 3 methods give approximately the same estimates for $\alpha^*_{i,0}$ and their standard deviations. For item 3, the coefficient of variation of the original ML estimates is slightly bigger than the corresponding bootstrap and jackknife parameter

estimates (0.62 compared to 0.52 and 0.44, respectively).

As in all preceding examples, in this case too the bootstrap distribution $\hat{\alpha}^*_{i,0}$, i=1,...,4, has an excellent fit by a normal distribution and most of the $\hat{\alpha}^*_{i,0}$'s and their standard deviations are equal to the original ML ones.

The approximately normal bootstrap distributions of the parameter estimates for item 4 are presented in Figures 3.18 to 3.20.

According to the bootstrap results, this example seems to indicate that the asymptotic variance matrix may not be trusted when one of the $\hat{\alpha}_{i,1}$ is large (3.40 or more), Table 3.18, even for a large sample size (1729). On the other hand, jackknife does not provide any warning about possible underestimation of the asymptotic standard deviations.

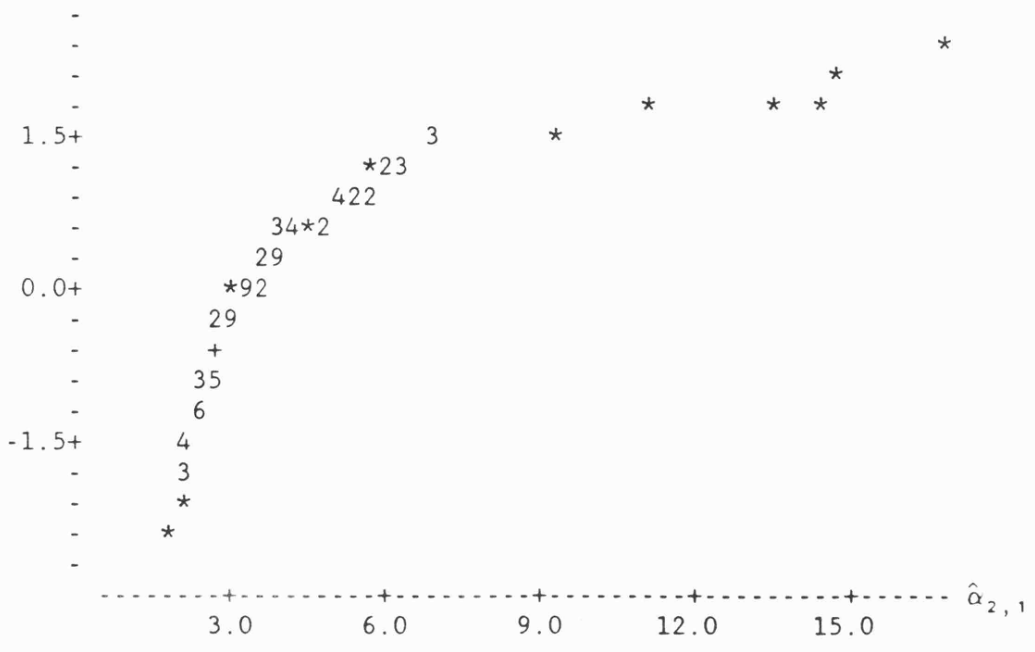


Figure 3.15- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}_{2,1}$ for the Lombard and Doering data (original ML $\hat{\alpha}_{2,1}$ = 3.40, bootstrap $\hat{\alpha}_{2,1}$ = 4.14 and R^2 = 67.4%).

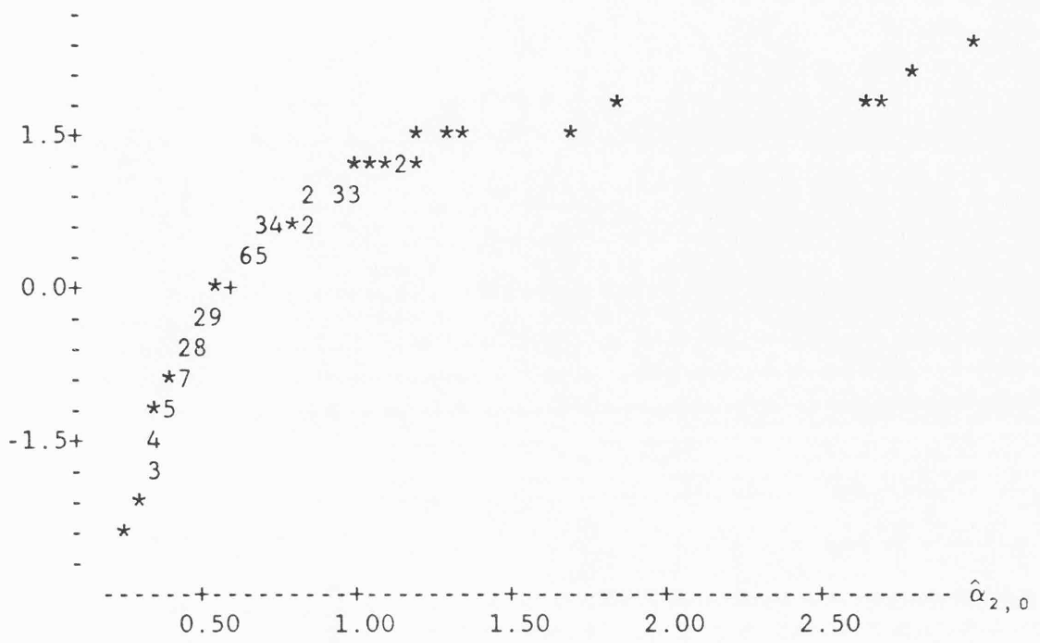


Figure 3.16- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}_{2,0}$ for the Lombard and Doering data (original ML $\hat{\alpha}_{2,0} = 0.60$, bootstrap $\hat{\alpha}_{2,0} = 0.74$ and $R^2 = 66.4\%$).

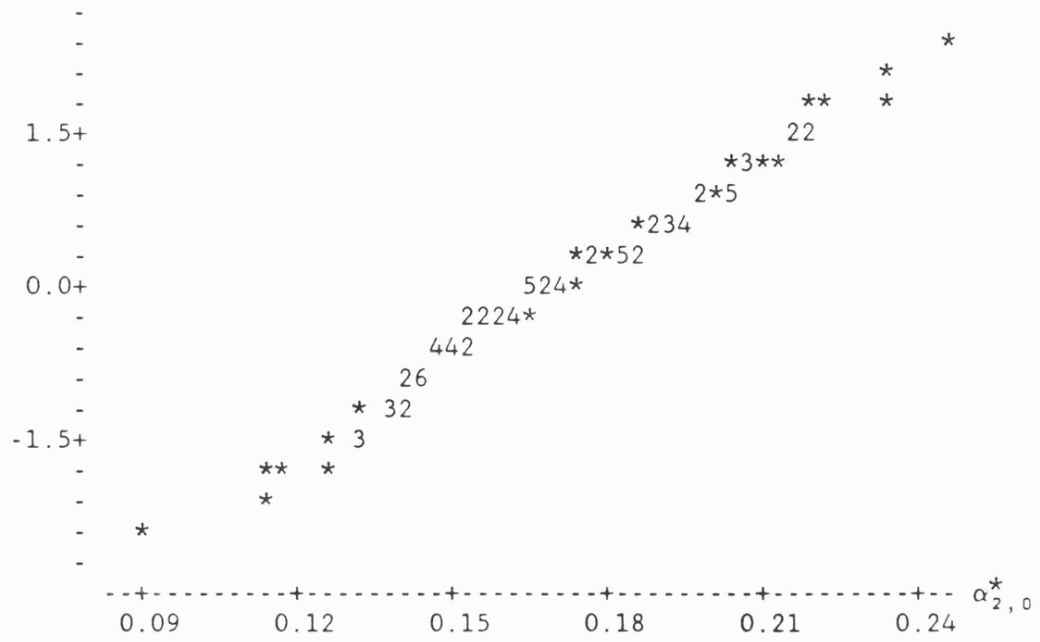


Figure 3.17- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}^*_{2,0}$ for the Lombard and Doering data (original ML and bootstrap $\hat{\alpha}^*_{2,0} = 0.17$, and $R^2 = 98.8\%$).

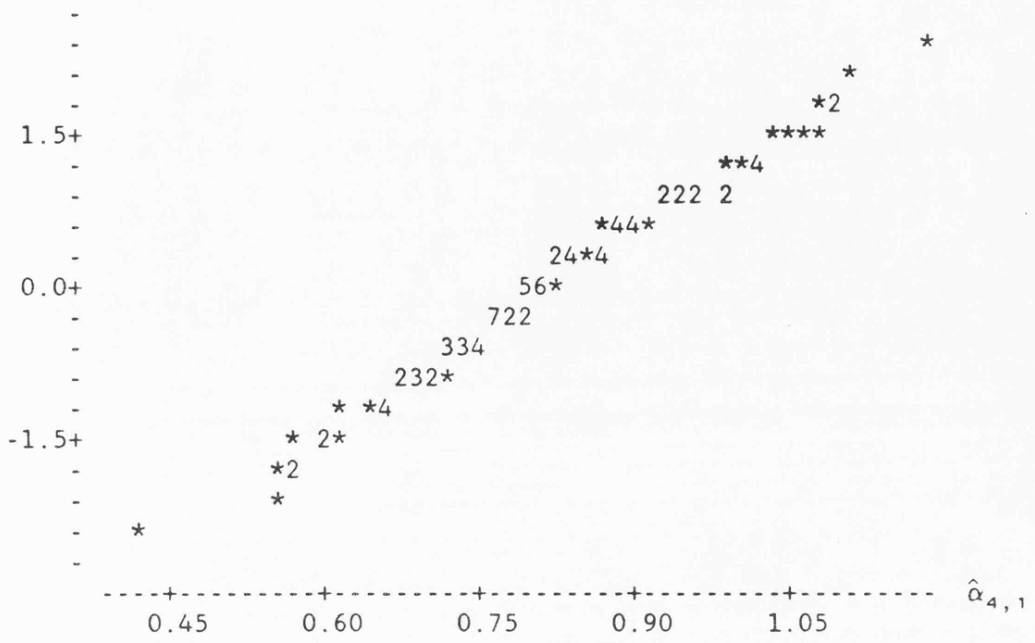


Figure 3.18- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}_{4,1}$ for the Lombard and Doering data (original ML $\hat{\alpha}_{4,1}$ = 0.77, bootstrap $\hat{\alpha}_{4,1}$ = 0.82 and R^2 = 99.2%).

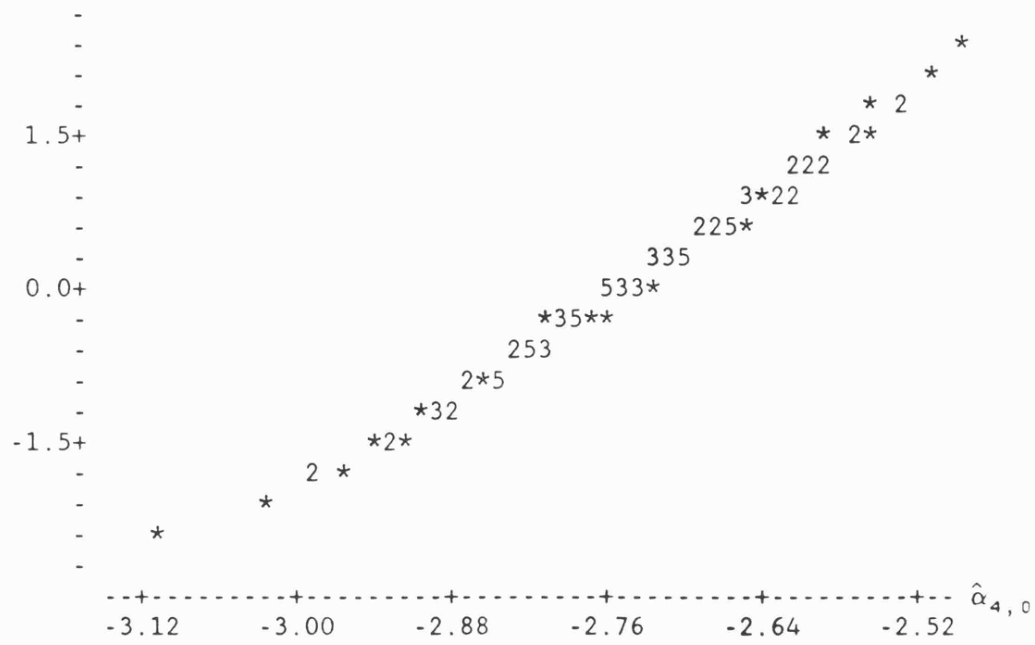


Figure 3.19- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}_{4,0}$ for the Lombard and Doering data (original ML and bootstrap $\hat{\alpha}_{4,0}$ equal to -2.75, and R^2 = 99.4%).

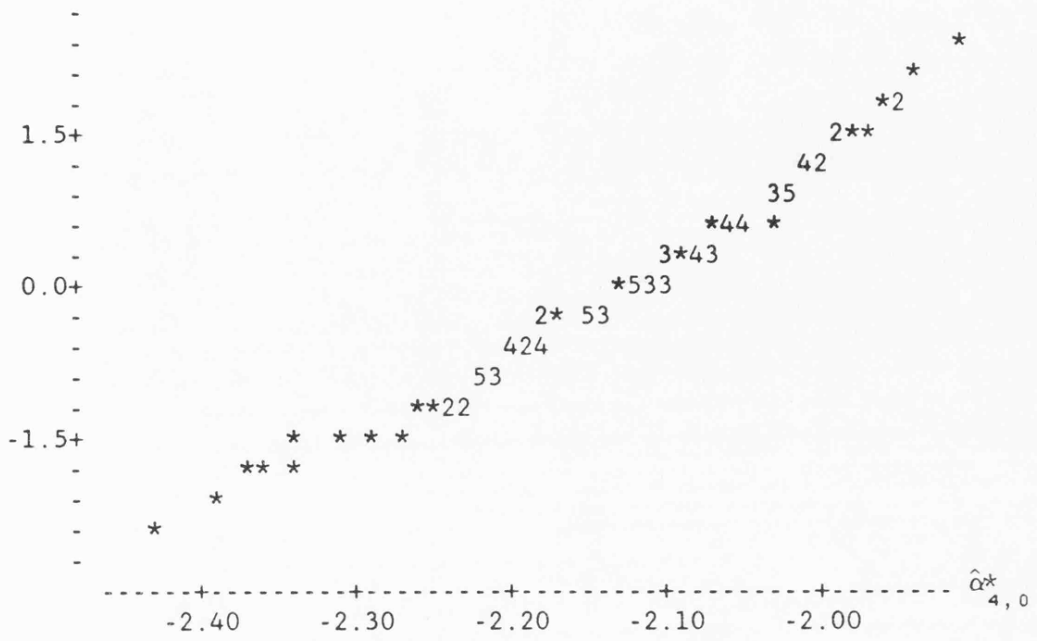


Figure 3.20- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}^*_{4,0}$ for the Lombard and Doering data (original ML $\hat{\alpha}^*_{4,0}$ = -2.18, bootstrap $\hat{\alpha}^*_{4,0}$ = -2.12 and R^2 = 98.1%).

In the following two tables we shall present and compare the bootstrap, the jackknife correlation matrices with asymptotic correlations between the parameter estimates.

Table 3.21- Correlations between the original ML parameter estimates based on the observed 2nd derivative matrix (under the diagonal) and on the information matrix (above the diagonal) for the Lombard and Doering data.

	$\hat{\alpha}_{1,1}$	$\hat{\alpha}_{2,1}$	$\hat{\alpha}_{3,1}$	$\hat{\alpha}_{4,1}$	$\hat{\alpha}_{1,0}$	$\hat{\alpha}_{2,0}$	$\hat{\alpha}_{3,0}$	$\hat{\alpha}_{4,0}$
$\hat{\alpha}_{1,1}$		-0.30	0.22	0.04	-0.44	-0.26	-0.03	-0.02
$\hat{\alpha}_{2,1}$	-0.22		-0.75	-0.16	0.14	0.82	0.10	0.11
$\hat{\alpha}_{3,1}$	0.13	-0.73		0.12	-0.09	-0.63	-0.12	-0.07
$\hat{\alpha}_{4,1}$	0.12	-0.35	0.27		-0.01	-0.15	-0.02	-0.65
$\hat{\alpha}_{1,0}$	-0.43	0.10	-0.05	-0.05		0.22	0.14	0.05
$\hat{\alpha}_{2,0}$	-0.20	0.81	-0.60	-0.30	0.19		0.29	0.15
$\hat{\alpha}_{3,0}$	0.13	0.10	-0.12	-0.04	0.13	0.30		0.08
$\hat{\alpha}_{4,0}$	0.12	0.24	-0.17	-0.67	0.07	0.26	0.10	

The similarity between the correlations based on the observed second derivative and the information matrix is good, since the larger difference between them is equal to 0.15 and corresponds to the correlation between $\hat{\alpha}_{1,1}$ and $\hat{\alpha}_{3,0}$ (0.13 and -0.03); $\hat{\alpha}_{4,1}$ and $\hat{\alpha}_{3,1}$ (0.27 and 0.12); and $\hat{\alpha}_{4,1}$ and $\hat{\alpha}_{2,0}$ (-0.30 and -0.15).

As in the preceding examples, most of the highest estimated correlations are between the discrimination and difficulty parameter estimates $\hat{\alpha}_{i,1}$ and $\hat{\alpha}_{j,0}$, i,j=1,..4.

Table 3.22- Bootstrap (under the diagonal) and jackknife (above the diagonal) estimates of correlations between the parameter estimates of the Lombard and Doering data.

	$\hat{\alpha}_{1,1}$	$\hat{\alpha}_{2,1}$	$\hat{\alpha}_{3,1}$	$\hat{\alpha}_{4,1}$	$\hat{\alpha}_{1,0}$	$\hat{\alpha}_{2,0}$	$\hat{\alpha}_{3,0}$	$\hat{\alpha}_{4,0}$
$\hat{\alpha}_{1,1}$		-0.17	0.06	0.21	-0.42	-0.17	-0.04	-0.13
$\hat{\alpha}_{2,1}$	-0.11		-0.74	-0.44	0.08	0.84	0.08	0.24
$\hat{\alpha}_{3,1}$	-0.11	-0.51		0.36	-0.06	-0.63	-0.11	-0.21
$\hat{\alpha}_{4,1}$	0.23	-0.13	0.30		-0.11	-0.37	-0.06	-0.68
$\hat{\alpha}_{1,0}$	-0.40	0.19	-0.18	-0.28		0.18	0.12	0.14
$\hat{\alpha}_{2,0}$	-0.12	0.97	-0.49	-0.11	0.25		0.28	0.25
$\hat{\alpha}_{3,0}$	0.02	-0.04	-0.15	0.03	0.15	0.03		0.12
$\hat{\alpha}_{4,0}$	-0.18	0.09	-0.16	-0.71	0.24	0.10	0.02	

The largest difference between the bootstrap estimates of the correlations and the original ML based on the observed second derivative matrix is equal to 0.27, between $\hat{\alpha}_{2,0}$ and $\hat{\alpha}_{3,0}$ (0.03 against 0.30), followed by a difference of 0.24 between $\hat{\alpha}_{1,1}$ and $\hat{\alpha}_{3,1}$ (-0.11 against 0.13), and 0.22 between $\hat{\alpha}_{2,1}$ and $\hat{\alpha}_{3,1}$ (-0.51 against -0.73) and between $\hat{\alpha}_{2,1}$ and $\hat{\alpha}_{4,1}$ (-0.13 against -0.35). These results are the same when comparing with the correlations based on the

information matrix, except for the correlation between $\hat{\alpha}_{1,1}$ and $\hat{\alpha}_{3,1}$, for which the difference is smaller, 0.09.

Comparisons between the jackknife estimates and the asymptotic correlations for the original ML parameter estimates, based on the observed second derivative matrix, show that the biggest difference is equal to 0.25 between $\hat{\alpha}_{1,1}$ and $\hat{\alpha}_{4,0}$ (-0.13 compared to 0.12), followed by 0.17 between $\hat{\alpha}_{1,1}$ and $\hat{\alpha}_{3,0}$ (-0.04 and 0.13). If we do the same comparison in relation to the correlations from the information matrix, these two differences will decrease to 0.11 and 0.01, respectively, and the remaining will change even less.

Therefore bootstrap and jackknife estimates of the correlations show the same degree of agreement with the asymptotic correlations between the original ML parameter estimates whether based on the observed 2nd derivative or on the information matrix.

6- *Arithmetic Reasoning Test on Black Women*

The last example corresponds to the data presented in Table 2.3 about the Arithmetic Reasoning Test (ART) on black women.

Table 3.23- Comparison between the bootstrap, original ML(in brackets) and the jackknife parameter estimates $\hat{\alpha}_{i,1}$ for the ART on black women.

i	$\hat{\alpha}_{i,1}$	$SD(\hat{\alpha}_{i,1})$	$CV(\hat{\alpha}_{i,1})$	$R^2(\hat{\alpha}_{i,1})$
1	6.79(14.39)38.82	7.16(67.78)6.77	1.05 (4.71) .17	83.8 95.4
2	1.63 (0.38) 0.32	3.32 (.22) .21	2.04 (0.58) .66	49.6 80.8
3	1.56 (0.37) 0.32	3.62 (.24) .22	2.32 (0.65) .69	47.3 91.4
4	0.14 (0.19) 0.16	2.94 (.24) .22	21.00 (1.26)1.38	44.0 93.2

Table 3.24- Comparison between the bootstrap, original ML(in brackets) and the jackknife parameter estimates $\hat{\alpha}_{i,0}$ for the ART on black women.

i	$\hat{\alpha}_{i,0}$	$SD(\hat{\alpha}_{i,0})$	$CV(\hat{\alpha}_{i,0})$	$R^2(\hat{\alpha}_{i,0})$
1	0.01 (0.24)130.10	1.27(4.63)3.05	127.00(18.52).02	93.2 76.0
2	-0.59(-0.33) -0.32	0.72 (.16) .16	1.22 (.48).50	54.3 75.8
3	-1.62(-0.96) -0.93	1.89 (.14) .18	1.17 (.14).19	45.7 70.8
4	-1.56(-1.08) -1.08	1.60 (.16) .18	1.02 (.15).17	36.7 67.7

This is an example of an extreme case where one $\hat{\alpha}_{i,1}$ dominates all the other items by its very large value (14.39), and the sample size (145) is very small.

Comparing the bootstrap with the original ML results in Tables 3.23 and 3.24 we can see some disagreement for all the results, the bootstrap estimates being larger than the original ML estimates, except for item 1. It seems that the dominating item 1 has affected all the other items, which show an even bigger discrepancy between the bootstrap and the original ML estimates.

The results also show that large values of the bootstrap parameter estimate $\hat{\alpha}_{i,1}$, that is, $\hat{\alpha}_{i,1} \geq 1.42$ for i=2,3,4, are always associated with small values of $\hat{\alpha}_{1,1}$, that is, $-0.38 \leq \hat{\alpha}_{1,1} \leq 0.77$.

While the bootstrap $\hat{\alpha}_{1,1}$'s are spread from -0.38 to 26.84, at least 90% of the bootstrap $\hat{\alpha}_{i,1}$ are concentrated between -1.48 and 1.48 for i=2,3,4. In items 2 and 3, up to 10% of the bootstrap $\hat{\alpha}_{i,1}$'s assume values from 1.48 to 17.96 while $\hat{\alpha}_{4,1}$ varies between -12.75 and 14.98 (Figure 3.24).

The performance of the bootstrap distribution of the parameter estimates $\hat{\alpha}_{i,1}$ with the corresponding $\hat{\alpha}_{i,0}$ is very similar, so that large values for $\hat{\alpha}_{i,1}$ are always associated with large $\hat{\alpha}_{i,0}$ in absolute value, as we can see in Figures 3.21, 3.22, 3.24 and 3.25 and Tables 3.23 and 3.24.

Figure 3.21 shows that the bootstrap distribution of $\hat{\alpha}_{1,1}$ either could be fitted by a mixture of two Normal distributions or by two different distributions: one normal and another with $\alpha_{1,1}$ equal to infinity. Although the normal probability plotting for the bootstrap distribution of $\hat{\alpha}_{1,0}$ provides R^2 equal to 93.2%, we may see a mixed of two normal distributions.

Figures 3.24 and 3.25 the normal probability plottings for $\hat{\alpha}_{i,1}$ and $\hat{\alpha}_{i,0}$ for $i \neq 1$. In these cases, it is more evident that most of the bootstrap estimates are fitted by a normal distribution, except for those sample for which $\hat{\alpha}_{i,1}$, $i \neq 1$, is very large. In this later case, it could be fitted by a distribution with $\alpha_{i,1}$, $i \neq 1$, equal to infinity.

The jackknife estimates and standard deviations are very close to the corresponding original ML ones, except for item 1 where jackknife $\hat{\alpha}_{1,1}$ and $\hat{\alpha}_{1,0}$ are bigger (38.82 and 130.10 compared to 14.39 and 0.24, respectively) with smaller coefficients of variation(0.17 and 0.20 compared to 4.71 and 18.52), probably underestimating the true standard deviations.

On the other hand, when comparing bootstrap with the original ML results we have seen that they disagree strongly and bootstrap gives a warning that the asymptotic theory probably is not working well.

The bootstrap estimates of bias of $\alpha_{i,1}$, i=1,...,4, are equal to -7.60, 1.25, 1.19 and -0.05, while the corresponding jackknife estimates are 24.60, -0.06, -0.05 and -0.03, respectively. Therefore jackknife has provided estimates equal or less biased than bootstrap, except for item 1.

The results for all items show great discrepancies between jackknife and bootstrap techniques. These discrepancies are related to the size of the parameter estimates, standard deviations and fit of a normal distribution.

Table 3.25- Comparison between the bootstrap, original ML(in brackets) and the jackknife parameter estimates $\hat{\alpha}^*_{i,0}$ for the ART on black women.

i	$\hat{\alpha}^*_{i,0}$	$SD(\hat{\alpha}^*_{i,0})$	$CV(\hat{\alpha}^*_{i,0})$	$R^2(\hat{\alpha}^*_{i,0})$
1	0.01 (0.02) 0.05	.14 (.24) .20	14.00(12.00)4.00	97.1 72.7
2	-0.31 (-0.31) -0.31	.15 (.13) .15	.48 (.42) .48	98.7 69.9
3	-0.85 (-0.90) -0.94	.20 (.11) .17	.24 (.12) .18	99.1 69.6
4	-1.04 (-1.06) -1.13	.20 (.14) .17	.19 (.13) .15	99.0 64.3

The bootstrap parameter estimates $\hat{\alpha}^*_{i,0}$ are not affected by the skewness of $\hat{\alpha}_{i,1}$ and $\hat{\alpha}_{i,0}$, i=1,...,4, showing substantial agreement with the original ML ones. This is to be expected, since they are very well approximated by a normal distribution, as we can see in Table 3.25 and Figures 3.23 and 3.27.

Jackknife results for $\alpha^*_{i,0}$ are very close to the original ML, except for item 1, which coefficient of variation is smaller (4.00 compared to 12.00). This is due to the fact that the reparametrization

depends on $\alpha_{i,1}$ and $\alpha_{i,0}$, for with the coefficient of variations are also smaller than those given by the asymptotic theory.

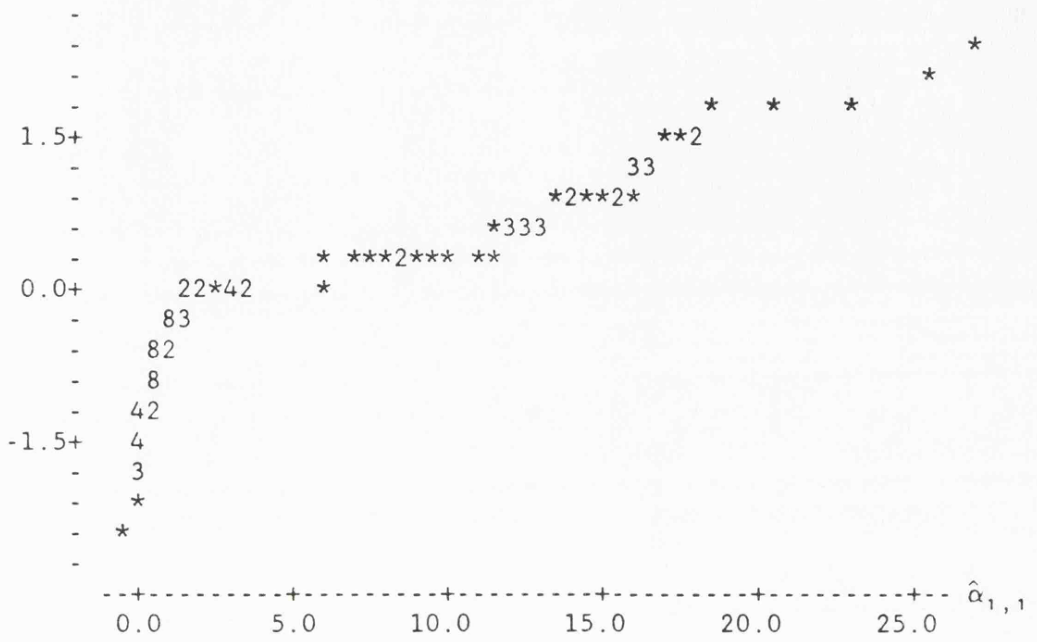


Figure 3.21- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}_{1,1}$ to the ART on black women (original ML $\hat{\alpha}_{1,1}$ = 14.39, bootstrap $\hat{\alpha}_{1,1}$ = 6.79 and R^2 = 83.8%).

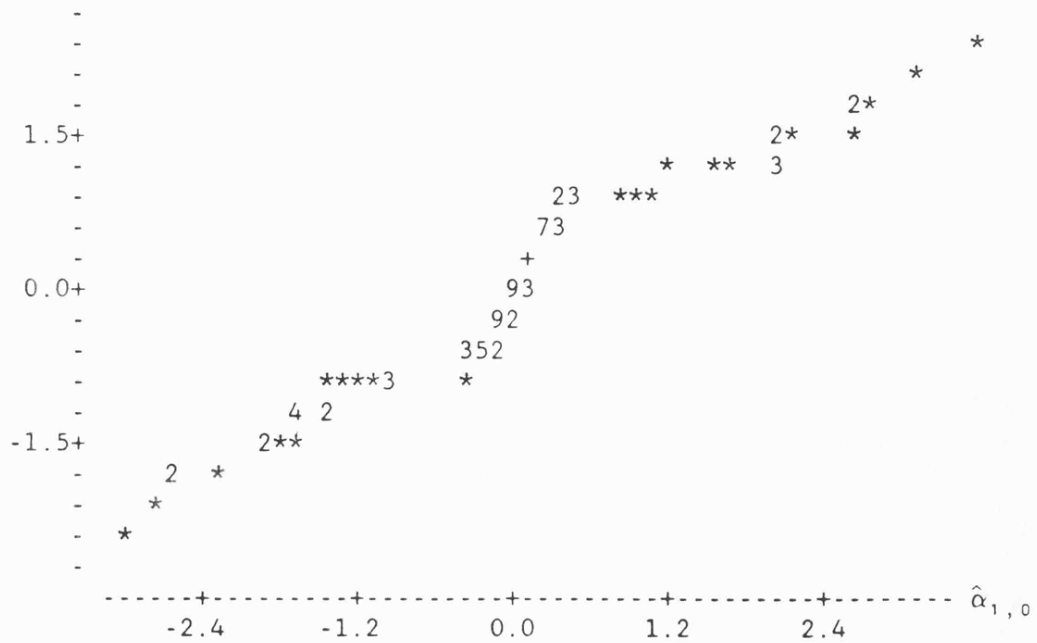


Figure 3.22- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}_{1,0}$ to the ART on black women (original ML $\hat{\alpha}_{1,0}$ = 0.24, bootstrap $\hat{\alpha}_{1,0}$ = 0.01 and R^2 = 93.2%).

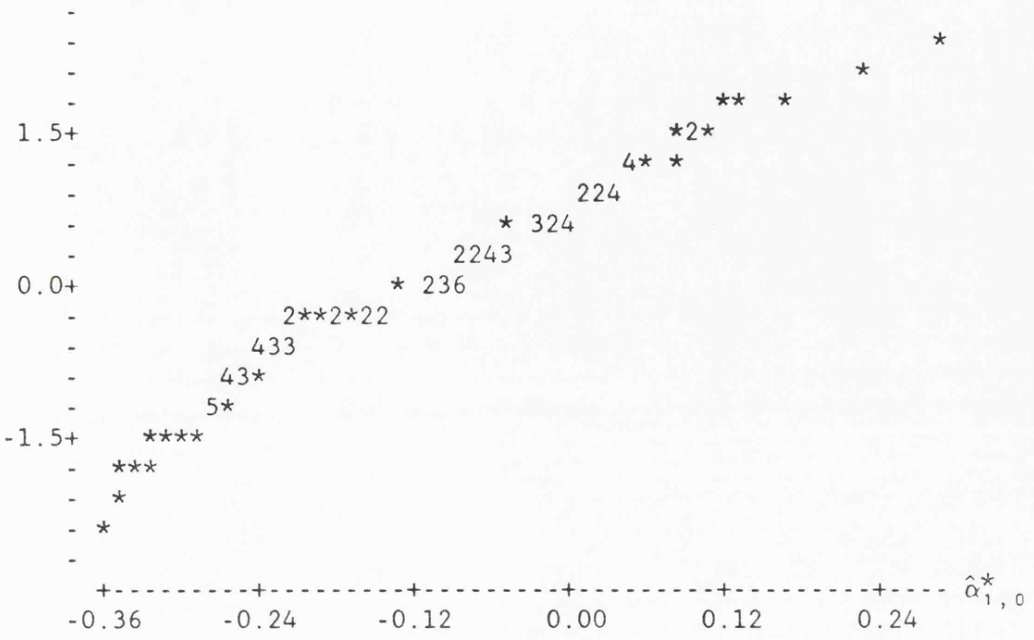


Figure 3.23- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}^*_{1,0}$ to the ART on black women (original ML $\hat{\alpha}^*_{1,0}$ = 0.02, bootstrap $\hat{\alpha}^*_{1,0}$ = 0.01 and R^2 = 97.1%).

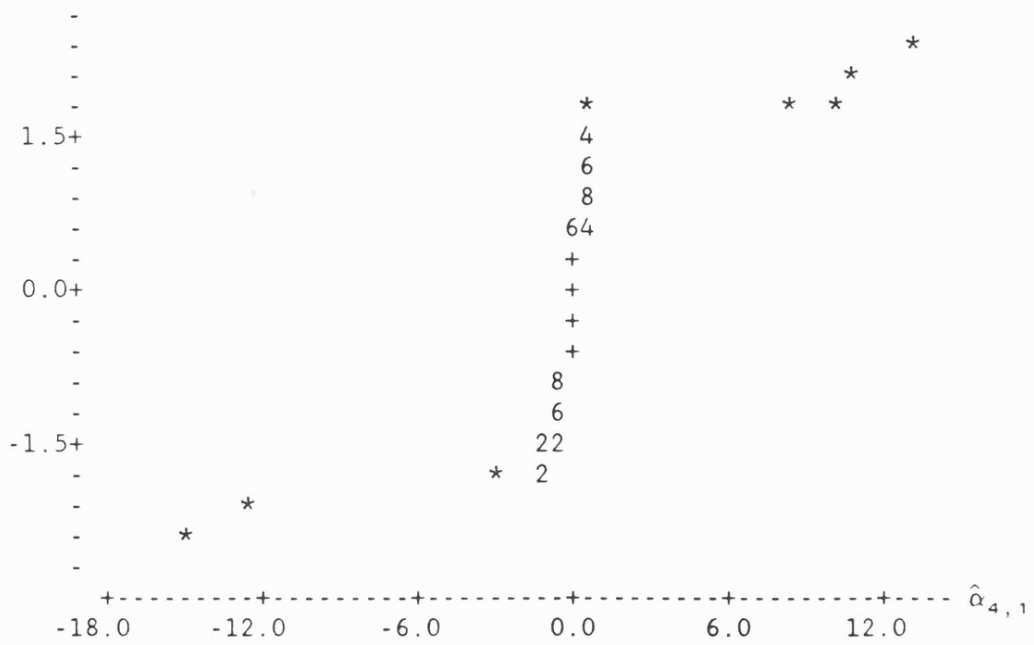


Figure 3.24- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}_{4,1}$ to the ART on black women (original ML $\hat{\alpha}_{4,1}$ = 0.19, bootstrap $\hat{\alpha}_{4,1}$ = 0.14 and R^2 = 44.0%).

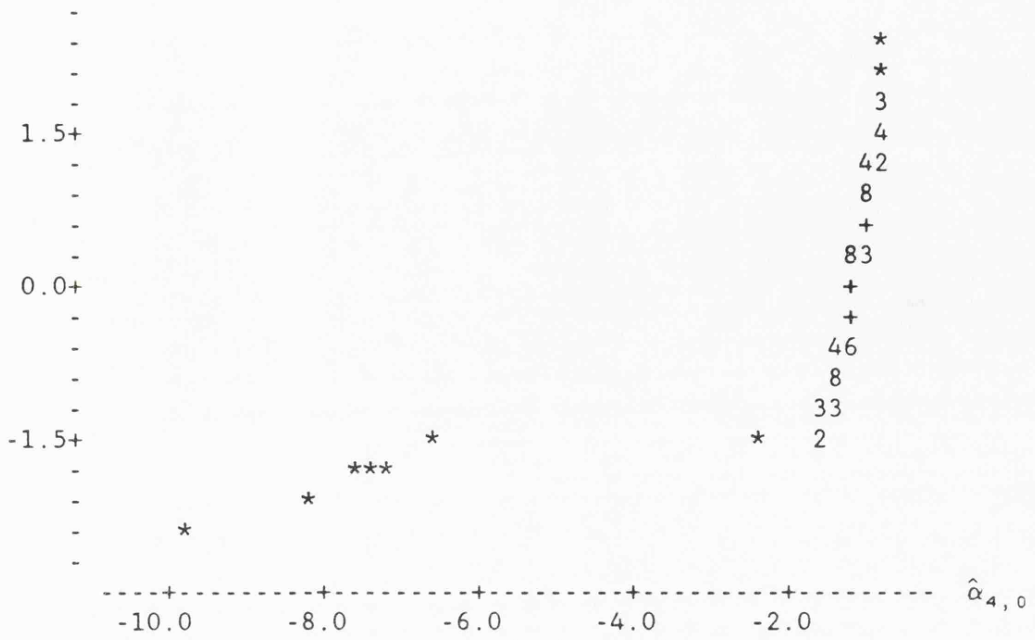


Figure 3.25- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}_{4,0}$ to the ART on black women (original ML $\hat{\alpha}_{4,0}$ = -1.08, bootstrap $\hat{\alpha}_{4,0}$ = -1.56 and R^2 = 36.7%).

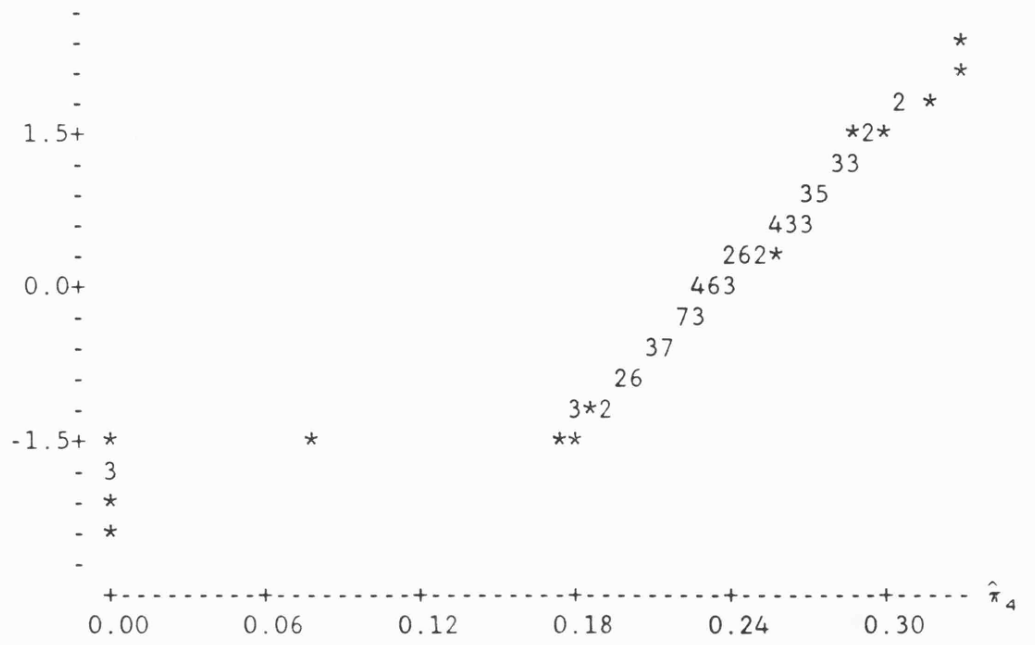


Figure 3.26- Normal probability plotting of the bootstrap parameter estimate $\hat{\pi}_4$ to the ART on black women (original ML and bootstrap $\hat{\pi}_4$ equal to 0.23, and R^2 = 77.0%)

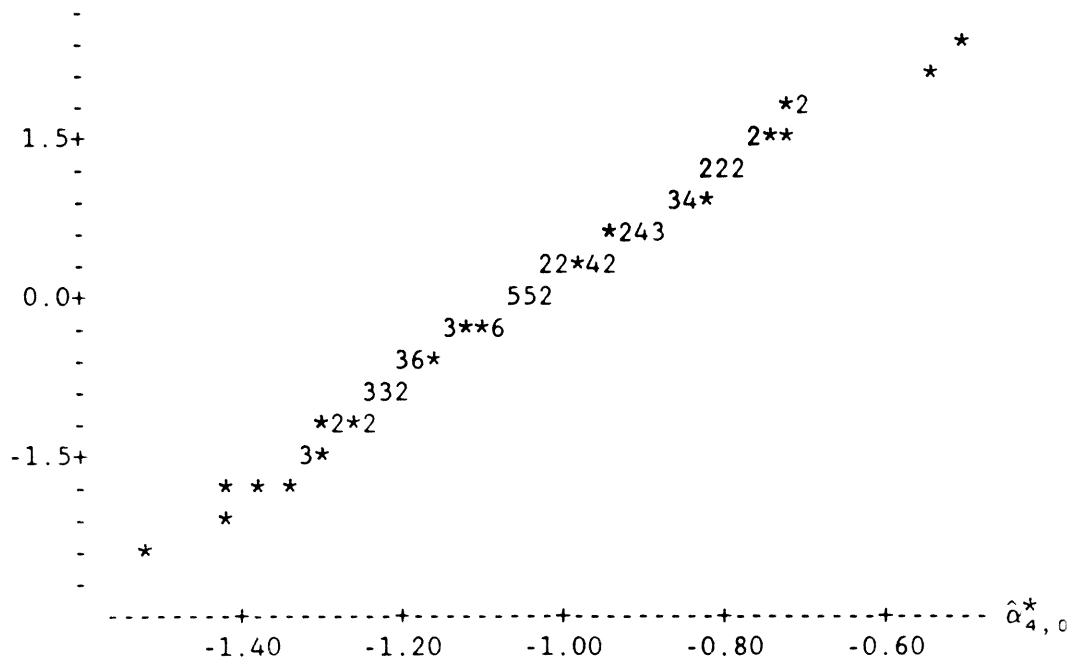


Figure 3.27- Normal probability plotting of the bootstrap parameter estimate $\hat{\alpha}^*_{4,0}$ to the ART on black women (original ML $\hat{\alpha}^*_{4,0}$ = -1.06, bootstrap $\hat{\alpha}^*_{4,0}$ = -1.04 and R^2 = 99.0%).

In order to complete the comparison between bootstrap,jackknife and the original ML parameter estimates we shall consider the correlation matrix given by Tables 3.26 and 3.27.

Table 3.26- Correlation matrix of the original ML parameter estimates based on the observed 2nd derivative matrix (under the diagonal) and on the information matrix (above the diagonal) for the ART on black women.

	$\hat{\alpha}_{1,1}$	$\hat{\alpha}_{2,1}$	$\hat{\alpha}_{3,1}$	$\hat{\alpha}_{4,1}$	$\hat{\alpha}_{1,0}$	$\hat{\alpha}_{2,0}$	$\hat{\alpha}_{3,0}$	$\hat{\alpha}_{4,0}$
$\hat{\alpha}_{1,1}$		-0.90	-0.87	-0.66	1.00	0.14	0.33	0.10
$\hat{\alpha}_{2,1}$	-0.17		0.78	0.59	-0.89	-0.16	-0.29	-0.09
$\hat{\alpha}_{3,1}$	-0.15	0.03		0.57	-0.87	-0.12	-0.37	-0.09
$\hat{\alpha}_{4,1}$	-0.01	-0.02	-0.02		-0.65	-0.09	-0.21	-0.15
$\hat{\alpha}_{1,0}$	0.73	0.10	0.10	0.04		0.16	0.34	0.10
$\hat{\alpha}_{2,0}$	0.01	-0.07	0.00	0.00	0.10		0.07	0.03
$\hat{\alpha}_{3,0}$	0.03	0.00	-0.19	0.01	0.10	0.03		0.04
$\hat{\alpha}_{4,0}$	0.00	0.00	0.00	-0.11	0.04	0.01	0.01	

There is strong disagreement between the correlation matrix of the original ML parameter estimates based on the observed second derivative and on the information matrix, with the later presenting equal or bigger values.

The largest discrepancies are related to the correlation between all the $\hat{\alpha}_{i,1}$ and $\hat{\alpha}_{j,1}$ and between $\hat{\alpha}_{1,0}$ and $\hat{\alpha}_{i,1}$, i=1,..,4. While the correlations, based on the observed second derivative matrix, between $\hat{\alpha}_{i,1}$ and $\hat{\alpha}_{j,1}$ range from -0.17 to 0.03, those from the information matrix range from -0.90 to -0.66 and from 0.57 to 0.78. Thus while in the former the parameter seems not to be correlated, in the latter some are even strongly correlated. Furthermore, while the correlations based on the observed second derivative matrix between $\hat{\alpha}_{1,0}$ and $\hat{\alpha}_{i,1}$, i=1,..,4, are equal to 0.73, 0.10, 0.10 and 0.04, which based on the information matrix are equal to 1.00, -0.89, -0.87 and -0.65, respectively.

The presence of the parameter estimates of item 1 in most of the largest discrepancies between both methods, may be associated with the large value assumed by $\hat{\alpha}_{1,1}$ (14.39) and the untrusted estimated standard deviations of $\alpha_{1,1}$ and $\alpha_{1,0}$ (67.78 and 4.63).

Table 3.27- Bootstrap (under the diagonal) and the jackknife (above the diagonal) estimates of correlations between the parameter estimate of the ART on black women.

	$\hat{\alpha}_{1,1}$	$\hat{\alpha}_{2,1}$	$\hat{\alpha}_{3,1}$	$\hat{\alpha}_{4,1}$	$\hat{\alpha}_{1,0}$	$\hat{\alpha}_{2,0}$	$\hat{\alpha}_{3,0}$	$\hat{\alpha}_{4,0}$
$\hat{\alpha}_{1,1}$		0.09	0.10	0.54	-0.11	-0.05	-0.03	-0.02
$\hat{\alpha}_{2,1}$	-0.30		0.01	-0.10	0.02	-0.07	-0.02	-0.01
$\hat{\alpha}_{3,1}$	-0.27	-0.18		-0.10	0.02	-0.02	-0.20	0.00
$\hat{\alpha}_{4,1}$	0.07	-0.07	-0.12		0.06	-0.01	0.00	-0.12
$\hat{\alpha}_{1,0}$	0.04	0.01	-0.02	0.00		0.11	0.10	0.01
$\hat{\alpha}_{2,0}$	0.26	-0.90	0.18	0.06	0.01		0.03	-0.05
$\hat{\alpha}_{3,0}$	0.29	0.15	-0.98	0.04	0.05	-0.16		-0.05
$\hat{\alpha}_{4,0}$	0.24	0.11	0.12	-0.11	-0.02	-0.08	-0.07	

The bootstrap estimates of the correlations between the parameter estimates assume values from -0.30 to 0.29 and from -0.90 to -0.98. the bootstrap estimate -0.90 corresponds to the correlation between $\hat{\alpha}_{2,1}$ and $\hat{\alpha}_{2,0}$, for which the observed second derivative provides a value equal to -0.07, the information matrix, -0.16, and the jackknife, -0.07. On the other hand, -0.98 is the correlation between $\hat{\alpha}_{3,1}$ and $\hat{\alpha}_{3,0}$, which is equal to -0.19 when based on the observed second derivatives, and equal to -0.37 on the information matrix, and finally -0.20 for the jackknife estimate.

The only large asymptotic correlation based on the observed second derivative matrix is between $\hat{\alpha}_{1,1}$ and $\hat{\alpha}_{1,0}$ and equal to 0.73, but it assumes a small value 0.04 in the bootstrap and -0.11 in the jackknife estimation.

Table 3.27 shows that jackknife estimates of the correlation assume values between -0.11 and 0.11 and one value 0.54 associated to $\hat{\alpha}_{1,1}$ and $\hat{\alpha}_{4,1}$. For the later estimates the asymptotic correlations are -0.01 (observed second derivatives) and -0.66 (information matrix), and 0.07 for the bootstrap estimate.

These results indicate that although there are some large discrepancies between the bootstrap and jackknife results compared with the asymptotic correlation matrix, they are closer to the correlation based on the observed second derivative than those based on the information matrix.

7- *Comparison of Bootstrap, Normal Bootstrap and ML Estimates*

The aim of this section is to investigate how close the (empirical) bootstrap parameter estimates $\hat{\alpha}_{i,1}$'s are to the corresponding normal bootstrap ones, in order to obtain more evidence which confirm the bootstrap results about the adequacy of the asymptotic variance-covariance matrix presented above.

We shall carried out this study considering 100 normal bootstrap samples for each one of the 5 sets of data, which will be drawn from a multinomial distribution with parameters $\pi_i(z)$, i=1,...,4, where $\pi_i(z)$ is the response function of a logit/probit model with parameters $\alpha_{i,1}$ and $\alpha_{i,0}$ equal to the ML estimates from the real data, and the latent variable Z is distributed as N(0,1).

We shall compare the bootstrap methods (empirical and normal) in relation to the mean, median, interquartile difference Q_3-Q_1 and standard deviation of the corresponding bootstrap distribution of $\hat{\alpha}_{i,1}$, i=1,...,4, We complement the analysis comparing both bootstrap estimates with the original ML results.

Table 3.28- Comparison between the bootstrap, original ML (in brackets) and the normal bootstrap parameter estimates $\hat{\alpha}_{i,1}$ to the Attitudes towards the U.S.Army.

i	$\hat{\alpha}_{i,1}$	$SD(\hat{\alpha}_{i,1})$	$Q_3 - Q_1$	Median
1	1.68 (1.64) 1.68	.25 (.24) .24	.34 (.32) .32	1.67 1.63
2	1.13 (1.12) 1.11	.15 (.14) .15	.20 (.19) .16	1.13 1.10
3	1.45 (1.41) 1.41	.20 (.19) .18	.31 (.26) .21	1.44 1.39
4	1.63 (1.60) 1.60	.20 (.22) .22	.27 (.30) .28	1.64 1.58

Table 3.28 shows an excellent agreement between the three procedures. The fitting of both bootstrap parameter estimates by a normal distribution ($R^2 \geq 92.1\%$) is shown by the similarity between mean and median, and as well by Q_3-Q_1.

Table 3.29- Comparison between the bootstrap, original ML (in brackets) and the normal bootstrap parameter estimates $\hat{\alpha}_{i,1}$ for the ART on white women.

i	$\hat{\alpha}_{i,1}$	$SD(\hat{\alpha}_{i,1})$	$Q_3 - Q_1$	Median
1	1.14 (1.04) 1.08	.42 (.32) .31	.41 (.43) .47	1.09 1.04
2	1.26 (1.24) 1.29	.40 (.39) .44	.45 (.42) .49	1.17 1.27
3	1.04 (1.00) 0.97	.34 (.30) .33	.44 (.40) .42	1.04 0.92
4	1.51 (1.44) 1.69	.56 (.45) .85	.73 (.61) .72	1.38 1.46

The similarity between both bootstrap methods and the original ML results is very good, except for item 4 where both bootstrap $\hat{\alpha}_{4,1}$ are slightly bigger than the ML estimate (1.51 and 1.69 compared to 1.44).

For item 4 the normal bootstrap median is closer to the ML parameter estimate than the mean. The interquartile differences Q_3-Q_1 are equal for both bootstrap methods, but they are different from the corresponding asymptotics.

Table 3.30- Comparison between the bootstrap, original ML (in brackets) and the normal bootstrap parameter estimates $\hat{\alpha}_{i,1}$ for the Stouffer and Toby data.

i	$\hat{\alpha}_{i,1}$	$SD(\hat{\alpha}_{i,1})$	$Q_3 - Q_1$	Median
1	1.19 (1.15) 1.24	.38 (.36) .37	.49 (.48) .54	1.12 1.25
2	1.82 (1.58) 1.69	.83 (.44) .64	.62 (.59) .62	1.61 1.56
3	1.44 (1.35) 1.34	.46 (.36) .36	.52 (.48) .48	1.36 1.33
4	2.72 (2.10) 2.90	2.99 (.66)2.23	.99 (.89)1.37	2.12 2.32

Items 1 and 3 present estimates nearly equal when comparing the 3 methods. Items 2 and 4 present some discrepancies, which are stronger for item 4.

The larger bootstrap estimates of $\alpha_{2,1}$ than the original ML is due to the occurence of some large values when $\hat{\alpha}_{4,1}$ was small. This shows some instability of the bootstrap distribution, probably because of the small sample size (216).

Both bootstrap estimates $\hat{\alpha}_{4,1}$ are closer to each other than to the corresponding ML estimate (2.72 and 2.90 compared to 2.21). The bootstrap medians are closer to the ML estimate than to the means due

to the skewness of the bootstrap distribution.

Comparing Q_3-Q_1, we can say that the only difference is for item 4, and the (empirical) bootstrap estimate is closer to the ML than the normal bootstrap.

The higher estimates for the normal bootstrap of item 4 is due to the variation in $\hat{\alpha}_{4,1}$, which assumes values from 1.14 to 14.75 with 25% of them bigger than 3.15. Although there are some small differences in item 4, we can still say that both bootstrap methods present very similar results.

Table 3.31- Comparison between the bootstrap, original ML (in brackets) and the normal bootstrap parameter estimates $\hat{\alpha}_{i,1}$ for the Lombard and Doering data.

i	$\hat{\alpha}_{i,1}$	$SD(\hat{\alpha}_{i,1})$	$Q_3 - Q_1$	Median
1	0.73 (0.72) 0.73	.09 (.09) .10	.13 (.12) .11	.71 0.72
2	4.14 (3.40) 3.79	2.71(1.14)1.92	2.07(1.54)1.28	3.29 3.40
3	1.39 (1.34) 1.38	.19 (.19) .18	.25 (.23) .24	1.37 1.38
4	0.82 (0.77) 0.78	.14 (.22) .13	.18 (.19) .18	.81 .77

We can see from Table 3.31 that all result are nearly equal, except for item 2. Both bootstrap methods present larger estimates $\hat{\alpha}_{2,1}$ than the original ML, though the normal bootstrap estimate is closer to the latter than to the empirical bootstrap (3.79, 4.14 compared to 3.40). The bootstrap medians for item 2 are closer to the ML estimates than the means.

The differences between both bootstrap methods for $\hat{\alpha}_{2,1}$ can be better understood is we look at their distributions.

The bootstrap distribution of $\hat{\alpha}_{2,1}$ assumes values between 1.67 and 16.90 with 25% of them smaller than 2.67 and others 25% bigger than 4.74. On the other hand, in the normal bootstrap $\hat{\alpha}_{2,1}$ ranges from 2.03 to 13.54, with Q_1 equal to 2.75 and Q_3 equal to 4.03. Therefore in the normal bootstrap the spread of the estimates $\hat{\alpha}_{2,1}$ is smaller.

Some instability observed in the Stouffer and Toby data reflected by item 4 not occur in this example, though the larger value of ML $\hat{\alpha}_{2,1}$ (3.40 compared to 2.10), as we can see by the strong similarity among the 3 procedures for items 1, 3 and 4. This is probably due to the larger sample size (1729) of the Lombard and Doering data.

Table 3.32- Comparison between the bootstrap, original ML (in brackets) and the Normal bootstrap parameter estimates $\hat{\alpha}_{i,1}$ for the ART on black women.

i	$\hat{\alpha}_{i,1}$	$SD(\hat{\alpha}_{i,1})$	$Q_3 - Q_1$	Median	
1	6.79(14.39)5.70	7.16(67.78)5.33	12.37(91.43)10.68	2.96	2.53
2	1.63 (.38) .71	3.32 (.14) .15	.20 (.19) .16	1.13	1.10
3	1.56 (.37)1.17	.20 (.19) .18	.31 (.26) .21	1.44	1.39
4	.14 (.19) .24	.20 (.22) . 22	.27 (.30) .28	1.64	1.58

The original ML parameter estimates $\hat{\alpha}_{i,1}$, i=2,3, are smaller than the bootstrap ones, while the ML estimate $\hat{\alpha}_{4,1}$ is similar, but with much smaller standard deviation.

Medians of both bootstrap methods are closer to the original ML parameter estimates than the corresponding means, except for item 1.

The interquartile differences Q_3-Q_1 are very similar for both bootstrap methods, but they are very different from the asymptotic approximation, specially for $\hat{\alpha}_{1,1}$.

The original ML parameter estimate $\hat{\alpha}_{1,1}$ and its standard deviation are much larger than the bootstrap estimates, 14.39 compared to 6.79 and 5.70, and 67.78 compared to 7.16 and 5.33, respectively.

As we have already pointed out in Chapter 2, when fitting a logit/probit model to the Arithmetic Reasoning Test on black women, the ML parameter estimate $\hat{\alpha}_{1,1}$ could be equal to any value bigger than 3 or 4,since the likelihood function is flat after this point.

It is worth saying that when carrying out the bootstrap methods we have always considered the same stopping rule for the iterative procedure of the estimation of the parameters. Hence, using the same stop rule, the normal bootstrap estimate of $\alpha_{1,1}$ is equal to 5.70, though the bootstrap samples are drawn from a distribution with $\alpha_{1,1}$ equal to 14.39.

The normal bootstrap distribution of $\hat{\alpha}_{1,1}$ assumes values between 0.09 and 15.79 with Q_1 equal to 0.69 and Q_3 equal to 11.37. The fitting by a normal distribution is the same as for the empirical bootstrap (83.6%). Furthermore, 52% of the parameter estimates are bigger than 3.0 and the median is 2.53. These results seem to indicate that the 'true' parameter could be equal to 3.0.

The disagreement between the bootstrap results for the remaining items are strongly due to the influence of item 1, since $\hat{\alpha}_{i,1}$, i=2,3,4, assume only large values in the bootstrap samples with $\hat{\alpha}_{1,1}$ small. That is, 8% of the normal bootstrap samples with $\hat{\alpha}_{1,1}<1.0$,

have one of the $\hat{\alpha}_{i,1}$'s, $i \neq 1$, bigger than 3.0.

The empirical bootstrap distribution of $\hat{\alpha}_{1,1}$ ranges from -0.39 to 26.84, with Q_1 equal to 0.55 and Q_3 equal to 12.92. Besides, the median is 2.96 and 28% of the bootstrap samples present $\hat{\alpha}_{1,1} < 3.0$ and $\hat{\alpha}_{i,1} > 3.0$, for some $i \neq 1$. This suggests that as in the normal bootstrap, item 1 is strongly affecting the remaining items, producing skewed distributions and larger estimates.

8- *Conclusions*

The results from the comparison between the bootstrap, jackknife and ML parameter estimates $\hat{\alpha}_{i,0}$, $\hat{\alpha}_{i,1}$ and $\hat{\alpha}^*_{i,0}$ and the corresponding variance-covariance matrix suggest that

(1) The results from the Attitudes towards the U.S.Army and Arithmetic Reasoning Test on white women data suggest that when $\hat{\alpha}_{i,1}$'s are nearly equal, the asymptotic variance matrix probably can be trusted, since the bootstrap standard deviations are very close to the asymptotic ones. Furthermore, this similarity increases as the sample size become larger.

(2) In the Stouffer and Toby data, n=216, the biggest values for $\hat{\alpha}_{i,1}$ are 1.58 and 2.10, usually not considered large, the discrepancy between the bootstrap standard deviations and that from the asymptotic theory is bigger than in the Lombard and Doering data, where n=1729 and $\max(\hat{\alpha}_{i,1})=3.14$. These results suggest that large values for $\hat{\alpha}_{i,1}$ are associated with skewed Normal distributions or a mixture of two

distributions, one Normal and another with $\alpha_{i,1}$ equal to infinity, and probably the asymptotic standard deviations of the parameter estimates are smaller than the true ones. There is also some evidence that the size of $\hat{\alpha}_{i,1}$ to be considered large depends on the sample size.

(3) The results from the Arithmetic Reasoning Test on black women suggest that when one of the $\hat{\alpha}_{i,1}$ is very large compared with the remaining $\hat{\alpha}_{i,1}$ and the sample size is small, we should probably not trust any estimates, since this item affects all the other.

(4) The bootstrap standard deviation and the coefficient of variation of $\hat{\alpha}_{i,1}$, $\hat{\alpha}_{i,0}$ and $\hat{\alpha}^{*}_{i,0}$ are always equal to or bigger than those obtained from the asymptotic theory.

(5) The better the bootstrap distribution of the parameter estimates $\hat{\alpha}_{i,1}$, $\hat{\alpha}_{i,0}$ and $\hat{\alpha}^{*}_{i,0}$ is fitted by a Normal distribution the better is the agreement between the bootstrap standard deviation and the asymptotic standard deviation, using the observed 2^{nd} derivatives.

(6) As the bootstrap distribution of the parameter estimates $\hat{\alpha}^{*}_{i,0}$ is fitted by a Normal distribution very well, most of the bootstrap results are equal to the corresponding original ML and their asymptotic variance parameter estimates. This shows that $\hat{\alpha}^{*}_{i,0}$ is not affected by the skewness of the bootstrap $\hat{\alpha}_{i,1}$ and $\hat{\alpha}_{i,0}$, though their variability is shown through large coefficient of variation of $\hat{\alpha}^{*}_{i,1}$.

(7) Jackknife parameter estimates $\hat{\alpha}_{i,1}$ and $\hat{\alpha}_{i,0}$ and their standard deviations tend to be very similar to the original ML ones, independent of the patterns of the $\hat{\alpha}_{i,1}$ and the sample size.

(8) The jackknife pseudovalue distribution tends not to fit a Normal distribution as well as the bootstrap distribution of the parameter estimates.

(9) Jackknife is not as good as bootstrap in warning about possible failings in the asymptotic standard deviations of $\hat{\alpha}_{i,1}$ and $\hat{\alpha}_{i,0}$. Therefore the bootstrap results should be trusted more than the jackknife ones.

(10) Regarding to the comparison between the asymptotic, bootstrap and jackknife estimates of correlations this study suggests that

Except for the Arithmetic Reasoning Test (ART) on black women, there is no difference between correlations based on the observed second derivative matrix and those obtained from the information matrix.

Except for the ART on black women, jackknife estimates of the correlations are equal or closer to the asymptotic ones than the corresponding bootstrap. Actually, the jackknife estimates of the correlations are closer to the asymptotic than to the bootstrap estimates only for the Lombard and Doering data, where the bootstrap largest difference in relation to the asymptotic estimate is 0.41, while for the jackknife it is 0.17. For the first 3 sets of data, the differences in relation to the asymptotic correlations are up to 0.20, and whether they are significant or not it is difficult to say.

For the ART on black women, there are strong discrepancies among all results, whether comparing the asymptotic correlations or those with the corresponding bootstrap or jackknife estimates of the correlations. These results suggest that the asymptotic correlations probably can not be trusted.

(11)- Considering the results from the comparison between the bootstrap, the normal bootstrap and the original ML discrimination parameter estimates $\hat{\alpha}_{i,1}$, this study suggests that

In general, when there is some difference between the bootstrap, normal bootstrap and the ML results for $\hat{\alpha}_{i,1}$, bootstrap estimates are closer to each other than to the ML estimates. The strongest similarity among them is related to the interquartile difference Q_3-Q_1, which could be expected since most or all of the estimates responsible for the skweness of the distributions are not considered.

The significant agreement between most of the (empirical) bootstrap and the normal bootstrap results is probably because the ML parameter estimates $\hat{\alpha}_{i,1}$ and $\hat{\alpha}_{i,0}$ are very close to the 'true' values.

At the same time, it is supporting the evidence that in some situations ($\hat{\alpha}_{i,1}$ very large) the asymptotic theory is likely underestimating the standard deviations and most of the estimates related to the ART on black women can not be trusted.

Chapter 4

RASCH MODEL

The main purpose of this chapter is to compare the fittings of one- and two-parameter logistic models, that is to compare the Rasch model and a logit/probit model. Parameters are estimated using the marginal maximum likelihood (MML) procedure through E-M algorithms.

First of all, we shall describe the MML procedure for the Rasch model, followed by a comparison between conditional maximum likelihood (CML) and MML estimation.

Essentially, the comparison between both models, Rasch and logit/probit, will be done using data sets, which are well-fitted by a logit/probit model and represent a broad range of patterns of the discrimination parameter $\alpha_{i,1}$.

Detailed investigation will be carried out for a data set where both models give a reasonable fit, even though, in the logit/probit model, one of the estimates of $\alpha_{i,1}$ is very large compared with the remaining ones and has a large standard deviation.

1- *Marginal Maximum Likelihood Estimation*

The Rasch model for one latent variable, according to the general model(1.5), Chapter 1, may be defined as

$$\text{logit}(\pi_i(z)) = \alpha_{i,0} + \alpha_1 z, \quad i=1,\ldots,p$$

where

$\pi_i(z)$ is the response function,

z is the latent variable distributed as N(0,1),

$\alpha_{i,0}$ is the difficulty parameter of item i and

α_1 is the discrimination parameter, the same for all items.

As the Rasch model is a special case of the general model(1.5), when all the slope parameters $\alpha_{i,1}$ are equal, we can estimate α_1 and $\alpha_{i,0}$, i=1,...,p, using the E-M algorithm for the marginal maximum likelihood (MML) procedure described in Chapter 1, section 3.2.2, just altering the maximization step.

Recall that even though the latent variable Z is distributed as N(0,1), we approximate by assuming that Z takes values $z_1, z_2, \ldots, z_k$ with probabilities $h(z_1), h(z_2), \ldots, h(z_k)$ which are chosen so that the joint probability function

$$f(x_s) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} g(x_s|z)\, h(z)dz, \quad s=1,\ldots,n$$

can be approximated with high accuracy by Gauss-Hermite quadrature as

$$f(x_s) = \sum_{t=1}^{k} g(x_s|z_t)\, h(z_t), \qquad s=1,\ldots,n$$

where z_t is a tabled quadrature point and $h(z_t)$ is the corresponding weight.

The parameter estimates $\hat{\alpha}_1$ and $\hat{\alpha}_{i,0}$, i=1,...,p, are obtained from the maximization of the likelihood function

$$L = \sum_{s=1}^{n} \log f(x_s)$$

The E-M algorithm used to maximise L is described in Chapter 1, section 3.2.2. The maximization step is changed, since the Rasch model has just one slope parameter to be estimated, instead of p.

The estimation of the parameters is performed by choosing any starting values for α_1 and $\{\alpha_{i,0}\}$, followed by repeated applications of E-M steps over the set of items until convergence is obtained. In detail we proceed as follows

E-step: Calculate the values of R_{it} and N_t given by

$$R_{it} = \sum_{s=1}^{n} x_{is}\, h(z_t|x_s) \quad \text{and}$$

$$N_t = \sum_{s=1}^{n} h(z_t|x_s)$$

where $h(z_t|x_s)$ is the posterior probability of z_t given x_s.

M-step: Obtain improved estimates of α_1 and $\{\alpha_{i,0}\}$, given by equations (4.1) and (4.2) below, using the values of R_{it} and N_t from the E-step.

In the following we justify the use of equations (4.1) and (4.2) in the maximization step of the E-M algorithm for the Rasch model.

As the latent variable Z assumes values $z_1, z_2, \ldots, z_k$, R_{it} is the observed number of individuals at z_t who answer item i positively and N_t is the expected number of individuals at z_t. The response function of the Rasch model is given by

$$\text{logit}(\pi_i(z_t)) = \alpha_{i,0} + \alpha_1 z_t , \qquad i=1,\ldots,p \text{ and } t=1,\ldots,k$$

and the estimation of the parameters corresponds to a standard logistic regression problem for binomial results.

Therefore the ML estimation may be done through weighted least squares iterations and

$$\hat{\alpha}_{i,0} + \hat{\alpha}_1 z_t + [\, N_t \hat{\pi}_i(z_t)\,(1-\hat{\pi}_i(z_t))\,]^{-1}\,(\,R_{it} - N_t \hat{\pi}_i(z_t)\,)$$

$$= \hat{\alpha}_{i,0} + \hat{\alpha}_1 z_t + \epsilon_i, \qquad i=1,\ldots,p \text{ and } t=1,\ldots,k$$

where ϵ_i's are independent with variance $[\, N_t \hat{\pi}_i(z_t)\,(1-\hat{\pi}_i(z_t))\,]^{-1}$.

Hence a routine method is to obtain an estimate of α_1 and $\alpha_{i,0}$, so an estimate of $\pi_i(z_t)$ for all i,t and then use a weighted least square fit to update one iteration at a time.

We have carried out this estimation procedure using the ANCOVAR technique, which consists of the following steps (i=1,..p and t=1,..,k):

(1) Take estimates $\hat{\alpha}_1^{(0)}$ and $\hat{\alpha}_{i,0}^{(0)}$ and obtain $\hat{\pi}_i(z_t)^{(0)}$.

(2) From $\mathrm{Var}(R_{it}^{(0)}) = N_t\, \hat{\pi}_i(z_t)^{(0)}\, (1-\hat{\pi}_i(z_t)^{(0)})$ obtain

$$\xi_{it}^{(0)} = [\mathrm{Var}(R_{it}^{(0)})]^{\frac{1}{2}} \left(\hat{\alpha}_{i,0}^{(0)} + \hat{\alpha}_1^{(0)} z_t \right) + [\mathrm{Var}(R_{it}^{(0)})]^{-\frac{1}{2}} \left(R_{it}^{(0)} - N_t\, \hat{\pi}_i(z_t)^{(0)} \right).$$

(3) Take $\hat{\alpha}_{i,0}^{(0v)} = \dfrac{\sum_{t=1}^{k} [\mathrm{Var}(R_{it}^{(0)})]^{\frac{1}{2}}\, \xi_{it}^{(0)}}{\sum_{t=1}^{k} \mathrm{Var}(R_{it}^{(0)})}$ and form

$$\hat{\varepsilon}_{it}^{(0v)} = \xi_{it}^{(0)} - [\mathrm{Var}(R_{it}^{(0)})]^{\frac{1}{2}}\, \hat{\alpha}_{i,0}^{(0v)}$$

(4) Put $\hat{a}_i^{(0v)} = \dfrac{\sum_{t=1}^{k} \mathrm{Var}(R_{it}^{(0)})\, z_t}{\sum_{t=1}^{k} \mathrm{Var}(R_{it}^{(0)})}$ and form

$$\hat{\eta}_{it}^{(0v)} = [\mathrm{Var}(R_{it}^{(0)})]^{\frac{1}{2}}\, [z_t - \hat{a}_i^{(0v)}]$$

(5) Then $$\hat{\alpha}_1^{(1)} = \frac{\sum_{i=1}^{p} \sum_{t=1}^{k} \hat{\varepsilon}_{it}^{(0v)}\, \hat{\eta}_{it}^{(0v)}}{\sum_{i=1}^{p} \sum_{t=1}^{k} [\hat{\eta}_{it}^{(0v)}]^2} \qquad (4.1)$$

and $$\hat{\alpha}_{i,0}^{(1)} = \hat{\alpha}_{i,0}^{(0v)} - \hat{\alpha}_1^{(1)}\, \hat{a}_i^{(0v)}, \quad i=1,\ldots,p \qquad (4.2)$$

Replace $\hat{\alpha}_1^{(0)}$ and $\hat{\alpha}_{i,0}^{(0)}$ by $\hat{\alpha}_1^{(1)}$ and $\hat{\alpha}_{i,0}^{(1)}$ and iterate again.

When we are using an E-M algorithm, in the next iteration of the maximization step, we must use the new value of R_{it} and N_t from the expectation step.

2- *Goodness-of-fit*

As the Rasch model is a particular case of the logit/probit model, the same considerations about the goodness-of-fit given in Chapter 1, section 7, hold here.

If a statistic chi-squared based on the observed and expected frequencies can be applied then, for the Rasch model, the number of degrees of freedom for the unpooled case is 2P-p-2.

However, in practice, very often the sample size (n) is small compared with 2P (number of possible score patterns). In this case, there will be many small expected frequencies so that pooling becomes necessary. Since the number of degrees of freedom in the pooled case is equal to (number of pooled categories with expected frequencies bigger or equal to 5)-p-1, there may no be degrees of freedom to judge the goodness-of-fit.

Furthermore, when the sample size is small compared with 2P, most of the expected frequencies will be very small, while the observed frequencies will be integer and almost always 0 and 1. This implies that the fitting of the model cannot be judged comparing the expected and observed frequencies of the score patterns.

3- *Applications of the Rasch Model*

We present below some applications of the Rasch model, which are carried out using a program based on Facone (Program used to fit the models we are working with), but substituting the maximization step of the E-M algorithm as described above and doing the necessary reformulations, in order to obtain the new variance-covariance matrix.

We shall analyse how the Rasch model fits the 5 sets of data , for which we have fitted before with a logit/probit model, with response function

$$\text{logit}(\pi_i(z)) = \alpha_{i,0} + \alpha_{i,1}\, z.$$

Attitudes towards the U.S.Army

In the following, we fit the Rasch model to the 'Attitudes towards the U.S.Army', which data was displayed in Table 3.1, Chapter 3. The fitting by a logit/probit model is discussed in section 2 of the same chapter.

Table 4.1- Parameter estimates and asymptotic standard deviations from fitting the Rasch model to the Attitudes towards the U.S.Army.

Item	1	2	3	4	
$\hat{\pi}_i$	0.69	0.33	0.24	0.19	
$\hat{\alpha}_{i,0}$	0.79	-0.71	-1.16	-1.46	$\hat{\alpha}_1$=1.41
$SD(\hat{\alpha}_{i,0})$	0.09	0.09	0.09	0.10	$SD(\hat{\alpha}_1)$=0.10

$\chi^2 = 12.07$ on 10 degrees of freedom ($p \approx 0.25$).

These results show that the Rasch model also fits the data well. This might be expected, since the $\hat{\alpha}_{i,1}$'s in the logit/probit model are very similar to each other. The loglikelihood values are also very similar, -2347.65 for the logit/probit and -2349.99 for the Rasch model.

Arithmetic Reasoning Test on white women

Table 4.2 presents the parameter estimates from fitting the Rasch model to the ART on white women. These data are described in Table 2.1, Chapter 2, followed by fitting a logit/probit model.

Table 4.2- Parameter estimates and asymptotic standard deviations from fitting the Rasch model to the ART on white women.

Item	1	2	3	4	
$\hat{\pi}_i$	0.65	0.63	0.48	0.38	
$\hat{\alpha}_{i,0}$	0.62	0.55	-0.06	-0.47	$\hat{\alpha}_1$=1.16
$SD(\hat{\alpha}_{i,0})$	0.17	0.17	0.17	0.17	$SD(\hat{\alpha}_1)$=0.16

χ^2 = 0.84 on 6 degrees of freedom ($p \approx 0.99$).

As in the preceding example, both Rasch and logit/probit fit the data very well, which could be expected since $\hat{\alpha}_{i,1}$ ranges from 1.00 to 1.44 in the logit/probit model. The loglikelihood values are nearly equal, -591.97 (logit/probit model) and -592.41 (Rasch model).

Attitudes under situations of conflict

These data are given by Stouffer and Toby(1951) and they are presented in Table 3.12, Chapter 3, followed by fitting a logit/probit model.

Table 4.3- Parameter estimates and asymptotic standard deviations from fitting the Rasch model to the Stouffer and Toby data.

Item	1	2	3	4	
$\hat{\pi}_i$	0.86	0.50	0.52	0.24	
$\hat{\alpha}_{i,0}$	1.86	0.01	0.09	-1.12	$\hat{\alpha}_1$=1.51
$SD(\hat{\alpha}_{i,0})$	0.24	0.19	0.19	0.21	$SD(\hat{\alpha}_1)$=0.1

χ^2 = 10.51 on 7 degrees of freedom ($p \approx 0.15$).

Note that although Stouffer and Toby's data are fitted by a logit/probit model in which $\hat{\alpha}_{i,1}$ assumes values between 1.14 and 2.10, Rasch model also fits well as measured by χ^2, though $\hat{\alpha}_{2,0}$ has an extremely large standard deviation. This suggests that there are no significant differences among the $\alpha_{i,1}$'s, which is also shown by the similarity between the loglikelihood values, -507.42 (logit/probit) and -508.53 (Rasch).

Cancer Knowledge

The data for this study on knowledge about cancer are given by Lombard and Doering (1947). They are displayed in Table 2.5, Chapter 2, followed by fitting a logit/probit model.

Table 4.4- Parameter estimates and asymptotic standard deviations from fitting the Rasch model to the Lombard and Doering data.

Item	1	2	3	4	
$\hat{\pi}_i$	0.18	0.58	0.46	0.04	
$\hat{\alpha}_{i,0}$	-1.52	0.33	-0.14	-3.09	$\hat{\alpha}_1$= 1.30
$SD(\hat{\alpha}_{i,0})$	0.08	0.06	0.06	0.11	$SD(\hat{\alpha}_1)$= 0.06

χ^2 = 67.13 on 10 degrees of freedom ($p \approx 0.001$).

As it would reasonable to expect, the Rasch model does not fit Lombard and Doering's data, since they are fitted well by a logit/probit model in which $\hat{\alpha}_{2,1}$ is clearly larger than the others (0.72; 3.40; 1.34; 0.77) for approximately the same coefficient of variation. The loglikelihood of the logit/probit model is -3622.68 compared to a smaller value -3651.75 for the Rasch model.

Arithmetic Reasoning Test on black women

Table 4.5 shows the fitting by the Rasch model to the ART on black women, displayed in Table 2.3, Chapter 2, followed by fitting a logit/probit model.

Table 4.5- Parameter estimates and asymptotic standard deviations from fitting the Rasch model to the ART on black women.

Item	1	2	3	4	
$\hat{\pi}_i$	0.50	0.42	0.27	0.25	
$\hat{\alpha}_{i,0}$	0.01	-0.34	-0.98	-1.13	$\hat{\alpha}_1$=0.50
$SD(\hat{\alpha}_{i,0})$	0.18	0.18	0.20	0.21	$SD(\hat{\alpha}_1)$=0.19

χ^2 = 11.81 on 6 degrees of freedom ($0.05<p<0.10$).

As we can see from Table 4.5 the ART on black women's data are reasonably fitted by the Rasch model, though the value of the loglikelihood statistic (11.81) judged as a chi-squared random variable is close to the 5% significance level (12.59). Even so, this is an unexpected result, since the same data are fitted by a logit/probit model with very large $\hat{\alpha}_{1,1}$ (14.39), while the remaining $\hat{\alpha}_{i,1}$'s are very similar. Moreover the loglikelihood values are nearly equal on the log-scale, -366.84 (Rasch) compared to -364.69 (logit/probit).

This result seems to indicate that item 1, in the logit/probit model, does not contain any information about the data, also showed through its large standard deviation. The untrustworthy result for item 1 can also be seen in the Rasch model through the large standard deviation of $\hat{\alpha}_{1,0}$, what provides a coefficient of variation equal to 18.

4- *Looking at the Behaviour of the Likelihood Function*

In order to obtain more information about how the ART on black women can also be fitted by a Rasch model, we shall look at the behaviour of its likelihood function. This will be done through the profile method proceeding, in the same way as described in Chapter 2 for a logit/probit model.

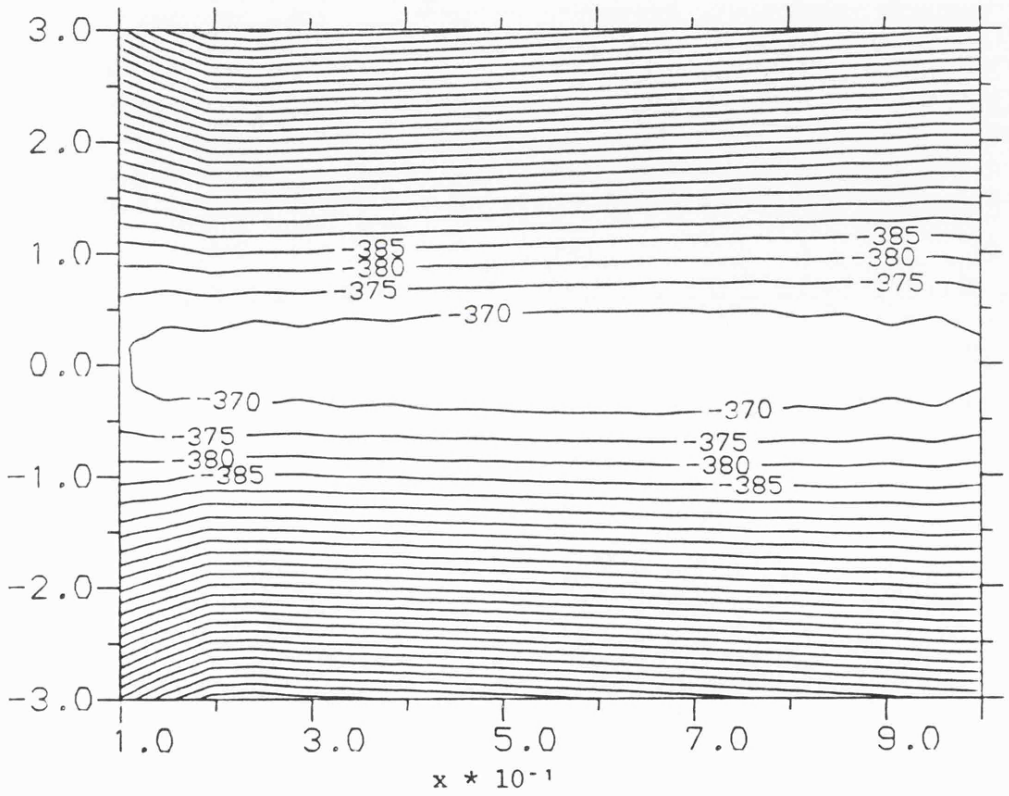


Figure 4.1- Loglikelihood values as a function of the parameter estimates $\hat{\alpha}_1$ and $\hat{\alpha}_{1,0}$ by fitting a Rasch model to the ART on black women.

The parallel lines in Figure 4.1 indicate that the values of the likelihood function are almost unchanging over all the range of $\hat{\alpha}_1$, for a fixed $\hat{\alpha}_{1,0}$. However, although the contouring shows a broad ridge going from West to East suggesting that $\hat{\alpha}_1$ is not a meaningful

parameter, there is a peak inside it, that can be seen from Figure 4.2. Therefore the likelihood function from fitting the Rasch model to the ART on black women behaves well, i.e, is unimodal though the increase is small, from -367.79 to -366.84.

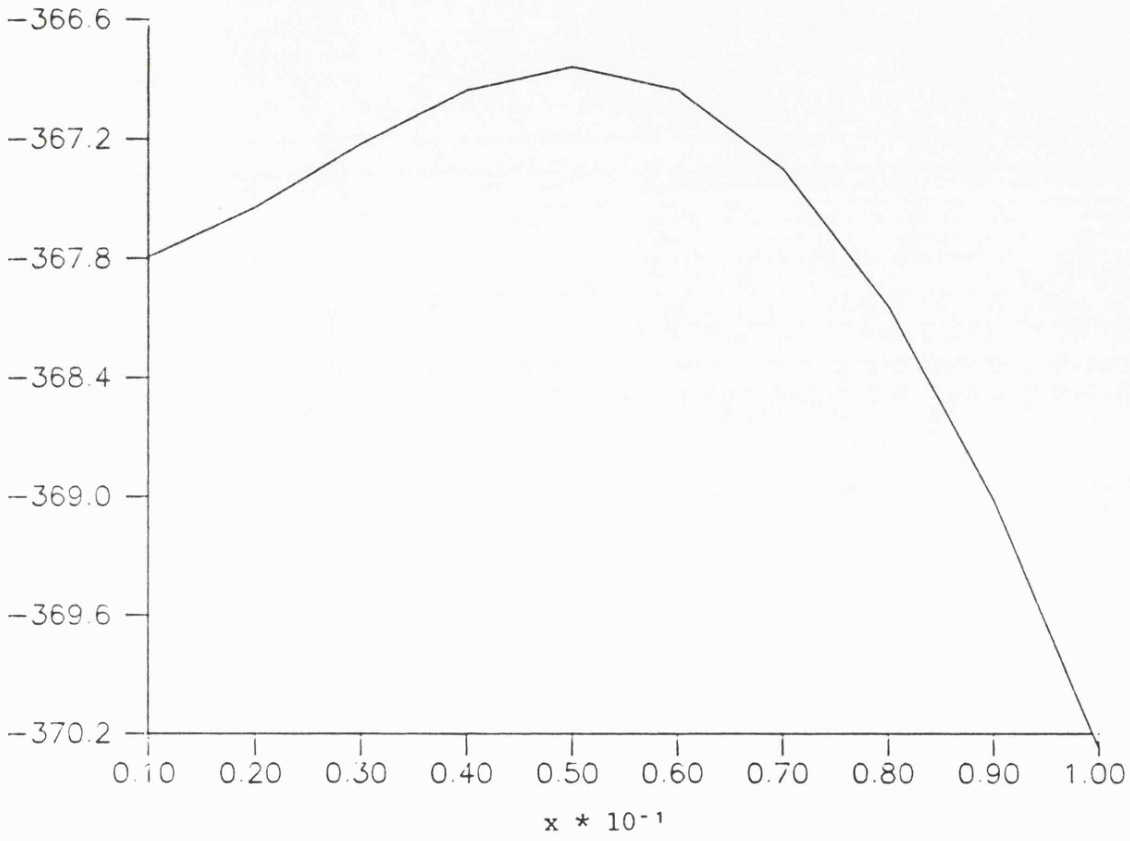


Figure 4.2- Maximum likelihood values over $\hat{\alpha}_{1,0}$ for each $\hat{\alpha}_1$ fixed by fitting a Rasch model to the ART on black women.

On the other hand, the likelihood function from fitting this data by a logit/probit model is not unimodal, since it continues increasing indefinitely as we found out in Chapter 2, Figures 2.3 and 2.10.

From these results we can conclude that the likelihood behaves better when ART on black women are fitted by the Rasch than by a logit/probit model. Therefore the Rasch model fits better this data than the logit/probit model.

In order to investigate whether one is likely to have a set of data with the same pattern of the $\hat{\alpha}_{i,1}$'s as in this example, fitted well by both the Rasch and a logit/probit model, we have done a study based on the normal bootstrap method for the ART on black women.

5- *Normal Bootstrapping*

We have generated 30 normal bootstrap samples from a multinomial distribution with parameters $\pi_i(z)$, i=1,..,4, the response function of a logit/probit model, assuming that the latent variable Z is distributed as N(0,1), and $\alpha_{i,1}$ and $\alpha_{i,0}$ are equal to the ML parameter estimates of the ART on black women's data (Tables 2.2, Chapter 2, where $\hat{\alpha}_{i,1}$, i=1,...,4, are equal to 14.39, 0.38, 0.39 and 0.19, respectively).

In Table 4.6 we give the parameter estimates from fitting a logit/probit model to each of the 30 normal bootstrap samples. We shall consider that the model fits the data well if the null hypothesis cannot be rejected on a significance level of 5%.

Table 4.6- Parameter estimates and asymptotic standard deviations(in brackets) from fitting two and one-parameter logistic (logit/probit and Rasch) models to each of the 30 normal bootstrap samples.

	item	logit/probit			Rasch	
S	i	$\hat{\alpha}_{i,1}$	$\hat{\alpha}_{i,0}$	$\hat{\pi}_i$	$\hat{\alpha}_{i,0}$	$\hat{\pi}_i$
1	1	0.50 (0.43)	0.13 (0.18)	0.53	0.13 (0.18)	0.53
	2	1.89 (3.30)	-0.90 (0.92)	0.29	-0.59 (0.19)	0.36
	3	-0.12 (0.30)	-0.80 (0.18)	0.31	-0.85 (0.19)	0.30
	4	0.65 (0.63)	-1.09 (0.26)	0.25	-1.07 (0.20)	0.26
2	1	10.95(51.31)	-0.80 (8.83)	0.31	-0.07 (0.18)	0.48
	2	0.37 (0.23)	-0.60 (0.18)	0.35	-0.62 (0.19)	0.35
	3	0.32 (0.24)	-0.99 (0.19)	0.27	-1.03 (0.20)	0.26
	4	0.27 (0.24)	-0.87 (0.18)	0.29	-0.92 (0.20)	0.28
3	1	0.65 (0.36)	0.08 (0.18)	0.52	0.08 (0.18)	0.52
	2	0.10 (0.23)	-0.18 (0.16)	0.46	-0.21 (0.19)	0.45
	3	0.59 (0.31)	-1.00 (0.21)	0.27	-1.03 (0.21)	0.26
	4	8.76(49.12)	-4.96(18.18)	0.01	-1.00 (0.21)	0.27
4	1	0.35 (0.23)	0.18 (0.17)	0.55	0.18 (0.17)	0.54
	2	-0.22 (0.23)	-0.32 (0.17)	0.42	-0.31 (0.17)	0.42
	3	12.10(45.45)	-5.44 (6.26)	0.01	-0.80 (0.18)	0.31
	4	-0.18 (0.26)	-1.12 (0.19)	0.25	-1.11 (0.19)	0.25
5	1	0.11 (0.22)	0.12 (0.17)	0.53	0.13 (0.18)	0.53
	2	0.52 (0.25)	-0.52 (0.18)	0.37	-0.52 (0.18)	0.37
	3	13.06(59.99)	-7.26(22.02)	0.00	-0.96 (0.20)	0.28
	4	0.39 (0.25)	-0.96 (0.19)	0.28	-0.99 (0.20)	0.27
6	1	1.15 (1.08)	-0.02 (0.21)	0.50	-0.01 (0.19)	0.50
	2	1.66 (2.04)	-0.60 (0.46)	0.35	-0.46 (0.19)	0.39
	3	-0.27 (0.43)	-1.24 (0.21)	0.22	-1.23 (0.20)	0.22
	4	-0.49 (0.34)	-1.09 (0.21)	0.25	-1.04 (0.19)	0.26
7	1	5.82(42.29)	-1.14 (8.60)	0.24	-0.31 (0.18)	0.42
	2	0.32 (0.30)	-0.42 (0.18)	0.40	-0.42 (0.18)	0.39
	3	0.14 (0.27)	-1.19 (0.20)	0.23	-1.24 (0.21)	0.22
	4	0.02 (0.25)	-1.07 (0.19)	0.26	-1.12 (0.20)	0.24
8	1	9.96(53.38)	-0.18 (2.90)	0.46	-0.01 (0.18)	0.50
	2	0.62 (0.30)	-0.32 (0.18)	0.42	-0.32 (0.18)	0.42
	3	0.52 (0.28)	-0.59 (0.19)	0.36	-0.62 (0.19)	0.35
	4	-0.26 (0.24)	-1.02 (0.14)	0.26	-1.09 (0.21)	0.25
9	1	1.03 (0.69)	-0.02 (0.20)	0.49	-0.02 (0.19)	0.49
	2	0.78 (0.51)	-0.56 (0.21)	0.36	-0.57 (0.19)	0.36
	3	0.83 (0.57)	-1.10 (0.27)	0.25	-1.11 (0.21)	0.25
	4	0.75 (0.46)	-0.68 (0.21)	0.33	-0.70 (0.20)	0.33

continue...

	item	logit/probit			Rasch	
S	i	$\hat{\alpha}_{i,1}$	$\hat{\alpha}_{i,0}$	$\hat{\pi}_i$	$\hat{\alpha}_{i,0}$	$\hat{\pi}_i$
10	1	0.60 (0.25)	0.10 (0.18)	0.53	0.10 (0.18)	0.52
	2	10.31(42.92)	-2.33(11.29)	0.09	-0.34 (0.18)	0.42
	3	-0.40 (0.25)	-0.83 (0.19)	0.70	-0.84 (0.19)	0.30
	4	0.16 (0.24)	-0.97 (0.19)	0.27	-1.01 (0.20)	0.27
11	1	11.37(52.26)	0.56 (7.12)	0.64	0.05 (0.19)	0.51
	2	0.50 (0.23)	-0.28 (0.18)	0.43	-0.31 (0.19)	0.42
	3	0.31 (0.25)	-1.25 (0.21)	0.22	-1.40 (0.23)	0.20
	4	0.89 (0.33)	-0.86 (0.21)	0.30	-0.85 (0.21)	0.30
12	1	11.43(45.24)	-1.64(11.31)	0.16	-0.16 (0.17)	0.46
	2	0.29 (0.23)	-0.65 (0.18)	0.34	-0.66 (0.18)	0.34
	3	0.42 (0.25)	-0.97 (0.19)	0.28	-0.96 (0.19)	0.28
	4	-0.31 (0.26)	-1.25 (0.21)	0.22	-1.26 (0.21)	0.22
13	1	0.30 (0.32)	-0.41 (0.18)	0.40	-0.44 (0.18)	0.39
	2	0.82 (0.88)	-0.34 (0.21)	0.42	-0.32 (0.18)	0.42
	3	-0.38 (0.31)	-0.96 (0.19)	0.28	-1.01 (0.20)	0.27
	4	3.40(11.99)	-2.53 (7.46)	0.07	-1.16 (0.21)	0.24
14	1	16.01(46.89)	-0.77 (9.24)	0.32	-0.02 (0.19)	0.49
	2	0.93 (0.28)	-0.21 (0.20)	0.45	-0.21 (0.19)	0.45
	3	0.45 (0.24)	-0.97 (0.20)	0.27	-1.06 (0.21)	0.26
	4	0.14 (0.23)	-1.00 (0.20)	0.27	-1.14 (0.22)	0.24
15	1	0.36 (0.72)	-0.04 (0.17)	0.49	-0.05 (0.18)	0.49
	2	0.60 (1.28)	-0.16 (0.17)	0.46	-0.16 (0.18)	0.46
	3	-0.06 (0.50)	-0.61 (0.18)	0.35	-0.66 (0.19)	0.34
	4	2.57(15.26)	-1.98 (8.28)	0.12	-1.11 (0.21)	0.25
16	1	11.13(43.92)	2.27(11.96)	0.91	0.27 (0.17)	0.57
	2	0.23 (0.22)	-0.47 (0.17)	0.38	-0.48 (0.18)	0.38
	3	0.46 (0.46)	-0.84 (0.19)	0.30	-0.82 (0.19)	0.30
	4	-0.20 (0.25)	-1.19 (0.20)	0.33	-1.22 (0.21)	0.33
17	1	0.54 (0.42)	-0.19 (0.18)	0.45	-0.20 (0.18)	0.45
	2	3.87(15.82)	-0.48 (1.74)	0.38	-0.20 (0.18)	0.45
	3	0.35 (0.48)	-0.66 (0.18)	0.34	-0.72 (0.19)	0.33
	4	0.49 (0.36)	-1.13 (0.21)	0.24	-1.20 (0.21)	0.23
18	1	4.44(25.81)	-0.46 (2.53)	0.39	-0.15 (0.17)	0.46
	2	-0.36 (0.37)	-0.30 (0.17)	0.42	-0.30 (0.17)	0.43
	3	0.18 (0.25)	-0.74 (0.18)	0.32	-0.75 (0.18)	0.32
	4	0.48 (0.51)	-1.32 (0.24)	0.21	-1.28 (0.21)	0.22
19	1	0.60 (0.40)	0.01 (0.18)	0.50	0.01 (0.18)	0.50
	2	0.87 (0.55)	-0.40 (0.21)	0.40	-0.39 (0.19)	0.40
	3	0.25 (0.36)	-0.98 (0.19)	0.27	-1.07 (0.21)	0.25
	4	1.35 (1.10)	-1.15 (0.44)	0.24	-0.96 (0.20)	0.28

continue...

	item	logit/probit			Rasch	
S	i	$\hat{\alpha}_{i,1}$	$\hat{\alpha}_{i,0}$	$\hat{\pi}_i$	$\hat{\alpha}_{i,0}$	$\hat{\pi}_i$
20	1	0.95 (1.04)	0.38 (0.23)	0.59	0.36 (0.19)	0.59
	2	0.53 (0.46)	-0.31 (0.18)	0.42	-0.32 (0.19)	0.42
	3	1.06 (1.17)	-1.06 (0.42)	0.26	-0.96 (0.20)	0.28
	4	0.39 (0.55)	-1.26 (0.23)	0.22	-1.34 (0.22)	0.21
21	1	5.92(36.26)	0.94 (6.35)	0.72	0.25 (0.18)	0.56
	2	0.32 (0.28)	-0.50 (0.17)	0.38	-0.53 (0.18)	0.37
	3	0.10 (0.27)	-1.04 (0.19)	0.26	-1.11 (0.21)	0.37
	4	0.38 (0.31)	-0.69 (0.18)	0.33	-0.72 (0.19)	0.33
22	1	7.13(39.14)	-0.24 (2.23)	0.44	-0.05 (0.19)	0.49
	2	0.68 (0.40)	-0.51 (0.19)	0.37	-0.53 (0.20)	0.37
	3	0.51 (0.33)	-0.85 (0.20)	0.30	-0.91 (0.21)	0.29
	4	0.26 (0.29)	-1.64 (0.23)	0.16	-1.82 (0.26)	0.14
23	1	0.60 (0.59)	-0.16 (0.18)	0.46	-0.16 (0.18)	0.46
	2	2.81 (8.63)	-1.14 (2.60)	0.24	-0.59 (0.19)	0.36
	3	-0.16 (0.27)	-0.87 (0.18)	0.31	-0.92 (0.20)	0.28
	4	0.53 (0.56)	-1.61 (0.28)	0.17	-1.62 (0.24)	0.16
24	1	0.96 (0.44)	-0.12 (0.20)	0.47	-0.12 (0.17)	0.47
	2	0.90 (0.48)	-0.82 (0.24)	0.30	-0.84 (0.18)	0.30
	3	0.60 (0.41)	-1.54 (0.26)	0.18	-1.68 (0.24)	0.16
	4	1.50 (0.89)	-1.31 (0.45)	0.21	-1.11 (0.20)	0.25
25	1	1.61 (1.35)	-0.14 (0.24)	0.46	-0.10 (0.18)	0.47
	2	1.40 (1.01)	-0.55 (0.29)	0.36	-0.43 (0.18)	0.39
	3	0.45 (0.31)	-0.83 (0.20)	0.30	-0.85 (0.19)	0.30
	4	-0.18 (0.32)	-1.27 (0.20)	0.22	-1.34 (0.22)	0.21
26	1	12.67(45.56)	1.73(11.61)	0.85	0.13 (0.17)	0.53
	2	0.12 (0.21)	-0.44 (0.17)	0.39	-0.44 (0.17)	0.39
	3	0.26 (0.24)	-0.84 (0.18)	0.30	-0.85 (0.19)	0.30
	4	0.34 (0.24)	-0.96 (0.19)	0.28	-0.95 (0.19)	0.28
27	1	0.63 (0.39)	-0.20 (0.18)	0.45	-0.20 (0.19)	0.45
	2	0.92 (0.54)	-0.34 (0.21)	0.41	-0.33 (0.19)	0.42
	3	1.49 (1.19)	-1.30 (0.55)	0.21	-1.04 (0.21)	0.27
	4	-0.22 (0.33)	-1.05 (0.19)	0.26	-1.10 (0.21)	0.25
28	1	8.64(45.73)	1.20 (9.65)	0.77	0.19 (0.18)	0.55
	2	0.08 (0.21)	0.01 (0.17)	0.50	0.01 (0.18)	0.50
	3	0.38 (0.26)	-0.99 (0.19)	0.27	-1.02 (0.20)	0.26
	4	0.37 (0.26)	-0.99 (0.19)	0.27	-1.02 (0.20)	0.26
29	1	14.56(47.13)	-1.91(12.26)	0.13	-0.10 (0.18)	0.47
	2	0.34 (0.22)	-0.41 (0.18)	0.40	-0.44 (0.18)	0.39
	3	0.70 (0.27)	-0.99 (0.21)	0.27	-0.97 (0.20)	0.27
	4	0.30 (0.23)	-0.88 (0.19)	0.29	-0.93 (0.20)	0.28

continue...

	item	logit/probit			Rasch	
S	i	$\hat{\alpha}_{i,1}$	$\hat{\alpha}_{i,0}$	$\hat{\pi}_i$	$\hat{\alpha}_{i,0}$	$\hat{\pi}_i$
30	1	11.96(45.80)	-1.56(11.40)	0.17	-0.13 (0.18)	0.47
	2	0.37 (0.22)	-0.39 (0.18)	0.40	-0.40 (0.18)	0.40
	3	0.37 (0.25)	-0.99 (0.19)	0.27	-1.02 (0.20)	0.26
	4	0.22 (0.25)	-1.16 (0.20)	0.24	-1.21 (0.21)	0.23

In order to complement the information from fitting a logit/probit and the Rasch model to each one of the 30 normal bootstrap samples obtained from the ART on black women data, we shall present Table 4.7, which displays for each model and sample the corresponding loglikelihood values and the observed chi-squared obtained when using the loglikelihood ratio statistics

$$\chi^2 = -2 \sum_i O_i \log \left[\frac{O_i}{E_i} \right]$$

where $\{O_i\}$ and $\{E_i\}$ are the observed and expected frequencies.

Table 4.7- Goodness-of-fit results from fitting a logit/probit and the Rasch model to each one of the 30 normal bootstrap samples obtained from ART on black women data.

Sample	Model	$\hat{\alpha}_1$	SD($\hat{\alpha}_1$)	loglik.	χ^2	d.f
1	LP	1.89	3.30	-366.28	4.36	3
	R	0.55	0.22	-368.24	6.60	6
2	LP	10.92	51.31	-365.57	8.12	4
	R	0.56	0.22	-367.02	12.77	8
3	LP	8.77	49.12	-367.24	7.04	5
	R	0.73	0.20	-370.25	12.11	9
4	LP	12.10	45.45	-367.54	4.77	4
	R	0.00	0.62	-369.67	8.53	7

continue...

Sample	Model	$\hat{\alpha}_1$	SD($\hat{\alpha}_1$)	loglik.	χ^2	d.f
5	LP	13.06	59.99	-366.01	7.25	5
	R	0.54	0.22	-368.71	8.92	6
6	LP	1.66	2.04	-352.75	6.66	3
	R	0.15	0.62	-359.25	18.46 *	6
7	LP	5.82	42.29	-356.70	1.86	4
	R	0.46	0.25	-357.17	3.48	7
8	LP	9.96	53.38	-371.82	3.56	3
	R	0.64	0.20	-376.59	11.14	7
9	LP	1.03	0.69	-370.14	3.96	6
	R	0.84	0.00	-370.24	5.73	9
10	LP	10.31	42.92	-368.79	7.34	3
	R	0.48	0.23	-373.31	15.48 *	7
11	LP	11.37	52.26	-358.88	4.88	4
	R	0.85	0.20	-363.16	14.16 *	6
12	LP	11.44	45.24	-354.28	3.17	1
	R	0.36	0.29	-357.33	11.15	6
13	LP	3.40	11.99	-359.39	12.26 *	3
	R	0.63	0.21	-363.06	19.77 *	8
14	LP	16.01	46.89	-360.62	5.25	3
	R	0.81	0.20	-366.25	19.44	9
15	LP	2.57	15.26	-374.92	3.11	4
	R	0.60	0.21	-376.01	4.46	8
16	LP	11.13	43.92	-361.85	1.42	1
	R	0.38	0.28	-364.40	10.23	6
17	LP	3.87	15.82	-370.28	7.34	5
	R	0.76	0.00	-371.30	11.02	8
18	LP	4.44	25.81	-363.93	1.83	2
	R	0.26	0.38	-366.86	5.09	6
19	LP	1.35	1.10	-368.07	7.71	4
	R	0.72	0.20	-368.82	9.61	8
20	LP	1.06	1.17	-359.92	4.57	2
	R	0.70	0.20	-360.49	6.02	5
21	LP	5.92	36.26	-369.46	3.32	4
	R	0.58	0.21	-370.16	5.08	7
22	LP	7.13	39.14	-344.92	1.71	2
	R	0.82	0.20	-347.48	5.80	6

continue...

Sample	Model	$\hat{\alpha}_1$	SD($\hat{\alpha}_1$)	loglik.	χ^2	d.f
23	LP	2.81	8.63	-347.58	3.31	2
	R	0.56	0.23	-350.24	9.94	6
24	LP	1.50	0.89	-341.90	1.76	3
	R	0.96	0.00	-342.41	2.24	6
25	LP	1.61	1.35	-356.68	3.33	1
	R	0.55	0.22	-362.89	13.97 *	6
26	LP	12.67	45.56	-370.71	12.28 *	4
	R	0.30	0.32	-372.59	16.97 *	8
27	LP	1.50	1.19	-363.72	8.00	4
	R	0.52	0.23	-367.50	15.01	8
28	LP	8.64	45.73	-368.72	2.60	5
	R	0.49	0.23	-370.29	5.37	8
29	LP	14.56	47.13	-366.81	5.23	5
	R	0.62	0.21	-371.18	12.14	7
30	LP	11.96	45.80	-360.50	4.81	4
	R	0.52	0.23	-362.62	9.35	7

$\hat{\alpha}_1 = \max(\hat{\alpha}_{i,1})$ for LP model *:the model does not fit(α=5%)

Both, the Rasch and the logit/probit models fit well twenty four (80%) of the bootstrap samples, from which sixteen have the $\max(\hat{\alpha}_{i,1})>3.0$ with coefficient of variation larger than 2.93.

Four bootstrap samples are fitted well by a logit/probit model, but not by the Rasch model. Although in 3 of them the observed chi-squares are very close to the tabulated value. The maximum values of $\hat{\alpha}_{i,1}$ in these 4 samples are 1.61, 1.65, 10.31 and 11.37.

Thus only 2 bootstrap samples are not fitted by the Rasch or a logit/probit model.

Therefore this study suggests that it is likely that the Rasch model fits well a set of data fitted by a logit/probit model when one of the $\hat{\alpha}_{i,1}$'s is very large compared with the remaining and it has large standard deviation.

On the other hand, if a logit/probit model fits the data well and none of the discrimination parameter estimates $\hat{\alpha}_{i,1}$ is different from the others then it is likely that the Rasch model also fits the data well.

6- *The Distribution of* $\hat{\pi}_i$

In searching for reasons that might explain why both models, Rasch and logit/probit, can fit the same data set when one of the discrimination parameter estimates $\hat{\alpha}_{i,1}$ is very large (bigger than 4.0), we decided to investigate the relation between difficulty parameter estimates $\hat{\alpha}_{i,0}$, i=1,...,4, from fitting both models to each one of the bootstrap samples.

As $\alpha_{i,0} = \text{logit}(\pi_i(z))$ at z=0, where π_i is then the probability of a positive response from the median individual, and since π_i has a rather more useful interpretation, making easier the comparison, we shall look at the frequency distribution of $\hat{\pi}_i$ for each item i, instead of $\hat{\alpha}_{i,0}$, for both models.

Table 4.8 displays the median, mean, standard deviation of the 30 estimates $\hat{\pi}_i$, i=1,...,4, obtained from fitting a logit/probit and the Rasch model to the 30 normal bootstrap samples refered in the preceding section.

Let min and max be the smallest and the largest of the 30 bootstrap estimates $\hat{\pi}_i$, for each item i.

Table 4.8- Frequency distribution of the parameter estimates $\hat{\pi}_i$, i=1,...,4, from fitting a logit/probit model and the Rasch model to each one of the 30 normal bootstrap samples (ART on black women).

Item	Model	Median	Mean	SD($\hat{\pi}_i$)	CV($\hat{\pi}_i$)	Min	Max
1	log/prob	0.48	0.48	0.18	0.38	0.13	0.91
	Rasch	0.50	0.50	0.04	0.08	0.39	0.59
2	log/prob	0.39	0.38	0.08	0.20	0.09	0.50
	Rasch	0.40	0.40	0.04	0.10	0.30	0.50
3	log/prob	0.27	0.26	0.08	0.30	0.00	0.36
	Rasch	0.28	0.27	0.04	0.15	0.16	0.35
4	log/prob	0.25	0.23	0.07	0.28	0.01	0.34
	Rasch	0.25	0.25	0.04	0.16	0.14	0.33

As we can see from Table 4.8, both models yield distributions that have approximately the same mean and median for all items. However the standard deviation of $\hat{\pi}_1$ is significantly higher from fitting a logit/probit model than the Rasch model to the data. This difference is because logit/probit yielded a more spread distribution in both directions, as $\hat{\pi}_1$ is assuming values from 0.13 to 0.91 compared with 0.39 to 0.59 from the Rasch model.

If we look at the spread of $\hat{\pi}_i$ for the remaining items, we can see that logit/probit model produced $\hat{\pi}_i$ values that are more spread to the left than those from the Rasch model. That is the minimum of $\hat{\pi}_i$ distribution, for each item i, from logit/probit model is always smaller than the corresponding one from Rasch, for the same maximum.

In the following we shall display in histograms for each item i the frequency distribution of $\hat{\pi}_i$, i=1,...,4, from fitting both models and we look closely at the relation between them through a bi-dimensional graph.

Item 1

Histogram of $\hat{\pi}_1$ from fitting a logit/probit model.

midpoint	count	
0.15	3	***
0.20	0	
0.25	1	*
0.30	2	**
0.35	0	
0.40	2	**
0.45	7	*******
0.50	6	******
0.55	3	***
0.60	1	*
0.65	1	*
0.70	1	*
0.75	1	*
0.80	0	
0.85	1	*
0.90	1	*

Histogram of $\hat{\pi}_1$ from fitting the Rasch model.

midpoint	count	
0.15	0	
0.20	0	
0.25	0	
0.30	0	
0.35	0	
0.40	2	**
0.45	7	*******
0.50	13	*************
0.55	7	*******
0.60	1	*
0.65	0	
0.70	0	
0.75	0	
0.80	0	
0.85	0	
0.90	0	

Figure 4.3 below shows two distinct groups, the first where there is a linear relation with slope 1.0 between $\hat{\pi}_i$'s obtained from both models. In this case $\hat{\pi}_i$ ranges from 0.40 to 0.60, which represents 60% of the sample distribution. The second group, where $\hat{\pi}_i$ from Rasch

model is approximately constant and equal to 0.50, while from logit/probit model $\hat{\pi}_i$ assumes values between 0.13 and 0.41 and from 0.61 and 0.91.

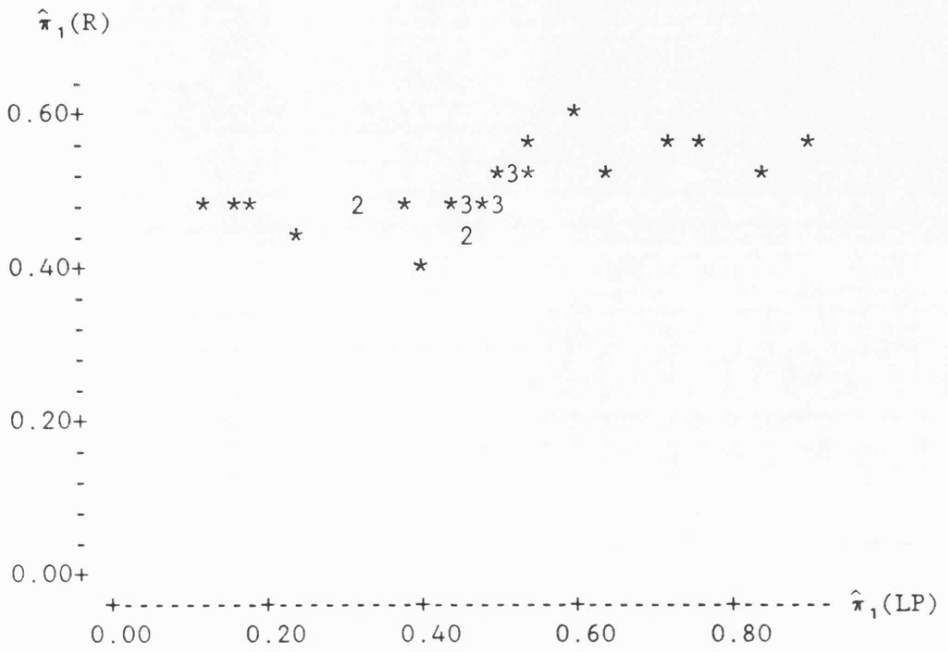


Figure 4.3- Comparison between $\hat{\pi}_1$'s from fitting a logit/probit and the Rasch model to 30 normal bootstrap samples from the ART on black women's data.

We can see from the histogram of $\hat{\pi}_1$ from fitting a logit/probit model that the second group is situated in the tails of the distribution of $\hat{\pi}_1$.

Fourteen out of these 30 bootstrap samples have $\hat{\alpha}_{1,1}$ bigger than 3.40 with very large standard deviation, from which only 3 samples belong to the first group. This means that the second group is formed by 11 samples with $\hat{\alpha}_{1,1}$ bigger than 3.40 and one sample with $\hat{\alpha}_{1,1}$ smaller than 2.0, which is located on the border of these two groups.

These results seem to indicate that the samples for which $\hat{\pi}_1$ assumes different values according to which model is fitted to the data, are those samples for which fitting a logit/probit model gives large $\hat{\alpha}_{1,1}$.

Item 2

Histogram of $\hat{\pi}_2$ from fitting a logit/probit model.

midpoint	count	
0.10	1	*
0.15	0	
0.20	0	
0.25	1	*
0.30	2	**
0.35	6	******
0.40	14	**************
0.45	5	*****
0.50	1	*

Histogram of $\hat{\pi}_2$ from fitting the Rasch model.

midpoint	count	
0.10	0	
0.15	0	
0.20	0	
0.25	0	
0.30	1	*
0.35	8	********
0.40	15	***************
0.45	5	*****
0.50	1	*

Figure 4.4 shows that if we take out the 3 points on the left, then we can fit a straight line passing through the origin. This means that, except for 3 samples, the Rasch and a logit/probit model give approximately equal $\hat{\pi}_2$ (the probability of positive response for a median individual).

The largest difference between $\hat{\pi}_2$'s from fitting both models comes from sample 10, where $\hat{\alpha}_{2,1}$ is equal to 10.31, $\hat{\pi}_2$ is equal to 0.42 from the Rasch(R) and 0.09 from the logit/probit (LP). This smaller $\hat{\pi}_2$ corresponds to the isolated point on the left of the distribution of $\hat{\pi}_2$ from fitting a logit/probit model displayed in the histogram above.

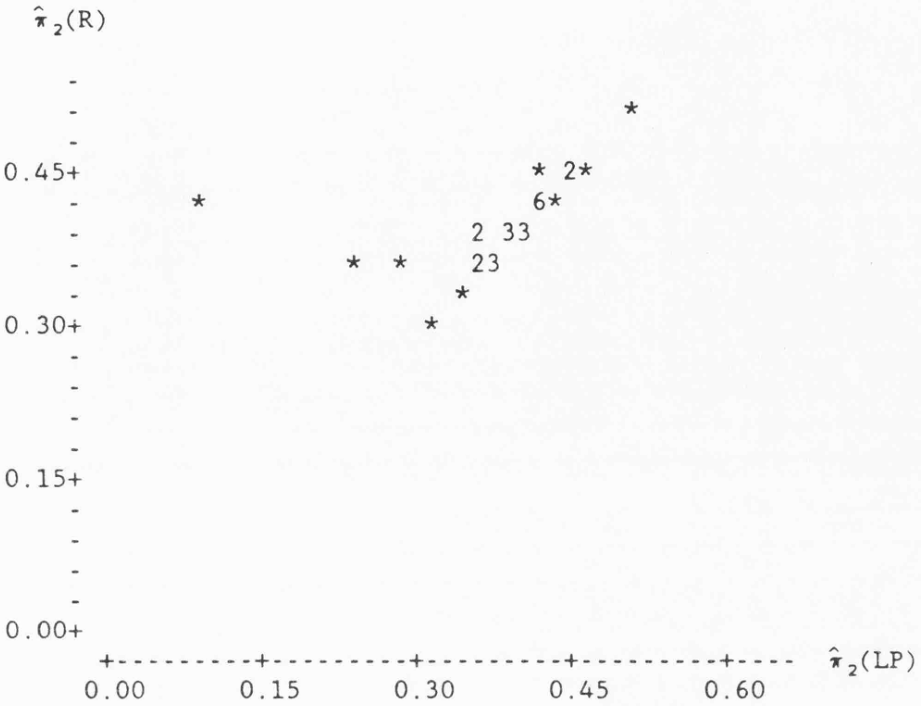


Figure 4.4- Comparison between $\hat{\pi}_2$'s from fitting a logit/probit and the Rasch model to 30 normal bootstrap samples from the ART on black women's data.

For the other two samples the difference between $\hat{\pi}_2$'s is smaller. For sample 23, $\hat{\pi}_2$(R)=0.24 and $\hat{\pi}_2$(LP)=0.36, where $\hat{\alpha}_{2,1}$is equal to 2.81. The next largest difference is from sample 1 where $\hat{\pi}_2$(R)=0.29 and $\hat{\pi}_2$(LP)=0.36 and $\hat{\alpha}_{2,1}$is equal to 1.89. These 3 samples are among the 4 that have the highest values of $\hat{\alpha}_{2,1}$. The fourth highest $\hat{\alpha}_{2,1}$ is from sample 17 with $\hat{\alpha}_{2,1}$ equal to 3.87 and $\hat{\pi}_2$(R)=0.45, $\hat{\pi}_2$(LP)=0.36.

Therefore for only sample 10 the estimates $\hat{\pi}_2$'s obtained from Rasch and logit/probit models are significantly different.

Item 3

Histogram of $\hat{\pi}_3$ from fitting a logit/probit model.

midpoint	count	
0.00	2	**
0.05	0	
0.10	0	
0.15	0	
0.20	4	****
0.25	12	************
0.30	9	*********
0.35	3	***

Histogram of $\hat{\pi}_3$ from fitting the Rasch model.

midpoint	count	
0.00	0	
0.05	0	
0.10	0	
0.15	1	*
0.20	2	**
0.25	12	************
0.30	12	************
0.35	3	***

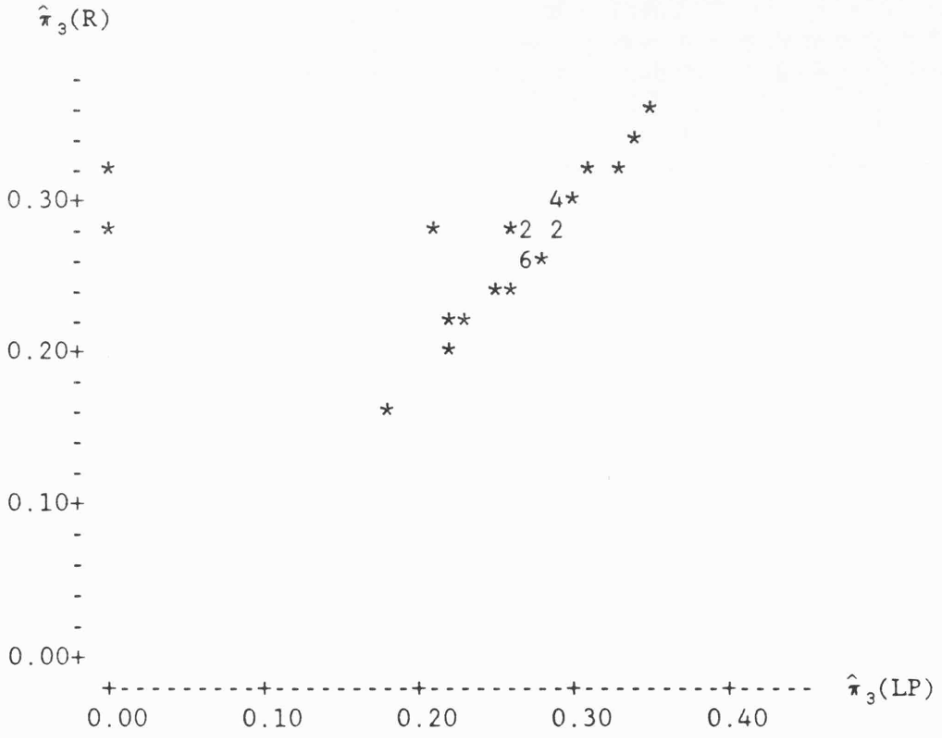


Figure 4.5- Comparison between $\hat{\pi}_3$'s from fitting a logit/probit and the Rasch model to 30 normal bootstrap samples from the ART on black women's data.

Looking at Figure 4.5 we can see that the relation between the parameter estimates $\hat{\pi}_3$ from fitting the Rasch and a logit/probit model can be expressed by a straight line passing through the origin; except

for samples 4 and 5. These samples are the only ones which present large values for $\hat{\alpha}_{3,1}$ (12.10 and 13.06) and $\hat{\pi}_3$'s equal to zero (shown in the histogram above).

Item 4

Histogram of $\hat{\pi}_4$ from fitting a logit/probit model.

midpoint	count	
0.00	1	*
0.05	1	*
0.10	1	*
0.15	2	**
0.20	5	*****
0.25	13	*************
0.30	5	*****
0.35	2	**

Histogram of $\hat{\pi}_4$ from fitting the Rasch model.

midpoint	count	
0.00	0	
0.05	0	
0.10	0	
0.15	2	**
0.20	4	****
0.25	17	*****************
0.30	5	*****
0.35	2	**

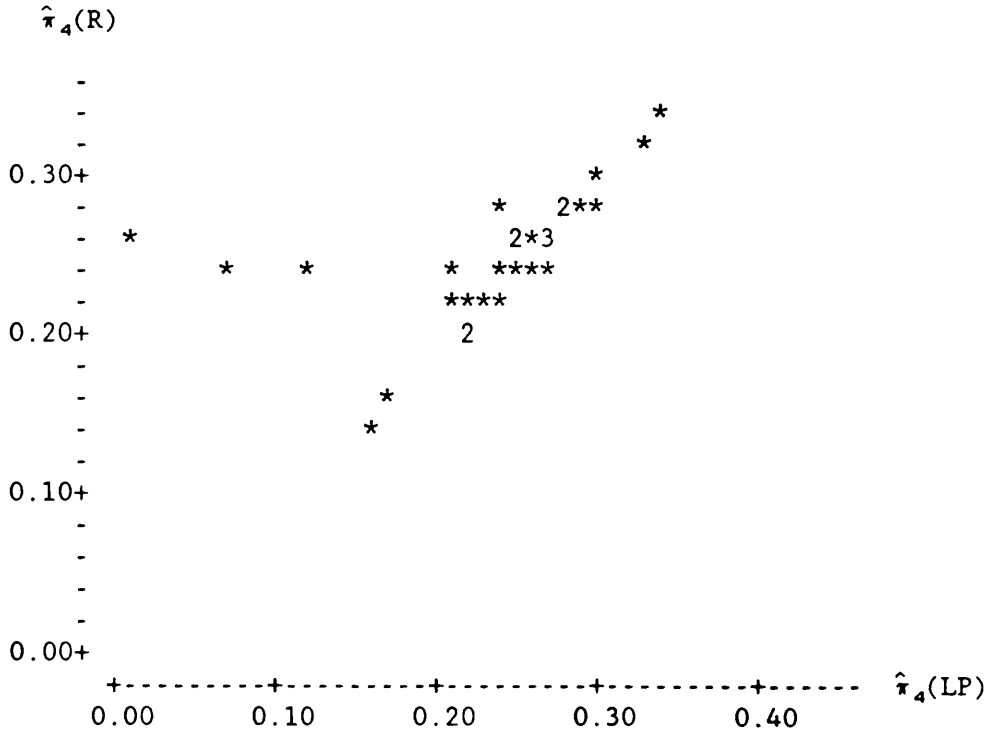


Figure 4.5- Comparison between $\hat{\pi}_4$'s from fitting a logit/probit and the Rasch model to 30 normal bootstrap samples from the ART on black women's data.

From Figure 4.6, item 4 presents the same relation as item 3 in terms of $\hat{\pi}_i$ obtained from fitting the Rasch or a logit/probit model to the same sample. Both models give approximately the same $\hat{\pi}_4$, except for the three samples (3,13,15) in which $\hat{\alpha}_{4,1}$ is bigger than 1.50, that is, 2.57, 3.40 and 8.76 and while $\hat{\pi}_4$ from the Rasch model assumes values around 0.25, the logit/probit model gives $\hat{\pi}_4$ ranging from 0.01 to 0.12 (the three points on the left of the distribution of $\hat{\pi}_4$ shown in the histogram above).

Conclusion

If the logit/probit model fits well, and the discrimination parameter estimate $\hat{\alpha}_{j,1}$ is large (>3.0) and has a large standard deviation, then in those cases where the Rasch model also fits well, there is agreement between the two models about the values $\hat{\pi}_i$, i=1,...,p, except sometimes for j=i.

These results also show that when one of the logit/probit discrimination parameter estimates $\hat{\alpha}_{i,1}$ is large and has large standard deviation, we cannot conclude in general that a model, in which $\pi_i(z)$ assume only two values 0 and 1, would fit the data any better than a Rasch model.

7- *Approximate Methods*

As we know from the literature that the ML estimation procedure seems to be prone to Heywood cases in factor analysis, we thought this might apply to the logit/probit model here and that approximate methods might possibly give better estimates for the Arithmetic Reasoning Test on black women.

We shall apply to the Arithmetic Reasoning Test on black women two approximate methods based directly on cross-product ratios $\Upsilon_{i,j}$, which stem from a result of Bartholomew (1980) for logit/logit and logit/probit models, and are described in Bartholomew (1987, Chapter 6).

Method 1

The first method is based on the idea that if we equate the expected and observed values of

$$\sum_{\substack{j=1 \\ j\neq i}}^{p} (\Upsilon_{ij} - 1), \quad i=1,\ldots,p$$

where $\Upsilon_{i,j}$ is the cross-product ratio between the variables i and j and the $\alpha_{i,1}$'s can be determined uniquely.

The solution of the system of equations

$$\alpha_{i,1} \sum_{\substack{j=1 \\ j\neq i}}^{p} \alpha_{j,1} = \sum_{\substack{j=1 \\ j\neq i}}^{p} (\hat{\Upsilon}_{i,j} - 1) = \hat{T}_i \qquad i=1,\ldots,p$$

where $\hat{\Upsilon}_{i,j}$ is the sample estimator of $\Upsilon_{i,j}$, can be done iteratively

using

$$\hat{\alpha}_{i,1}^{(r+1)} = \left[\hat{T}_i + \left\{ \hat{\alpha}_{i,1}^{(r)} \right\}^2 \right] \div$$

$$\div \left\{ \left| \sum_{i=1}^{p} \hat{T}_i \right| + \sum_{i=1}^{p} \left\{ \hat{\alpha}_{i,1}^{(r)} \right\}^2 \right\}^{\frac{1}{2}} \qquad (4.3)$$

Method 2

The second method is essentially the 'minres' method of Harman and Jones (1966), which is based on minimizing

$$SS = \sum_{\substack{i=1 \\ i \neq j}}^{p} \sum_{j=1}^{p} (r_{ij} - 1 - \alpha_{i,1} \alpha_{j,1})^2 .$$

The solution of the equations by setting the derivatives equal to zero can be done through the following iterative formula

$$\hat{\alpha}_{i,1}^{(r+1)} = \left[\sum_{\substack{j=1 \\ j \neq i}}^{p} (\hat{r}_{ij} - 1) \hat{\alpha}_{j,1}^{(r)} + \left\{ \hat{\alpha}_{i,1}^{(r)3} \right\} \right] \div$$

$$\div \left\{ \sum_{i=1}^{p} \left[\hat{\alpha}_{i,1}^{(r)} \right]^2 \right\}. \qquad (4.4)$$

The criterion used to stop the iterative process carried out during the approximate methods of estimation of $\hat{\alpha}_{i,1}$ was to stop when the difference between consecutive iterations r and r+1 was smaller than 0.00001 for

$$\hat{\alpha}_{i,1}^{(r+1)} - \hat{\alpha}_{i,1}^{(r)} \qquad \text{for all } i, \text{ in method 1}$$

and

$$SS^{(r+1)} - SS^{(r)} \qquad \text{in method 2,}$$

where $$SS^{(r)} = \sum_{\substack{i=1 \\ i \neq j}}^{p} \sum_{j=1}^{p} \left[\hat{\Upsilon}_{i,j} - 1 - \hat{\alpha}_{i,1}^{(r)} \hat{\alpha}_{j,1}^{(r)} \right]^2$$

and $\hat{\Upsilon}$ is the estimated cross product ratio between variables i and j, $i \neq j$.

We have applied both approximate methods described above to the ART on black women's data. Even starting from many different points, there is no convergent solution for $\alpha_{1,1}$, since it continues increasing indefinitely as the iterative procedure is carried out.

When applied to Lombard and Doering's data (Bartholomew, 1987, p.161), these two approximate methods both yield discrimination parameter estimates which are similar to ML ones in the relative importance assigned to the four items.

We have also applied approximate methods 1 and 2 to the 30 normal bootstrap samples we have been working with. This allows a check on whether these approximate methods would give the same pattern of $\hat{\alpha}_{i,1}$ or would confirm statements in the literature that maximum likelihood estimation often gives larger estimates than other methods.

Table 4.9- Discrimination parameter estimates $\hat{\alpha}_{i,1}$ from MLE, approximate methods 1 and 2, respectively, for the 30 normal bootstrap samples obtained from the ART on black women data.

	$\hat{\alpha}_{1,1}$	$\hat{\alpha}_{2,1}$	$\hat{\alpha}_{3,1}$	$\hat{\alpha}_{4,1}$		$\hat{\alpha}_{1,1}$	$\hat{\alpha}_{2,1}$	$\hat{\alpha}_{3,1}$	$\hat{\alpha}_{4,1}$
1	0.50	1.89	0.12	0.65	16	11.13	0.23	0.46	0.20
	0.34	1.44	0.12	0.96		43.87	0.00	0.01	0.00
	0.52	1.42	0.16	0.73		4.30	0.11	0.26	0.09
2	10.95	0.37	0.32	0.27	17	0.54	3.87	0.35	0.49
	3.88	0.16	0.19	0.16		0.59	1.54	0.68	0.55
	0.00	0.23	0.01	-0.08		0.63	1.77	0.44	0.58
3	0.65	0.10	0.59	8.76	18	4.44	0.36	0.18	0.48
	1.14	0.21	0.48	1.90		2.02	0.28	0.08	0.55
	0.64	0.09	0.57	2.75		1.66	0.42	0.16	0.60
4	0.35	0.22	12.10	0.18	19	0.60	0.87	0.25	1.35
	0.01	0.00	54.76	-0.01		0.83	0.87	0.35	1.04
	0.19	0.11	4.14	0.11		0.57	0.92	0.21	1.37
5	0.11	0.52	13.06	0.39	20	0.95	0.53	1.06	0.39
	0.01	0.10	15.23	0.05		1.22	0.50	0.91	0.42
	0.07	0.37	3.57	0.26		1.25	0.45	0.85	0.50
6	1.15	1.66	0.27	0.49	21	5.92	0.32	0.10	0.38
	1.39	1.31	0.32	0.50		1.18	0.41	0.36	0.58
	1.20	1.62	0.24	0.47		2.08	0.31	0.11	0.38
7	5.82	0.32	0.14	0.02	22	7.13	0.68	0.51	0.26
	0.62	0.23	0.78	0.49		2.73	0.63	0.49	0.17
	0.10	-0.03	2.43	0.36		2.50	0.71	0.48	0.21
8	9.96	0.62	0.52	0.26	23	0.60	2.81	0.16	0.53
	3.15	0.55	0.43	0.11		0.58	1.87	0.10	0.63
	2.38	0.66	0.53	0.21		0.70	1.68	0.13	0.59

continue...

	$\hat{\alpha}_{1,1}$	$\hat{\alpha}_{2,1}$	$\hat{\alpha}_{3,1}$	$\hat{\alpha}_{4,1}$		$\hat{\alpha}_{1,1}$	$\hat{\alpha}_{2,1}$	$\hat{\alpha}_{3,1}$	$\hat{\alpha}_{4,1}$
9	1.03	0.78	0.83	0.75	24	0.96	0.90	0.60	1.50
	1.02	0.86	0.92	0.73		0.91	1.15	0.69	1.32
	0.95	0.90	0.92	0.75		1.05	0.94	0.61	1.45
10	0.60	10.31	0.40	0.16	25	1.61	1.40	0.45	0.18
	0.06	28.78	0.02	0.00		1.08	1.08	1.07	0.66
	0.40	3.90	0.24	0.09		1.56	1.50	0.44	0.23
11	11.37	0.50	0.31	0.89	26	12.67	0.12	0.26	0.34
	2.79	0.34	0.22	1.09		52.54	0.00	0.00	0.00
	2.65	0.42	0.26	1.06		4.40	0.05	0.13	0.18
12	11.43	0.29	0.42	0.31	27	0.63	0.92	1.49	0.22
	29.42	0.02	0.02	0.02		0.83	0.88	1.21	0.27
	4.16	0.14	0.24	0.16		0.59	0.99	1.42	0.18
13	0.30	0.82	0.38	3.40	28	8.64	0.08	0.38	0.37
	0.20	1.10	0.24	2.27		13.15	0.01	0.07	0.05
	0.32	1.01	0.44	1.93		3.37	0.04	0.24	0.24
14	16.01	0.93	0.45	0.14	29	14.56	0.34	0.70	0.30
	3.43	0.82	0.37	0.15		21.78	0.01	0.10	0.04
	4.68	0.67	0.25	0.06		4.44	0.18	0.45	0.14
15	0.36	0.60	0.06	2.57	30	11.96	0.37	0.37	0.22
	0.02	0.06	-0.01	25.52		26.54	0.02	0.04	0.01
	0.34	0.64	0.07	1.90		3.87	0.21	0.21	0.12

Since for sample 2, the approximate Method 2 did not converge, the results below are for 29 bootstrap samples.

In 66% of the samples, a logit/probit model and both approximate methods give the same pattern for the discrimination parameter estimates. That is, in 9 samples, all $\hat{\alpha}_{i,1}$'s were smaller than 3.0 for the same relative importance assigned to all items, while in 10 samples, one of the $\hat{\alpha}_{i,1}$'s was large ($\geq$3.4).

In nine (31%) samples, one of the ML estimates of $\alpha_{i,1}$ was bigger than 4.0, while the approximate methods estimates were small.

Finally, for only one sample, the approximate method 1 has produced one very large estimate of $\hat{\alpha}_{i,1}$ while for the other methods all were smaller than 2.6.

These results seem to indicate that the ML procedure is more likely to produce large estimates of the discrimination parameter than these 2 approximate methods. Even so it might happen that one of these approximate methods will produce large $\hat{\alpha}_{i,1}$ the corresponding ML estimate would be small.

8- *Comparison between Marginal and Conditional Maximum Likelihood Estimation*

The estimation procedure developed above, using an E-M algorithm for the Rasch model with one latent variable, and defined as a special case of the general model(1.5), is equivalent to the marginal maximum likelihood (MML) estimation for the one parameter logistic model (Rasch model) given by Thissen (1982).

As we have already discussed in Chapter 1, section 3.1, in this paper Thissen points out that if we assume that the distribution of the latent variable is N(0,1) then the MML procedure is similar to a combination of the conditional maximum likelihood (CML) estimation of the item (difficulty) parameters with estimation of mean and variance of the population (latent) distribution as described by Andersen and Madsen (1977). The mean of the difficulty parameters is equivalent to Andersen and Madsen's population mean and the square of the estimated discrimination parameter is the same as the variance of the population distribution for conventionally standardized MML estimates.

Recall that in the context of CML the response function of the Rasch model is defined as

$$\pi_i(y) = \frac{\exp(y-b_i)}{1 + \exp(y-b_i)}, \quad i=1,\ldots,p$$

where b_i is the difficulty parameter of item i and y is a latent ability parameter normally distributed with mean μ and variance σ^2.

The CML procedure used to estimate b_i is conditioned on the observed number of positive responses given to item i and it does not require any assumption about the distribution of the latent variable. Furthermore, since only the differences between y and b_i appear in the model, adding a constant to all difficulty parameters and 'person parameters' does not affect the model at all. This implies that one must impose a constraint, typically, that Σb_i is equal to zero in order to obtain a unique solution.

As the MML in this thesis assumes that the latent variable is distributed as N(0,1) and the response function of the Rasch model is defined as

$$\text{logit}(\pi_i(z)) = \alpha_{i,0} + \alpha_1 z, \quad i=1,\ldots,p$$

then the standardization of the MML difficulty parameter estimates $\hat{\alpha}_{i,0}$ is done by setting

$$b_i^* = -\alpha_{i,0}/\alpha_1 \tag{4.5}$$

and multiplying by the same correction factor k

$$b_i^* - \bar{b}^*, \quad i=1,\ldots,p \tag{4.6}$$

where $\bar{b}^*$ is the arithmetic mean of b_i^* and

$$\text{where} \quad k = \frac{\sum_{i=1}^{p} | b_i^* - \bar{b}^* |}{\sum_{i=1}^{p} | b_i |}$$

Note that the similarity between MML and CML estimation combined with the estimation of the latent distribution as described by Andersen and Madsen (1977) is not in terms of equality of the corresponding standardized MML and the CML difficulty parameter estimates, but in terms of equality of $k(b^*_i-\bar{b}^*)$ ((4.5) and (4.6)) and α_1 to the mean and standard deviation of the latent distribution.

In the following we shall consider as examples the Law School Admission Test (LSAT) data, sections 6 and 7, in order to compare the corresponding difficulty parameter estimates obtained from fitting the Rasch model, using CML and MML procedures.

The results for the CML estimation will be taken from Andersen and Madsen (1977) and Andersen (1980), and the MML estimates will be obtained by applying the E-M algorithm described at the begining of this chapter.

Table 4.10- Difficulty parameter estimates from fitting the Rasch model to the LSAT, section 6.

item	MML	MML*	CML
1	2.729	-1.255	-1.256
2	0.999	0.476	0.475
3	0.240	1.235	1.236
4	1.306	0.169	0.168
5	2.099	0.624	0.623
	$\bar{b}^*$ =-1.97	mean=1.47	mean=1.47
	$\hat{\alpha}_1$ = 0.75	variance=0.56	variance=0.55

MML*: standardized MML

According to Andersen and Madsen (1977), the Rasch model fits these data very well (i.e.,on significance level 0.05), which agrees with the results obtained from using the MML procedure, since in this case the observed chi-squared is equal to 17.90 with 17 degrees of freedom.

The standardized MML difficulty parameter estimates match with the corresponding CML estimates closely, and the estimated mean difficulty and discrimination match with the mean and standard deviation of the latent distribution obtained from the CML estimates.

Table 4.11- Difficulty parameter estimates from fitting the Rasch model to the LSAT, section 7.

item	MML	MML*	CML
1	1.868	-0.541	-0.641
2	0.791	0.535	0.583
3	1.461	-0.134	-0.134
4	0.522	0.804	0.758
5	1.993	-0.666	-0.566
	$\bar{b}^*$ =-1.31	mean=1.33	
	$\hat{\alpha}_1$ = 1.01	variance=1.02	

MML*:standardized MML

Section 7 of the LSAT data is not fitted by the Rasch model on 0.05 significance level, either using CML (Andersen (1973b)) or MML methods (observed chi-squared equal to 36.24 with 22 degrees of freedom).

The similarity between the corresponding standardized MML and CML difficulty parameter estimates is not as close as in the first example, where the data were fitted well by the Rasch model.

From these examples and many others that we have used to compare CML and MML methods, there is some evidence that the standardized MML difficulty parameter estimates are likely to be very similar to the corresponding CML estimates when the Rasch model fits the data well, but they can be quite different when the discrimination parameters are not the same for all items. This result could be expected since the estimation of the CML difficulty parameter of item i is based on the total number of positive responses to this item , while the MML estimates is obtained taking into account the score patterns. Hence when the Rasch model does not fit the data there is a source of variation in the data which tends to increase the difference between CML and MML estimates.

9- *Comparison between Rasch and Logit/Probit Models in terms of the Likelihood Ratio Statistic*

Our main objective in this section is to compare a logit/probit model with response function given by

$$\text{logit}(\pi_i(z)) = \alpha_{i,0} + \alpha_{i,1}\, z \qquad i=1,\ldots,p$$

with the Rasch model, defined by

$$\text{logit}(\pi_i(z)) = \alpha'_{i,0} + \alpha_1\, z \qquad i=1,\ldots,p.$$

That is, compare a logit/probit with the Rasch model testing

$$H_0: \alpha_{1,1} = \ldots = \alpha_{p,1} = \alpha_1$$

using the likelihood ratio (LR) test.

Let $L = \prod_{s=1}^{n} f(x_s)$ be the likelihood function.

Then the LR statistic may be taken as

$$l = \frac{L \text{ (Rasch } | \text{ p+1 parameters)}}{L \text{ (logit/probit } | \text{ 2p parameters)}}$$

If H_0 is true then the asymptotic distribution of $-2\ln(l)$ is chi-squared with p-1 degrees of freedom (Kendall and Stuart, 1979, page 247).

Therefore, using a LR test is asymptotically equivalent to basing a test on the ML estimators of the parameter tested. However these results only hold if the conditions of the asymptotic normality and efficiency of the ML estimators are satisfied.

As we have already pointed out in this chapter, section 2, in practice very often we cannot use the chi-squared statistic as a goodness-of-fit for either Rasch or logit/probit models, since the number of degrees of freedom is negative, for instance.

If the conditions of the asymptotic distribution of the LR statistic hold, then we can compare the fitting of the Rasch model with the logit/probit model based on the difference between the number of parameters, which is fixed. If the difference between the 2 models is significant then at least one of the $\hat{\alpha}_{i,1}$'s is different from the

others. This implies that a logit/probit model will fit better than the Rasch model.

Application

Now we use the LR to compare the logit/probit and the Rasch model on the data from Test 11A of the NFER (20 items). We assume that the conditions of asymptotic normality and efficiency of the ML estimator are satisfied.

Let us consider the case when the sample was stratified according to the location of the school: England(342), Wales(86) and Ireland(73). Chapter 5 , section 2, contains an extensive discussion about the pattern of the parameter estimates $\hat{\alpha}_{i,1}$'s when fitting a logit/probit model to those subtests obtaining by deleting some items from Test 11A. Here, using the LR test, we compare the fittings of a logit/probit and the Rasch model to those subtests, which vary in number of items, sample sizes and pattern of $\hat{\alpha}_{i,1}$.

It is convenient to remember that the parameter estimates $\hat{\alpha}_{i,1}$ (≥ 0.50) and the standard deviations are approximately linearly related so that larger estimates have larger standard deviations.

Table 4.12- Comparison between Rasch and logit/probit models in terms of the LR statistic, when fitting subtests of Test 11A - England.

Items	$\hat{\alpha}_{i,1}$ (min,max)	Loglikelihood Rasch	Loglikelihood Log/Prob	df	-2ln(l)
1 to 20	0.84;3.15	-3435.37	-3390.92	19	88.90*
del 7,12	0.84;3.44	-3123.95	-3080.76	17	86.38*
1 to 10	0.70;3.95	-1790.21	-1746.36	9	87.70*
1to6,8to10	0.71;3.09	-1649.99	-1623.18	8	53.62*
1 to 7	0.78;4.76	-1223.09	-1197.63	6	50.92*
1 to 6	0.79;3.22	-1062.86	-1049.58	5	26.56*
1 to 5	0.83;2.67	-919.37	-911.62	4	15.50*
2 to 6	1.37;3.64	-835.58	-828.80	4	21.56*
3 to 7	1.30;6.14	-885.75	-866.53	4	38.44*
11 to 20	1.04;3.98	-1743.93	-1718.91	9	50.04*
11,13to20	1.08;4.36	-1563.85	-1541.63	8	44.44*
11 to 15	1.63;3.97	-762.78	-756.60	4	12.36*
11,13to15	1.60;7.12	-590.68	-583.63	3	14.10*

*: H_0 is rejected with $0<p<0.05$.

For all subtests Table 4.12 shows that H_0 is rejected, that is, at least one of the $\hat{\alpha}_{i,1}$'s is different from the others. This implies that the logit/probit model fits the data better than the Rasch model.

The minimum values assumed by $\hat{\alpha}_{i,1}$ range from 0.70 to 1.63 and the maximum values from 2.67 and 4.76. Therefore the subtest of items 1 to 5 represents the smallest variation (from 0.83 to 2.67) among the values assumed by $\hat{\alpha}_{i,1}$, for which at least on of the $\hat{\alpha}_{i,1}$'s is different from the others.

Table 4.13- Comparison between Rasch and logit/probit models in terms of the LR statistic, when fitting subtests of Test 11A-Wales.

Items	$\hat{\alpha}_{i,1}$ (min,max)	Loglikelihood Rasch	Log/Prob	df	-2ln(l)
1 to 20	0.25;2.73	-911.86	-893.23	19	37.26*
del 7,12	0.20;2.12	-831.18	-817.26	17	27.84*
1 to 10	0.85;4.59	-470.32	-459.36	9	21.92*
1to6,8to10	0.87;4.41	-431.65	-427.34	8	8.62
1 to 7	0.88;3.34	-327.35	-323.61	6	7.48
1 to 6	0.88;2.58	-282.98	-281.16	5	3.64
1 to 5	0.74;4.38	-238.07	-235.10	4	5.94
2 to 6	1.00;1.97	-231.59	-231.12	4	0.94
3 to 7	0.82;4.76	-250.10	-244.38	4	11.44*
11 to 20	0.46;2.61	-463.68	-455.52	9	16.32
11,13to20	0.47;2.95	-421.32	-414.82	8	13.00
11 to 15	0.99;16.03	-209.91	-205.22	4	9.38
11,13to15	0.64;1.87	-164.66	-163.44	3	2.44

*: H_0 is rejected with $0<p<0.05$.

The equality among all $\hat{\alpha}_{i,1}$'s is rejected for four subtests, for which the number of items ranges from 20 to 5. For these subtests, the smallest variation for the $\hat{\alpha}_{i,1}$'s is when deleting items 7 and 12 from the original test, in which $\hat{\alpha}_{i,1}$ assumes values between 0.20 and 2.12; the biggest variation is for the subtest of items 3 to 7, in which $\hat{\alpha}_{i,1}$ ranges from 0.82 to 4.76.

Comparison between the pattern of the $\hat{\alpha}_{i,1}$'s and the results from the LR test show some contradictions, which are probably due to the violation of the assumptions of this test.

Thus, for instance, Table 5.8 in Chapter 5, shows that the subtest of items 1 to 10 have two large $\hat{\alpha}_{i,1}$'s, 3.72 and 4.59 for items 6 and 7. When deleting item 7 from this subset, the pattern of $\hat{\alpha}_{i,1}$'s practically does not change and $\hat{\alpha}_{6,1}$ is equal to 4.41. As H_0 is

rejected for the subtest of items 1 to 10, but accepted when deleting item 7, this suggests that the difference among the $\hat{\alpha}_{i,1}$'s for the first case is due to item 7.

On the other hand, H_0 is accepted for the subset of items 1 to 7, but it is rejected for items 3 to 7. The comparison between the patterns of $\hat{\alpha}_{i,1}$ for the subset of items 1 to 7 and 3 to 7 shows that the minimum $\hat{\alpha}_{i,1}$ are nearly the same, 0.88 and 0.82, and both have two $\hat{\alpha}_{i,1}$'s bigger than 3.0, for items 6 and 7. In the first subset these estimates and standard deviations (in parenthesis) are 3.21 (1.43) and 3.34 (1.56), while for items 3 to 7 they are equal to 3.39 (1.70) and 4.76 (3.97). As the largest $\hat{\alpha}_{7,1}$ has also the largest coefficient of variation, one could expect the same result when applying the LR test, instead of rejecting only one of them.

Table 4.14- Comparison between Rasch and logit/probit models in terms of the LR statistic, when fitting subtests of Test 11A -Ireland.

Items	$\hat{\alpha}_{i,1}$ (min,max)	Loglikelihood Rasch	Log/Prob	df	-2ln(l)
1 to 20	0.29;16.24	-748.87	-723.56	19	50.62*
del 7,12	0.29; 2.43	-693.34	-682.02	17	22.64
1 to 10	1.00;16.72	-386.62	-370.43	9	32.38*
1to6,8to10	1.01;16.38	-356.23	-350.23	8	12.00
1 to 7	1.34;16.29	-264.92	-260.44	6	8.96
1 to 6	1.20; 2.73	-231.36	-229.96	5	2.80
1 to 5	0.99; 3.23	-198.30	-196.37	4	3.86
2 to 6	1.25; 2.40	-191.80	-190.84	4	1.92
3 to 7	1.16;16.58	-190.21	-185.40	4	9.62*
11 to 20	0.42;16.36	-372.67	-358.24	9	28.86*
11,13to20	0.37; 2.67	-347.25	-337.89	8	18.72*
11 to 15	1.82;16.13	-158.36	-155.67	4	5.38
11,13to15	1.75; 3.26	-132.30	-131.80	3	0.80

*: H_0 is rejected with $0<p<0.05$.

As for Wales, most of the hypothesis H_0 are not rejected. In this case $\hat{\alpha}_{i,1}$ can often assume very large values both in subtests, for which H_0 is accepted and in subtests for which it is rejected.

Table 5.5, Chapter 5, shows that $\hat{\alpha}_{i,1}$ assumes two values bigger than 3.0 that is, 5.32 and 16.24, for items 7 and 12 respectively. When deleting these two items, although the $\hat{\alpha}_{i,1}$'s change very little, H_0 is no longer rejected. This would lead one to infer that the rejection of equality of all the $\hat{\alpha}_{i,1}$'s in the original test is due to $\hat{\alpha}_{7,1}$ and $\hat{\alpha}_{12,1}$.

Only for Wales, the comparison between the values of $\hat{\alpha}_{i,1}$'s and the results of the LR test shows some contradictions. The same discussion about accepting or rejecting H_0, when comparing the pattern of $\hat{\alpha}_{i,1}$'s from the subset of items 1 to 10 with 1 to 6, 8 to 10, and 1 to 7 with 3 to 7, holds here, though $\hat{\alpha}_{6,1}$ and $\hat{\alpha}_{7,1}$ assume very large values.

Conclusion

The comparison between the Rasch and the general logit/probit model was done testing the hypothesis H_0: all $\hat{\alpha}_{i,1}$'s are equal, using the LR statistic, for different sample sizes, pattern of $\hat{\alpha}_{i,1}$'s and number of items. The results from this study give evidence that for a large sample size, like England(342), the LR statistic tends to reject H_0, while for small sample sizes (Wales(84) and Ireland(73)), it tends not to reject H_0. It is not clear whether this effect is due to the larger power of a larger sample size, or whether the asymptotic chi-squared distribution is not applicable.

Chapter 5

STABILITY of the DISCRIMINATION PARAMETER ESTIMATES $\hat{\alpha}_{i,1}$

1- *Effect on the Size of $\hat{\alpha}_{i,1}$ of Decreasing the Number of Items for a Fixed Sample Size*

We shall investigate the effect on $\hat{\alpha}_{i,1}$ when fitting a logit/probit model to Test 11A and Test 12 of the NFER data, of varying the number of items by deleting some of the original ones. The sample sizes (501 and 502, respectively) are kept fixed.

The subsets are formed so that we can measure the effect on the pattern of $\hat{\alpha}_{i,1}$'s when the 2 items with the largest $\hat{\alpha}_{i,1}$'s in the original test are included or are not included in the subset, as the number of items decreases.

The questionnaries will be also analysed in order to find out whether the occurrence of a large $\hat{\alpha}_{i,1}$ is associated with the kind of question.

We start by analysing each test, following with a comparison of the main results.

1.1- Test 11A

Tables 5.1 and 5.2 display the parameter estimates $\hat{\alpha}_{i,1}$ and their asymptotic standard deviations from fitting a logit/probit model to different sets of items drawn from Test 11A.

Table 5.1- Parameter estimates $\hat{\alpha}_{i,1}$ and asymptotic standard deviations (in brackets) from fitting a logit/probit model to Test 11A.

i	1 to 20	1 to 15	1 to 10	5 to 9
1	0.92 (0.14)	0.90 (0.14)	0.75 (0.13)	
2	1.37 (0.18)	1.35 (0.19)	1.29 (0.20)	
3	1.50 (0.20)	1.40 (0.19)	1.21 (0.18)	
4	1.68 (0.20)	1.78 (0.22)	1.76 (0.23)	
5	1.18 (0.15)	1.23 (0.16)	1.06 (0.15)	0.80 (0.13)
6	2.33 (0.28)	2.72 (0.35)	4.28 (0.73)	9.74 (28.49)
7	2.53 (0.30)	2.91 (0.38)	4.34 (0.77)	16.12 (513.86)
8	1.35 (0.16)	1.49 (0.18)	1.77 (0.22)	2.13 (0.29)
9	1.14 (0.15)	1.18 (0.16)	1.27 (0.18)	1.42 (0.19)
10	2.04 (0.24)	2.10 (0.25)	1.84 (0.22)	
11	1.50 (0.18)	1.41 (0.18)		
12	2.13 (0.25)	2.12 (0.26)		
13	2.08 (0.27)	1.99 (0.26)		
14	1.32 (0.17)	1.27 (0.17)		
15	2.62 (0.36)	2.40 (0.33)		
16	0.83 (0.13)			
17	1.18 (0.15)			
18	1.70 (0.20)			
19	0.97 (0.14)			
20	1.54 (0.18)			

Table 5.2- Parameter estimates $\hat{\alpha}_{i,1}$ and asymptotic standard deviations (in brackets) from fitting a logit/probit model to Test 11A.

i	1 to 20	1,2,8 to 20	11 to 20	13 to 17
1	0.92 (0.14)	0.87 (0.14)		
2	1.37 (0.18)	1.33 (0.19)		
3	1.50 (0.20)			
4	1.68 (0.20)			
5	1.18 (0.15)			
6	2.33 (0.28)			
7	2.53 (0.30)			
8	1.35 (0.16)	1.06 (0.15)		
9	1.14 (0.15)	1.03 (0.15)		
10	2.04 (0.24)	1.85 (0.22)		
11	1.50 (0.18)	1.68 (0.21)	1.76 (0.23)	
12	2.13 (0.25)	2.32 (0.30)	2.22 (0.30)	
13	2.08 (0.27)	2.32 (0.31)	2.41 (0.35)	2.60 (0.51)
14	1.32 (0.17)	1.41 (0.18)	1.46 (0.20)	1.45 (0.24)
15	2.62 (0.36)	3.21 (0.50)	3.41 (0.59)	4.54 (1.61)
16	0.83 (0.13)	0.89 (0.14)	0.92 (0.15)	1.01 (0.19)
17	1.18 (0.15)	1.20 (0.16)	1.23 (0.17)	1.17 (0.21)
18	1.70 (0.20)	1.82 (0.22)	1.81 (0.24)	
19	0.97 (0.14)	1.02 (0.15)	1.02 (0.15)	
20	1.54 (0.18)	1.78 (0.22)	1.85 (0.24)	

From Table 5.1 we can see that when fitting a logit/probit model to the whole set of items of Test 11A, the parameter estimates $\hat{\alpha}_{i,1}$ assume values from 0.83 to 2.62. The three largest estimates are equal to 2.62, 2.53 and 2.33 for items 15, 7 and 6, respectively. All $\alpha_{i,1}$'s are not correlated or weakly correlated, since the maximum correlation estimate is equal to 0.22, between items 6 and 7.

The comparison between the parameter estimates $\hat{\alpha}_{i,1}$ from fitting a logit/probit model to all 20 items and those obtained by decreasing the number of items to 15 leads to

i)approximately the same values of $\hat{\alpha}_{i,1}$, when deleting items 16 to 20,

ii) bigger $\hat{\alpha}_{15,1}$ (3.21 compared to 2.62) when deleting items 3 to 7, while the estimates $\hat{\alpha}_{i,1}$ practically do not change for the remaining items.

Decreasing the number of items to 10 and 5, we observe the following changes, when fitting a logit/probit model to

i)items 1 to 10

$\hat{\alpha}_{6,1}$ and $\hat{\alpha}_{7,1}$ increase to from 2.33 to 4.28 and from 2.53 to 4.34 with standard deviation equal to 0.73 and 0.77, respectively.

When considering only five items, i=5 to 9, then $\hat{\alpha}_{i,1}$ for items 6 and 7, increase even more to assume very large values, 9.74 and 16.12 with large standard deviations, 28.49 and 513.86, respectively.

ii)items 11 to 20

$\hat{\alpha}_{15,1}$ increases from 2.62 to 3.41 with an increase in their standard deviation.

Also, when restricting to items 13 to 17, $\hat{\alpha}_{15,1}$ becomes even bigger, 4.54 with standard deviation equal to 1.61.

Results from Test 11A show that as the number of items decreases from 20 to 15,10 and 5 items, the biggest parameter estimates $\hat{\alpha}_{i,1}$, i=6,7,15, also increase , assuming large values especially when the number of items is small (5). There is an approximate linear relation

between $\hat{\alpha}_{i,1}$ and corresponding standard deviation with larger estimates having larger standard deviations.

At the same time, the remaining items show great stability, since they assume approximately the same values under different subtests and number of items.

We tried stratifying these samples into two groups according to the location of the school in a metropolitan (125) or nonmetropolitan (376) area. We observe the same pattern of $\hat{\alpha}_{i,1}$ as we decrease the number of items as described above for the whole test, although the smaller sample (metropolitan) presented larger coefficients of variation due to larger standard deviations.

Analysis of the items in Test 11A

Test 11A corresponds to a story entitled 'King Lion'. This story is written as a fable, in which a small animal (a squirrel) outwits a more powerful one (King Lion).

The lion announces that in order to save the animals work in fetching his food, he will eat one of them every day, in an order they choose. They are left to decide how to put his suggestions into practice. The squirrel saves everybody's life by leading the lion to a deep pool where, he alleges, a strange creature is waiting for him. On seeing his reflexion on the water, the lion jumps into the pool and is drowned.

In order to find out whether the larger size of $\hat{\alpha}_{i,1}$ for items 6, 7, 12 and 15 is associated with the type of question asked. We have analysed all questions in terms of whether the questions were clearly formulated or not and if the answers were explicitly given or they should be infered. The results from this analysis were related to the $\hat{\alpha}_{i,1}$ values . In the following, we present the main results.

While the animals were deciding which one will be the lion's first food, there emerged 2 plans, which would not work since there was something wrong with each one.

Item 6 asks for a description of the first plan, which was clearly given in the story. Item 7 asks what was wrong with this plan, for which the subjects had to infer.

Items 8 and 9 correspond to questions 6 and 7 for the second plan. The only difference between them is that the answer for item 9 was given in the text. However this difference should not be responsible for the occurrence of a large $\hat{\alpha}_{7,1}$, since there were other items similar to 7.

When the squirrel offers himself to be the first food for the lion, he asked: 'Do you have any objections?' and the other animals answered hurriedly: 'Not at all'. Question 12 asks why the animals answered hurriedly. The answer was not explicit in the text, but was very clear from the story.

Question 15 is about what reason the squirrel gave to the lion for being late. The answer was explicitly given.

Therefore from the previous analysis, there is no evidence that the larger $\hat{\alpha}_{i,1}$ for items 6,7,12 and 15 are associated with the nature of the questions.

1.2- Test 12

As a second example, we shall investigate the effect on the size of the discrimination parameter estimates $\hat{\alpha}_{i,1}$ as the number of items decrease for a fixed sample size obtained from fitting a logit/probit model to test 12-NFER data.

Table 5.3- Parameter estimates $\hat{\alpha}_{i,1}$ and asymptotic standard deviations (in brackets) from fitting a logit/probit model to Test 12.

i	1 to 18	delet.15	delet.16	delet.15,16
1	1.17 (0.18)	1.18 (0.18)	1.16 (0.18)	1.15 (0.18)
2	1.62 (0.19)	1.69 (0.20)	1.70 (0.20)	1.73 (0.21)
3	1.26 (0.15)	1.30 (0.16)	1.29 (0.16)	1.30 (0.16)
4	1.61 (0.18)	1.64 (0.19)	1.63 (0.19)	1.61 (0.19)
5	2.08 (0.23)	2.22 (0.26)	2.19 (0.26)	2.25 (0.27)
6	1.34 (0.17)	1.40 (0.18)	1.41 (0.18)	1.44 (0.19)
7	1.49 (0.17)	1.51 (0.18)	1.51 (0.18)	1.50 (0.18)
8	2.20 (0.26)	2.29 (0.23)	2.34 (0.29)	2.35 (0.30)
9	1.49 (0.17)	1.49 (0.18)	1.48 (0.18)	1.45 (0.18)
10	0.87 (0.13)	0.87 (0.13)	0.88 (0.13)	0.88 (0.13)
11	0.62 (0.12)	0.64 (0.12)	0.65 (0.12)	0.66 (0.12)
12	2.02 (0.22)	2.16 (0.25)	2.16 (0.25)	2.21 (0.26)
13	1.24 (0.18)	1.30 (0.19)	1.30 (0.19)	1.34 (0.20)
14	1.65 (0.23)	1.64 (0.23)	1.62 (0.23)	1.59 (0.23)
15	4.50 (0.83)		2.72 (0.35)	
16	4.39 (0.70)	2.84 (0.35)		
17	1.75 (0.20)	1.78 (0.21)	1.76 (0.21)	1.75 (0.21)
18	1.58 (0.18)	1.60 (0.18)	1.60 (0.19)	1.58 (0.19)

The fitting to the whole set of items shows two items, 15 and 16, having large estimates $\hat{\alpha}_{i,1}$, 4.50 and 4.39, and standard deviations equal to 0.83 and 0.70, respectively. The correlation between these

parameter estimates is equal to 0.45, while the remaining elements of the correlation matrix are approximately equal to zero.

In order to find out the effect of the presence of items 15 and 16 on the size of the discrimination parameter estimates for the other items, we have considered 3 situations: deleting just item 15, deleting just item 16 and deleting both from the original test.

Table 5.3 shows that the parameter estimates $\hat{\alpha}_{i,1}$ for $i \neq 15,16$, are very stable, since these estimates and their standard deviations are nearly equal to the original ones, independently of whether items 15 or 16 are present or not in the set of items under consideration.

Deleting only one of these items, for example 15, yields the same effect on $\hat{\alpha}_{16,1}$ (decreases from 4.39 to 2.84 with standard deviation equal to 0.35) as deleting item 16, on the estimate $\hat{\alpha}_{15,1}$ (decreases from 4.50 to 2.72 with standard deviation equal to 0.35). Although, in these situations $\hat{\alpha}_{i,1}$, i=15 or 16, is smaller , the sequence of the score patterns in increasing order of the component scores is approximately the same, that is, we can still see two fairly distinct groups according to whether they have answered item 15 or 16 wrong followed by those that have answered it right.

Table 5.4- Parameter estimates $\hat{\alpha}_{i,1}$ and asymptotic standard deviations (in brackets) from fitting a logit/probit model to Test 12.

i	5 to 16	4 to 15	4 to 14,16	3 to 14
1				
2				
3				1.16 (0.15)
4		1.56 (0.19)	1.59 (0.19)	1.51 (0.19)
5	1.97 (0.22)	2.32 (0.29)	2.36 (0.29)	2.37 (0.30)
6	1.20 (0.16)	1.34 (0.18)	1.33 (0.18)	1.38 (0.19)
7	1.44 (0.17)	1.48 (0.18)	1.48 (0.18)	1.51 (0.19)
8	2.20 (0.27)	2.65 (0.36)	2.55 (0.35)	2.61 (0.37)
9	1.52 (0.18)	1.48 (0.18)	1.49 (0.18)	1.41 (0.18)
10	0.79 (0.13)	0.80 (0.13)	0.80 (0.13)	0.81 (0.13)
11	0.63 (0.12)	0.63 (0.12)	0.62 (0.12)	0.64 (0.12)
12	1.87 (0.23)	2.21 (0.27)	2.22 (0.27)	2.28 (0.29)
13	1.39 (0.20)	1.52 (0.22)	1.50 (0.22)	1.54 (0.23)
14	1.70 (0.25)	1.60 (0.24)	1.61 (0.24)	1.58 (0.23)
15	9.22 (2.71)	2.73 (0.38)		
16	6.78 (0.60)		2.74 (0.36)	
17				
18				

i	9 to 14	10 to 15	10 to 14,16	11 to 16
9	1.27 (0.20)			
10	0.72 (0.14)	0.81 (0.14)	0.81 (0.14)	
11	0.77 (0.15)	0.67 (0.13)	0.67 (0.13)	0.55 (0.11)
12	2.09 (0.39)	1.91 (0.31)	1.91 (0.31)	1.39 (0.17)
13	1.68 (0.30)	1.43 (0.24)	1.43 (0.24)	1.08 (0.17)
14	1.65 (0.30)	1.97 (0.37)	1.97 (0.37)	1.97 (0.35)
15		3.19 (0.88)		16.01(106.81)
16			3.19 (0.88)	14.40(179.71)

The effect on the pattern of $\hat{\alpha}_{i,1}$ of decreasing the number of items to 12, was analysed using 5 different sets of items, some excluding item 15 or 16 and others including both. As we can see from Table 5.4 the size and pattern of the discrimination parameter estimates behave in the same way as observed in the previous analysis.

Finally we have considered 4 sets of 6 items. When excluding both items 15 and 16, the remaining $\hat{\alpha}_{i,1}$ show great stability as in the preceding investigations; including item 15 (16) and excluding item 16 (15), the parameter estimate $\hat{\alpha}_{15,1}$ ($\hat{\alpha}_{16,1}$) decreases to 3.19 having standard deviation equal to 0.88. However when both, items 15 and 16, are among the 6 items, $\hat{\alpha}_{15,1}$ and $\hat{\alpha}_{16,1}$ become very large (16.01 and 14.46) and have large standard deviations (106.81 and 179.71, respectively); while the remaining $\hat{\alpha}_{i,1}$ do not change significantly.

This example shows two items for which $\hat{\alpha}_{i,1}$ are correlated (0.45) and they are larger when included in the same subtest. That is, correlated parameters might lead to larger values of $\hat{\alpha}_{i,1}$ when the number of items is decreased, which does not, however, imply a small $\hat{\alpha}_{i,1}$ when just one of these items is present. On the other hand, $\hat{\alpha}_{i,1}$ for the remaining items shows great stability, even when the number of items was small (6).

We have also analysed Test 12 under all these different combinations of items, when the sample size was stratified in 2 groups, according to the location of the school in a metropolitan (n=127) or nonmetropolitan (n=375) area. Overall the different situations, the fitting of logit/probit model to both areas have lead to the same pattern of $\hat{\alpha}_{i,1}$ that we found when considering

the whole sample size. In this case, even the smaller sample size (127) has not produced larger $\hat{\alpha}_{i,1}$ and larger standard deviations than those from the whole sample.

Analysis of the items in Test 12

Test 12 corresponds to the story 'That Sinking Feeling' by Betsy Byars, which is described in the form of a first-person narrative. The story depicts how a little girl gets revenge on her older brothers and their friends who refuse to let her play with them, by pulling the plugs out of a makeshift raft of oil drums.

In order to find out whether the large size of $\hat{\alpha}_{i,1}$ for items 15 and 16, that is, 4.50 and 4.39, were associated with the kind of item, we analysed all the questions. This analysis was carried out in terms of whether the questions were clearly formulated or not and if the answers were given explicitly or should be infered from the story. The results from these analysis were related to the values assumed by $\hat{\alpha}_{i,1}$.

In the following the main results from the analysis of the questions.

Item 15 asks what was the sex of the narrator, which answer was not explicitly given in the story, but established by linguistic and circumstantial evidence. For example, her brother says: 'I'll kill her' and her mother advises: 'I shouldn't push your luck, madam'.

On the other hand, item 10 asks how many of the storyteller's brothers were on the raft, which was also not explicitly given in the story, but for which $\hat{\alpha}_{10,1}$ is small (0.87).

Item 16 asks the approximate age of the storyteller, which was not explicitly given. The children have to realise that she was younger than most of the boys, and guess an age supported, for example, by a picture given in the text. Evidence to answer this item were also given in question 3, which asks: 'How do you know that the storyteller is younger than most of the boys at the quarry?'. Therefore questions 16 and 3 are correlated and their answers come from the same source, but $\hat{\alpha}_{3,1}$ is much smaller than $\hat{\alpha}_{16,1}$ (1.26 compared with 4.39).

This previous analysis about the nature of items 15 and 16 in relation to the remaining items in the test, leads us to conclude that the large values for $\hat{\alpha}_{15,1}$ and $\hat{\alpha}_{16,1}$ are not associated with the kind of question.

Conclusion

We have analysed the effect on the size of $\hat{\alpha}_{i,1}$ using two examples, with the same order of test length (18 and 20) and sample size (502 and 501). Although the two biggest estimates of $\hat{\alpha}_{i,1}$, 4.39 and 4.50, in test 12 were larger than the three biggest $\hat{\alpha}_{i,1}$, 2.33, 2.53 and 2.62 in test 11A , and in the former they were correlated, the main results are the same:

(1) as the number of items decreases , the largest $\hat{\alpha}_{i,1}$ increases and becomes very large when the test length is small (5 or 6 items).

(2) parameter estimates $\hat{\alpha}_{i,1}$ and standard deviations are approximately linear related so that larger estimates have larger standard deviations.

(3) the remaining items show great stability of $\alpha_{i,1}$ estimates, even when the sample size is small or when deleting an item with large $\hat{\alpha}_{i,1}$.

(4) the large parameter estimate $\hat{\alpha}_{i,1}$ for some items does not seem to be associated with the type of question asked.

2- *Effect on the Size of* $\hat{\alpha}_{i,1}$

as the Number of Items and the Sample Size Decrease

In order to investigate the effect on $\hat{\alpha}_{i,1}$ as the number of items and sample size decrease, we consider Test 11A again, when the population is stratified according to the country where the school is located: England, Wales and Ireland.

Since the sample size of children in English schools is not small (342) while the samples from Wales and Ireland are small (86 and 73), we will be able to compare the effect on the pattern of $\hat{\alpha}_{i,1}$'s under different sample size and different test length.

Furthermore, as the component score is a linear function of $X_1, X_2, \ldots, X_p$ with coefficients equal to $\alpha_{1,1}, \alpha_{2,1}, \ldots, \alpha_{p,1}$, that is, $X = \sum \alpha_{i,1} X_i$ then a change on the pattern of $\alpha_{i,1}$ when deleting items, might affect the component scores too.

As the component score is a function of the number of items, larger tests will tend to produce larger component scores. Since we are comparing different test lengths, what matters is the order of the new component scores and not their estimated values. This effect on the component scores of deleting items will be measured by the Spearman correlation.

The investigation of the effect on $\hat{\alpha}_{i,1}$ of deleting items will be complemented with an analysis of the questions, in order to find out whether a large discrimination parameter estimate is associated with the kind of question.

In the following we shall analyse the results from each country individually, starting with Ireland, and finally comparing them.

2.1- Ireland

Tables 5.5 and 5.6 display the parameter estimates $\hat{\alpha}_{i,1}$ and standard deviations from fitting a logit/probit model to the Ireland data for different sets of items.

Table 5.5- Parameter estimates $\hat{\alpha}_{i,1}$ and asymptotic standard deviations(in brackets) from fitting a logit/probit model to Test 11A- Ireland.

i	1 to 20	delet.7,12	1 to 10	1to6,8to10
1	1.19 (0.39)	1.30 (0.44)	1.01 (0.36)	1.01 (0.43)
2	1.49 (0.51)	1.40 (0.47)	1.20 (0.42)	1.34 (0.52)
3	1.90 (0.57)	2.22 (0.78)	1.36 (0.44)	1.45 (0.56)
4	1.78 (0.54)	1.55 (0.52)	1.80 (0.63)	1.80 (0.77)
5	1.36 (0.41)	1.29 (0.42)	1.21 (0.40)	1.30 (0.50)
6	2.79 (0.77)	1.99 (0.65)	16.72(606.19)	16.38(701.67)
7	5.32 (3.15)		12.65(34.45)	
8	1.71 (0.48)	1.33 (0.45)	2.61 (0.70)	2.25 (0.83)
9	1.15 (0.38)	1.25 (0.45)	1.00 (0.38)	1.10 (0.42)
10	1.40 (0.43)	1.11 (0.40)	1.79 (0.52)	2.05 (0.81)
11	1.19 (0.37)	1.28 (0.42)		
12	16.24(519.01)			
13	1.13 (0.44)	1.43 (0.51)		
14	1.42 (0.46)	1.30 (0.44)		
15	1.59 (0.54)	1.89 (0.66)		
16	0.29 (0.26)	0.29 (0.27)		
17	1.32 (0.46)	1.40 (0.51)		
18	2.12 (0.63)	2.43 (0.82)		
19	0.99 (0.36)	1.08 (0.41)		
20	1.93 (0.57)	2.04 (0.64)		

In the whole set of items, $\hat{\alpha}_{i,1}$ assumes values from 0.29 to 16.24, with standard deviation being larger for larger estimates. The three biggest estimates are 2.79, 5.32 and 16.24 for items 6,7 and 12, respectively. The maximum correlation between $\alpha_{i,1}$ and $\alpha_{j,1}$, $i \neq j$, is equal to 0.25, for items 7 and 8.

Deleting items 7 and 12, some estimates remain the same, while others increase or decrease slightly. The strongest change is related to item 6, for which $\hat{\alpha}_{6,1}$ decreases to 1.99 (29% smaller).

Decreasing to a half the initial number of items, deleting items 11 to 20, affects mainly items 6 and 7. The estimates $\hat{\alpha}_{i,1}$, $i=6,7$, assume even larger values 16.72 and 12.65 compared to 2.79 and 5.32, respectively, with a larger standard deviations. Very small variations can be observed in the values of $\hat{\alpha}_{i,1}$ for the other items. Actually, it does not make any difference to the pattern of $\hat{\alpha}_{i,1}$'s whether $\hat{\alpha}_{7,1}$ is equal to 5.32 or 12.65, but the same is not true for item 6.

Now, consider the subset of items, 1 to 10, for which $\hat{\alpha}_{6,1}$ and $\hat{\alpha}_{7,1}$ are very large, then

(i) deleting item 7 alone, there is no significant difference between the corresponding estimates , since the largest change is the decrease of $\hat{\alpha}_{8,1}$ from 2.61 to 2.25.

(ii) deleting items 8 to 10 (Table 5.6), the parameter estimates $\hat{\alpha}_{6,1}$ and $\hat{\alpha}_{7,1}$ are still larger, but the remaining ones are closer to those from the original test than those from the subset of items 1 to 10.

(iii) deleting items 7 to 10 (Table 5.6), all $\hat{\alpha}_{i,1}$'s increase, except $\hat{\alpha}_{6,1}$, which decreases to 1.81 so that its relative importance in the test also decreases.

Table 5.6- Parameter estimates $\hat{\alpha}_{i,1}$ and asymptotic standard deviations (in brackets) from fitting a logit/probit model to Test 11A- Ireland.

i	1 to 7	1 to 6	1 to 5	2 to 6	3 to 7
1	1.39 (0.46)	2.08 (0.78)	1.96 (0.77)		
2	1.34 (0.49)	1.20 (0.50)	0.99 (0.45)	1.25 (0.55)	
3	1.81 (0.59)	2.73 (1.21)	2.24 (0.91)	2.40 (1.19)	1.31(0.45)
4	1.98 (0.78)	2.24 (1.00)	3.23 (2.19)	1.57 (1.50)	1.64(0.60)
5	1.43 (0.49)	1.51 (0.56)	1.65 (0.61)	1.46 (0.60)	1.16(0.41)
6	16.29(696.26)	1.81 (0.70)		1.82 (0.82)	16.58(755.8)
7	4.35 (1.75)				10.16(11.62)

i	11 to 20	11,13to20	11 to 15	11,13to15
11	1.43 (0.44)	1.28 (0.46)	2.15 (0.78)	1.75 (0.76)
12	16.36(617.56)		16.13(579.26)	
13	1.60 (0.57)	1.70 (0.62)	1.83 (0.69)	2.34 (1.13)
14	1.49 (0.47)	1.39 (0.50)	1.82 (0.61)	1.84 (0.81)
15	2.49 (0.92)	2.43 (0.95)	3.06 (1.23)	3.26 (2.07)
16	0.42 (0.28)	0.37 (0.29)		
17	2.18 (0.75)	2.67 (1.14)		
18	2.19 (0.69)	2.59 (1.04)		
19	1.06 (0.40)	1.16 (0.45)		
20	2.33 (0.84)	2.41 (0.90)		

Decreasing the number of items to 5 and considering the subsets of

(i) items 1 to 5, then most of the estimates $\hat{\alpha}_{i,1}$ and the corresponding standard deviations increase. Item 4 has the largest change in terms of $\hat{\alpha}_{i,1}$ (from 1.78 to 3.23) and coefficient of variation (from 0.30 to 0.68).

(ii) items 2 to 6, the largest differences are related to $\hat{\alpha}_{3,1}$, increasing and assuming the biggest value, 2.40, followed by $\hat{\alpha}_{6,1}$ decreasing to 1.82.

(iii) items 3 to 7, the estimates $\hat{\alpha}_{i,1}$ and coefficient of variations are nearly equal to those from the subtest 1 to 10, and the coefficients of variation for the remaining items are just slightly bigger than the original ones.

Decreasing to a half the original test length, deleting items 1 to 10, some estimates of $\alpha_{i,1}$ increase up to 65% ($\hat{\alpha}_{17,1}$ from 1.32 to 2.18, followed by $\hat{\alpha}_{15,1}$ from 1.59 to 2.49). The coefficients of variation do not change.

Consider the subset of items 11 to 20, in which there is one very large $\hat{\alpha}_{i,1}$ (16.36 for item 12), then

(i) deleting item 12, the estimates of $\alpha_{i,1}$ practically do not change.

(ii) deleting items 16 to 20, all $\hat{\alpha}_{i,1}$ increase, except $\hat{\alpha}_{12,1}$, which was already very large.

Furthermore, deleting item 12 from this set of items (11 to 15) then $\hat{\alpha}_{13,1}$ and the standard deviation of $\hat{\alpha}_{15,1}$ increase, while $\hat{\alpha}_{11,1}$ decreases.

As we have already pointed out in Chapter 2, when one of the $\hat{\alpha}_{i,1}$'s is very large, if we scale the individuals using the component scores then we can see 2 distinct groups: first, individuals who answered item i negatively, followed by those who answered positively. How large $\hat{\alpha}_{i,1}$ must be in order to produce two groups depends on the size of the remaining estimates $\hat{\alpha}_{i,1}$.

For example, for the subset of items 1 to 5, the largest $\hat{\alpha}_{i,1}$ is 3.25, but without distributing the individuals into 2 groups. At the same time, a maximum $\hat{\alpha}_{i,1}$ equal to 3.40 for the Lombard and Doering data produced 2 groups (Chapter 3).

However, for a pattern of $\hat{\alpha}_{i,1}$ as given by Test 11A, in which there is 2 large estimates (5.32 and 16.24), while the remaining are smaller than 2.80, we could expect a partition of the individuals into at least 2 groups. Actually, in this case, there are clearly 3 groups: first of those who answered wrong both items, followed by those who answered one item right and the other wrong, and a third of those who answered both right.

In the following we discuss the main results from the correlation matrix for the component scores (Table 5.7). This will be integrated with the previous analysis of the effect on $\hat{\alpha}_{i,1}$ when deleting one or more items, in especial when these items have a large estimates.

Table 5.7 - Correlations between the component scores from fitting a logit/probit model to different subsets of items from Test 11A-Ireland.

Items	1 to 20	del 7,12	1 to 10	1to6,8to10	1 to 7
del 7,12	0.98				
1 to 10	0.90	0.85			
1to6,8to10	0.89	0.85	0.97		
1 to 7	0.89	0.87	0.94	0.92	
1 to 6	0.84	0.85	0.81	0.83	0.91
1 to 5	0.78	0.81	0.73	0.76	0.84
2 to 6	0.85	0.85	0.86	0.89	0.93
3 to 7	0.89	0.86	0.94	0.91	0.99

Items	1 to 20	del 7,12	11 to 20	11,13to20	11to15
11 to 20	0.85	0.88			
11,13to20	0.80	0.85	0.98		
11 to 15	0.70	0.72	0.78	0.71	
11,13to15	0.60	0.62	0.69	0.66	0.93

Deleting items 7 and 12 from Test 11A, in which they assume large $\hat{\alpha}_{i,1}$, practically does not affect the pattern of $\hat{\alpha}_{i,1}$'s. The Spearman correlation between the component scores is very high (0.98). This means that the order of the component scores is nearly the same, whether items 7 and 12 are present or not, although we cannot see any groups as before.

The same result is also observed when deleting item 7 from the subset of items 1 to 10, and when deleting item 12 from the subset of items 11 to 20, in terms of high Spearman correlation between component scores and unchanged estimates $\hat{\alpha}_{i,1}$.

When deleting item 7 from the subset of items 1 to 7, the pattern of $\hat{\alpha}_{i,1}$'s changes slightly and this might be responsible for the lower (0.91) correlation between the component scores. The same kind of result is observed when deleting item 12 from the subset of items 11 to 15, for a Spearman correlation equal to 0.93.

Table 5.7 shows that the Spearman correlation between the component scores from the whole set of items and subsets of 10 or 5 items can be equal or smaller depending on which items are included in the subtest. For example, it is equal to 0.90 for a test length of ten items (i=1 to 10) and 0.89 for five items (i=3 to 7), while it decreases to 0.78 for the subset of items 1 to 5.

Both the pattern of $\hat{\alpha}_{i,1}$'s and the order of the component scores in the subset of items 3 to 7 are closer to those from the original test than for items 1 to 5.

When considering a test length of 5 items, in the first part of Table 5.7, we can see that the correlation is higher when item 7 is included in the subset.

Therefore if we delete only item 7 or 12 from the original test or from the subset of items 1 to 10 or 11 to 20, for which $\hat{\alpha}_{i,1}$, i=7 or 12, is large, the Spearman correlation of the component scores is high (0.97).

At the same time the Spearman correlation between the component scores from the original test and smaller set of items with equal length (5 items) is bigger when items 7 or 12 are included in the subset.

These two results are highly correlated with the pattern of $\hat{\alpha}_{i,1}$'s (Tables 5.5 and 5.6), since for a small test length (5 items) if items 7 or 12 are present then the pattern of $\hat{\alpha}_{i,1}$ is closer to the original ones; and when deleting just items 7 or 12 from sets of 20 or 10 items, the new estimates are also very similar to the original ones.

Therefore at the same time that there is evidence that an item with large $\hat{\alpha}_{i,1}$ does not give any extra information about the latent variable, when the number of items is small(5), it contains more information than some other items, in terms of producing higher correlation between the component scores.

2.2- Wales

Table 5.8- Parameter estimates $\hat{\alpha}_{i,1}$ and asymptotic standard deviations(in brackets) from fitting a logit/probit model to Test 11A- Wales.

i	1 to 20	delet.7,12	1 to 10	1to6,8to10
1	1.04 (0.34)	1.10 (0.36)	0.86 (0.33)	0.96 (0.36)
2	1.20 (0.48)	1.39 (0.55)	1.10 (0.50)	1.10 (0.52)
3	1.24 (0.41)	1.26 (0.43)	1.20 (0.42)	1.24 (0.47)
4	0.97 (0.33)	1.08 (0.36)	0.85 (0.33)	0.98 (0.38)
5	1.20 (0.36)	1.11 (0.36)	0.93 (0.34)	0.87 (0.34)
6	2.17 (0.62)	2.10 (0.63)	3.72 (1.43)	4.41 (2.75)
7	2.73 (0.80)		4.59 (2.42)	
8	1.50 (0.43)	1.35 (0.41)	1.91 (0.58)	1.90 (0.64)
9	1.61 (0.47)	1.40 (0.44)	1.56 (0.50)	1.43 (0.48)
10	1.50 (0.43)	1.48 (0.44)	1.49 (0.46)	1.34 (0.45)
11	0.99 (0.33)	0.98 (0.33)		
12	1.90 (0.52)			
13	1.87 (0.62)	1.69 (0.58)		
14	1.20 (0.38)	1.16 (0.39)		
15	1.90 (0.69)	2.12 (0.82)		
16	0.25 (0.25)	0.20 (0.25)		
17	1.31 (0.40)	1.38 (0.42)		
18	1.71 (0.47)	1.89 (0.54)		
19	0.93 (0.33)	0.93 (0.33)		
20	1.80 (0.50)	1.85 (0.54)		

When fitting a logit/probit model to Test 11A for Wales' data, $\hat{\alpha}_{i1}$ assumes values from 0.25 to 2.73 with larger estimates having a larger standard deviations, except for $\hat{\alpha}_{16,1}$ (the smallest estimate with the largest coefficient of variation). Items 6 and 7 have the largest estimates $\hat{\alpha}_{i,1}$ (2.17 and 2.73). The maximum correlation between $\alpha_{i,1}$

and $\alpha_{j,1}$, $i \neq j$, is 0.18 for items 6 and 7.

Deleting items 7 and 12, both, the estimates of $\alpha_{i,1}$ and their coefficients of variation are very stable, that is, nearly equal to the original ones.

Decreasing to a half the number of items, and considering items 1 to 10, seven estimates decrease slightly (up to 20%), while $\hat{\alpha}_{6,1}$ and $\hat{\alpha}_{7,1}$ increase to 3.72 and 4.59, respectively.

If from the subset of items 1 to 10, we delete

(i) item 7, the strongest change is for item 6, for which $\hat{\alpha}_{6,1}$ increases to 4.41 with higher coefficient of variation, although in practice this difference is not significant.

(ii) items 8 to 10, the estimates of $\alpha_{6,1}$ and $\alpha_{7,1}$ decrease to 3.21 and 3.34, while the remaining ones do not change significantly, as in the previous subsets of items.

(iii) items 7 to 10, the parameter estimates $\hat{\alpha}_{i,1}$, i=1 to 4, increase while $\hat{\alpha}_{6,1}$ decreases to 1.13, so that its relative importance in this subtest also decreases.

The effect on the pattern of $\hat{\alpha}_{i,1}$'s of deleting item 7 from the subset of items 1 to 10 and from 1 to 7 is very different. While in the former it does not affect the pattern of $\hat{\alpha}_{i,1}$ estimates, in the later it does strongly, even changing the relative importance of some items.

Table 5.9- Parameter estimates $\hat{\alpha}_{i,1}$ and asymptotic standard deviations (in brackets) from fitting a logit/probit model to Test 11A- Wales.

i	1 to 7	1 to 6	1 to 5	2 to 6	3 to 7
1	1.10 (0.40)	1.70 90.65)	1.52 (0.64)		
2	1.35 (0.62)	2.07 (1.12)	2.23 (1.44)	1.73 (1.08)	
3	1.31 (0.48)	2.58 (1.39)	4.38 (5.50)	1.97 (1.22)	0.96 (0.40)
4	0.88 (0.36)	1.05 (0.43)	0.95 (0.41)	1.17 (0.52)	0.82 (0.36)
5	0.94 (0.38)	0.88 (0.37)	0.74 (0.35)	1.00 (0.46)	1.05 (0.38)
6	3.21 (1.43)	1.13 (0.47)		1.29 (0.63)	3.39 (1.70)
7	3.34 (1.56)				4.76 (3.97)

i	11 to 20	11,12 to 20	11 to 15	11,13 to 15
11	1.18 (0.40)	1.07 (0.39)	0.99 (0.38)	0.64 (0.55)
12	2.61 (0.95)		16.03 (82.91)	
13	1.61 (0.62)	1.34 (0.56)	1.73 (0.68)	1.62 (1.46)
14	1.13 (0.42)	1.01 (0.41)	1.08 (0.42)	0.90 (0.67)
15	2.37 (1.06)	2.95 (1.63)	1.47 (0.62)	1.87 (1.76)
16	0.46 (0.28)	0.47 (0.29)		
17	1.40 (0.47)	1.47 (0.53)		
18	1.56 (0.50)	1.79 (0.64)		
19	1.17 (0.40)	1.07 (0.40)		
20	2.15 (0.72)	2.24 (0.84)		

Decreasing the test length to 5 items, and considering the subsets of

(i) items 1 to 5

The parameter estimate $\hat{\alpha}_{4,1}$ remains the same, $\hat{\alpha}_{5,1}$ decreases while the remaining ones increase, especially $\hat{\alpha}_{3,1}$ from 1.24 to 4.38 with a larger coefficient of variation.

(ii) items 2 to 6

The new estimates of $\alpha_{i,1}$ are slightly larger that the original ones and those from the subtest 1 to 10, except for $\hat{\alpha}_{6,1}$ which decreases to 1.29 also altering its relative importance in this test.

(iii) items 3 to 7

The estimates $\hat{\alpha}_{i,1}$ are very close to those from the subset of items 1 to 10 , what means a significant increase on $\hat{\alpha}_{i,1}$, for i=6 and 7, values when compared to the original ones. Therefore, in this case, the effect on $\hat{\alpha}_{i,1}$ of decreasing the test length to 10 or 5 items is the same for the subset of items 1 to 10 and 3 to 7.

Decreasing the number of items to 10, considering items 11 to 20, most of the $\hat{\alpha}_{i,1}$ slightly change in both directions. The strongest change is an increase for $\hat{\alpha}_{12,1}$ and $\hat{\alpha}_{15,1}$ both from 1.90 to 2.61 and 2.37.

Consider the subset of items 11 to 20, then deleting

(i) item 12

The parameter estimate $\hat{\alpha}_{15,1}$ increases to 2.95 with a larger coefficient of variation, while for some items, the estimates increase or decrease slightly.

(ii) item 16 to 20

The strongest change is for items 12 and 15, since $\hat{\alpha}_{12,1}$ assumes a very large value (16.03) and $\hat{\alpha}_{15,1}$ decreases to 1.47, being even

smaller than the original one (1.90). The remaining estimates of the discrimination parameter are stable.

Furthermore, if item 12 is deleted from this subset (11 to 15) then the estimate $\hat{\alpha}_{15,1}$ is equal to the original one, while the remaining ones are slightly smaller, although all coefficients of variation are bigger.

In the following, Table 5.10 displays the Spearman correlations between the component scores from fitting a logit/probit model to subsets of items from Test 11A.

Table 5.10- Correlations between the component scores from fitting a logit/probit model to different subsets of items from Test 11A- Wales.

Items	1 to 20	del 7,12	1 to 10	1to6,8to10	1 to 7
del 7,12	0.98				
1 to 10	0.91	0.85			
1to6,8to10	0.90	0.88	0.96		
1 to 7	0.87	0.81	0.95	0.90	
1 to 6	0.76	0.78	0.76	0.81	0.86
1 to 5	0.68	0.72	0.66	0.71	0.76
2 to 6	0.78	0.79	0.80	0.84	0.86
3 to 7	0.85	0.77	0.93	0.87	0.95

Items	1 to 20	del 7,12	11 to 20	11,13to20	11 to 15
11 to 20	0.87	0.90			
11,13to20	0.86	0.89	0.96		
11 to 15	0.76	0.77	0.87	0.75	
11,13to15	0.72	0.72	0.78	0.75	0.89

Although the sample size of Wales is of the same order as for Ireland (86 compared to 73), none of its parameter estimates $\hat{\alpha}_{i,1}$ in the original test assumes large values.

On the other hand, items 6 and 7 have the biggest $\hat{\alpha}_{i,1}$ (2.17 and 2.73), and the effect on the pattern of $\hat{\alpha}_{i,1}$ of deleting items is very similar to those observed for Ireland. The Spearman correlation matrix for the component scores for Wales is very close to that for Ireland. As the same results obtained there are also valid here, and we will not repeat them.

The similarity between the correlations matrix of the component scores for Ireland and Wales is likely to be due to the fact that the order of the component scores is highly correlated with the effect on $\hat{\alpha}_{i,1}$ of deleting items, since both countries present approximately the same kind of change in the pattern of $\hat{\alpha}_{i,1}$.

2.3- England

In the following, we consider the same subset of items from Test 11A for England as for Wales and Ireland, although in this case the three largest $\hat{\alpha}_{i,1}$ values are not associated to the same items as before.

Table 5.11- Parameter estimates $\hat{\alpha}_{i,1}$ and asymptotic standard deviations(in brackets) from fitting a logit/probit model to Test 11A– England.

i	1 to 20	delet.7,12	1 to 10	1to6,8to10
1	0.84 (0.16)	0.84 (0.16)	0.70 (0.15)	0.71 (0.16)
2	1.41 (0.23)	1.45 (0.24)	1.37 (0.25)	1.50 (0.27)
3	1.48 (0.24)	1.47 (0.24)	1.18 (0.22)	1.32 (0.25)
4	1.95 (0.28)	1.95 (0.29)	2.28 (0.36)	2.69 (0.48)
5	1.17 (0.18)	1.17 (0.18)	1.09 (0.18)	1.27 (0.21)
6	2.32 (0.33)	2.02 (0.19)	3.95 (0.78)	3.09 (0.60)
7	2.34 (0.34)		3.83 (0.76)	
8	1.26 (0.19)	1.14 (0.18)	1.53 (0.24)	1.31 (0.22)
9	1.04 (0.18)	0.97 (0.17)	1.30 (0.22)	1.16 (0.21)
10	2.57 (0.37)	2.43 (0.36)	1.95 (0.29)	1.99 (0.32)
11	1.72 (0.24)	1.68 (0.24)		
12	2.03 (0.29)			
13	2.46 (0.39)	2.46 (0.40)		
14	1.35 (0.20)	1.37 (0.21)		
15	3.15 (0.54)	3.44 (0.64)		
16	1.31 (0.21)	1.32 (0.22)		
17	1.13 (0.18)	1.18 (0.19)		
18	1.60 (0.23)	1.67 (0.24)		
19	0.98 (0.17)	1.08 (0.18)		
20	1.47 (0.21)	1.56 (0.23)		

In the whole set of items, the parameter estimate $\hat{\alpha}_{i,1}$ assumes values from 0.84 to 3.15 with about the same coefficient of variation. Items 6, 7 and 12 are among the six for which $\hat{\alpha}_{i,1}$ is bigger than 2.0, although the largest estimate is $\hat{\alpha}_{15,1}$. The maximum correlation between $\alpha_{i,1}$ and $\alpha_{j,1}$, $i \neq j$, is equal to 0.25 for items 6 and 7.

Deleting items 7 and 12, the pattern of $\hat{\alpha}_{i,1}$'s remains practically the same, since the increase is up to 9% for $\hat{\alpha}_{15,1}$ (3.15) and the decrease is 13% for $\hat{\alpha}_{6,1}$ (2.02).

Decreasing to a half the initial number of items, deleting items 11 to 20, affects mainly items 6, 7 and 10. The estimates of $\alpha_{i,1}$, i=6,7, increase to 3.95 and 3.83, while $\hat{\alpha}_{10,1}$ decreases to 1.95 (24% smaller). It is interesting to observe that $\hat{\alpha}_{10,1}$ is the second largest estimate in the original test.

Table 5.12- Parameter estimates $\hat{\alpha}_{i,1}$ and asymptotic standard deviations (in brackets) from fitting a logit/probit model to Test 11A- England.

i	1 to 7	1 to 6	1 to 5	2 to 6	3 to 7
1	0.78 (0.17)	0.79 (0.13)	0.83 (0.19)		
2	1.56 (0.29)	1.79 (0.34)	1.82 (0.38)	1.84 (0.35)	
3	1.36 (0.26)	1.52 (0.32)	1.34 (0.30)	1.46 (0.31)	1.30 (0.25)
4	2.48 (0.44)	3.22 (0.77)	2.67 (0.68)	3.64 (1.03)	2.33 (0.41)
5	1.18 (0.21)	1.47 (0.27)	1.82 (0.40)	1.37 (0.25)	1.06 (0.19)
6	4.76 (1.41)	2.37 (0.46)		2.38 (0.47)	6.14 (2.68)
7	2.90 (0.54)				3.10 (0.59)

continue...

i	11 to 20	11,13 to 20	11 to 15	11,13 to 15
11	1.95 (0.31)	1.68 (0.28)	2.04 (0.36)	1.77 (0.36)
12	1.91 (0.31)		2.70 (0.57)	
13	2.92 (0.53)	2.97 (0.56)	2.87 (0.58)	2.81 (0.61)
14	1.54 (0.24)	1.46 (0.24)	1.63 (0.28)	1.60 (0.30)
15	3.98 (0.88)	4.36 (1.09)	3.97 (1.06)	7.12 (4.40)
16	1.28 (0.23)	1.35 (0.25)		
17	1.12 (0.19)	1.16 (0.20)		
18	1.80 (0.28)	1.88 (0.30)		
19	1.04 (0.18)	1.08 (0.19)		
20	1.78 (0.28)	1.88 (0.30)		

If from the subset of items 1 to 10, we delete

(i) item 7, then occurs same slightly changes in both directions. The main changes are a larger $\hat{\alpha}_{4,1}$ (2.69) and smaller $\hat{\alpha}_{6,1}$ (3.09), which is still bigger than the original one.

(ii) items 8 to 10, then the parameter estimates $\hat{\alpha}_{4,1}$ and $\hat{\alpha}_{6,1}$ assume larger values, 2.48 and 4.76, while $\hat{\alpha}_{7,1}$ decreases to 2.90, which is still bigger than the original one.

(iii) items 7 to 10, the estimate $\hat{\alpha}_{4,1}$ increases and assumes the largest value in this set of items (3.22), while $\hat{\alpha}_{6,1}$ decreases to 2.37 and becomes nearly equal to the original one. These changes make the relative importance of these items in this set also change.

Decreasing the test length to 5 items, and considering the subset of

(i) items 1 to 5

Except for $\hat{\alpha}_{1,1}$, the remaining estimates increase up to 36%, for which $\hat{\alpha}_{4,1}$ is the one that has the highest increase (from 1.95 to 2.67).

(ii) items 2 to 6

As in the former set of items, the maximum increase is for $\hat{\alpha}_{4,1}$, which assumes a larger value (3.64). The estimate $\hat{\alpha}_{6,1}$ (2.38) is very close to the original one.

(iii) items 3 to 7

The parameter estimates $\hat{\alpha}_{i,1}$, i=3,5, are slightly smaller while the remaining estimates are bigger than the original ones. The strongest increase is for $\hat{\alpha}_{6,1}$, which assumes very large value (6.14) with a larger coefficient of variation.

Decreasing the number of items to 10, considering items 11 to 20, then most of the $\hat{\alpha}_{i,1}$'s change in both directions. The strongest change is for $\hat{\alpha}_{15,1}$, which increases from 3.15 to 3.98 keeping about the same coefficient of variation.

Considering the subset of items 11 to 20, then deleting

(i) item 12

All estimates practically do not change. The parameter estimate $\hat{\alpha}_{15,1}$ was already large, 3.98, so that an increase to 4.36 is not significant.

(ii) items 11 to 15

All estimates of $\alpha_{i,1}$ increase up to 33% in relation to the original ones and the set of items 11 to 20, for about the same coefficient of variation. The strongest increase is for $\hat{\alpha}_{12,1}$ (from 2.03 to 2.70) followed by $\hat{\alpha}_{15,1}$ (from 3.15 to 3.97).

Going further, deleting item 12 from this set, then $\hat{\alpha}_{15,1}$ becomes very large (7.12) with a larger coefficient of variation. The remaining estimates of $\alpha_{i,1}$ are very close to those from the subset of items 11 to 20.

In the following, we analyse the Spearman correlation between component scores from fitting a logit/probit model to some subset of items from Test 11A-England.

Table 5.13 - Correlations between the component scores of fitting a logit/probit model to different subsets of items from Test 11A-England.

Items	1 to 20	del 7,12	1 to 10	1to6,8to10	1 to 7
del 7,12	0.98				
1 to 10	0.88	0.84			
1to6,8to10	0.89	0.88	0.96		
1 to 7	0.83	0.79	0.92	0.91	
1 to 6	0.77	0.78	0.83	0.90	0.94
1 to 5	0.71	0.73	0.74	0.83	0.84
2 to 6	0.75	0.75	0.80	0.88	0.91
3 to 7	0.81	0.77	0.90	0.89	0.98

Items	1 to 20	del 7,12	11 to 20	11,13to20	11 to 15
11 to 20	0.89	0.89			
11,13to20	0.85	0.89	0.97		
11 to 15	0.78	0.73	0.86	0.75	
11,13to15	0.72	0.73	0.82	0.80	0.89

Deleting items 7 and 12 from Test 11A, for which $\hat{\alpha}_{i,1}$ is equal to 2.34 and 2.03, practically does not affect the pattern of $\hat{\alpha}_{i,1}$'s. The Spearman correlation between the component scores is high (0.98), which means that the order of the component scores is nearly equal in both tests.

Two other high correlations for the component scores, 0.96 and 0.97, are observed between the set of items 1 to 10 and 1 to 6, 8 to 10; and when deleting item 12 from the subset of items 11 to 20.

Nevertheless, in this case, the estimate $\hat{\alpha}_{7,1}$ is equal to 3.83, but $\hat{\alpha}_{12,1}$ is not so large (1.93).

The three previous high Spearman correlations for the component scores agree with the similarity of the pattern of $\hat{\alpha}_{i,1}$'s between the corresponding subset of items.

When deleting item 7 from the subset of items 1 to 7, some changes occur in the values of $\hat{\alpha}_{i,1}$, for items 4 and 6, which alter their relative importance in the calculation of the component scores. These changes might have affected the order of the new component scores, producing a smaller Spearman correlation (0.94).

The same previous changes in the relative importance of items 4 and 6, is also observed for item 11, when deleting item 12 from the subset of items 11 to 15, although now the correlation between the component scores is smaller (0.89).

The Spearman correlations between the component scores from Test 11A and those from the subset of items 1 to 5 and the subset of items 3 to 7 are equal to 0.71 and 0.74, respectively. However when Test 11A is replaced by the subset of items 1 to 10, the correlation coefficients increase to 0.81 and 0.90.

In relation to the whole test, the correlations for the component scores are approximately the same, whether items 1 and 2 or items 6 and 7 are included. However, in relation to the subset of items 1 to 10, the correlation is bigger when items 6 and 7 are included in the test instead of items 1 and 2.

In the subset of items 1 to 10, the estimates $\hat{\alpha}_{i,1}$ for items 6 and 7 assume the largest values (3.95 and 3.83), while $\hat{\alpha}_{1,1}$ and $\hat{\alpha}_{2,1}$ are equal to 0.70 and 1.37. Therefore the later higher correlation (0.90)

might be partially explained by the fact that items 6 and 7 have more weight than items 1 and 2 on the determination of the component scores for the subset of items 1 to 10.

2.4- Conclusions

The results of this extensive investigation of the effect on $\hat{\alpha}_{i,1}$ of deleting items for a variety of sample sizes lead to the following conclusions:

(1) The effect on $\hat{\alpha}_{i,1}$ of deleting items from tests with 10 or fewer items, is approximately the same for these 3 sets of data, although the sample sizes are different (73 and 86 compared with 342). The sample sizes are not so small as to make estimation of a model with 10 items unreliable.

If, on the other hand, the number of items is large, say 20 items, then increasing the sample size from 73 to 342 tends to increase the number of different score patterns with frequencies bigger than zero, with more effect on the pattern of $\hat{\alpha}_{i,1}$'s.

Therefore the magnitude of the sample size must be judged in relation to the number of items, when considering the effect of deleting items.

(2) If ones deletes only items with large $\hat{\alpha}_{i,1}$'s from tests with 20,18 or even 10 items, the $\hat{\alpha}_{j,1}$ of the remaining items show great stability. This is also apparent in the high Spearman correlations ($\geq$0.96) between the corresponding component scores before and after deletion.

(3) For a small sub-test of 5 items the pattern of $\hat{\alpha}_{i,1}$'s is closer to those of the test with all items, if the items with large $\hat{\alpha}_{i,1}$ are present in the test. This is also shown by the larger Spearman correlations between the corresponding component scores. This pattern is observed for England when considering 10 items instead of the whole test.

(4) From (2) and (3) we can conclude that an item with a large $\hat{\alpha}_{i,1}$ may not give any additional information about the latent variable in a large test length(20), but when the number of items is small(5), it may contain more information than other items.

(5) The occurrence of a large $\hat{\alpha}_{i,1}$ seems to depend more on which items are included in the test then on the sample size and test length. This result is extensively investigated in the next chapter.

(6) As the number of items decreases, the largest $\hat{\alpha}_{i,1}$ tends to increase and become very large when the test length is small.

(7) Parameter estimates $\hat{\alpha}_{i,1}(\geq 0.50)$ and standard deviations are approximately linear related so that larger estimates have larger standard deviations.

(8) Except for Test 12 (whole set of items), the occurrence of a large $\hat{\alpha}_{i,1}$ was not due to correlated estimates $\hat{\alpha}_{i,1}$ and $\hat{\alpha}_{j,1}$, $i \neq j$, since the correlations were always zero or nearly equal to zero.

(9) A large parameter estimate $\hat{\alpha}_{i,1}$ for some items does not seem to be associated with the type of question asked.

3- *Tests with Large Number of Items*

It has been pointed out in the literature that the occurrence of Heywood cases in Factor Analysis might be due to the small number of variables (items) or small sample size (Chapter 1, section 5). It is also expected that for the same sample size, if the number of items increases then the probability of the occurrence of an Heywood case decreases.

In order to investigate the relation between the number of items and the occurrence of a large $\hat{\alpha}_{i,1}$, we fitted a logit/probit model to 11 tests, used to measured the reading ability of children aged 11 in 1983, by the National Foundation for Educational Research (NFER).

Firstly, we shall discuss the main results from fitting a logit/probit model to Test 8, which has a large number of items (43), considering the whole sample size (527) and when the sample is stratified according to the country where the school is located.

Finally we present a summary of the maximum and minimum values assumed by the discrimination parameter estimates, the asymptotic standard deviations for all tests, relating them to the sample size and test length.

Test 8

The fitting of a logit/probit model to Test 8 yields parameter estimates $\hat{\alpha}_{i,1}$ from 0.08 to 2.65, among which 39 estimates are smaller than 2.0. Question 42 has the biggest $\hat{\alpha}_{i,1}$ (2.65) with standard deviation equal to 0.43. A larger $\hat{\alpha}_{i,1}$ has a larger standard deviation, except for estimates smaller than 0.30.

We also fitted a logit/probit model when the sample was stratified according to the location of the school, that is, England (365), Wales (86) and Ireland (76), for which the main results are displayed in Table 14. As for Test 11A, the England sample size is larger than for Wales and Ireland, and we are interested in finding out whether with a large number of items, a large $\hat{\alpha}_{i,1}$ will still occur.

Table 5.14- Frequency distribution of the parameter estimates $\hat{\alpha}_{i,1}$ from fitting a logit/probit model to Test 8, according to the location of the school.

	Ireland	Wales	England
$0.03 \leq \hat{\alpha}_{i,1} \leq 2.0$	34	38	39
$2.00 < \hat{\alpha}_{i,1} \leq 3.0$	5	2	1
$\hat{\alpha}_{i,1} > 3.0$	4	3	3

In the following we present the items for which the discrimination parameter estimates are bigger than 3.0 with the corresponding standard deviations in brackets.

Items	Ireland	Wales	England
4		3.43 (0.83)	
7	3.44 (2.13)	3.09 (0.79)	
8		3.80 (1.01)	
9	13.42 (857.39)		
40	16.90 (1330.41)		3.17 (0.69)
41			3.51 (0.81)
42	3.14 (2.09)		4.06 (1.20)

Therefore, although Wales and Ireland have nearly equal sample size, when fitting a logit/probit model the later yielded 2 very large $\hat{\alpha}_{i,1}$'s. Moreover the parameter estimates bigger than 2.0 from Ireland's data have larger coefficient of variation than those from Wales' data. On the other hand the pattern of $\hat{\alpha}_{i,1}$'s and their coefficients of variation for Wales are very similar to those for England.

We have already given two examples with different test lengths, Test 11A and Test 8, in which Ireland and Wales have about the same sample size, but the former produced two very large $\hat{\alpha}_{i,1}$'s. It seems that there is something different about the distribution and nature (configuration) of the Ireland score pattern that produces very large estimates $\hat{\alpha}_{i,1}$ even when the test length is 43 items.

It is convenient to point out that so far we have not obtained more than 2 large estimates $\hat{\alpha}_{i,1}$'s for the same test.

Analysis of the items in Test 8

In order to find out whether there was something special about the items with large $\hat{\alpha}_{i,1}$, we have studied carefully all the questions. In the following we discuss those questions for which $\hat{\alpha}_{i,1}$ is large.

Test 8 corresponds to the booklet on the topic of 'Space', which was organised into 5 sections and one index, one for each page.

For all 43 questions there was a choice of a single answer among 2, 5 or 6 options. Most of the questions were asked on specific pages (or sections), but there were also some questions that involved children refering from one passage to another, to locate,interpret and relate information given.

In the second section of Test 8, a story entitle 'Target Mars' about a space landing on a planet was described in a sequence of 6 boxes. Items 4 to 8 correspond to sentences about what was happening in each box. For each one of these items, the children were asked to choose a number between 1 and 6, which corresponds to the box in which the sentence occurred.

Items 4,7,8 for Wales and item 7 for Ireland have $\hat{\alpha}_{i,1}$ bigger than 3.0. However we have not found anything special about these items that would make them different from items 5 and 6.

The following three items (9 to 11) were about the same passage, but refering to the whole context of the story. For Ireland's data, item 9 has a very large $\hat{\alpha}_{9,1}$ (13.42), and was asking on which planet the spaceship was landing. Since for item 9 as well as for 10 and 11 the questions were very clear and the answers were written in the

booklet, the large estimate does not seem to be related with the kind of question.

Finally, to answer the last four questions (40 to 43), the children had to use an index in order to indicate which page contained information about Pluto, Marty, Mercury and the solar system, respectively. The index diplayed the names exactly as in the questions, followed by the page number.

Items 40, 41 and 42 have $\hat{\alpha}_{i,}$, bigger than3.0 for England, while for Ireland items 40 and 42 have $\hat{\alpha}_{i,}$, equal to 16.90 and 3.14, respectively. As in the preceding analysis it seems that there is nothing special about these questions that make them different from answering about Mercury (item 43).

Eleven tests with different tests length and sample sizes

In Table 5.15, we present some results from fitting a logit/probit model to the eleven tests, used by the NFER in 1983, to measure the reading ability of children aged 11.

Table 5.15- Parameter estimates and asymptotic standard deviations (in brackets) from fitting a logit/probit model to Reading Ability Tests (NFER data).

test	n.i.	s.s	$\min(\hat{\alpha}_{i,1})$		$\max(\hat{\alpha}_{i,1})$		$\min(\hat{\pi}_i)$	$\max(\hat{\pi}_i)$
13A	24*	440	0.29	(0.11)	2.01	(0.28)	0.42	0.97
4	25	236	0.20	(0.15)	2.39	(0.44)	0.14	0.95
11	29*	270	0.33	(0.19)	2.50	(0.57)	0.25	0.99
11A	20	501	0.83	(0.13)	2.62	(0.36)	0.27	0.97
8	43	497	0.08	(0.09)	2.65	(0.43)	0.39	0.99
3	42	304	0.25	(0.14)	2.93	(0.56)	0.23	0.99
5	39	498	0.94	(0.14)	3.23	(0.52)	0.41	0.99
6	55	507	0.61	(0.10)	4.02	(0.56)	0.09	0.97
12	18	502	0.62	(0.12)	4.50	(0.83)	0.15	0.98
13	40*	498	0.31	(0.10)	4.70	(0.71)	0.12	0.97
1	37	495	0.26	(0.09)	5.25	(0.97)	0.17	0.99

n.i.:number of items s.s.:sample size

* :test length after deleting one item, for which all children answered it right or wrong.

Complementing the information given by Table 5.15, we observed that the largest coefficients of variation of $\hat{\alpha}_{i,1}$ were always associated with estimates smaller than 0.50, while for the remaining $\hat{\alpha}_{i,1}$, they were approximately the same, especially for samples bigger than 400.

Generally, items with the largest $\hat{\alpha}_{i,1}$ have also the largest $\hat{\pi}_i$, although $\hat{\pi}_i$ bigger than 0.90 were also associated with $\hat{\alpha}_{i,1} \geq 0.73$.

Table 5.15 shows that among the six tests for which the maximum $\hat{\alpha}_{i,1}$ is smaller than 3.0, four correspond to tests with the smallest sample sizes and different lengths.

In tests four and eleven the maximum $\hat{\alpha}_{i,1}$ are equal to 2.39 and 2.50 for about the same number of items (25 and 29) and sample size (236 and 270). Test three, length 42 items and sample size 304, have maximum $\hat{\alpha}_{i,1}$ equal to 2.93. Test 11A has the smallest maximum (2.01) for a length of 24 items and 440 observations.

From Table 5.15 we can see that seven tests have sample sizes of order 500, while the remaining ones are smaller with a minimum of 236 observations. The five tests, for which at least one parameter estimate $\hat{\alpha}_{i,1}$ is bigger than 3.0, have the largest sample size.

Tests thirteen, five and one have 40, 39 and 37 items, from which 6,1 and 5 items, respectively, have $\hat{\alpha}_{i,1} \succ 3.0$. Tests six and twelve have 55 and 18 items with 3 and 2 items for which $\hat{\alpha}_{i,1}$ is bigger than 3.0, respectively.

Fitting a logit/probit model to tests of length 20 or fewer items (tests 11A, 12 and subtests analysed in the previous sections), there was a maximum of 2 items for which $\hat{\alpha}_{i,1}$ was bigger than 3.0. However, for the same sample size, but double test length, Table 5.15 shows tests with 0 to 6 estimates $\hat{\alpha}_{i,1}$ bigger than 3.0. Tests like test thirteen, in which 6 out of 40 items have $\hat{\alpha}_{i,1} \succ 3.0$, are probably not desirable.

Consider again tests with sample size of order 500. The comparison between the test length and the maximum value assumed by $\hat{\alpha}_{i,1}$ shows that

(1) For two tests, 11A and 12, with twenty and eighteen items, the maximum values assumed by $\hat{\alpha}_{i,1}$ are 2.62 and 4.50 for approximately the same coefficient of variation (0.14 and 0.18).

(2) For the four tests with length of about 40 items (tests 8, 5, 13 and 1), the maximum $\hat{\alpha}_{i,1}$'s are equal to 2.65, 3.23, 4.70 and 5.25, respectively, for about the same coefficient of variation (about 0.16).

(3) For Test 6 with length of 55 items, the maximum $\hat{\alpha}_{i,1}$ is equal to 4.02 with coefficient of variation 0.14.

Therefore there was not any improvement in terms of decreasing the occurrences of large $\hat{\alpha}_{i,1}$ (>3.0), increasing the test length from 20 to 40 or more items, for sample size of order 500.

These results seems to give evidence that a test with 20 items is not more likely to have at least one $\hat{\alpha}_{i,1}$ bigger than 3.0 than a test with double length, for sample size of order 500.

Chapter 6

AN INVESTIGATION of the CONDITIONS giving rise to LARGE $\hat{\alpha}_{i,1}$

In order to understand the circumstances when large discrimination parameter estimates $\hat{\alpha}_{i,1}$'s arise, we shall first study the configuration of the score patterns and its relation to the size of $\hat{\alpha}_{i,1}$. We start from a simple data set with two items, sample size two and score patterns '10' and '01' each with frequency 1.

1- *Introductory Examples*

1.2- Two Variables: a Theoretical Result

Suppose that a test with two items is answered by two individuals, for whom the score patterns are '10' and '01'.

Then the likelihood function for the probit model may be written

$$L = \int_{-\infty}^{\infty} \pi_1(z)\ [1-\pi_2(z)]\ d\Phi(z) \qquad \int_{-\infty}^{\infty} \pi_2(z)\ [1-\pi_1(z)]\ d\Phi(z).$$

More generally, with expected values taken over some unspecified latent variable distribution

$$L = E\left[\pi_1(1-\pi_2)\right]\ E\left[\pi_2(1-\pi_1)\right].$$

Now,

$$(1-\pi_2)(1-\pi_2) = \frac{1}{2} + \frac{1}{2}(\pi_1-\pi_2) - \frac{1}{2}\pi_1\,\pi_2 - \frac{1}{2}(1-\pi_1)$$

Now,

$$\pi_1(1-\pi_2) = \frac{1}{2} + \frac{1}{2}(\pi_1-\pi_2) - \frac{1}{2}\pi_1\pi_2 - \frac{1}{2}(1-\pi_1)(1-\pi_2)$$

$$\leq \frac{1}{2} + \frac{1}{2}(\pi_1-\pi_2).$$

Similarly,

$$\pi_2(1-\pi_1) \leq \frac{1}{2} + \frac{1}{2}(\pi_2-\pi_1).$$

Thus,

$$E[\pi_1(1-\pi_2)]\ E[(1-\pi_1)\pi_2] \leq$$

$$\leq \frac{1}{2}[1-E(\pi_2-\pi_1)]\ \frac{1}{2}[1+E(\pi_2-\pi_1)]$$

$$= \frac{1}{4}\left\{1-(E(\pi_2-\pi_1))^2\right\} \leq \frac{1}{4}$$

But the upper bound is achieved iff

$$E(\pi_1\pi_2) = E[(1-\pi_1)(1-\pi_2)] = 0 \quad \text{and} \quad E(\pi_2-\pi_1) = 0,$$

that is, the upper bound is achieved iff

$$E(\pi_1) = E(\pi_2) = \frac{1}{2} \qquad \text{and} \qquad E(\pi_1\pi_2) = 0,$$

that is, when the first and second order margins match the observed data.

$$E(\pi_1\pi_2)=0 \iff (\pi_1>0 \Rightarrow \pi_2=0) \text{ and } (\pi_2>0 \Rightarrow \pi_1=0).$$

Therefore there are many possible ways of choosing a response function to maximise the likelihood, but within logit, probit and logit/probit models, they must be threshold models, eg,

$$\pi_1(z)=0, \quad z < 0 \qquad\qquad \pi_2(z)=1, \quad z < 0$$

and

$$\pi_1(z)=1, \quad z \geq 0 \qquad\qquad \pi_2(z)=0, \quad z \geq 0$$

This is a response function with a threshold at z=0, and it can be obtained as a limiting case of the general model(1.5), Chapter 1, as α_1 tends to infinity with α_0/α_1 fixed at zero.

Although this example involves only two variables, it shows rigorously that a threshold model may be the MLE.

1.2- Three Variables: Simulated Data

Due to the complexity of the algebra for even 3 variables, we shall study the relation between the size of $\hat{\alpha}_{i,1}$ and the configuration of the score patterns through some examples. We start from a simple and artificial data set with 3 items, which have the same first and second order margins, and sample size 240. Then, when fitting a logit/probit model to this data, the possible differences between the size of $\hat{\alpha}_{i,1}$'s will be due to the effect of the third order iterations.

Table 6.1 displays the frequencies of the score patterns and ratios of the conditional frequencies, which were obtained as described below.

As the frequencies of the score patterns for 231 and 312 are the same, only one of them is displayed.

Table 6.1- Distribution of the score patterns and ratios of the conditional frequencies.

123	n	ratio	231	n	ratio	logit/probit
000	70		000	70		$\hat{\alpha}_{1,,1}$= 9.41
100	10	(7.0)	100	20	(3.5)	SD($\hat{\alpha}_{1,,1}$)= 12.50
001	20		001	10		
101	20	(1.0)	101	20	(0.5)	
010	20		010	20		$\hat{\alpha}_{i,,1}$= 1.69, i=2,3
110	20	(1.0)	110	10	(2.0)	SD($\hat{\alpha}_{i,,1}$)= 0.36
011	10		011	20		
111	70	(0.1)	111	70	(0.3)	

Consider the score patterns for items 2 and 3, first when item one is answered negatively (0), and second when it is answered positively (1). We can calculate the ratios of the frequencies of the corresponding score patterns in these two cases. We call these ratios as 'ratios for item 1'.

For instance, when x_1=0 the frequency of x_2=0 and x_3=0 is 70 and when x_1=1 the corresponding frequency is 10, giving a ratio equal to 7.0.

We repeat this procedure for items 2 and 3.

Then the distribution of ratios for every item is split into two parts: ratios smaller than 1 and ratios equal to or bigger than 1. The analysis of the distribution is done comparing the values assumed by the ratios on the left with those on the right side. The closer to zero the smallest ratio and the bigger the largest one for an item, the more extreme is the distribution of the ratios for this item.

Finally we compare the patterns of ratios for all items with the size of $\hat{\alpha}_{i,,1}$'s.

Now we return to Table 6.1 in order to analyse the main results.

If we compare the ratios for item 1 with those for items 2 or 3, we can see that in the former they are more extreme (7.0 and 0.1) than in the latter (3.5 and 0.3). At the same time, $\hat{\alpha}_{1,,1}$ is very large (9.41) while $\hat{\alpha}_{2,,1}$ and $\hat{\alpha}_{3,,1}$ are equal to 1.69.

Furthermore the ratios for item 1 indicate that if an individual answers negatively (positively) to items 2 and 3 then it is more probable that he will also answer negatively (positively) to item 1.

From the original set of data (Table 6.1), we derive 7 new sets by altering the frequencies of some score patterns. As a consequence the sample size of the new set of data might change, assuming a value between 190 and 240. For four of these seven new sets, all items have the same first and second order margins. The same analysis carried out for the original data is extended to all sets. The results are used to measure the effect on the size of $\hat{\alpha}_{i,,1}$ of changing the pattern of the ratios.

Thus, for example, if the score pattern '100' for items 123 has frequency 5 instead of 10, then the new ratio is equal to 14.0. This change makes the pattern of ratios for item 1 more extreme, while those for items 2 and 3 almost do not change (0.25 instead of 0.50). Moreover this change also yields a larger estimate $\hat{\alpha}_{1,,1}$ (32.65 with standard deviation 1340.49) for nearly the same $\hat{\alpha}_{2,,1}$ and $\hat{\alpha}_{3,,1}$ (1.78 with standard deviation 0.32).

The next set of data to be analysed corresponds mainly to an interchange between items 1 and (2 and 3) from the original test(Table 6.1) of the frequencies for the first and last ratios. Therefore, as in that case, all items have the same first and second order margins.

Table 6.2- Distribution of the score patterns and ratios of the conditional frequencies.

123	n	ratio	231	n	ratio
000	70		000	70	
100	20	(3.5)	100	10	(7.0)
001	10		001	20	
101	10	(1.0)	101	10	(2.0)
010	10		010	10	
110	10	(1.0)	110	20	(0.5)
011	20		011	10	
111	70	(0.3)	111	70	(0.1)

$\hat{\alpha}_{1,,1}$= 1.72
$SD(\hat{\alpha}_{1,,1})$= 0.34

$\hat{\alpha}_{i,,1}$= 3.98, i=2,3
$SD(\hat{\alpha}_{i,,1})$= 1.42

Table 6.2 shows that the interchanging of frequencies between items 1 and 2,3 yields $\hat{\alpha}_{1,,1}$ nearly equal to $\hat{\alpha}_{i,,1}$, i=2,3, from the original test. At the same time items 2 and 3 have also a large $\hat{\alpha}_{i,,1}$ (3.98), as item 1 had a large $\hat{\alpha}_{1,,1}$ in the original test.

Consider now the substitution of all frequencies 10 in Table 6.2 by frequencies 5. Then the ratios for item 1 remain the same, while they become more extreme for items 2 and 3 (14.00; 4.00; 0.25; 0.07). The results is that $\hat{\alpha}_{2,,1}$ and $\hat{\alpha}_{3,,1}$ assume larger values, 6.8 with standard deviation 3.09, while $\hat{\alpha}_{1,,1}$ changes very little (1.77 with standard deviation 0.34).

Again in relation to Table 6.2, consider the substitution of the frequency 20 by 30 for the score patterns '100' and '011' for items 123. Then the ratios for item 1 are closer to each other than before and $\hat{\alpha}_{1,,1}$ decreases to 1.05 with standard deviation 0.22. The new ratios for items 2 and 3 are 3.0 and 0.3, instead of 2.0 and 0.5, for nearly the same parameter estimates $\hat{\alpha}_{i,,1}$ (4.38 with standard deviation 2.29).

The following set of data (Table 6.3) was generated in such way that all items have the same corresponding ratios, which are equal to the ratios for item 1 in the original test. The sample size is 200 instead of 240, since the frequencies 20 were replaced by 10.

Table 6.3- Distribution of the score patterns and ratios of the conditional frequencies.

123	n	ratio	231	n	ratio
000	70		000	70	
100	10	(7.0)	100	10	(7.0)
001	10		001	10	
101	10	(1.0)	101	10	(1.0)
010	10		010	10	
110	10	(1.0)	110	10	(1.0)
011	10		011	10	
111	70	(0.1)	111	70	(0.1)

$\hat{\alpha}_{1,,1} = 3.56$
$SD(\hat{\alpha}_{1,,1}) = 0.91$

$\hat{\alpha}_{i,,1} = 3.56,\ i=2,3$
$SD(\hat{\alpha}_{i,,1}) = 0.91$

As expected all items have the same $\hat{\alpha}_{i,,1}$, which is large (3.56), but smaller than $\hat{\alpha}_{1,,1}$ in the original test, where just item 1 has a extreme distribution of ratios.

Furthermore, Table 6.3 shows that if an individual answers negatively (positively) to any two items then it is more probable that he will also answer negatively (positively) to the remaining one.

We now present an example in which the distribution of ratios for all items is extreme, but not equal as in the preceding example. In this case the distribution of ratios for item 1 is less extreme than for items 2 and 3.

Table 6.4- Distribution of the score patterns and ratios of the conditional frequencies.

123	n	ratio	231	n	ratio
000	70		000	70	
100	10	(7.0)	100	5	(14.0)
001	5		001	10	
101	10	(0.5)	101	10	(1.0)
010	5		010	5	
110	10	(0.5)	110	10	(0.5)
011	10		011	10	
111	70	(0.1)	111	70	(0.1)

$\hat{\alpha}_{1,,1} = 3.45$
$SD(\hat{\alpha}_{1,,1}) = 0.76$

$\hat{\alpha}_{i,,1} = 4.69$, $i=2,3$
$SD(\hat{\alpha}_{i,,1}) = 1.36$

From the distribution of ratios in Table 6.4 we can say that if an individual answers positively to any of two items than is more probable that he will answer positively to the third one too. The same is also true for answering negatively, but in this case, the probability is higher when the known answers for the first two items include item 1. Besides the parameter estimate $\hat{\alpha}_{1,,1}$ is slightly smaller than $\hat{\alpha}_{2,,1}$ and $\hat{\alpha}_{3,,1}$.

Consider Table 6.4 when the score patterns '100' and '011' for the sequence of items 123 have their frequencies 10 replaced by 20. Then the ratios for item 1 are less extreme, (3.5; 0.5; 0.5; 0.3), while they became slightly more extreme for items 2 and 3, (14.0, 2.0, 0.2, 0.1). Fitting a logit/probit model to these data, $\hat{\alpha}_{1,,1}$ decreases to 1.74 with standard deviation 0.34 while $\hat{\alpha}_{2,,1}$ and $\hat{\alpha}_{3,,1}$ are larger and equal to 5.14 with standard deviation 2.18.

Conclusions

Under the conditions of this investigation, that is, tests with 3 items in which the smallest and biggest ratio for all items were always associated with the same score patterns 00 and 11, we can conclude that

(1) When fitting a logit/probit model the occurrence of a large $\hat{\alpha}_{i,1}$ is strongly associated with an extreme distribution of the ratios for item i, which depends on the whole score pattern. How extreme this distribution must be depends on the pattern of the ratios for the remaining items. The most extreme distribution of ratios produces the largest $\hat{\alpha}_{i,1}$.

(2) As the distribution of the ratios for item i becomes more extreme, the frequencies of the score patterns give more information about the prediction of item i from the remaining items.

1.3- Four Variables: Real and Simulated Data

In all eight tests with 3 items that we have analysed, the relation between the size of $\hat{\alpha}_{i,1}$ and the pattern of the ratios for each item was very clear and consistent. The analysis was simplified due to the small number of items and the allocation of the extreme ratios to the same score patterns (00 and 11).

Increasing the number of items to 4, the possible number of different score patterns increases to 16 and the extreme ratios for each item will not probably be associated with the same score patterns. Hence the analysis of the pattern of the ratios and its relation with the size of $\hat{\alpha}_{i,1}$ become more complex.

These considerations lead us to search for a more precise measure of the distribution of ratios for every item, that is, for a more precise measure of the predictability of an item from the others.

Comparing empirically the distribution of ratios for numerous data sets with 4 items, we found out that the size of $\hat{\alpha}_{i,1}$ also depends on for which score pattern the most extreme ratios occur. For example, if the same extreme ratios for item 1 are from the score patterns '000' and '111' for items 234 then $\hat{\alpha}_{1,1}$ will be larger than if they are from the score patterns '010' and '100' or '110' and '011'.

The most satisfactory simple measure of the predictability of item i from the other items was found to be the slope of the line obtained from regressing the ln(ratios for item i) on the number of positive responses. Thus, for instance, when regressing the ln(ratio) for item 1, the score patterns '001', '010' and '100' for items 234 have the

same value 1 for the independent variable, since they have the same number of positive responses.

We expect that the item, which has the biggest slope (absolute value), is that item the most predictable from all the others and that consequently may have the largest $\hat{\alpha}_{i,1}$. At the same time, if the biggest slope is much larger than the others then the corresponding item will probably have a large $\hat{\alpha}_{i,1}$, while for the remaining ones the $\hat{\alpha}_{i,1}$'s will be much smaller.

In the following we shall illustrate the application of this procedure to 3 examples, which represent the three kinds of results we found so far. These results are supported by an analysis of a significant number of data sets, in which all the score patterns occur, which are not reported here. We shall discuss later the case when some score patterns have frequencies of zero.

Bootstrap Sample 5 from the ART on Black Women

Table 6.5 displays the distribution of the score patterns and ratios for the conditional frequencies for the normal bootstrap sample 5 from the ART on black women data (Chapter 4, section 5).

Table 6.5- Distribution of the score patterns and ratios of the conditional frequencies for the normal bootstrap sample 5 from the ART on black women data.

	1234		2341		3412		4123	
0000	23		23		23		23	
1000	31	(0.7)	11	(2.1)	6	(3.8)	12	(1.9)
0001	12		31		11		6	
1001	4	(3.0)	13	(2.4)	7	(1.6)	2	(3.0)
0010	6		12		31		11	
1010	5	(1.2)	3	(4.0)	5	(6.2)	4	(2.8)
0100	11		6		12		31	
1100	13	(0.8)	7	(0.8)	2	(6.0)	4	(7.8)
0011	2		4		13		7	
1011	7	(0.3)	5	(0.8)	8	(1.6)	3	(2.3)
0101	4		5		4		5	
1101	5	(0.8)	8	(0.6)	3	(1.3)	7	(0.7)
0110	7		2		4		13	
1110	8	(0.9)	3	(0.7)	7	(0.6)	5	(2.6)
0111	3		7		5		8	
1111	4	(0.8)	4	(1.8)	4	(1.2)	4	(2.0)
Slope	-0.18		-0.30		-0.61		-0.22	
R^2(%)	7.00		15.70		45.70		9.10	

The absolute values of the slopes show that item 3 is the most predictable from the others, since its slope is double the size of the largest one among the remaining ones. This gives evidence that when fitting a logit/probit model to this data, $\hat{\alpha}_{3,1}$ will probably be much larger than the others.

Fitting a logit/probit model to the data in Table 6.5, the discrimination parameter estimates and standard deviations (in brackets) are

$\hat{\alpha}_{1,1}$	$\hat{\alpha}_{2,1}$	$\hat{\alpha}_{3,1}$	$\hat{\alpha}_{4,1}$
0.11 (0.22)	0.52 (0.25)	13.06 (59.99)	0.39 (0.25)

for a statistic chi-square equal to 7.24 and 5 degrees of freedom. ($p<0.05$).

The comparison between the parameter estimates $\hat{\alpha}_{i,1}$ and the pattern of the ratios for every item shows that the slopes (absolute value) give the order of the $\hat{\alpha}_{i,1}$'s, and that item 3 has the large estimate $\hat{\alpha}_{3,1}$ as predicted.

The investigation (not reported here) of several tests with 4 items, in which all score patterns occur, shows that an item with large $\hat{\alpha}_{i,1}$ always corresponds to the largest slope, but sometimes the ratio between its slope value and the next biggest one was smaller than two. This means that sometimes there is not enough information from the pattern of the slopes to predict the occurrence of a large $\hat{\alpha}_{i,1}$. On the other hand, the slopes very often give the order of the $\hat{\alpha}_{i,1}$'s, especially when the $\hat{\alpha}_{i,1}$'s are not very similar.

It is also usually true that an item with a large $\hat{\alpha}_{i,1}$ has an R^2 much bigger than the other items.

Very often, a joint analysis of the slopes and the R^2's gives some information as to whether the Rasch model fits the data or whether the more general logit/probit model is required. One cannot expect too much from the analysis of the slopes and R^2 values because they do not change with the sample size in the same way as a likelihood ratio test. For instance, doubling the observed frequencies leaves slopes and R^2 unchanged.

The value of R^2 is strongly dependent on the range of different values assumed by the ln(ratio) for all those score patterns with the same number of positive responses. In general when a logit/probit model fits the data but Rasch does not fit and, for instance, $\hat{\alpha}_{1,1}$ is very large, then the logarithm of the ratios for item 1 corresponding to score patterns '100', '010' and '001' for items 234 are much closer to each other than the logarithm of the ratios for the remaining items. As a consequence R^2 for item 1 tends to be much bigger than for the remaining ones.

When both models fit the data, and one of the $\hat{\alpha}_{i,1}$'s is very large compared with the others, and all $\hat{\alpha}_{i,1}$'s have large asymptotic standard deviations,then R^2 for the largest $\hat{\alpha}_{i,1}$ tends to be bigger than the others, but not so much greater than when only the logit/probit model fits the data. This effect can be seen in Table 6.5 where the low R^2's seem to indicate that the items are giving contradictory information.

Macready and Dayton's Data

Macready and Dayton's data (Bartholomew, 1987, page 127) corresponds to the results of a test on 4 items selected at random from a domain of items. Each item consisted of a multiplication of a two-digit number by a three-or-four digit number involving 'carry' operations.

Table 6.6- Distribution of the score patterns and ratios of ratios of the conditional frequencies for the Macready and Dayton's data.

	1234		2341		3412		4123	
0000	41		41		41		41	
1000	13	(0.3)	6	(0.2)	1	(0.02)	4	(0.1)
0100	6		1		4		13	
1100	7	(1.2)	2	(2.0)	4	(1.0)	6	(0.5)
0001	4		13		6		1	
1001	6	(1.5)	7	(0.5)	2	(0.3)	4	(4.0)
0010	1		4		13		6	
1010	3	(3.0)	5	(1.2)	3	(0.2)	5	(0.8)
0101	5		3		5		3	
1101	23	(4.6)	7	(2.3)	4	(0.8)	1	(0.3)
0110	2		4		6		7	
1110	7	(3.5)	4	(1.0)	1	(0.2)	23	(3.3)
0011	4		6		7		2	
1011	1	(0.25)	23	(3.8)	7	(1.0)	4	(2.0)
0111	4		1		23		7	
1111	15	(3.8)	15	(15.0)	15	(0.6)	15	(2.1)
Slope	0.60		1.31		0.92		0.80	
R^2(%)	23.90		79.00		41.80		32.80	

In Table 6.6 the pattern of the slopes shows that item 2 is the most predictable item from the others, since it has the largest slope and the largest R^2. The small difference between the two largest slopes, 1.31 and 0.92, does not allow one to say whether $\hat{\alpha}_{2,1}$ will be much larger than the others.

Fitting a logit/probit model to Macready and Dayton's data, the discrimination parameter estimates and standard deviations (in brackets) are

$\hat{\alpha}_{1,1}$	$\hat{\alpha}_{2,1}$	$\hat{\alpha}_{3,1}$	$\hat{\alpha}_{4,1}$
1.59 (0.42)	4.04 (2.17)	1.16 (0.36)	1.92 (0.55)

for a statistic chi-square equal to 6.91 with 2 degrees of freedom ($0.05<p<0.08$).

As expected from the analysis of the distribution of ratios, item 2 has the largest $\hat{\alpha}_{i,1}$ (4.04), which is not considered an occurrence of only one very large $\hat{\alpha}_{i,1}$, since it is followed by $\hat{\alpha}_{2,1}$ equal to 1.92.

The Rasch model also fits this data, for which the loglikelihood is very similar to that for the logit/probit model (-336.64 compared with -334.30) and the statistic chi-sqared is equal to 11.98 for 6 degrees of freedom. This gives evidence that the $\hat{\alpha}_{i,1}$'s are equal or nearly equal and it might explain that the slopes do not give the order of the $\hat{\alpha}_{i,1}$'s.

Arithmetic Reasoning Test (ART) on White Women

The ART on white women is described in Chapter 2, Table 2.1, where fitting by a logit/probit model is also discussed.

Table 6.7- Distribution of the score patterns and ratios of the conditional frequencies for the ART on white women data.

	1234		2341		3412		4123	
0000	20		20		20		20	
1000	23	(1.2)	20	(1.0)	14	(0.7)	8	(0.4)
0100	20		14		8		23	
1100	18	(0.9)	11	(0.8)	2	(0.2)	8	(0.4)
0001	8		23		20		14	
1001	8	(1.0)	18	(0.8)	11	(0.6)	2	(0.1)
0010	14		8		23		20	
1010	9	(0.6)	5	(0.6)	9	(0.4)	5	(0.2)
0101	5		9		5		9	
1101	15	(3.0)	20	(2.2)	7	(1.4)	6	(0.7)
0110	11		2		8		18	
1110	20	(1.8)	7	(3.5)	6	(0.8)	15	(0.8)
0011	2		8		18		11	
1011	6	(3.0)	15	(1.9)	20	(1.1)	7	(0.6)
0111	7		6		15		20	
1111	42	(6.0)	42	(7.0)	42	(2.8)	42	(2.1)
Slope	0.69		0.79		0.60		0.70	
R^2(%)	71.20		73.90		54.40		61.20	

The similarity of the slopes ranging from 0.60 to 0.79 indicates that none of the items is more predictable than the others, which is also confirmed by similar R^2. Therefore we can expect that the $\hat{\alpha}_{i,1}$'s

will be very similar, and do not expect that the slopes will give the order.

A logit/probit model fits well with the discrimination parameter estimates and standard deviations (in brackets) equal to

$\hat{\alpha}_{1,1}$	$\hat{\alpha}_{2,1}$	$\hat{\alpha}_{3,1}$	$\hat{\alpha}_{4,1}$
1.04 (0.32)	1.24 (0.39)	1.00 (0.30)	1.44 (0.45)

The ranges of values assumed by $\hat{\alpha}_{i,1}$ agrees with the expected pattern of $\hat{\alpha}_{i,1}$ based on the comparison between the predictability of an item from all the others.

Conclusions

The results from the investigation of the relation between the pattern of the $\hat{\alpha}_{i,1}$'s and the predictability of one item from the others give evidence that

(1) If none of the items is more predictable than the other items, which is shown by very similar slopes, then probably all the $\hat{\alpha}_{i,1}$'s will be very similar. (ART on white women data)

(2) If one of the items is the most predictable, but its slope is not much larger than the others then this item will have the largest $\hat{\alpha}_{i,1}$, which might or not might be very large in relation to the remaining ones. Furthermore, the item with large $\hat{\alpha}_{i,1}$ may have an R^2 much greater than the others and this happens more often when a

(3) If one of the items is much more predictable than the other ones, which is shown by a much larger slope, then probably this item will have a very large $\hat{\alpha}_{i,1}$ while for the others $\hat{\alpha}_{i,1}$ will be small. (Sample 5 from the ART on black women data).

(4) The slopes can not be used with certainty to predict the order of the $\hat{\alpha}_{i,1}$'s, though they very often give the right order.

2- *Generating a (p+1)th Item with any Fixed* $\hat{\alpha}_{p+1,1}$ *and* $\hat{\alpha}_{p+1,0}$

The empirical findings of the previous section were the inspiration for the next step in searching for reasons of the occurrence of large discrimination parameter estimates when fitting a logit/probit model.

If we are able to generate a set of data in which one of the items has large $\hat{\alpha}_{i,1}$ then we will be better able to understand the occurrence of a large discrimination parameter when fitting a logit/probit model.

We now present a procedure under which we can add a (p+1)th variable with any fixed $\hat{\alpha}_{p+1,0}$ and $\hat{\alpha}_{p+1,1}$, to each score pattern $\mathbf{x}_s$ without altering the previous ML estimates of $\alpha_{i,0}$ and $\alpha_{i,1}$, i=1,...,p.

2.1- Maximum Likelihood Estimation

Suppose that L(p) is the loglikelihood function for p items, $f_p(x|x_s)$ is the conditional probability of score x for the (p+1)th item if x_s is the score pattern for the first p items and θ_{sx} is the observed frequency of (x_s,x), x=0,1. Then

$$L(p+1) = \sum_s [\theta_{s0} \ln(f_p(0|x_s)f(x_s)) + \theta_{s1} \ln(f_p(1|x_s)f(x_s))]$$

$$= \sum_s [\theta_{s0} \ln(f_p(0|x_s)) + \theta_{s1} \ln(f_p(1|x_s))] + L(p),$$

where $L(p) = \sum_s [\theta_{s0} \ln(f(x_s)) + \theta_{s1} \ln(f(x_s))]$

So, taking the supremum over all possible choices for the parameters $\alpha_{i,1}$ and $\alpha_{i,0}$ for i=1,...,p+1,

$$\sup L(p+1) \leq \sup L(p) + \sup \{ \sum_s [\theta_{s0} \ln(f_p(0|x_s)) + \theta_{s1} \ln(f_p(1|x_s))]\}$$

$$\leq \sup L(p) + \theta_{s0} \ln \frac{\theta_{s0}}{\theta_s} + \theta_{s1} \ln \frac{\theta_{s1}}{\theta_s} .$$

Consequently, the ML estimation for (p+1) items is achieved if $\alpha_{i,v} = \hat{\alpha}_{i,v}$, for i=1,...,p and v=0,1 and

$$f_p(1|x_s) = \frac{\theta_{s1}}{\theta_s} \quad \text{for all } x_s. \tag{6.1}$$

where

$$f_p(1|x_s) = \frac{f(x_s,1)}{f(x_s)} = \frac{\int_{-\infty}^{\infty} f(x_s|z)\, f(1|z)\, h(z)\, dz}{\int_{-\infty}^{\infty} f(x_s|z)\, h(z)\, dz}.$$

For a suitable θ_{s0} and θ_{s1} this will allow any $\hat{\alpha}_{p+1,0}$ and $\hat{\alpha}_{p+1,1}$ to be the MLE for the (p+1)th item, while leaving $\hat{\alpha}_{i,0}$ and $\hat{\alpha}_{i,1}$, i=1,...,p, as

for MLE for the first p items.

If (6.1) holds, then, we prove below that the elements of the asymptotic covariance matrix of $\hat{\alpha}_{i,v}$'s for i=1,...,p and v=0 or 1, do not increase when the $(p+1)^{th}$ variable is added.

Let $D_{u_r} = \dfrac{\partial}{\partial \alpha_{u,r}}$, for u=1,...,p+1 and r=0 or1. Then

$$D_{u_r}D_{v_t}L(p+1) = D_{u_r}D_{v_t}L(p) +$$

$$+ D_{u_r}D_{v_t} \sum [\, \theta_{s0}\ln(f_p(0|x_s)) + \theta_{s1}\ln(f_p(1|x_s))\,]$$

That is,

$$D_{u_r}D_{v_t}L(p+1) = D_{u_r}D_{v_t}L(p) +$$

$$\sum_s \left\{ \frac{\theta_{s0}}{f_p(0|x_s)} \left[D_{u_r}D_{v_t}f_p(0|x_s)\right] - \frac{\theta_{s0}}{f^2(0|x_s)} \left[D_{u_r}f_p(0|x_s)\right] \left[D_{v_t}f_p(0|x_s)\right] \right.$$

$$\left. + \frac{\theta_{s1}}{f_p(1|x_s)} \left[D_{u_r}D_{v_t}f_p(1|x_s)\right] - \frac{\theta_{s1}}{f^2(1|x_s)} \left[D_{u_r}f_p(1|x_s)\right] \left[D_{v_t}f_p(1|x_s)\right] \right\}$$

Suppose now that $f_p(0|x_s) = \dfrac{\theta_{s0}}{\theta_s}$ for all x_s. Then

$$D_{u_r}D_{v_t}L(p+1) = D_{u_r}D_{v_t}L(p) + \sum_s \left\{ \theta_s D_{u_r}D_{v_t}\left[f_p(0|x_s) + f_p(1|x_s)\right] \right.$$

$$\left. - \theta_s \left[\frac{1}{f_p(0|x_s)} \right] \left[D_{u_r}f_p(0|x_s)\right] \left[D_{v_t}f_p(0|x_s)\right] + \right.$$

$$+ \quad \frac{1}{f_p(1|x_s)} \left[D_{u_r} f_p(1|x_s) \right] \left[D_{v_t} f_p(1|x_s) \right] \Bigg] \Bigg\}$$

ie, $-D_{u_r}D_{v_t}L(p+1) = -D_{u_r}D_{v_t}L(p) +$

$$\sum_s \theta_s \left\{ \frac{\left[D_{u_r} f_p(0|x_s)\right] \left[D_{v_t} f_p(0|x_s)\right]}{f_p(0|x_s)} + \frac{\left[D_{u_r} f_p(1|x_s)\right] \left[D_{v_t} f_p(1|x_s)\right]}{f_p(1|x_s)} \right\}$$

So the observed information matrix for the (p+1) items in this case is no less than (in the sense of positive definiteness of the difference) the observed information matrix for the first p items.

Considering the information matrices for p+1 and p items, we have

$$\begin{bmatrix} {}_{2p}A_{2p} & {}_{2p}C'_{2} \\ {}_{2}C_{2p} & {}_{2}D_{2} \end{bmatrix} \geq \begin{bmatrix} {}_{2p}B_{2p} & 0 \\ 0 & 0 \end{bmatrix}$$

(information matrix for p+1 items)

E: information matrix for p items.

Consider also that

$$\begin{bmatrix} A & C \\ C' & D \end{bmatrix}^{-1} = \begin{bmatrix} \tilde{A} & \tilde{C} \\ \tilde{C}' & \tilde{D} \end{bmatrix}$$

Then for $\epsilon > 0$

$$\begin{bmatrix} A & C \\ C' & D \end{bmatrix} + \epsilon I \geq \begin{bmatrix} B & 0 \\ 0 & 0 \end{bmatrix} + \epsilon \begin{bmatrix} 0 & 0 \\ 0 & I \end{bmatrix}$$

i.e.,

$$\begin{bmatrix} A + \epsilon I & C \\ C' & D + \epsilon I \end{bmatrix} \geq \begin{bmatrix} B & 0 \\ 0 & \epsilon I \end{bmatrix}$$

Therefore

$$\begin{bmatrix} A + \epsilon I & C \\ C' & D + \epsilon I \end{bmatrix}^{-1} \leq \begin{bmatrix} B^{-1} & 0 \\ 0 & \epsilon^{-1} I \end{bmatrix}$$

i.e.,

$$\begin{bmatrix} \tilde{A}(\epsilon) & \tilde{C}(\epsilon) \\ \tilde{C}'(\epsilon) & \tilde{D}(\epsilon) \end{bmatrix} \leq \begin{bmatrix} B^{-1} & 0 \\ 0 & \epsilon^{-1} I \end{bmatrix}$$

So for all $\epsilon > 0$, $\tilde{A}(\epsilon) \leq \tilde{B}^{-1}$.

Since $\tilde{A}(\epsilon)$ has elements continuous in ϵ, and $\lim_{\epsilon \downarrow 0} \tilde{A}(\epsilon) = \tilde{A}$, it follows that $\tilde{A} \leq B^{-1}$. Therefore the asymptotic covariances as estimated for $\hat{\alpha}_{i,0}$ and $\hat{\alpha}_{i,1}$, i=1,...,p, does not increase when the (p+1)th is added.

It also follows from (6.1) that we can bring into $f_p(1|x_s)$ any $f(1|z)$ distribution over Z. For instance, for the threshold model

$$f(1|z) = \begin{cases} 0 & \text{if } z \leq z_{t'} \\ 1 & \text{if } z > z_{t'} \end{cases}$$

for every score pattern x_s, and thus

$$\theta_{s1} = \theta_s f_p(1|x_s) = \theta_s \int_{z_{t'}}^{\infty} \frac{f(x_s|z)\, h(z)}{f(x_s)}\, dz$$

$$= \theta_s \int_{z_{t'}}^{\infty} h(z|x_s)\, dz \qquad (6.2)$$

for every score pattern x_s.

This means that we chop the posterior distribution of Z given x_s at the same point z_t, for every x_s.

2.2- An E-M Algorithm

Recall that, in this thesis, the parameters of a one latent variable logit/probit or probit model are estimated using an E-M algorithm for the MML procedure, as described in Chapter 1, section 3.2.2. That is, the parameters $\alpha_{i,v}$ for i=1,...,p and v=0,1, are estimated maximising the loglikelihood

$$\ln L = \sum_{s=1}^{n} \ln f(x_s)$$

and setting the following partial derivatives equal to zero,

$$\frac{\partial \log L}{\partial \alpha_{i,v}} = \int_{-\infty}^{\infty} \frac{\partial \pi_i(z)}{\partial \alpha_{i,v}} \left\{ \frac{R_i - N \pi_i(z)}{\pi_i(z)\ (1-\pi_i(z))} \right\} dz \qquad (6.3)$$

where θ_s is the observed frequency of x_s,

$$R_i = \sum_s x_{is}\ \theta_s\ h(z|x_s) \qquad \text{and} \qquad (6.4)$$

$$N = \sum_s \theta_s\ h(z|x_s) \qquad \text{for i=1,..,p and v=0,1.} \qquad (6.5)$$

Therefore when adding a (p+1)th variable with any fixed $\hat{\alpha}_{p+1,0}$ and $\hat{\alpha}_{p+1,1}$ to each response vector x_s, if (6.1) holds, i.e,

$$f_p(1|x_s) = \frac{\theta_{s1}}{\theta_s} \qquad \text{for all } x_s$$

the equations (6.3) remain the same, for i=1,...,p, which is equivalent to obtaining the same R_i and N. That is, if (6.1) holds, then (6.4) and (6.5) are equal to (6.6) and (6.7), i.e,

$$R_i = \sum_s x_{is} [\, \theta_{s0}\, h(z|x_{s0}) + \theta_{s1}\, h(z|x_{s1}) \,] \quad \text{and} \qquad (6.6)$$

$$N = \sum_s [\, \theta_{s0}\, h(z|x_{s0}) + \theta_{s1}\, h(z|x_{s1}) \,] \qquad (6.7)$$

for i=1,...,p, where $x_{s0}=(x_s,0)$ and $x_{s1}=(x_s,1)$.

2.3- Relation between R_{p+1} and N

Similarly to (6.6), for the (p+1)th variable R_{p+1} may be given by

$$R_{p+1} = \sum_{s=1}^{n} [\, x_{p+1,s0}\, \theta_{s0}\, h(z|x_{s0}) + x_{p+1,s1}\, \theta_{s1}\, h(z|x_{s1}) \,]$$

where $x_{p+1,s0}=0$ and $x_{p+1,s1}=1$, and thus

$$R_{p+1} = \sum_{s=1}^{n} \theta_{s1}\, h(z|x_{s1}) \qquad (6.8)$$

From (6.1)

$$h(z|x_{s0}) = \frac{f(x_{s0}|z)\ \ h(z)}{f(x_{s0})}$$

$$= \frac{\theta_s}{\theta_{s0}} \ \frac{f(x_{s0}|z)\ h(z)}{f(x_s)} \qquad (6.9)$$

and similarly

$$h(z|x_{s_1}) = \frac{\theta_s}{\theta_{s_1}} \quad \frac{f(x_{s_1}|z)\, h(z)}{f(x_s)} \tag{6.10}$$

Substituting (6.10) in (6.8)

$$R_{p+1} = \sum_{s=1}^{n} \theta_s \quad \frac{f(x_{s_1}|z)\, h(z)}{f(x_s)}$$

and substituting (6.9) and (6.10) in (6.7)

$$N = \sum_{s=1}^{n} \theta_s \left[\frac{f(x_{s_0}|z)\, h(z)}{f(x_s)} + \frac{f(x_{s_1}|z)\, h(z)}{f(xs_)} \right]$$

$$= \sum_{s=1}^{n} \theta_s \quad \frac{h(z)}{f(x_s)} \quad \left[f(x_{s_0}|z) + f(x_{s_1}|z) \right]$$

Since $f(x_{s_1}|z) = \pi_{p+1}\, f(x_s|z)$ and $f(x_{s_0}|z) + f(x_{s_1}|z) = f(x_s)$,

$$\frac{R_{p+1}}{N} = \frac{\sum_{s=1}^{n} \theta_s \quad \frac{h(z)}{f(x_s)} \quad \pi_{p+1}(z)\, f(x_s|z)}{\sum_{s=1}^{n} \theta_s \quad \frac{h(z)}{f(x_s)} \quad f(x_s|z)}$$

and therefore

$$\frac{R_{p+1}}{N} = \pi_{p+1}(z). \tag{6.11}$$

We can take the threshold model as a limiting case.

3.1- Algorithm for Generating a (p+1)th Item

As our main goal is to add a (p+1)th item to a set of data, for which $\hat{\alpha}_{p+1,1}$ is very large, when fitting a logit/probit model, we shall only describe in detail this situation.

Recall that, when fitting a logit/probit model to data sets in this thesis, the parameters are estimated using the MML procedure, in which the marginal probability function

$$f(x_s) = \int_{-\infty}^{\infty} f(x_s|z)\, h(z)\, dz \qquad s=1,\ldots,n$$

is approximated by Gauss-Hermite quadrature, i.e.,

$$f(x_s) = \sum_{t=1}^{k} f(x_s|z_t)\, h(z_t) \qquad s=1,\ldots,n$$

where z_t is one of the tabulated quadrature points, which are chosen to best approximate $f(x_s)$,

$$g(x_s|z_t) = \prod_{i=1}^{p} \pi_i(z_t)^{x_{is}} \left[1-\pi_i(z_t)\right]^{1-x_{is}}$$

where

$\pi_i(z_t)$ is the response function of variable i at z_t,

$h(z_t)$ is the weight of the quadrature point z_t, which are approximately the normalized values of the probability density of a N(0,1) random variable at the points z_t.

The MML estimation procedure is carried out through an E-M algorithm as described in Chapter 1, section 3.2.2. As in the continuous case, equations (6.4) and (6.5), it involves the calculation of two main quantities

$$R_{it} = \sum_{s=1}^{n} x_{is}\ \theta_s\ h(z_t|x_s), \qquad i=1,\ldots,p \text{ and } t=1,\ldots,k$$

$$N_t = \sum_{s=1}^{n} \theta_s\ h(z_t|x_s)$$

where θ_s is the observed frequency of x_s and $h(z_t|x_s)$ is the posterior probability of z_t given x_s.

Therefore N_t is the expected number of individuals at z_t and R_{it} is the expected number of positive responses to item i among those individuals at z_t.

In the following we concentrate discussion on the posterior distribution of z_t given the score pattern x_s.

As the quantity $h(z_t|x_s)$ is the probability that an individual with response vector x_s is located at z_t then

N_{ts} = (observed frequency of x_s) $h(z_t|x_s)$

is the expected number of individuals with score pattern x_s at z_t.

This means that the observed frequency of the score pattern x_s is distributed over the k points z_t, t=1,...,k. Therefore, there is a set of values N_{ts}, t=1,...,k, for each score pattern x_s.

We shall use the distribution of N_{ts}, t=1,...,k, of each score pattern x_s to add a new variable to x_s, for all s, for which $\hat{\alpha}_{p+1,1}$ is very large.

Suppose that $x = (x_1,...,x_p)$ and the new zero-one variable is x_{p+1}.

We use the following steps for the generation of an additional variable x_{p+1}, which will have a large discrimination parameter estimate when the new set of data is fitted by a logit/probit model:

(1) Obtain the distribution of N_{ts}, t=1,...,k, for every score pattern x_s.

(2) Chop the distribution of N_{ts}, t=1,...,k, of each score pattern at the same point $z_{t'}$, for $1<t'<k$.

(3) For each score pattern x_s, add all N_{ts} for which $t \geq t'$, that is,

$$\sum_{t=t'}^{k} N_{ts} = \theta_s \sum_{t=t'}^{k} h(z_t|x_s) \qquad (6.12)$$

where θ_s is the observed frequency of x_s.

(4) Obtain the nearest integer to (6.12) and take it as the observed frequency θ_{s1} of the new score pattern $x_{s0} = (x_s,1)$, for each s.

It is only for practical reasons that we take the nearest integer to (6.12) as the observed frequency of the score pattern x_{s1}.

(5) Set the difference between the observed frequency of x_s and x_{s1}, that is, the difference between θ_s and θ_{s1} equal to the observed frequency of $x_{s0} = (x_s,0)$.

Therefore for every score pattern x_s, we generate two new score patterns x_{s0} and x_{s1}, which differ just in the values '0' and '1' assumed by the variable x_{p+1}.

In the following we present some examples of the application of this procedure to generate a new variable, for which the discrimination parameter estimate is large.

3.2- Simulated Data

We start from an example of adding a third variable to a set of data with sample size 145, for which the fitting by a logit/probit model will yield a large $\hat{\alpha}_{3,1}$.

The distribution of the score pattern (frequencies in brackets) of the initial set of data is

00 (28) 01 (38) 10 (33) 11 (46) for s equal 1,2,3,4

respectively.

Table 6.8 displays the distribution of N_{ts}, the expected number of individuals with score patterns x_s at z_t, for all the 4 score patterns and 16 quadrature points.

Table 6.8- Frequency distribution of N_{ts}, the expected number of individuals with score pattern x_s (s=1,..,4) at z_t (t=1,..,16).

t	Z_t	N_{t1}	N_{t2}	N_{t3}	N_{t4}
1to3	≤ -4.49	0.00	0.00	0.00	0.00
4	-3.60	0.03	0.02	0.01	0.01
5	-2.76	0.32	0.32	0.23	0.21
6	-1.95	1.85	2.00	1.45	1.56
7	-1.16	5.43	6.46	5.06	6.01
8	-0.39	8.54	11.22	9.42	12.37
9	0.39	7.36	10.67	9.61	13.91
10	1.16	3.47	5.50	5.36	8.58
11	1.95	0.87	1.56	1.60	2.83
12	2.76	0.11	0.23	0.24	0.48
13	3.60	0.01	0.02	0.02	0.04
14to16	≥ 4.49	0.00	0.00	0.00	0.00
Total	-	28	38	33	46

From Table 6.8 the sum of N_{ts}, t≥t', for every score pattern x_s is

	score pattern			
t'	1	2	3	4
9	11.82	17.98	16.83	25.84
10	4.46	7.31	7.22	11.93
11	0.99	1.81	1.86	3.35

After obtaining the distribution of N_{ts} for every score pattern x_s, as given in Table 6.8, a common point $z_{t'}$ is chosen at which to

chop the posterior distribution into two pieces. For every different t', a new set of data can be generated from adding all the frequencies N_{ts}, $t \geq t'$, for each score pattern x_s.

Table 6.9 shows the frequency distribution of the new data set when choosing t' equal to 10 and taking the nearest integer of (6.12) as the observed frequency of x_3 equal to '1' for each score pattern.

Table 6.9- Frequency distribution of the data generated by chopping the distribution of N_{ts} at t'equal to 10 for every score pattern x_s.

s	x_s	n	x_{s1}	n	x_{s0}	n
1	00	28	001	4	000	24
2	01	38	011	7	010	31
3	10	33	101	7	100	26
4	11	46	111	12	110	34

The discrimination parameter estimates and standard deviations (in brackets) from fitting a logit/probit model to this new set of data are

$\hat{\alpha}_{1,1}$	$\hat{\alpha}_{2,1}$	$\hat{\alpha}_{3,1}$
0.29 (0.55)	0.18 (0.38)	3.80 (31.33)

Therefore following the steps described in the previous section, we have generated a new set of data by adding a variable i to one already known, for which $\hat{\alpha}_{i,1}$ is large. For this new set of data,the distribution of ratios is

	ratios					
item	00	01	10	11	slope	R^2(%)
1	1.1	1.8	1.0	1.7	0.22	35.4
2	1.3	1.3	1.7	1.7	0.13	50.0
3	6.0	4.4	3.7	2.8	-0.38	95.0

In this example, the slopes indicate that item 3, which has the largest $\hat{\alpha}_{i,1}$, is the most predictable. They also gives the order of the other two estimates.

An R^2 equal to 95% for item 3 indicates almost perfect correlation between the number of positive responses to items 1 and 2 and the ln(ratios for item 3). This could be expected since item 3 was artificially generated. In this case we cannot relate the pattern of R^2 with the fitting of both, Rasch and logit/probit, models as in the preceding section. If we take a large sample size, item 3 will still be generated in the same way and the R^2 will still be larger than the others, but the Rasch model will not be accepted anymore.

If instead of t'=10 we chop the distribution of N_{ts} at t'=11, and procede as before, we have another new set of data. Fitting a logit/probit model to these data (t'=11), the discrimination parameter estimates and standard deviations (in brackets) are

$\hat{\alpha}_{1,1}$	$\hat{\alpha}_{2,1}$	$\hat{\alpha}_{3,1}$
0.28 (0.53)	0.20 (0.43)	4.38 (35.98)

The analysis in terms of predictability of an item from the others gives the following results

	ratios					
item	00	01	10	11	slope	R^2(%)
1	0.9	0.5	0.8	0.7	-0.13	16.3
2	0.8	0.7	0.5	0.7	-0.03	2.4
3	27.0	18.0	15.5	14.3	-0.32	84.3

As expected, these results agree with the preceding ones for t'=10 in terms of the relation between the size of $\hat{\alpha}_{i,1}$ and the predictability of an item from the others.

Consider the set of data generated from t'=10 as the initial data set. Then we can generate a fourth variable, repeating the same procedure as before.

Fitting a logit/probit model to this new set of data, when chopping the posterior distribution of z_t given x_s at t'=9, then items 3 and 4 have a very large $\hat{\alpha}_{i,1}$, 16.24 and 15.64, while $\hat{\alpha}_{1,1}$ and $\hat{\alpha}_{2,1}$ are equal to 0.30 and 0.33. The difference between the estimates for items 1 and 2 in the initial set of data and these ones is due to the approximation used to obtain the observed frequencies. In this example, four score patterns have frequencies zero, and items 3 and 4 have very extreme distribution of ratios.

3.3- Cancer Knowledge

The original data for the second example was obtained by deleting the second variable, for which $\hat{\alpha}_{2,1}$ was equal to 3.40, from the Lombard and Doering data (Chapter 2, Table 2.5). As we can see below none of the $\hat{\alpha}_{i,1}$ is large, when fitting a logit/probit model to this reduced data set.

Table 6.10- Frequency distribution of the score patterns of the Lombard and Doering's data (Table 2.5) after deleting item 2.

s	134	n	s	134	n
1	000	708	5	101	19
2	100	157	6	110	201
3	001	25	7	011	56
4	010	528	8	111	35

i	$\hat{\alpha}_{i,1}$	$SD(\hat{\alpha}_{i,1})$
1	0.78	0.16
3	0.84	0.19
4	1.45	0.38

$\chi^2 = 3.02$ with 1 d.f.

In the following, Tables 6.11 and 6.12 give the results of applying to the data in Table 6.9 the steps (1) to (4) described in section 3.1 of this chapter, in order to generate item 2.

Table 6.11- Distribution of N_{ts}, the expected number of individuals with score pattern x_s, s=1,..,8 at z_t, t=1,..,16.

t	N_{t1}	N_{t2}	N_{t3}	N_{t4}	N_{t5}	N_{t6}	N_{t7}	N_{t8}
1,2	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.86	0.01	0.00	0.04	0.00	0.00	0.00	0.00
5	11.25	0.34	0.01	0.97	0.00	0.03	0.00	0.00
6	66.03	3.78	0.14	11.26	0.01	0.69	0.03	0.00
7	184.61	19.52	1.15	60.94	0.19	6.90	0.48	0.04
8	247.30	47.88	4.75	156.48	1.43	32.46	3.84	0.62
9	153.28	54.19	9.05	185.37	4.95	70.22	13.99	4.14
10	40.56	26.26	7.41	94.04	7.40	65.22	21.94	11.86
11	3.97	4.75	2.27	17.81	4.20	22.82	13.04	13.03
12	0.12	0.27	0.22	1.07	0.78	2.57	2.53	4.75
13	0.00	0.00	0.00	0.02	0.04	0.09	0.15	0.54
14	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02
15,16	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Total	708	157	25	528	19	201	56	35

As in the preceding example, the next step is to choose $z_{t'}$, which chops the distribution N_{ts}, and add the values for $t \geq t'$. Table 6.11 displays the results for 3 possible chopping points, from which 3 new set of data with 4 variables can be generated.

Table 6.12- Sum of N_{ts}, $t \geq t'$, for every score pattern x_s, s=1,...,8.

t'	1	2	3	4	5	6	7	8
8	444.58	134.35	23.70	454.79	18.80	193.38	55.50	34.61
9	197.28	85.47	18.94	298.31	17.37	160.92	51.66	34.51
10	44.65	31.28	9.89	112.94	12.42	90.70	37.65	30.19

Taking the nearest integer to the frequencies in Table 6.12 and setting them equal to the observed frequencies of x_2 equal to '1', three new set of data are generated, which are displayed in Table 6.13.

Table 6.13- Frequencies of the data set generated by chopping the posterior distribution of z_t given x_s at t' for every score pattern x_s.

s	134	2134	t'=10	t'=9	t'=8
1	000	0000	664	511	263
		1000	44	197	445
2	100	0100	126	72	23
		1100	31	85	134
3	001	0001	15	6	1
		1001	10	19	24
4	010	0010	415	230	73
		1010	113	298	455
5	101	0101	7	2	0
		1101	12	17	19
6	110	0110	110	40	8
		1110	91	161	193
7	011	0011	18	4	1
		1011	38	52	55
8	111	0111	5	1	0
		1111	30	34	35

These three new sets of data are reasonably fitted by a logit/probit model as we can see from Table 6.14, which also displays the parameter estimates.

Table 6.14- Parameter estimates and asymptotic standard deviations (in brackets) from fitting a logit/probit model to the data sets presented in Table 6.12.

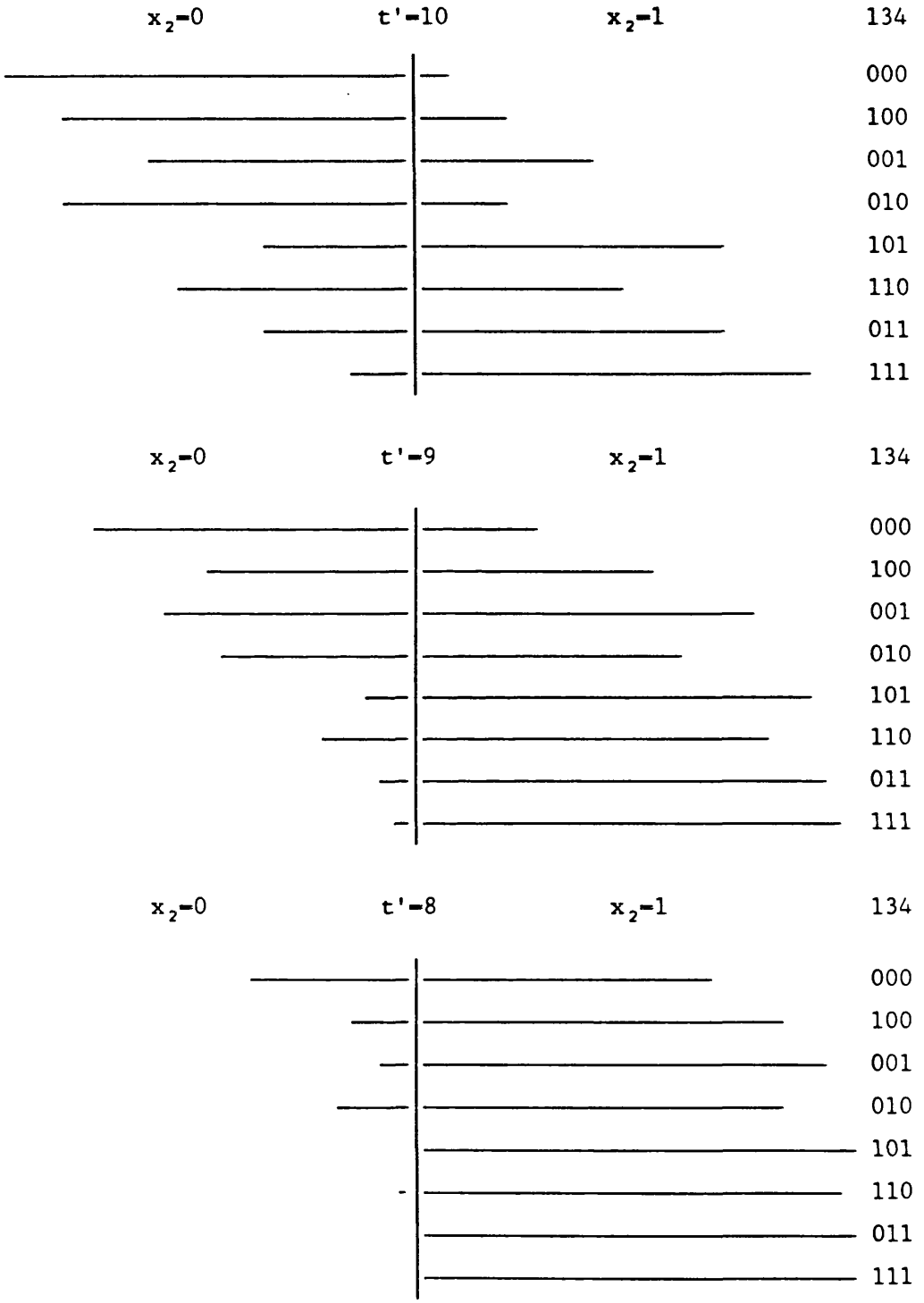
	t'=10		t'=9		t'=8	
i	$\hat{\alpha}_{i,1}$	$\hat{\pi}_i$	$\hat{\alpha}_{i,1}$	$\hat{\pi}_i$	$\hat{\alpha}_{i,1}$	$\hat{\pi}_i$
1	0.78 (0.09)	0.21	0.78 (0.09)	0.21	0.81 (0.11)	0.21
2	13.85(38.50)	0.00	11.23(21.20)	0.45	13.48(40.27)	1.00
3	0.85 (0.09)	0.47	0.85 (0.09)	0.47	0.86 (0.09)	0.47
4	1.45 (0.16)	0.04	1.45 (0.19)	0.04	1.46 (0.25)	0.04
	χ^2=3.08 with 3 d.f.		χ^2=2.96 with 6 d.f.		χ^2=4.08 with 4 d.f.	

As expected, item 2 (generated) has a very large $\hat{\alpha}_{2,1}$, while $\hat{\alpha}_{i,1}$ and $\hat{\pi}_i$ for the remaining items and t'=8, 9 and 10 are the same as for the original data set.

3.4- Relation between the Generation of an Item with Large $\hat{\alpha}_{i,1}$ and its Predictability from the Other Items

We shall discuss the connection between the generation of an item with large $\hat{\alpha}_{i,1}$ and its predictability from the other items, as discussed in the previous section, with the following diagram

Diagram 6.1- Proportion of the observed frequencies of the score patterns $x_S=(x_1,x_3,x_4)$, which is designed to generate x_2 equal to '0' and '1'.



From Diagram 6.1 we can see that since the right side of the posterior distribution of z_t given $x_s=(x_1,x_3,x_4)$ was set to give the observed frequencies of x_2 equal to '1', for t'=10 most of the frequency of $x_s=(0,0,0)$ is allocated to $x_2=0$, on the left side.

As the number of positive responses of x_s increases, the posterior distribution gradually shifts to the right, increasing the proportion of positive responses to item 2, so that when $x_s=(1,1,1)$ most of the frequency is allocated to x_2 equal to 1. This implies that if we take the ratios p/(1-p), where p is the proportion of positive response to item 2 given x_s, some of them will be very extreme. In other words item 2, for which $\hat{\alpha}_{2,1}$ is large, is highly predictable from the remaining items.

Therefore the generation of an item, which has a very large $\hat{\alpha}_{i,1}$, as described in this section is based on the same idea of predictability investigated in the preceding one.

As the chopping point moves to the left, the part of the posterior distribution designed to give the proportion of positive responses to item 2, moves to the right, so that for t'=8 most of the frequencies are allocated to x_2 equal to 1, which produces some score patterns with frequencies zero.

In order to complement the connection between this section and the preceding one, Table 6.15 displays the slopes and R^2 used to measure the predictability of an item from the others. As we have already pointed out in section 1.3 of this chapter, the interpretation of R^2 is not equivalent, since item 2 was generated artificially.

Table 6.15- Comparison between the pattern of $\hat{\alpha}_{i,1}$ and the predictability of an item from the others.

	t'=10			t'=9			t'=8		
i	$\hat{\alpha}_{i,1}$	slope	R^2(%)	$\hat{\alpha}_{i,1}$	slope	R^2(%)	$\hat{\alpha}_{i,1}$	slope	R^2(%)
1	0.78	0.43	41.6	0.78	0.56	53.3	0.81	0.67	41.3
2	13.85	1.38	89.4	11.23	1.29	75.4	13.48	1.26	82.1
3	0.85	0.44	32.5	0.85	0.52	43.2	0.86	0.67	64.2
4	1.45	0.87	48.5	1.45	1.00	50.5	1.46	1.35	83.4

Table 6.15 shows that for t' equal to 9 and 10, item 2, which has a very large $\hat{\alpha}_{2,1}$ and the largest slope, is the most predictable.

For the data generated from taking t'=8, Table 6.15 shows two large slopes indicating that item 2 is as predictable as item 3, even though $\hat{\alpha}_{2,1}$ is very large. This result does not contradict the previous findings, since they were restricted to data sets in which all score patterns occur.

When t'=8, two score patterns have frequencies zero, and in order to obtain the ratio we have to replace 'zero' by another number, which could lead to very different ratios, and therefore, could lead to different slopes. In Table 6.15 for t'=8, the two ratios were obtained replacing 1/2 for the frequency zero and subtracting 1/2 from the other frequency.

Although the ratios cannot be measured precisely, Diagram 6.1 suggests that for t'=8, the distribution of ratios for item 2 is less extreme, since most of the frequencies are allocated to x_2 equal to one.

When extending this study to other data sets, in which at least one of the score patterns had frequency zero, we found the same kind of results as when all score patterns occur.

3.5- Conclusions

In this section we have presented a procedure, based on equation (6.1), under which we can add a (p+1)th variable with any fixed $\hat{\alpha}_{p+1,0}$ and $\hat{\alpha}_{p+1,1}$ to each response vector x_s without altering the previous estimates of $\alpha_{i,0}$ and $\alpha_{i,1}$, i=1,...,p. Using this procedure the covariance matrix of $\hat{\alpha}_{i,v}$'s for i=1,...,p and v=0 or 1, does not increase when the (p+1)th variable is added.

As a particular case of (6.1), we can generate an item with large $\hat{\alpha}_{p+1,1}$, by chopping the posterior distribution of Z given x_s at the same point z_t, for all x_s and applying (6.2). This implies that if we take the ratios, p/(1-p), where p is the proportion of positive response to item p+1 given x_s, some of them will be very extreme. In other words, item p+1, is highly predictable from the remaining items.

Chapter 7

MEASUREMENT of the LATENT VARIABLE

1- *Introduction*

In the general latent trait model (Chapter 1, equation 1.5) it is assumed that the probability of a positive response to an item in a test is a monotonic function of a latent variable Y, representing the trait in question. In fixed effects versions of the model each individual's position on the latent scale is represented by a parameter; in the random effects versions, individuals are supposed to be sampled at random from some population so that their latent position is the value of a random variable.

Considering Y either as a parameter or a variable has given rise to different procedures when looking for more information about Y, after the model has been fitted.

In Educational Testing, where Y is usually treated as a parameter, some work has been done in estimating Y for a given individual and estimating the parameters of the latent distribution; see for example, Andersen and Madsen (1977), Samanthanan and Blumenthal (1978), Lord (1983) and Mislevy (1984).

On the other hand, Bartholomew (1980), treating Y as a random variable uniformly distributed on (0,1) has dealt with the situation by scaling the latent variable, i.e., locating the individuals in a Y-space on the basis of their observed response patterns **x**.

In this chapter, using the latter approach, we concentrate the discussion on the measure of the latent variable, when fitting logit/probit or logit/logit models.

We start by presenting the main results about scaling the latent variable Y in a logit/logit model given by Bartholomew (1980, 1981, 1984). After this, considering the response function for the logit/logit or logit/probit model given by

$$\text{logit}[\pi_i(z)] = \alpha_{i,0} + \alpha_{i,1}z,$$

where $z=H^{-1}(y)$ is logistic or normally distributed, we present some new theoretical results about the relation between the posterior density $h(z|\mathbf{x})$, its mean $E(Z|\mathbf{x})$ and the component score $c_1(\mathbf{x}) = \sum \alpha_{i,1}x_i$. Some findings complement and others contradict Bartholomew's results, depending on the pattern of the $\hat{\alpha}_{i,1}$'s.

Finally we investigate the shape of the posterior density $h(z|\mathbf{x})$ when at least one of the $\hat{\alpha}_{i,1}$ is very large, and we suggest a cluster analysis in the latent-space based on $h(z|\mathbf{x})$.

2- *Theoretical Results for the Relation between $E(Z|\mathbf{x})$ (or $E(Y|\mathbf{x})$) and $\sum \alpha_{i,1}x_i$*

According to Bartholomew (1980, 1981) the scaling of the latent variable Y should be done via the posterior density of y given the score pattern $\mathbf{x}$. Thus, for example, he suggests the mean $E(Y|\mathbf{x})$ (or $E(Z|\mathbf{x})$), which may not be particularly appropriate when the posterior density $h(z|\mathbf{x})$ is highly skewed. He argues that since Y is assumed to be uniform a priori, an individual's y value may be interpreted as his

quantile in the population and the posterior expectation $E(Y|x)$ then seems a natural way of comparing individuals.

Bartholomew (1980) pointed out that for the logit/logit model $E(Y|x)$ is an approximately linear function of the component score $c_1(x) = \sum \alpha_{i,1} x_i$, which can be justified by a Taylor expansion if all $\alpha_{i,1}$'s are small. At the same time, when all $\alpha_{i,1}$'s are equal to 1 and π_i's are equal to 0.5 then the exact value of $E(Y|x)$ is $(1 + c_1(x)) / (2 + A)$, where $A = \sum \alpha_{i,1}$. He also found out from empirical work that the relationship between $E(Y|x)$ and $c_1(x)$ is approximately linear well outside the range of the validity of this later result. We show that this is often false when at least one of the $\hat{\alpha}_{i,1}$ is large (say $\geq 3/\sigma$, where σ is the standard deviation of the latent distribution).

For the logit/logit model, Bartholomew (1984) shows that when π_i and $\alpha_{i,1}$ are fixed, the posterior density $h(y|x)$ depends on x only through the component score $c_1(x)$. And therefore, under this conditions $c_1(x)$ is a Bayesian sufficient statistic of y. This property is not shared, for example, by the probit model used by Bock and Liberman (1970). We shall show that $h(y|x)$ is a function of x only through $\sum \alpha_{i,1} x_i$ if no $\alpha_{i,1}$ is infinity.

Bartholomew (1984) also shows that $E(\Phi(y)|c_1(x))$ is a nondecreasing function of $c_1(x)$ for every nondecreasing function of $\Phi(y)$. In particular, $E(Y|c_1(x))$ or $E(Z|c_1(x))$ is an increasing function of $c_1(x)$. This means that the component score induces a stochastic ordering of the posterior distributions. Thus, for example, the rank of individuals given by $c_1(x)$ is the same as given by $E(Y|x)$ and $E(Z|x)$. Therefore, if we are only interested in the ranking of the individuals on the latent scale, we can use any one of these three

measures, from which the component score is the easiest to be obtained.

Now we give three results, which summarise our findings and are valid for both logit/logit and logit/probit models, that is when Z is logistic or normally distributed.

Result 1

If no $\alpha_{i,1}$ is infinity and two score patterns have the same posterior mean $E(Z|x)$ then they have the same component score $\sum \alpha_{i,1} x_i$ and the same posterior density $h(z|x)$.

Let $h(z|x) = \dfrac{g(x|z)\, h(z)}{f(x)}$.

Then from (1.3)

$$g(x|z) = \prod_{i=1}^{p} \left[\pi_i(z) \right]^{x_i} \left[1-\pi_i(z) \right]^{1-x_i}$$

$$= \prod_{i=1}^{p} \left[\frac{\pi_i(z)}{1-\pi_i(z)} \right]^{x_i} \left[1 - \pi_i(z) \right]$$

$$= \prod_{i=1}^{p} \left\{ \exp \left[\ln \left[\frac{\pi_i(z)}{1-\pi_i(z)} \right] \right] \right\}^{x_i} \left[1 - \pi_i(z) \right]$$

$$= \prod_{i=1}^{p} \left[\exp \left(\alpha_{i,0} + \alpha_{i,1} z \right) x_i \right] \left[1 - \pi_i(z) \right] \qquad (7.1)$$

$$= \frac{\exp(c_1(x)z)\, \exp(c_0(x))\, f(0,z)}{f(x)} \qquad (7.2)$$

where $c_0(x) = \sum_{i=1}^{p} \alpha_{i,0}\, x_i$ and $c_1(x) = \sum_{i=1}^{p} \alpha_{i,1}\, x_i$.

And thus

$$f(x) = f(0)\, \exp(c_0(x))\, M_{Z|0}(c_1(x)) \tag{7.3}$$

where $M_{Z|0}(c_1(x))$ is the moment generating function of the latent variable Z given a zero response on all items $c_1(x)$.

Substituting (7.3) in (7.2), we obtain that the posterior density of z given the score pattern x is

$$h(z|x) = \frac{\exp(c_1(x)z)\, h(z|0)}{M_{Z|0}(c_1(x))} \quad \text{for every score pattern } x. \tag{7.4}$$

From (7.4) and for every score pattern x, the moment generating function of the posterior distribution of Z given x is

$$M_{Z|x}(t) = \frac{M_{Z|0}(c_1(x)+t)}{M_{Z|0}(c_1(x))} \tag{7.5}$$

Therefore from (7.5), the posterior density $h(z|x)$ is a function of x only through the component score $c_1(x)$, if no $\alpha_{i,1}$ is infinity. This result was first given by Bartholomew (1984), when assuming π_i and $\alpha_{i,1}$ are fixed for the logit/logit model.

Furthermore from (7.5)

$$E(Z|x) = \frac{M'_{Z|0}(c_1(x))}{M_{Z|0}(c_1(x))} = \frac{\partial}{\partial t}\left[\log M_{Z|0}(c_1(x))\right]$$

And therefore

$$E(Z|x) = \frac{\partial}{\partial t} K_{Z|0}(t)\Big|_{t=c_1(x)} \tag{7.6}$$

and

$$Var(Z|x) = \frac{\partial^2}{\partial t^2} K_{Z|0}(c_1(x)+t)\Big|_{t=0} \tag{7.7}$$

where $K_{Z|0}$ is the cumulant generating function of $Z|0$ and

$$c_1(x) = \sum_{i=1}^{p} \alpha_{i,1}\, x_i.$$

But

$$\frac{\partial^2}{\partial t^2} K_{Z|0}(t) = \frac{E(e^{zt})\,E(z^2 e^{zt}) - E(z\,e^{zt})\,E(z\,e^{zt})}{E(e^{zt})^2} > 0 \tag{7.8}$$

Since $E(e^{zt})^2 > 0$ and from the Cauchy inequality

$$E\left[ze^{\frac{1}{2}zt}\; e^{\frac{1}{2}zt}\right]^2 < E\left[z^2 e^{zt}\right]\; E\left[e^{zt}\right],$$

since Z is a random variable.

It follows from (7.6) that $E(Z|x)$ is increasing in $\sum_{i=1}^{p} \alpha_{i,1}\, x_i$.

Therefore if for two score patterns x_1 and x_2

$$E(Z|x_1) = E(Z|x_2) \xrightarrow{(7.7)} K'_{z|0}(c_1(x_1)) = K'_{z|0}(c_1(x_2))$$

$$\xrightarrow{(7.6)} c_1(x_1) = c_1(x_2) \quad \text{since } E(Z|x) \text{ is increasing in } c_1(x).$$

Finally, result 1 follows from (7.5) and (7.6).

Result 2

If the posterior density $h(z|x_s)$ is normal, then its mean $E(Z|x_s)$ is linear in the component score $c_1(x_s)$.

If the mean $E(Z|x_s)$ is linear in the component score $c_1(x_s)$, then the posterior density $h(z|x_s)$ will be close to the normal distribution.

Proof:

If $h(z|x_s)$ is normal then the mean $E(Z|x_s)$ is linear in the component score $c_1(x_s)$ from (7.6).

If, on the other hand, the posterior mean is linear in the component score, then for some fixed a_0 and a_1,

$E(Z|x_s) = a_0+a_1c_1(x_s)$, so from (7.6) for all score patterns x_s

$$a_0+a_1c_1(x_s) = \frac{\partial}{\partial t} K_{z|0}(t) \Big|_{t=c_1(x_s)}.$$

For typical choices of the parameters $\alpha_{i,1}$ this will mean that there are distinct values t_s, $s=1,\ldots,2P$ for which

$$a_0+a_1t_s = \frac{\partial}{\partial t} K_{z|0}(t) \Big|_{t=t_s}$$

This does not quite amount to the property of linearity in t which would imply a normal distribution for the posterior $h(z|x_s)$, but it comes as close as is possible to that with $K_{z|0}(t)$ determined only at a finite number of values. Fixing $K'_{z|0}(t_s)$ leads to $K_{z|0}(t_s)$ having the value appropriate for a normal distribution, for all s. If p is large, the posterior distribution is therefore constrained to be close to the normal distribution.

Bartholomew(1984) shows that if Z has a standard logistic distribution, then in some circumstances the relation between the posterior mean and the component score is linear. He conjectures that an approximate linear relation is often valid for such prior distribution of Z. The results here show that one may think of normality of posterior distributions instead of linearity.

An application of this result can be seen when fitting a logit/probit model to the Law School Admission Test, section 6, (LSAT VI), as shown below.

Law School Admission Test, Section 6

LSAT VI consists of 5 items taken by 1000 individuals designed to measure a single latent variable. This data set is fitted well (α=5%) by a logit/probit model with parameter estimates $\hat{\alpha}_{i,1}$, i=1,...,5, equal to 0.83, 0.72, 0.89, 0.69 and 0.66, respectively.

As we have already analysed in Chapter 4, section 8, it is also fitted by the Rasch model.

Figure 7.1 shows clearly that the posterior mean E(Z|x) is a linear function of the component score $c_1(x) = \sum \alpha_{i,1} x_i$.

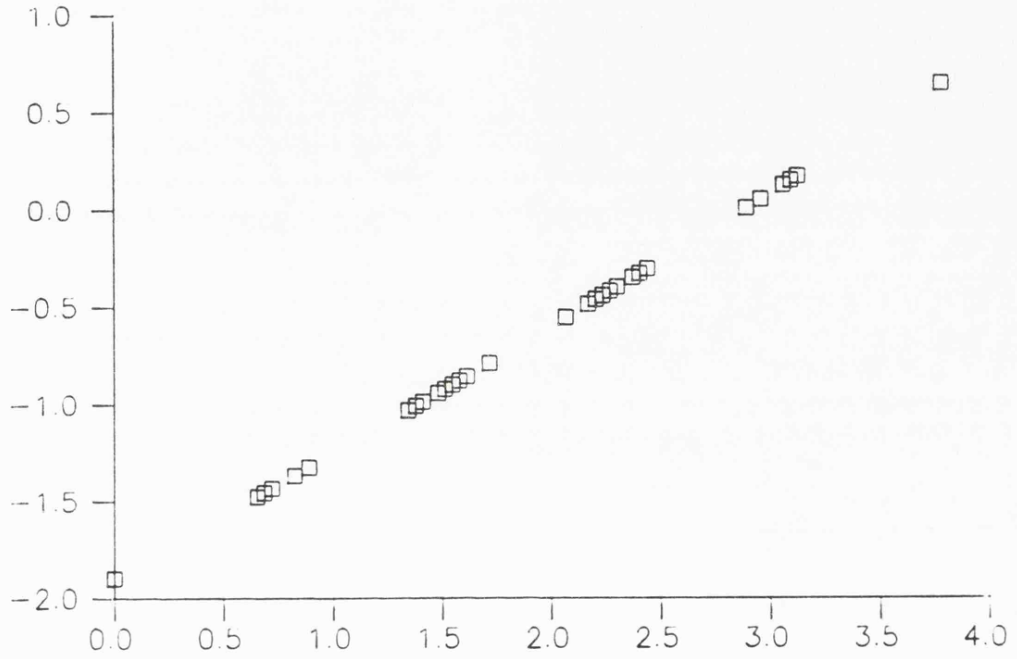


Figure 7.1- Relation between E(Z|x) and $\sum \alpha_{i,1} x_i$ when fitting a logit/probit model to the LSAT VI.

Table 7.1- Estimates of E(Y|x), E(Z|x) and the component score $\sum \alpha_{i,1} x_i$ when fitting a logit/probit model to the LSAT VI.

group	E(Y|x)	E(Z|x)	$\sum \alpha_{i,1} x_i$	$\sum x_i$
1	0.007	-1.90	0.00	0
2	0.12 to 0.15	-1.47 to -1.32	0.66 to 0.89	1
3	0.21 to 0.27	-1.03 to -0.79	1.34 to 1.72	2
4	0.33 to 0.41	-0.55 to -0.30	2.07 to 2.44	3
5	0.50 to 0.55	0.01 to 0.17	2.89 to 3.13	4
6	0.69	0.64	3.79	5

From Figure 7.1 we can see that the score patterns are distributed into 6 groups along the line $-1.92 + 0.67\ c_1(\mathbf{x})$. Table 7.1 shows that they correspond to the 6 different values assumed by $\sum x_i$. As the number of positive responses increases by one unit, both posterior means, $E(Z|x)$ and $E(Y|x)$, and the component score $c_1(\mathbf{x})$ jump to higher values.

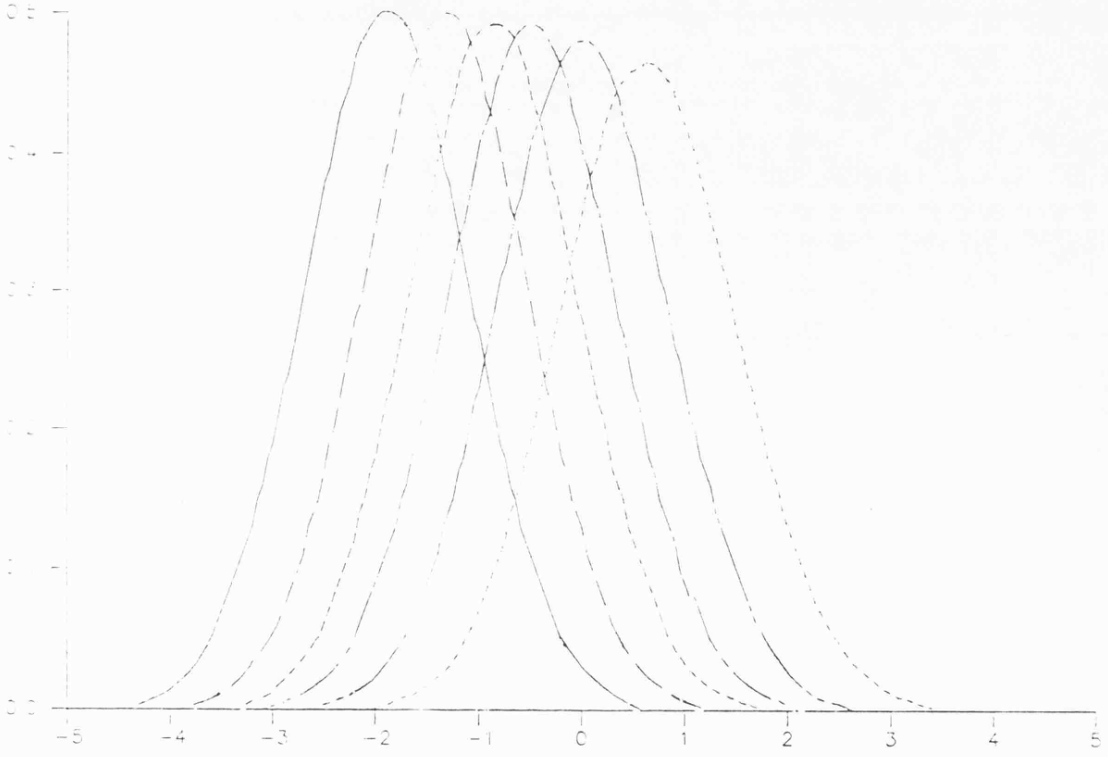


Figure 7.2- Posterior densities $h(z|x)$ when fitting a logit/probit model to the LSAT VI, for the score patterns '00000', '01000', '00101', '01101', '10111' and '11111'.

As $E(Z|x)$ is a linear function of $c_1(\mathbf{x})$ then from result 2 and for every score pattern $\mathbf{x}_s$, s=1,...,32, the posterior density $h(z|\mathbf{x}_s)$ is approximately normal. Besides as the discrimination parameter estimates $\hat{\alpha}_{i,1}$ are nearly the same for all items, the posterior distributions have approximately the same variance (Figure 7.2).

Result 3

For the logit/probit (or logit/logit) model the posterior density $h(z|x)$ is not a function of x through $\sum_{i=1}^{p} \alpha_{i,1} x_i$ if at least one of the $\alpha_{i,1}$'s is equal to infinity.

Proof:

Assume that $\alpha_{1,1}$ is equal to infinity so that

$$\pi_1(z) = \begin{cases} 0 & \text{if } z \leqslant z_0 \\ 1 & \text{if } z > z_0 \end{cases}$$

Then

$$g(x|z) = \prod_{i=1}^{p} [\, \pi_i(z) \,]^{x_i} [\, 1 - \pi_i(z) \,]^{1-x_i}$$

$$= \prod_{i=1}^{p} [\, \pi_i(z) \,]^{x_i} [\, 1 - \pi_i(z) \,]^{1-x_i} * \begin{cases} 0 & \text{if } \begin{cases} z \leqslant z_0 \text{ and } x_1=1 \\ z > z_0 \text{ and } x_1=0 \end{cases} \\ 1 & \text{if } \begin{cases} z \leqslant z_0 \text{ and } x_1=0 \\ z > z_0 \text{ and } x_1=1 \end{cases} \end{cases}$$

From (7.1), $g(x|z)$ can also be written as

$$= \exp\left[\sum_{i=2}^{p} \alpha_{i,0} x_i + z \sum_{i=2}^{p} \alpha_{i,1} x_i \right] \prod_{i=2}^{p} [1-\pi_i(z)] * \begin{cases} 0 & \text{if } \begin{cases} z \leqslant z_0 \text{ and } x_1=1 \\ z > z_0 \text{ and } x_1=0 \end{cases} \\ 1 & \text{if } \begin{cases} z \leqslant z_0 \text{ and } x_1=0 \\ z > z_0 \text{ and } x_1=1 \end{cases} \end{cases} \tag{7.10}$$

Substituting $g(\mathbf{x}|z)$ given by (7.10) in $h(z|\mathbf{x}) = \dfrac{g(\mathbf{x}|z)\, h(z)}{f(\mathbf{x})}$,

it follows that $h(z|\mathbf{x})$ is <u>not</u> a function of $\mathbf{x}$ through $\sum_{i=1}^{p} \alpha_{i,1}\, x_i$ if at least one of the $\alpha_{i,1}$'s is equal to infinity.

3- Applications showing the *Relation between E(Z|x) and* $\sum \alpha_{i,1}x_i$, *when at least one of the* $\hat{\alpha}_{i,1}$*'s is large*

One of the consequences of result 3 is that the relation between the posterior mean $E(Z|\mathbf{x})$ and the component score $c_1(\mathbf{x})$ may not be linear, if at least one of the $\hat{\alpha}_{i,1}$'s is large (say $\geq 3|\sigma$, where σ is the standard deviation of the latent distribution). This situation is illustrated using four tests with 6 to 40 items, and different number of large $\hat{\alpha}_{i,1}$.

3.1- Test 11A (Ireland, items 1 to 6, 8 to 10)

The data for this example was obtained from Test 11A (Chapter 5, section 2.1), when considering only items 1 to 6, 8 to 10 and children studying in Irish schools. The sample size is 73 and the number of different score patterns is 45.

The parameter estimates $\hat{\alpha}_{i,1}$ when fitting a logit/probit model are given in Table 5.5. Recall that all $\hat{\alpha}_{i,1}$'s are smaller than 2.26, except $\hat{\alpha}_{6,1}$, which is equal to 16.38.

The relation between the component score $c_1(\mathbf{x})$ and $E(Z|\mathbf{x})$, for all different score patterns, is given in Figure 7.3 below.

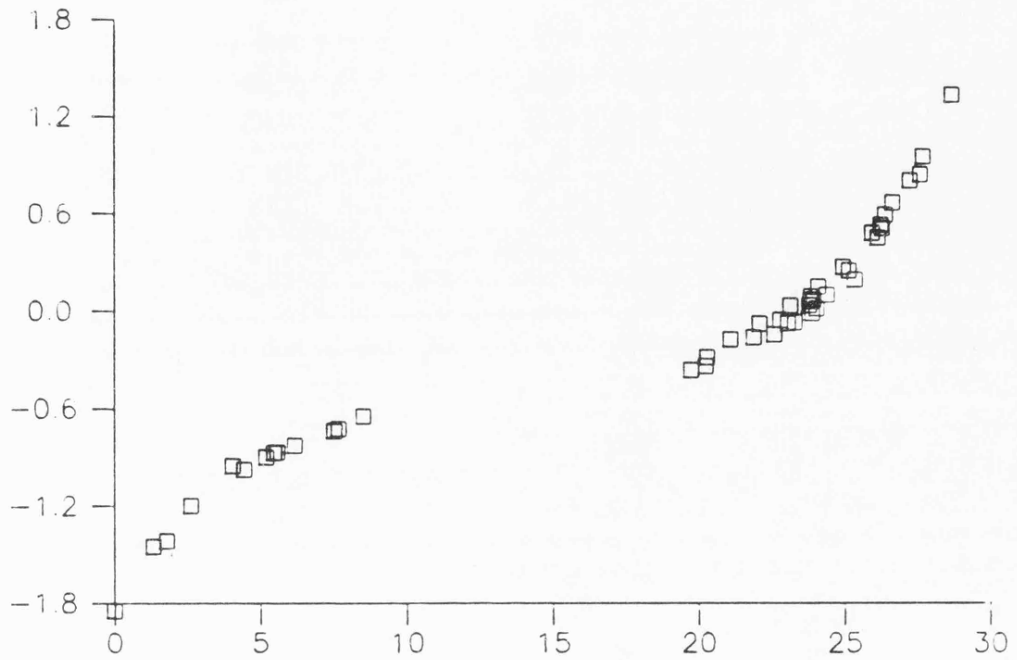


Figure 7.3- Relation between E(Z|x) and $\sum \alpha_{i,1} x_i$ when fitting a logit/probit model to Test 11A (Ireland items 1 to 6, 8 to 10).

The ranking of the individuals according to to the component score or E(Z|x) (not presented here) shows 2 distinct groups: the first one formed by those individuals, who have answered '0' to item 6, followed by the remaining ones, who have answered '1' to this item. This separation is shown in Figure 7.1, where the component score jumps from 8.51 to 19.74, while E(Z|x) practically does not change (-0.50 and -0.47). The same is true for E(Y|x), which assumes values from 0.31 to 0.32. This means that according to the expected values item 6 does not contain any additional information about the latent variable, which is not already in the remaining items. On the other hand, the component scores, which are strongly dependent on $\hat{\alpha}_{i,1}$ instead of h(z|x), give the opposite information.

Therefore, for this test in which one of the items has large $\hat{\alpha}_{i,1}$, the component score does not provide reliable information about the clustering of the individuals in the latent scale, except for the right ranking.

3.2- Test 11A (Ireland, items 1 to 10)

This data set corresponds to the previous one, when including item 7. As before, the parameter estimates $\hat{\alpha}_{i,1}$, when fitting a logit/probit model are given in Chapter 5, Table 5.5. In this case, items 6 and 7 are the only ones with large $\hat{\alpha}_{i,1}$ (16.72 and 12.65).

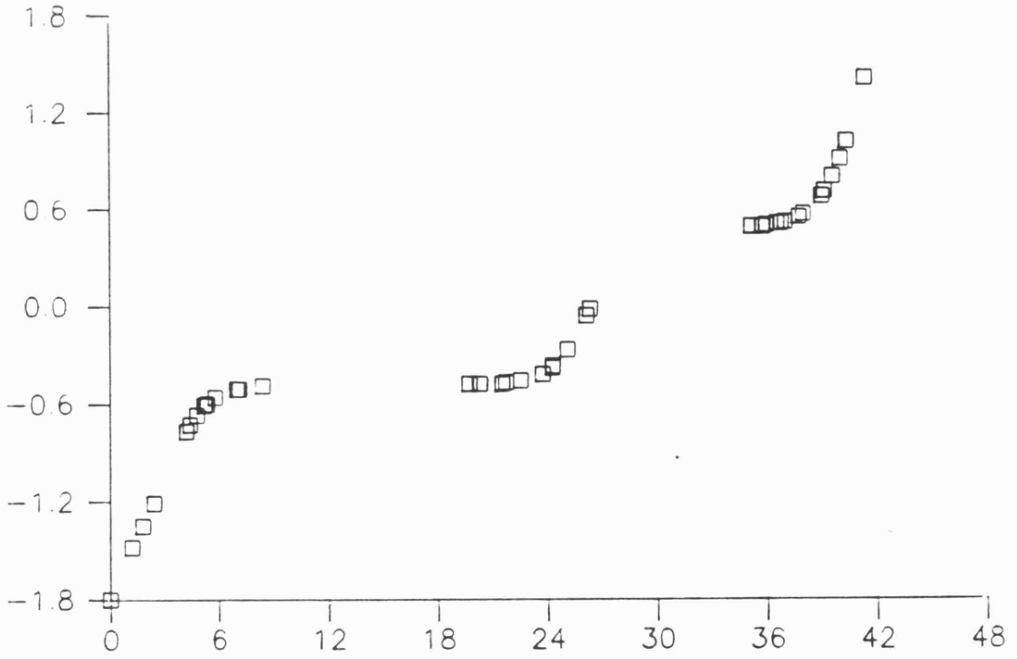


Figure 7.4- Relation between E(Z|x) and $\sum \alpha_{i,1} x_i$ when fitting a logit/probit model to Test 11A (Ireland, items 1 to 10).

The ranking of the individuals according to the component score or $E(Z|x)$ shows 3 groups, depending on the answers to items 6 and 7. The first group is formed by the 17 individuals who have answered '00'; it is followed by the 12 individuals who have answered '10', and finally the third group of the remaining 44 individuals who have answered '11' to items 6 and 7.

The separation between groups is shown through the two sudden great increases on the values of the component scores (from 8.41 to 19.73 and from 26.34 to 35.19). On the other hand, $E(Z|x)$ presents two different results: practically does not change (-0.49 to -0.48) between the first and second groups, but jumps from -0.02 to 0.49 between the second and the third groups.

These results suggest that item 7 contains information about the latent variable, which is not in the remaining items. This is shown by a much higher position on the latent scale according to the posterior mean $E(Z|x)$ for the individuals who have answered '1' instead of '0' to item 7. At the same time, the fact that $E(Z|x)$ practically does not change when x_6=0 changes to x_6=1 (separation between the first and second groups), this suggests that item 6 does not contain any additional information about the latent variable. These results give evidence that although both items 6 and 7 have large $\hat{\alpha}_{i,1}$'s, they contain different amount of information about the latent variable.

Nevertheless, there are some contradictions:

(1) The probability of a positive response to item 7 for a median individual, $\hat{\pi}_7(z=0)$, is equal to 0.04 while the observed proportion of positive responses is 0.58. Usually, we neither expect nor observe such a big difference between these two quantities. For item 6,

$\hat{\pi}_6(z=0)$ is equal to 0.99, while the observed proportion of positive responses is equal to 0.74, which is within the expected difference.

(2) As shown in Chapter 5, section 2.1, deleting item 7 from this set of data the pattern of $\hat{\alpha}_{i,1}$ almost does not change. This suggests that item 7 does not contain any additional information about the latent variable.

We suspect that these contradictory results are due to sampling error, since the sample size is small (73) for a test with 9 or 10 items.

Figure 7.4 also shows that the relation between $E(Z|\mathbf{x})$ is not even linear within the groups. The flat parts of the curve corresponds to clusters of individuals with nearly the same expected value $E(Z|\mathbf{x})$, but different component scores. As the responses to items 6 and 7 within each group are fixed, the differences between the expected values $E(Z|\mathbf{x})$ or the component scores are due to the remaining items.

3.3- Test 12

The description of Test 12 and the parameter estimates $\hat{\alpha}_{i,1}$ when fitted by logit/probit model are given in Chapter 5, Table 5.3. Recall that Test 12 is formed by 18 items, for which all $\hat{\alpha}_{i,1} \leq 2.10$, except $\hat{\alpha}_{15,1}$ and $\hat{\alpha}_{16,1}$ (4.50 and 4.39) and the sample size is 502.

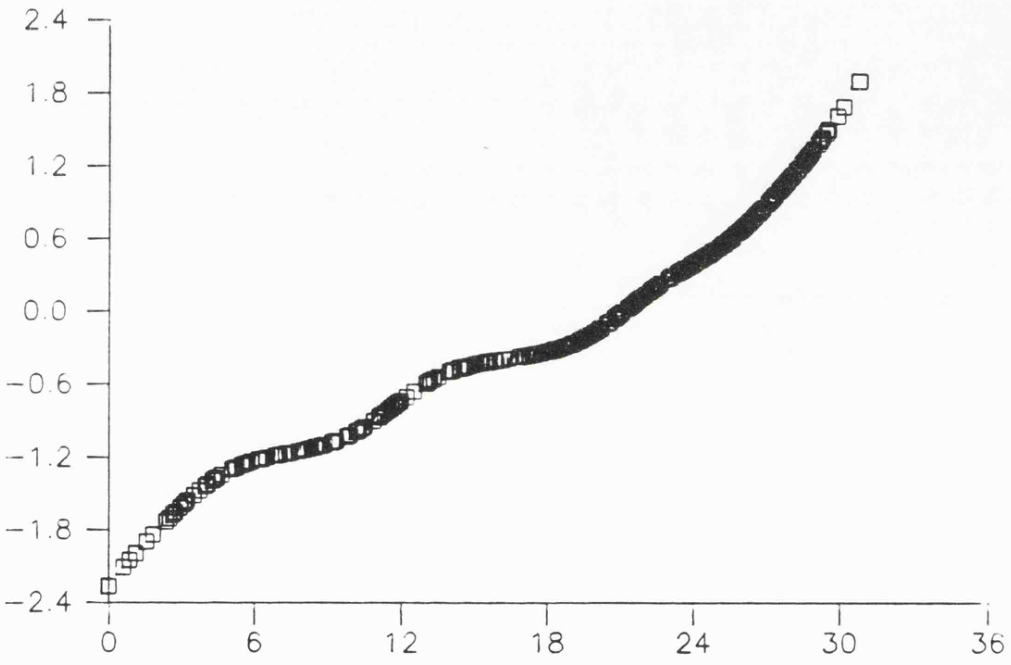


Figure 7.5- Relation between $E(Z|x)$ and $\sum \alpha_{i,1} x_i$ when fitting a logit/probit model to Test 12.

The relationship between posterior mean $E(Z|x)$ and the component score $c_1(x)$ is shown in Figure 7.4, which is complemented by Table 7.2. The dark parts of the curve represent great concentration of individuals with different score patterns in a small range of $E(Z|x)$ and $c_1(x)$ values.

Table 7.2- Estimates of E(Y|x), E(Z|x) and the component score $\sum \alpha_{i,1} x_i$ when fitting a logit/probit model to Test 12.

E(Y|x)	E(Z|x)	$\sum \alpha_{i,1} x_i$	$\sum x_i$	n (%)
[0.02;0.10)	[-2.26;-1.35)	[0.00; 4.65)	0 to 4	51 (10)
[0.10;0.20)	[-1.35;-0.90)	[4.65;10.91)	3 to 7	55 (11)
[0.20;0.30)	[-0.90;-0.55)	[10.91;13.36)	5 to 10	25 (5)
[0.30;0.40)	[-0.55;-0.26)	[13.36;19.13)	6 to 12	71 (14)
[0.40;0.50)	[-0.26; 0.00)	[19.13;21.16)	9 to 12	45 (9)
[0.50;0.60)	[0.00; 0.27)	[21.16;22.97)	10 to 14	34 (7)
[0.60;0.70)	[0.27; 0.57)	[22.97;25.30)	12 to 15	77 (15)
[0.70;0.80)	[0.57; 0.94)	[25.30;27.22)	14 to 16	65 (13)
[0.80;0.90)	[0.94; 1.48)	[27.22;30.91)	15 to 17	67 (13)
[0.90;0.95]	[1.48; 1.91]	[30.91;32.88]	17 to 18	12 (2)

Figure 7.5 shows that the relation between E(Z|x) and $c_1(x)$ is linear only for a partition of E(Z|x) in 5 specific sections. Each one of the first 4 sections corresponds approximately to the first 4 intervals for E(Z|x) given in Table 7.2.

In the first interval ($-2.26 \leqslant E(Z|x) < -1.35$) we observe that 98% of the individuals have answered '0' to both items 15 and 16.

In the second and fourth intervals, there is a greater change in $c_1(x)$ than in E(Z|x), which is shown by two slightly flat sections. The highest proportion of answers to items 15 and 16, in the second interval 52.71% to '00', while in the fourth interval is 70.4% to '11'.

In the third interval, all individuals answered '1' to at least one of the items 15 and 16, and the higher proportion of patterns is 44% to '11'.

Considering all the intervals together, for which $E(Z|\mathbf{x}) \geq -0.26$ or $X \geq 19.13$ the relation between these two measures is linear and 98.3% of the individuals have answered '11' to items 15 and 16.

From these results we can conclude that the non-linearity between the posterior mean $E(Z|\mathbf{x})$ and the component score $c_1(\mathbf{x})$ over all values assumed by them is due to the whole score patterns, instead of only due to the items with large $\hat{\alpha}_{i,1}$ (i=15,16).

Consider that the actual values of $\hat{\alpha}_{15,1}$ and $\hat{\alpha}_{16,1}$ are infinity, and therefore, $g(\mathbf{x}|z)$ can be written as (7.10). Now Figure 7.6, instead of Figure 7.5, shows the relation between $E(Z|\mathbf{x})$ and the component score, which is also not linear.

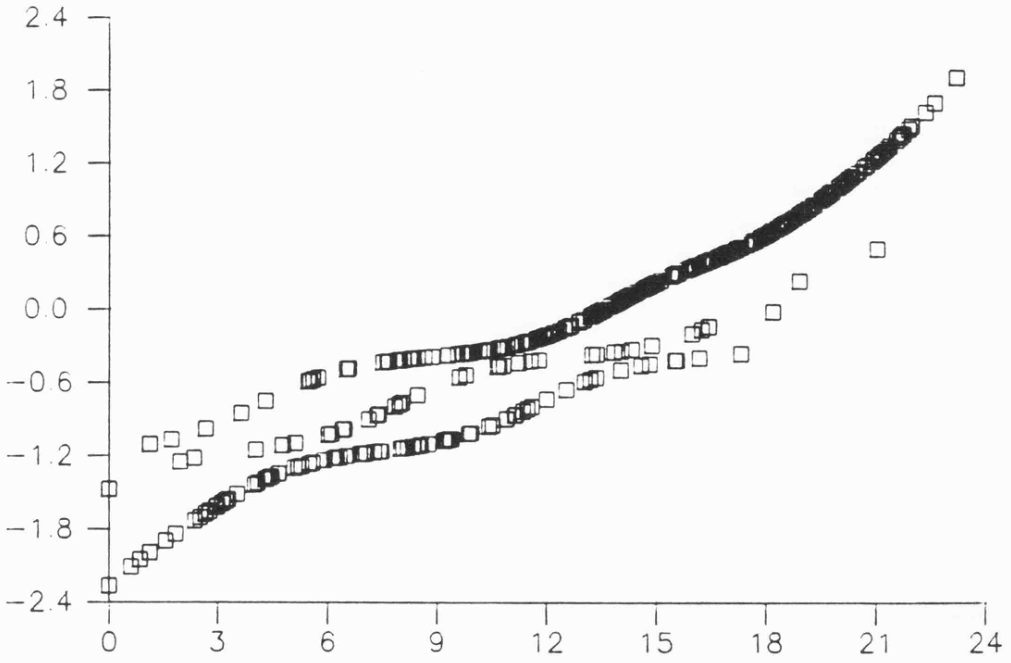


Figure 7.6- Relation between $E(Z|\mathbf{x})$ and $\sum \alpha_{i,1} x_i$, assuming $\alpha_{15,1}$ and $\alpha_{16,1}$ equal to infinity, when fitting a logit/probit model to Test 12.

Figure 7.6 shows roughly three curves, which one corresponding to an specific pattern for '$x_{15}x_{16}$', the answers to the items with $\alpha_{i,1}$ equal to infinity. From the top to the bottom, the first curve is given by the 359 score patterns with '$x_{15}x_{16}$' = '11', the second one by the 39 and 2 score patterns with '10' and '01', and finally, the last one for the 106 patterns with '00'.

Comparing Figures 7.5 and 7.6 we can conclude that for a specific answer to the items with $\alpha_{i,1}$ equal to infinity, the relation between $E(Z|x)$ and the component score $c_1(x)$ is closer to linearity than when taking $c_1(x)$ over all items and $\alpha_{i,1}$'s not equal to infinity.

3.4- Test 13

Test 13 was also applied by the National Foundation of Education Research in order to measure the reading ability of pupils of aged 11 in 1983. The sample size was 498 and the test length 40 items. The distribution of the discrimination parameter estimates $\hat{\alpha}_{i,1}$, i=1,...,40, when fitting a logit/probit model may be given by

$\hat{\alpha}_{i,1}$	count
[0.31; 1.00)	10
[1.00; 2.00)	20
[2.00; 3.00)	4
> 3.00	6

Therefore the fitting of a logit/probit model to Test 13 (length 40) provides six parameter estimates $\hat{\alpha}_{i,1}$ bigger than 3.0.

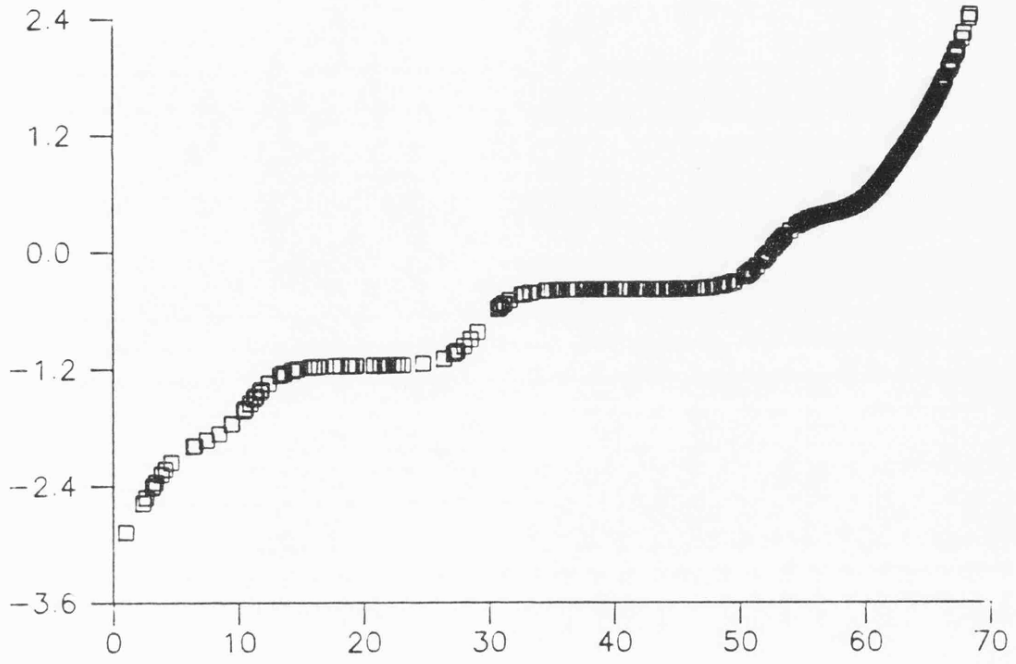


Figure 7.7- Relation between E(Z|x) and $\sum \alpha_{i,1} x_i$ when fitting a logit/probit model to Test 13.

Figure 7.7 shows that the relation between E(Z|x) and the component score $c_1(x)$ is not linear. As in Figure 7.5 , the dark parts of the curve represent great concentration of individuals with different score patterns in a small range of E(Z|x) and $c_1(x)$ values.

Table 7.3 below was constructed in such way that it reflects the different aspects of the relationship between E(Z|x) and $c_1(x)$ displayed in Figure 7.7. Thus, for example, the second and fourth intervals represent the two flat parts of the curve, in which E(Z|x) remains approximately constant, while $c_1(x)$ increases significantly. In the second interval, for 21 individuals with different score patterns, E(Z|x) ranges only from -1.18 to -1.09, while $c_1(x)$ increases significantly from 15.72 to 26.42. In the fourth interval,

for a large number of individuals (75) with different score patterns, $E(Z|x)$ remains almost constant (-0.40 to -0.38), while $c_1(x)$ increases from 34.61 to 46.86.

Table 7.3- Estimates of $E(Y|x)$, $E(Z|x)$ and the component score $\sum \alpha_{i,1} x_i$ when fitting a logit/probit model to Test 13.

$E(Y\|x)$	$E(Z\|x)$	$\sum \alpha_{i,1} x_i$	$\sum x_i$	n (%)
[0.005;0.120)	[-2.87;-1.18)	[1.02;15.72)	1 to 13	30 (6)
[0.120;0.143)	[-1.18;-1.09)	[15.72;26.42)	9 to 18	21 (4)
[0.143;0.346)	[-1.09;-0.40)	[26.42;34.61)	13 to 21	21 (4)
[0.346;0.353)	[-0.40;-0.38)	[34.61;46.86)	17 to 30	75 (15)
[0.353;0.400)	[-0.38;-0.26)	[46.86;50.49)	21 to 30	15 (3)
[0.400;0.510)	[-0.26; 0.02)	[50.49;52.71)	22 to 27	20 (4)
[0.510;0.601)	[0.02; 0.26)	[52.71;54.62)	24 to 27	10 (2)
[0.601;0.701)	[0.26; 0.56)	[54.62;60.26)	25 to 32	118 (24)
[0.701;0.801)	[0.56; 0.92)	[60.26;62.56)	30 to 36	85 (17)
[0.801;0.900)	[0.92; 1.42)	[62.56;65.10)	31 to 37	63 (13)
[0.900;0.982]	[1.42; 2.43]	[65.10;68.53]	34 to 39	40 (8)

The curve also changes its slope significantly, but is less flat than in the second and fourth intervals, when $E(Z|x)$ ranges from 0.26 to 0.56 and $c_1(x)$ from 54.62 to 60.26. In this interval, there is a great concentration of individuals (24% against the expected 10%), all of them with different score patterns.

The investigation of reasons why flat parts occur led us to look at the relation between the distribution of the number of positive responses given to the 6 items with large $\hat{\alpha}_{i,1}$ ($\geq$3.0) and the slope of the curve. A selection of the results is displayed in Table 7.4.

Table 7.4- Frequency distribution of the number of positive responses given to the six items with $\hat{\alpha}_{i,1} \geq 3.0$ for some intervals of E(Z|x).

E(Z|x)	0	1	2	3	4	5	6	total
[-2.87;-1.18)	25	5						30
[-1.18;-1.09)	9	8	3	1				21
[-1.09;-0.40)	5	4	2	9	0	1		21
[-0.40;-0.38)	0	3	4	49	4	9	6	75
[0.26; 0.56)	0	0	0	2	21	44	51	118

Table 7.4 shows that there is a great combination of possible results for the 6 items with large $\hat{\alpha}_{i,1}$, even in the flat parts of the curve. Thus for example, where 75 score patterns have approximately the same E(Z|x), -0.40 to -0.38, the only possible result that does not happen is all 6 items answered '0'. This means that the score patterns are not concentrated on a specific configuration for the items with large $\hat{\alpha}_{i,1}$.

Moreover, in the flat parts of the curve we found score patterns with the same response to the 6 items with large $\hat{\alpha}_{i,1}$, have component score values significantly different. For example, two score patterns, in which all these 6 items were answered '0', were associated to either a component score equal to 15.71 or 24.77 for nearly the same expected value (-1.18 and -1.14).

This implies, that at least for these score patterns, the greater relative difference between the component scores than between the posterior means E(Z|x) is not due to the items with large $\hat{\alpha}_{i,1}$.

For an expected number of individuals equal to 10% and E(Z|x) between 0.26 and 0.56, it was observed 23.7% of the sample, of whom 90.5% have answered '1' to five or to the six items with large $\hat{\alpha}_{i,1}$. From this point E(Z|x) and $\sum \alpha_{i,1}x_i$ increases faster and most of the individuals have answered '1' to the 6 items with large $\hat{\alpha}_{i,1}$.

These results combined with those from Table 7.4 indicate that when at least one of the $\hat{\alpha}_{i,1}$'s is very large, for some score patterns it may occur that the posterior mean practically does not change while the component score increases significantly, even when the response to the items with large $\hat{\alpha}_{i,1}$ is fixed.

4- *Distribution of the individuals on the latent scale according to h(z|x)*

Very often, in practice, we are not only interested in the ranking of the individuals, which is obtained either from the component scores or from the posterior means E(Z|x) or E(Y|x). Thus, for example, in Educational Testing, we may be interested in comparing the lower with the higher abitity group of individuals. The criterion for the distribution (allocation) of the respondents in groups is usually based on an arbitrary percentage, for example 20%.

If we know the distribution of the individuals along the latent scale, then we can use this information to partition the sample in groups. One way to do this is to use the information given by the posterior density h(z|x) or even the mean E(Z|x) (or E(Y|x)).

If we intend to use the mean E(Z|x) (or E(Y|x)) as the measure of comparison between the position of the individuals on the latent scale

then we must have information about the shape of $h(z|x)$, at least in terms of skweness and spread.

Let us consider two individuals with different score patterns x_1 and x_2 and the posterior densities $h(z|x_1)$ and $h(z|x_2)$, which are not skew and have nearly the same dispersion. If $h(z|x_1)$ and $h(z|x_2)$ have roughly the same mean then x_1 and x_2 lead to the same beliefs about the value of Z.

In these situations the mean $E(Z|x)$ (or $E(Y|x)$ is a reliable measure to compare individuals according to their position on the latent scale.

The main goal of this section is to present the results from the investigation of the shape of the $h(z|x)$ we have found so far in practice. This will be done using two real data sets for tests with 18 and 40 items, for which the fittings of a logit/probit model yield two and six $\hat{\alpha}_{i,1}$'s bigger than 3.0.

4.1- Test 12

As mentioned in the previous section, Test 12 has 18 items and was answered by 502, which have provided 417 different score patterns. Therefore for each one of these 417 score patterns there is one posterior density $h(z|x)$.

In order to investigate the shape of the posterior densities $h(z|x)$ and how they are distributed along Z, we have we have selected a representative sample of observed $h(z|x)$'s, from which we have chosen to display here the following three sets (Figures 7.8 to 7.10).

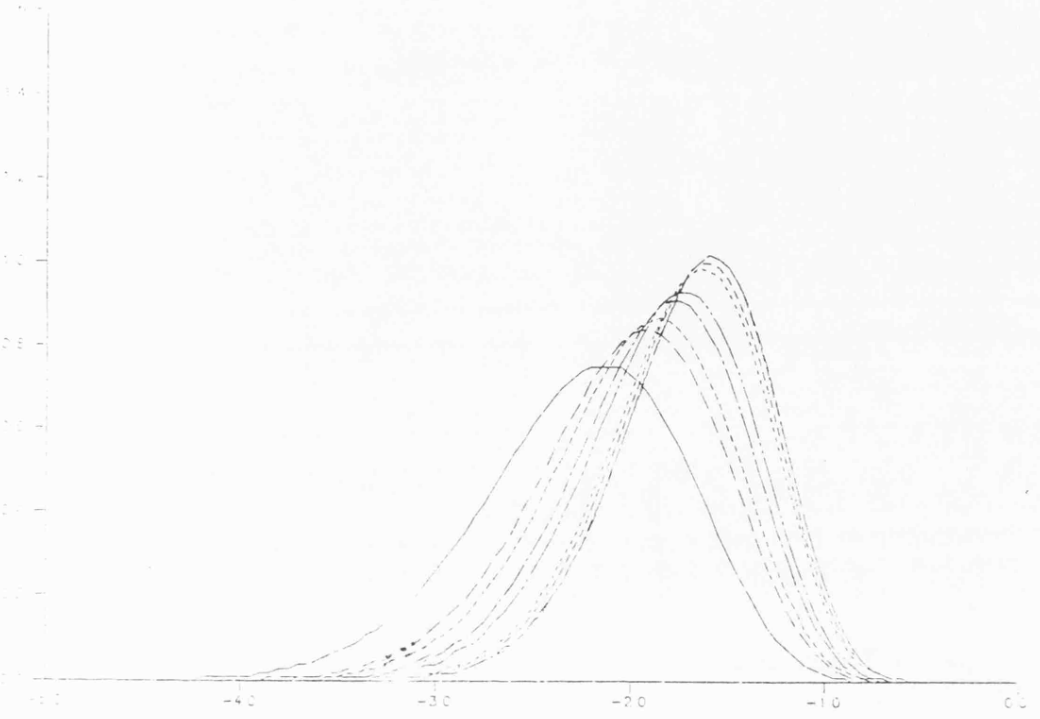


Figure 7.8- Posterior densities h(z|x) for the first ten different score patterns of test 12, for which $-2.26 \leqslant E(z|x) \leqslant -1.67$.

Figure 7.8 displays the posterior distributions h(z|x) for the first 10 different score patterns with the smallest E(Y|x) (or E(Z|x)). For these sets of h(z|x), the mean E(Z|x) assumes values from -2.26 to -1.67 while E(Y|x) ranges from 0.02 to 0.06. For these score patterns most of the items were answered '0', including items 15 and 16 for which $\hat{\alpha}_{i,1}$ are large.

The continuous line represents h(z|x) when an individual has answered '0' to all items and corresponds to the lowest observed ability. It also presents the biggest dispersion and is skewed to the left. As E(Y|x) (or E(Z|x)) increases, h(z|x) becomes less skew, less spread and similar posterior means represent individuals with nearly the same h(z|x).

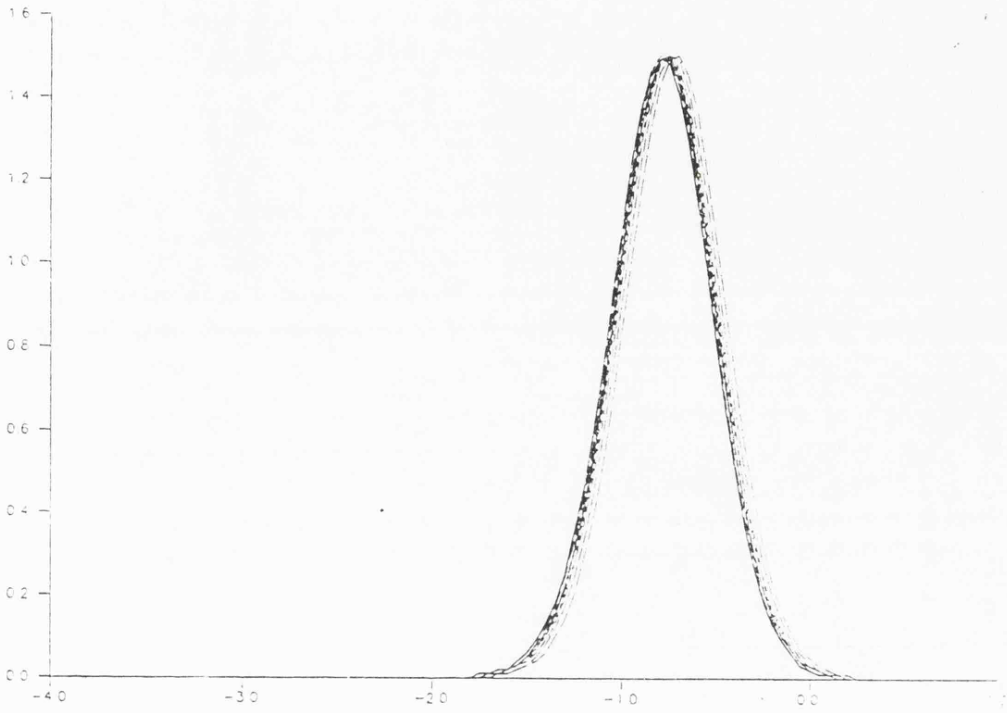


Figure 7.9- Posterior densities h(z|x) for some score patterns of Test 12, for which -0.81 ⩽ E(Z|x) ⩽ -0.66.

Figure 7.9 displays the h(z|x) for ten different score patterns, for which E(Z|x) assumes values from -0.81 to -0.66 (or E(Y|x) ranges from 0.22 to 0.27). The normal probability plots have shown that h(z|x)'s are approximately normal distributions with the same dispersion. This implies that the difference between h(z|x) is only in terms of location and these individuals lead to approximately the same beliefs about the value of Z. This was also found to be true for score patterns with similar posterior means, but which are not located in the ten higher observed positions on the latent scale.

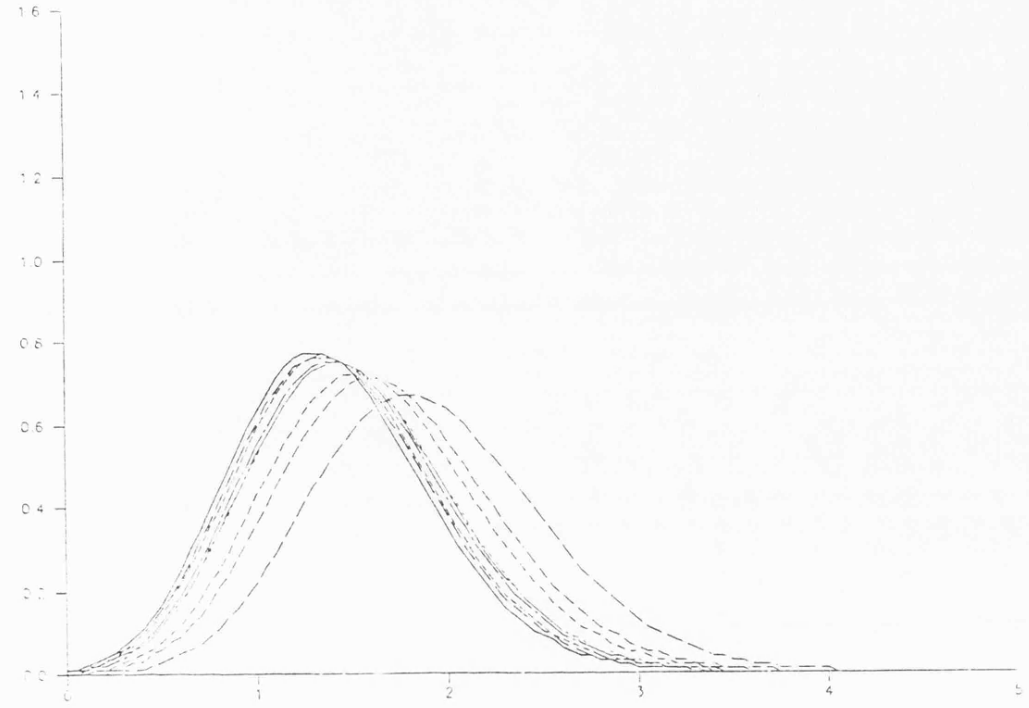


Figure 7.10- Posterior densities h(z|x) for the last different ten score patterns of Test 12, for which $1.39 \leqslant E(Z|x) \leqslant 1.90$.

Figure 7.10 displays the ten observed score patterns, which provide the ten different largest posterior means ($1.39 \leqslant E(Z|x) \leqslant 1.90$ and $0.89 \leqslant E(Y|x) \leqslant 0.95$). For these score patterns most of the items were answered '1', including as expected the items with large $\hat{\alpha}_{i,1}$. Now the posterior densities h(z|x) are slightly skew to the right and the dispersion is increasing as E(Y|x) (or E(Z|x)) increases.

From Figures 7.8 to 7.10 and many others not represented here, we can conclude that the means E(Z|x) (or E(Y|x)) represent very well the position of the individuals on the latent scale, since the posterior distribution is approximately normal. There are some restrictions on

the extremes, where h(z|x) is slightly skew to the left or to the right depending on the responses to the items with large $\hat{\alpha}_{i,1}$.

These results lead us to conclude that we do not need to determine all the h(z|x)'s to have a clear idea about the distribution of h(z|x) along Z. Instead, we can select a representative sample of h(z|x), selecting x so that the whole set of values assumed by the E(Y|x) (or E(Z|x)) is covered. Using this criterion we shall determine the h(z|x) for the two score patterns, which provide the observed lowest and highest position on the latent scale and for those which corresponds E(Y|x) equal to 0.10, 0.20,.....,0.90.

Figure 7.11 (next page) displays a representative collection of the 417 observed posterior densities h(z|x). Based on this figure we detect groups of score patterns (or individuals) who have nearly the same posterior densities h(z|x), differing only on the location parameter.

Thus, for example, the position on the latent scale of individual with E(Z|x) from -1.35 to 0.00 can be measured more precisely than those with E(Z|x) ranging from -2.26 to -1.35 or from 0.57 to 1.91.

Therefore if we desire to make groups of individuals according to their distribution on the latent scale, we can combine the information obtained from Figure 7.11, which gives the shape of h(z|x) and its location along Z, with the results from Table 7.3, which provides the observed frequency distribution of those h(z|x).

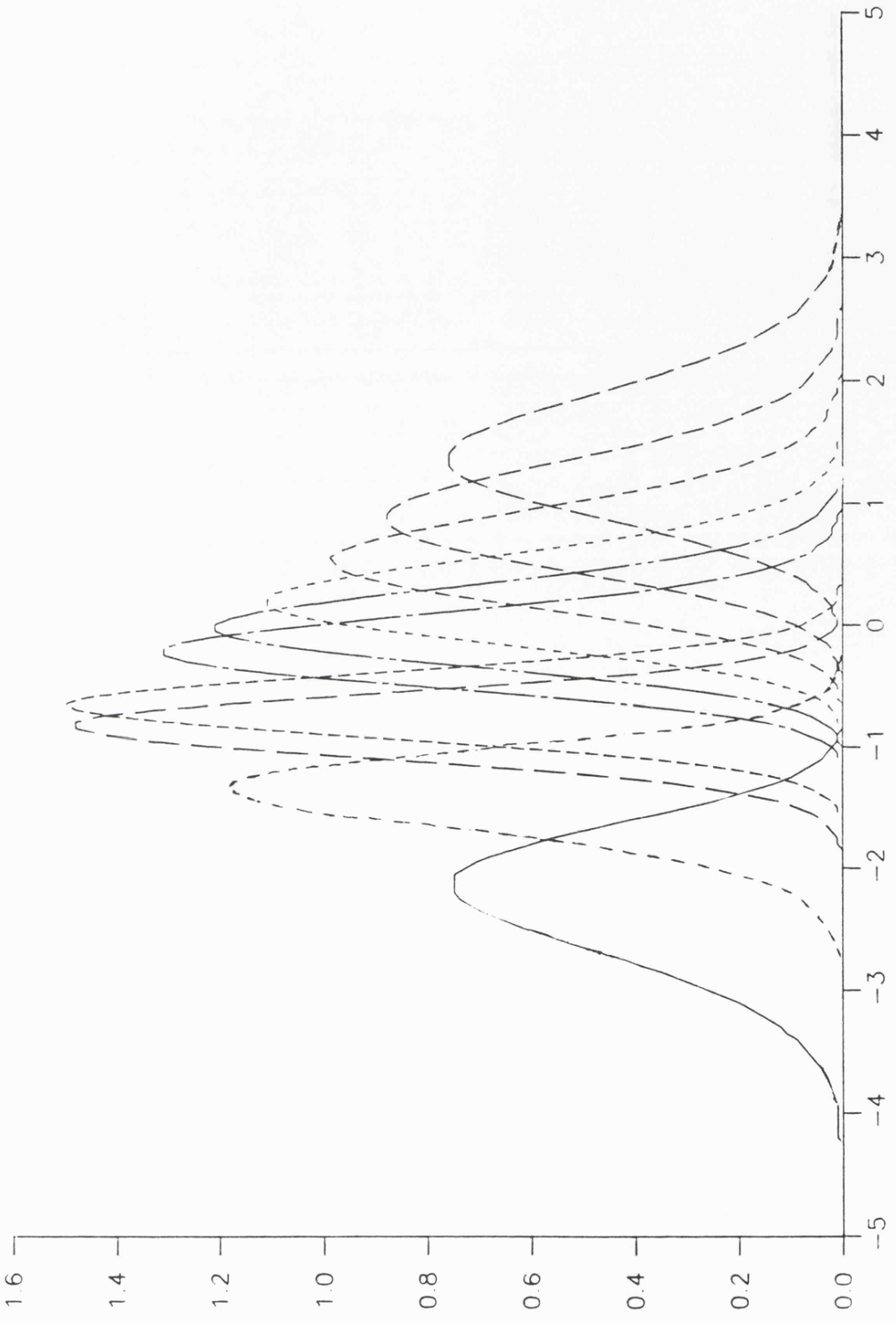


Figure 7.11- Representative collection of posterior densities h(z|x) for the observed score patterns of Test 12.

4.2- Test 13

As described in the previous section, Test 13 has 40 items and the fitting by a logit/probit model yielded 6 items with $\hat{\alpha}_{i,1} \geq 3.0$. The 498 individuals who answered the test provided 488 different score patterns to which one corresponds one posterior density h(z|x).

As for Test 12, we have determined the posterior densities h(z|x) for a significant number of observed score patterns, so that the E(Z|x)'s are distributed along the whole latent scale Z. More precisely, for each interval of E(Z|x) in Table 7.3 we have selected at least 10 different score patterns and we have determined their h(z|x)'s.

The results agree with those from Test 12 in terms of

(1) the equivalence between similar E(Z|x)'s (or E(Y|x)) and nearly equal h(z|x)'s. Similar E(Z|x)'s (or E(Y|x)'s) come from nearly equal h(z|x)'s, specially for those individuals who are not located in the extrems left and right of the latent scale;

(2) representativity of the whole set of h(z|x) through few h(z|x), taking into account the two score patterns which are in the lowest and highest position of the latent scale and at least 9 score patterns, for which their E(Y|x) are distributed along (0,1).

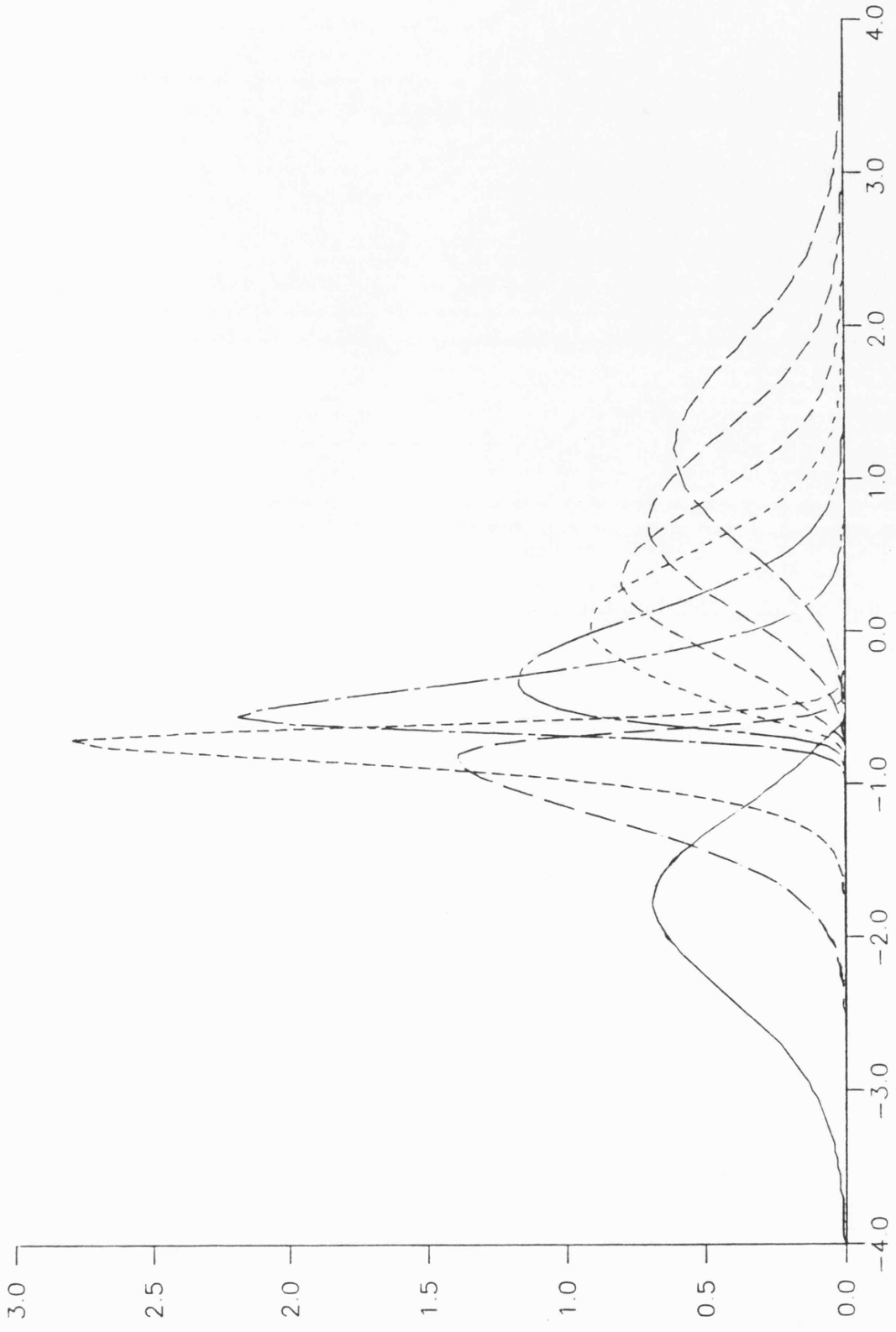


Figure 7.12- Representative collection of posterior densities h(z|x) for the observed score patterns of Test 13.

Figure 7.12 displays a representative collection of observed posterior densities h(z|x) for Test 13, for which E(Y|x) are equal to 0.05, 0.10, 0.20, 0.35, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90 and 0.98 or E(Z|x) are equal to -2.97, -1.35, -0.90, -0.38, -0.25, 0.00, 0.25, 0.56, 0.92, 1.42 and 2.43.

Looking at Figure 7.12 we can see how h(z|x) changes in terms of shape and dispersion along Z. It also provides a measure of precision for comparing individuals located in different points of along the latent scale Z.

In Figure 7.12, although most of the consecutive means E(Y|x) are equidistant, the posterior densities h(z|x) corresponding to E(Y|x) equal to 0.60 is closer(more similar) to that with mean 0.50 than 0.70.

Furthermore, combining the information given in Table 7.3 and Figure 7.12 we will be able to make groups using the information given by the h(z|x)'s. Thus, for example, the 21 (or 75) individuals who have E(Y|x) between -1.18 and -1.09 (or -0.40 and -0.38) should belong to the same group, since they have nearly the same h(z|x) and therefore they lead to the same beliefs about the latent variable.

We have also look at diagrams like Figures 7.11 and 7.12 for many tests with smaller number of items (10 or less), which yielded one or two large $\hat{\alpha}_{i,1}$ ($\geq$3.0) when fitted by a logit/probit model. The posterior densities tend to be skew to the left or to the right depending on the responses to the items with large $\hat{\alpha}_{i,1}$ and the variances of the distributions differ. When the $\hat{\alpha}_{i,1}$ are close to each other the posterior distributions are approximately normal distributed, which confirm result 2.

5- *Conclusions*

The investigations carried out in this chapter for the logit/logit and logit/probit models lead to the following conclusions:

(1) If no $\alpha_{i,1}$ is infinity and two score patterns have the same mean $E(Z|x)$ then they have the same component score $\sum \alpha_{i,1}x_i$ and the same posterior density $h(z|x)$.

(2) If the posterior density $h(z|x)$ is normal, then its mean $E(Z|x)$ is linear in the component score $c_1(x)$. If the mean $E(Z|x)$ is linear in $c_1(x)$, then the posterior density $h(z|x)$ will be close to the normal distribution.

(3) The posterior density $h(z|x)$ is <u>not</u> a function of x through the component score $c_1(x)$ if at least one of the $\alpha_{i,1}$'s is equal to infinity.

(4) The relation between the posterior mean, $E(Y|x)$ or $E(Z|x)$, and the component score is unlikely to be linear when at least one of the $\hat{\alpha}_{i,1}$ is large (say $\geq 3|\sigma$, where σ is the standard deviation of the latent distribution). This may be due to the fact that the component scores are strongly dependent on the values of $\alpha_{i,1}$ while $E(Y|x)$ (or $E(Z|x)$) depends on π_i, which is nearly the same for all $\hat{\alpha}_{i,1} \geq 3|\sigma$, independently of $\hat{\alpha}_{i,0}$.

(5) The greater the test length, the greater the possible number of different score patterns and configuration of $\hat{\alpha}_{i,1}$'s can occur and the less likely the linearity between the posterior mean and the component score seems to be.

(6) Significant differences between component scores do not always reflect different positions on the latent scale, according to the E(Y|x) or E(Z|x). They are shown through flat sections or jumps in the curve obtained when plotting the component scores against the means E(Y|x) (or E(Z|x)).

(7) The occurrence of flat sections seems to depend on the number of items with large $\hat{\alpha}_{i,1}$ and test length. At the same time, we expect that the effect of 2 large $\hat{\alpha}_{i,1}$ in a test with 40 items is smaller than in a test with 20 items. Usually, they do not present a specific pattern for the items with large $\hat{\alpha}_{i,1}$.

(8) Consider a test for which the sample size not small compared with the number of items, for example Test 12 and 13. It seems that even though when fitting a logit/probit model some items have large $\hat{\alpha}_{i,1}$, similar E(Z|x)'s (or E(Y|x)'s) come from nearly equal h(z|x)'s, which are approximately normal distributions, specially for those individuals who are not located at the extreme left and right of the latent scale. For a smaller number of items, h(z|x) tends to be skew to the left or to the right depending on the responses to the items with large $\hat{\alpha}_{i,1}$, and the variances are different.

Therefore the general pattern that emerges is that as the number of items increases, the posterior distributions look more normal and less skew, though with different variances. This is even true if there are several $\alpha_{i,1}$'s estimated as large, and the relation between the posterior means and the component scores is far from linear.

(9) We do not need to determine all the h(z|x)'s to have a clear idea about the distribution of h(z|x) along the latent scale Z. Instead, we can select a representative sample of h(z|x), selecting the score pattern x so that the whole set of values assumed by E(Y|x) (or E(Z|x)) is covered.

(10) If we desire to make groups of individuals according to their distribution on the latent scale, we can combine the information obtained from the shape of h(z|x)'s for all x (Figure 7.12, for example) with the observed frequency distribution of these h(z|x) (Table 7.3, for example).

CONCLUSIONS

Most of the results in this thesis were obtained for the logit/probit model for binary response data given by Bartholomew (1980), even though they also hold for other common binary response models. Large discrimination parameter estimates correspond to $\hat{\alpha}_{i,1} \geq 3/\sigma$, where σ is the standard deviation of the prior latent distribution. In summary the main results are

Chapter 2 - BEHAVIOUR of the LIKELIHOOD

The investigation of the behaviour of the likelihood function using an approximate method provides results equivalent to the profile method. Both suggest that large $\hat{\alpha}_{i,1}$ probably indicates bad behaviour of the likelihood, which will be shown by the presence of a long ridge. In this case the second derivative matrix or the information matrix are not good guides to the variability of these estimates.

If $\hat{\alpha}_{i,1}$ is not large, the first order asymptotic theory is appropriate.

Among the several reparametrizations we tried only the one given by

$$\hat{\alpha}^{*}_{i,0} = \hat{\alpha}_{i,0} / (1 + \hat{\alpha}^{2}_{i,1})^{\frac{1}{2}}$$

provided a better behaviour of the likelihood, independent of the size of the parameter estimates.

This reparametrization corresponds to the probit of the expected value of the response function of a probit model, that is,

$$\hat{\alpha}^{*}_{i,0} = \Phi^{-1}(E(\alpha_{i,0} + \alpha_{i,1} z)) = \Phi^{-1}(E(P(X_i=1|z))) = \Phi^{-1}(P(X_i=1))$$

Chapter 3 - ADEQUACY of the ASYMPTOTIC VARIANCE-COVARIANCE MATRIX using BOOTSTRAP and JACKKNIFE

The more closely the bootstrap distribution of the parameter estimates is fitted by a normal distribution, the better is the agreement between the bootstrap and the asymptotic standard deviation.

If $\hat{\alpha}_{i,1}$ is not large, the asymptotic variance matrix can probably be trusted, since the bootstrap estimates and standard deviations are very close to the ML estimates and to the asymptotic standard deviations. Furthermore, this similarity increases as the sample size becomes larger.

Large values for $\hat{\alpha}_{i,1}$ are associated with skewed distributions or a mixture of two distributions, one normal and another with $\alpha_{i,1}$ equal to infinity. Probably the asymptotic standard deviations of the parameter estimates are smaller than the true ones.

If the sample size is small and one of the items has very large $\hat{\alpha}_{i,1}$ while the remaining ones are small, all with relative large standard deviations then it is likely that most of the estimates can not be trusted.

In summary, although the bootstrap distribution must underestimate the variation in the true sampling distribution, there is strong evidence that it gives a better guide than the usual first order normal approximation. Bootstrapping methods seem to be very useful for investigating the adequacy of the normal approximation in doubtful cases. When the discrimination parameters are small the asymptotic theory works well, but when they get large it is inadequate.

Jackknife parameter estimates and their standard deviations tend to be very similar to the original ML ones, independent of the pattern of the $\hat{\alpha}_{i,1}$'s and the sample size. Therefore, jackknife is not as good as bootstrap in warning about possible inadequacy of the asymptotic standard deviations. This undesirable result for the jackknife method may be due to the small number of different jackknife pseudovalues (16 in the case examined), and a larger number of items would provide more satisfactory results.

Chapter 4 - RASCH MODEL

It is likely that the Rasch model fits well a set of data fitted by a logit/probit model when all $\hat{\alpha}_{i,1}$'s are very similar to each other or one of the $\hat{\alpha}_{i,1}$'s is very large compared with the remaining ones, and all estimates have relative large standard deviations. In this case the likelihood function for the Rasch model behaves better than the one for the logit/probit, and thus Rasch fits the data better.

The standardized marginal ML difficulty parameter estimates are likely to be very similar to the corresponding conditional ML estimates when the Rasch model fits the data well. However they can be quite different when the discrimination parameter are not the same for all items, since in this case there is a source of variation in the data which tends to increase the difference between conditional and marginal ML estimates.

The rejection of the Rasch model in favour of the general logit/probit model using the likelihood ratio statistic may be due to the good power properties or to the lack of applicability of the asymptotic chi-squared distribution. We think the latter is the most likely.

Chapter 5-STABILITY of the DISCRIMINATION PARAMETER ESTIMATES $\hat{\alpha}_{i,1}$

When considering the effect of deleting items, the magnitude of the sample size must be judged in relation to the number of items.

An item with a large $\hat{\alpha}_{i,1}$ may not give any additional information about the latent variable in a test with 10 or more items, but for smaller length, for example 5, it may contain more information than the remaining items.

The occurence of a large $\hat{\alpha}_{i,1}$ seems to depend more on which items are included in the test then on the sample size and test length. For the data we have analysed, large $\hat{\alpha}_{i,1}$ values were not associated to the type of question asked.

As the number of items decreases, the largest $\hat{\alpha}_{i,1}$ tends to increase and become very large, when the test length is small.

Parameter estimates $\alpha_{i,1}$ (≥ 0.50) and standard deviations are approximately linearly related so that larger estimates have larger standard deviations.

The probability of the occurence of a large $\hat{\alpha}_{i,1}$ does not increase as the number of items decreases for sample size of order 500, as it is often said to happen for Heywood cases in factor analysis.

Chapter 6 - AN INVESTIGATION of the CONDITIONS giving rise to LARGE $\hat{\alpha}_{i,1}$

The more predictable is one item from all the remaining ones, the larger is its discrimination parameter estimate $\hat{\alpha}_{i,1}$.

We have presented a procedure, based on equation (6.1), under which we can add a (p+1)th variable with any fixed $\hat{\alpha}_{p+1,0}$ and $\hat{\alpha}_{p+1,1}$ to each response vector without altering the previous estimates of $\alpha_{i,0}$ and $\alpha_{i,1}$, i=1,...,p. The resulting covariance matrix of $\hat{\alpha}_{i,v}$'s for i=1,...,p and v=0,1 does not increase when the (p+1)th variable is added. In particular we can generate an item with large $\hat{\alpha}_{i,1}$ with patterns similar to those in real data sets. For this configuration of data, a threshold response may be the ML estimated response function, and it is seems more reasonable to accept them as legitimate, and not to seek to remove them by adding or dropping items, as has often been suggested.

If no $\alpha_{i,1}$ is infinity and two score patterns have the same mean $E(Z|x)$ then they have the same component score $\sum \alpha_{i,1}x_i$ and the same posterior density $h(z|x)$.

If the posterior density $h(z|x)$ is normal, then its mean $E(Z|x)$ is linear in the component score $c_1(x)$. If the mean $E(Z|x)$ is linear in $c_1(x)$, then the posterior density $h(z|x)$ will be close to the normal distribution.

The posterior density $h(z|x)$ is <u>not</u> a function of x through the component score $c_1(x)$ if at least one of the $\alpha_{i,1}$'s is equal to infinity.

The relation between the posterior mean, $E(Y|x)$ or $E(Z|x)$, and the component score is unlikely to be linear when at least one of the $\hat{\alpha}_{i,1}$ is large. This may be due to the fact that the component scores are strongly dependent on the values of $\alpha_{i,1}$ while $E(Y|x)$ (or $E(Z|x)$) depends on π_i, which is nearly the same for all $\hat{\alpha}_{i,1} \geq 3|\sigma$, independently of $\hat{\alpha}_{i,0}$.

The greater the test length, the greater the possible number of different score patterns and configuration of $\hat{\alpha}_{i,1}$'s can occur and the less likely the linearity between the posterior mean and the component score seems to be.

Significant differences between component scores do not always reflect different positions on the latent scale, according to the $E(Y|\mathbf{x})$ or $E(Z|\mathbf{x})$. They are shown through flat sections or jumps in the curve obtained when plotting the component scores against the means $E(Y|\mathbf{x})$ (or $E(Z|\mathbf{x})$).

The occurrence of flat sections seems to depend on the number of items with large $\hat{\alpha}_{i,1}$ and test length. At the same time, we expect that the effect of 2 large $\hat{\alpha}_{i,1}$ in a test with 40 items is smaller than in a test with 20 items. Usually, they do not present a specific pattern for the items with large $\hat{\alpha}_{i,1}$.

As the number of items increases, the posterior distributions look more normal and less skew, though with different variances. This is even true if there are several $\alpha_{i,1}$'s estimated as large, and the relation between the posterior means and the component scores is far from linear.

We do not need to determine all the $h(z|\mathbf{x})$'s to have a clear idea about the distribution of $h(z|\mathbf{x})$ along the latent scale Z. Instead, we can select a representative sample of $h(z|\mathbf{x})$, selecting the score pattern $\mathbf{x}$ so that the whole set of values assumed by $E(Y|\mathbf{x})$ (or $E(Z|\mathbf{x})$) is covered.

If we desire to make groups of individuals according to their distribution on the latent scale, we can combine the information obtained from the shape of $h(z|\mathbf{x})$'s for all $\mathbf{x}$ (Figure 7.12, for example) with the observed frequency distribution of these $h(z|\mathbf{x})$ (Table 7.3, for example).

REFERENCES

Andersen,E.B.(1970) Asymptotic properties of conditional maximum likelihood estimators. J. Roy. Statist. Soc., B, 34, 283-301.

Andersen,E.B.(1972) The numerical solution of a set of conditional estimation equations. J. Roy. Statist. Soc., B, 34, 42-54.

Andersen,E.B.(1973a) Conditional inference for multiple-choice questionnaries. Br. J. Math. Statist. Psychol., 26, 31-44.

Andersen,E.B.(1973b) A goodness of fit test for the Rasch model. Psychometrika, 38, 123-140.

Andersen,E.B. and Madsen,M.(1977) Estimating parameters of the latent population distribution. Psychometrika, 42, 357-374.

Andersen,E.B.(1980) Discrete Statistical Models with Social Science Applications. Amsterdam: North-Holland Publishing Company.

Anderson,J.C. and Gerbing,D.W.(1984) The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. Psychometrika, 49, 155-173.

Baker,F.B.(1988) Methodology review: item parameter estimation under one-, two-, and three-parameter logistic models. Appl. Psychol. Measurement, 11, 111-141.

Bartholomew,D.J.(1980) Factor analysis for categorical data. J. Roy. Statist. Soc., B, 42, 293-321.

Bartholomew,D.J.(1981) Posterior analysis of the factor model. Br. J. Math. and Statist. Psychol., 34, 93-99.

Bartholomew,D.J.(1983) Latent variable models for ordered categorical data. J. Econometrics, 22, 229-243.

Bartholomew,D.J.(1984) Scaling binary data using a factor model. J. Roy. Statist. Soc., B, 46, 120-123.

Bartholomew,D.J.(1987) Latent Variable Models and Factor Analysis. London: Charles Griffin & Company Ltd.

Bartholomew,D.J.(1988) The sensitivity of latent trait analysis to choice of prior distribution. Br. J. Math. and Statist. Psychol., 41, 101-107.

Beran,R. and Srivastava,M.S.(1985) Bootstrap tests and confidence regions for functions of a covariance matrix. Ann. Statist., 13, 95-115.

Birnbaum,A.(1968) Some latent trait models and their use in infering an examinee's ability. In Lord,F.M. and Novick,M.R., Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley, Chapters 17-20.

Bock,R.D.(1972) Estimating item parameters and latent ability when responses are scored in two or more nominal categories. Psychometrika, 37, 29-51.

Bock,R.D. and Aitkin,M.(1981) Marginal maximum likelihood estimation of item parameters: application of an E-M algorithm. Psychometrika, 46, 443-459.

Bock,R.D. and Lieberman,M.(1970) Fitting a response model for n dichotomously scored items. Psychometrika, 35, 179-197.

Boomsma, A.(1985) Nonconvergence, improper solutions and starting values in LISREL maximum likelihood estimation. Psychometrika, 50, 229-242.

Chatterjee,S.(1984) Variance estimation in factor analysis: An application of the bootstrap. Br. J. Math. Statist. Psychol., 37, 252-262.

Clarkson,D.B.(1979) Estimating the standard errors of rotated factor loadings by jackknifing. Psychometrika, 44, 297-314.

Cox,D.R.(1966) A simple example of a comparation envolving quantal data. Biometrika, 53, 215-220.

Dempster,A.P.,Laird,N.M. and Rubin,D.B.(1977) Maximum likelihood from incomplete data via the E-M algorithm. J. Roy. Statist. Soc., B, 39, 1-38.

Diciccio,T. and Tibshirani(1987) Bootstrap confidence intervals and bootstrap approximations. J. Amer. Statist. Assoc., 82, 163-170.

Dinero,T. and Haertel,E.(1977) Applicability of the Rasch model with varying item discriminations. Appl. Psychol. Measurement, 1, 581-592.

Efron,B.(1979) Bootstrap methods: another look at the jackknife. Ann. Statist., 7, 1-26.

Efron,B.(1984) Better bootstrap confidence intervals. Tech. Rep. Stanford University Dept. Statist.

Efron,B. and Tibshirani,R.(1986) Bootstrap method for standard errors, confidence intervals and other measures of statistical accuracy. Statist. Science, 1, 54-75.

Efron,B.(1987) Better bootstrap confidence intervals. J. Amer. Statist. Assoc., 82, 171-185.

Fachel,J.M.G.(1986) The C-type Distribution as an Underlying Model for Categorical Data and its use in Factor Analysis. Ph.D. Thesis, University of London.

Fisher,G.H.(1981) On the existence and uniqueness of maximum likelihood estimates in the Rasch model. Psychometrika, 46, 59-77.

Goldstein,H.(1980) Dimensionality, bias, independence and measurement scale problems in latent trait test score models. Br. J. Math. Statist. Psychol., 33, 234-246.

Grönroos,M.(1985) Boothstrapping in Factor Analysis. 45th Biennal Session of the International Statistical Institute, 12-22.

Gruijter,D.N.M.(1985) A note on the assymptotic variance-covariance matrix of item parameter estimates in the Rasch model. Psychometrika, 50, 247-249.

Gustafsson,J.E.(1980a) A solution of the conditional estimation problem for long tests in the Rasch model for dichotomous. Educ. and Psychol. Measurement, 40, 377-385.

Gustafsson,J.E.(1980b) Testing and obtaining fit of data to the Rasch model. Br. J. Math. Statist. Psychol., 33, 205-233.

Haberman,S.J.(1977) Maximum likelihood estimates in exponencial response models. Ann. Statist., 5, 815-841.

Hambleton,R.K. and Swaminathan,H.(1985) Item Response Theory: Principles and Applications. Boston: Kluwer-Nijhoff.

Holland,P.W.(1981) When are item response models consistent with observed data? Psychometrika, 46, 79-92.

Hulin,C., Lissak,R. and Drasgow,F.(1982) Recovery of two and three-parameter logistic item characteristic curves: a Monte Carlo study. Appl. Psychol. Measurement, 6, 249-260.

Jorgensen,M.A.(1987) Jackknifing fixed points of iterations. Biometrika, 74, 207-211.

Jöreskog,K.G. and Sörbom,D.(1984) LISREL VI user's guide. Mooresville, Indiana: Scientific Software, Inc.

Kelderman,H.(1984) Loglinear Rasch model tests. Psychometrika, 49, 223-245.

Kendall,M.G. and Stuart,A.(1979) The Advanced Theory of Statistics, vol.2, New York: Hafner.

Lord,F.M.(1952) A theory of test scores. Psychometrika Monograph, number 7, 17.

Lord,F.M.(1968a) An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. Educ. and Psychol. Measurement
28, 989,1020.

Lord,F.M. and Novick,M.R.(1968b) Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley.

Lord,F.M.(1983a) Statistical bias in maximum likelihood estimators of item parameters. Psychometrika, 48, 425-435.

Lord,F.M.(1983b) Maximum likelihood estimation of item response parameters when some responses are omitted. Psychometrika, 48, 477-482.

Lord,F.M. and Wingersky,M.S.(1983) Sampling variances and covariances of parameter estimates in item response theory. In D.J.Weiss(Ed.), Proceedings of the 1982 Item Response Theory and Computadorized Adaptative Testing Conference. Minneapolis: University of Minnesota, Department of Psychology, Computerized Adaptative Testing Laboratory. (p.69-88).

Louis,T.A.(1982) Finding the observed information matrix when using the E-M algorithm. J. Roy. Statist. Soc., B, 44, 226-233.

Mantel,N. and Haenszel,W.(1959) Statistical aspects of the analysis of data from retrospective studies of disease. J. Nat. Cancer Inst., 22, 719-748.

McFadden,D.(1982) Qualitative response models. Advances in Econometrics, W. Hildenbrand, Cambridge: Univ. Press, 1-37.

McKinley,R.L. and Mills,C.N.(1985) A comparison of several of goodness-of-fit statistics. Appl. Psychol. Measurement, 9, 49-57.

Miller,R.G.(1974) The jackknife _ a review. Biometrika, 61, 1-15.

Mislevy,R.J.(1984) Estimating latent distributions. Psychometrika, 49, 359-381.

Mislevy,R.J.(1985) Estimation of latent group effects. J. Am. Statist. Assoc., 80, 993-997.

Molenaar,I.W.(1983) Some improved diagnostics for failure of the Rasch model. Psychometrika, 48, 49-72.

Pennell,R.(1972) Routinely computable confidence intervals for factor loadings using the 'jackknife'. Br. J. Math. Statist. Psychol., 25, 107-114.

Quenouille,M.H.(1956) Notes on bias estimation. Biometrika, 43, 353-360.

Ramsey,J.(1975) Solving implicit equations in psychometric data analysis. Psychometrika, 40, 337-360.

Rasch,G.(1960) Probabilistic models for some intelligence and attainment tests. Copenhagen: Paedagogiske Institut.

Rosenbaum,P.R.(1984) Testing the conditional independence and monotonicity assumptions of the item response. Psychometrika, 49, 425-435.

Sanathanan,L. and Blumenthal,S.(1978) The logistic model and estimation of latent structure. J. Amer. Statist. Assoc., 73, 794-799.

Samejima,F.(1974) Normal ogive model on the continuos response level in the multi-dimensional latent space. Psychometrika, 39, 111-121.

Shea,B.L.(1984) FACONE: A computer program for fitting the logit latent model by maximum likelihood. Department of Statistics, London School of Economics.

Stouffer,S. and Toby,J.(1951) Role conflit and personality. J. Sociology, 56, 395-406.

Stouffer,S., Guttman,L., Suchman,E., Lazarsfeld,P., Star,S. and Clausen,J.(1950) Measurement and prediction, Volume 4 of Studies in Social Psychology during World War. Princeton, NJ: Princeton University Press. Reprinted, 1966, New York: John Wiley.

Stroud,A.H. and Secrest,D.(1966) Gaussian Quadrature Formulas. Prentice-Hall.

Tatsuoka,K.K.(1984) Caution indices based on item response theory. Psychometrika, 49, 95-110.

Thissen,D.(1982) Marginal maximum likelihood estimation for the one-parameter logistic model. Psychometrika, 47, 175-186.

Thissen,D. and Wainer,H.(1982) Some standard errors in item response theory. Psychometrika, 47, 141-147.

Tsutakawa,R.K.(1984) Estimation of two-parameter logistic item response curves. J. Educ. Statist., 9, 263-276.

Tukey,J.W.(1958) Bias and confidence in not quite large samples. Ann. Math. Statist., 29,614 [δ1]

van de Driel,O.P.(1978) On various causes of improper solutions in maximum likelihood factor analysis. Psychometrika, 43, 225-243.

van de Vijner,F.J.(1986) The robustness of Rasch estimates. Appl. Psychol. Measurement, 10, 45-57.

van den Wollenberg,A.L.(1982) Two new test statistics for the Rasch model. Psychometrika, 47, 123-140.

van den Wollenberg,A.L., Wierda,F.W. and Jansen,P.G.W. (1988) Consistency of Rasch model parameter estimation: a simulation study. Appl. Psychol. Measurement, 12, 307-313.

Wood,R.C.;Wingersky,M.S. and Lord, F.M.(1976) Logistic- a computer program for estimating examinee ability and item characteristic curve parameters (Research Memorandum 76-6) Princeton NJ: Educational Testing Service.

Wright,B. and Douglas,G.A.(1977) Conditional versus unconditional procedures for sample-free item analysis. Educ. and Psychol. Measurement, 37, 573-586.

Wright,B. and Panchapakesan,W.(1969) A procedure for sample-free item analysis. Educ. and Psychol. Measurement, 29, 23-48.

Wright,B.D. and Mead,R.J.(1978) Bical: Calibrating Items and Scales with the rasch Model (research Memorandum N^0 23A). Chicago: University of Chicago, Statistical Laboratory.

Wu,C.F.J.(1983) On the convergence properties of the EM algorithm. Ann. Statist., 3, 95-103.

Yen,W.M.(1981) Using simulation results to choose a latent trait model. Appl. Psychol. Measurement, 5, 245-262.

Yen,W.M.(1985) Increasing item complexity; a possible cause of scale shrinkage for unidimensional item response theory. Psychometrika, 50, 399-410.