Bootstrapping latent variable models for binary response

M. T. Albanese
Universidade do Federal do Rio Grande de Sul, Brazil
M. Knott
London School of Economics and Political Science

August 11, 2000

Abstract

Estimated asymptotic variances for the estimates of the parameters in a logit-probit model for binary response data are unreliable for moderate sized samples. We show how bootstrapping gives a better idea of the sampling distribution of the estimators, and can also allow an assessment of the reliability of the scoring of individuals on the latent scale.

Keywords: LATENT VARIABLE; BINARY RESPONSE; LOGIT–PROBIT; COMPONENT SCORE; BOOTSTRAPPING; CONDITIONAL MEAN RANKING

Bootstrapping latent variable models for binary response

Abstract

Estimated asymptotic variances for the estimates of the parameters in a logit-probit model for binary response data are unreliable for moderate sized samples. We show how bootstrapping gives a better idea of the sampling distribution of the estimators, and can also allow an assessment of the reliability of the scoring of individuals on the latent scale.

Keywords: LATENT VARIABLE; BINARY RESPONSE; LOGIT–PROBIT; COMPONENT SCORE; BOOTSTRAPPING; CONDITIONAL MEAN RANKING

1. Introduction

When a one factor logit-probit model is fitted by maximum likelihood to binary (0,1) response data, one very often wishes to score the observed response patterns on the scale of the latent variable. We prefer to score with the estimated conditional mean of the latent variable for the given response pattern, though Bartholomew (1984) has suggested scoring the response patterns by their component scores: these are defined as the sum of the estimated discrimination parameters for items responding 1. A difficulty with both methods of scoring is that little is known about the sampling properties of the estimators of the discrimination parameters. Usually one depends on the first order asymptotic distribution theory for maximum likelihood estimators which suggests that the discrimination parameters have a sampling distribution which is asymptotically normal. The covariance matrix of the estimators is estimated from the observed information matrix. For logit-probit models for binary response with their usually large number of parameters it is not clear that sample sizes are great enough in practice to justify the use of this procedure. One would like to know whether the distributions of the estimators are approximately normal. One also needs a guide to the reliability of the scoring of the response patterns, and this is difficult to obtain from standard asymptotic theory.

A common way to investigate questions of this type when exact mathematical results are not feasible is to generate bootstrap samples, and use these as a substitute for the sampling distributions of the estimators. Bootstrap techniques were described in Efron (1982), while Hall (1992) is a more recent reference.

In Section 2 we give the definition of the model, and notation followed in Section 3 by some interpretation of the parameters of the model and a discussion of scoring respondents in the latent variable. Section 4 describes the bootstrap methods, which are then applied to the estimation of discrimination parameters for four data sets in Section 5. The results on scoring,

illustrated for two of the data sets are in Section 6.

We may shortly summarise our conclusions as being that the asymptotic normal distribution theory is inadequate for any large discrimination parameter when samples are of the size often used in practice, but that in any case the scorings are well-behaved.

2. The model and its estimation

Suppose that $X_1, ..., X_p$ are p binary variables taking values 0 and 1. Let x_{is} be the value of the i'th variable X_i for the s'th individual, $s = 1, ..., n$. The row vector $x_s^t = (x_{1s}, ..., x_{ps})$ is referred to as the **response pattern** for the s'th individual. Let Z be a single latent variable.

We shall assume conditional independence, that is the joint conditional probability $g(x|z)$ is

$$g(x|z) = \prod_{i=1}^{p} g_i(x_i|z) \tag{1}$$

where $g_i(x_i|z)$ is the conditional probability of response x_i for the i'th item. Conditional independence is the assumption that the latent variable Z is sufficient to explain all the association between the responses given to different items by an individual. As the X's are binary,

$$g(x|z) = \prod_{i=1}^{p} (\pi_i(z))^{x_i} (1 - \pi_i(z))^{1-x_i} \tag{2}$$

for $i = 1, ..., p$, where $\pi_i(z) = P[X_i = 1|z]$ is called the **response function**.

Consequently, from (1) and (2) the joint density function of x can be written as

$$f(x) = \int_{R_z} h(z) \prod_{i=1}^{p} \pi_i(z)^{x_i} (1 - \pi_i(z))^{1-x_i} dz \tag{3}$$

where R_z is the range space of Z and h(.) is the prior density of Z.

One may obtain many different models by specification of the latent variable prior density $h(.)$ and the shape of the response function $\pi_i(z)$. In this paper we are concerned with the **logit-probit** response function defined as

$$\log \frac{\pi_i(z)}{1 - \pi_i(z)} = \alpha_{0i} + \alpha_{1i} z \tag{4}$$

where Z is distributed as $N(0, 1)$ and the parameters α_{0i} and α_{1i} are referred to as **difficulty** and **discrimination** parameters respectively. These parameters determine the position and the shape of the response function. The parameters α_{0i} and α_{1i} are estimated by marginal maximum likelihood (MML) using a modified E-M procedure (Bock and Aitkin 1981) available as a Fortran program TWOMISS (Albanese and Knott 1992).

The parameter α_{1i} is called a discrimination parameter because as a coefficient of Z its size determines the effect which a given change in Z

has on the probability of a positive response. Taking $\alpha_{1i} > 0$ implies that increasing Z increases the probability of a positive response, which is desirable in most of the applications. The difficulty parameter α_{0i} defines the probability of a positive response to item i for a median individual ($Z = 0$).

The numerical analysis involved in such estimation is extensive, and there is still a margin of choice and discretion in the final estimate, because of slow convergence of iterations and a problem in the choice of the number of 'quadrature points'. Large scale global Monte-Carlo investigation of the sampling properties of estimators for these models is not possible because of the computing time needed and the large number of different possible configurations for the parameter values. It is in these circumstances that one expects Bootstrap methods to be at their most valuable.

3. Model interpretation and scoring

We have, in this paper, emphasized the role of the discrimination parameter α_{1i}, which is important both for the interpretation of the model and in the scoring procedures which result from it. One will often find that the values of the estimates for α_{1i} are large – that is to say that they have values greater than, say, 2.5. A large value for α_{1i} does not imply that the model does not fit, but rather that the response function has a threshold at $-\alpha_{0i}/\alpha_{1i}$. In this case the response for item i is 0 for Z values smaller than the threshold and 1 for Z larger than the threshold. Any value for $\alpha_{1i} \geq 2.5$ effectively gives a threshold response for that item, so that a large range of values α_{1i} are equivalent. This means that the data will not distinguish between α_{1i}'s which are greater than 2.5.

It is possible to reparametrise the model to avoid this effect, see Albanese (1990), by taking $\alpha^*_{1i} = \alpha_{1i}/(1 + \alpha^2_{1i})^{\frac{1}{2}}$. We have given results for α^*_{1i} in Section 5 below.

The discrimination parameters α_{1i} can be used in scoring the respondents on the latent scale. Bartholomew (1984) suggests that the *component score* $\sum \hat{\alpha}_{1i} x_i$ should be used to score respondents on the latent variable Z, and shows that the ranking of respondents is the same as when each respondent is scored using the estimated conditional mean of Z given the response, $E(Z|x)$. In spite of the problem with the large and fairly indeterminate values for $\hat{\alpha}_{1i}$ mentioned above, we show in Section 6 that scorings of respondents by their conditional means are well behaved.

4. Bootstrapping

There are two types of bootstrap sampling that one might try for these models. Empirical bootstrap sampling takes a random sample of size n with replacement of the n individuals who respond to the items. For that sample one finds the estimates of the parameters of interest and records

the scores $E(Z|x)$ on the latent scale. Each such random sample is treated as if it were a replication of the original experiment. We take 100 of these pseudo-replications.

For parametric bootstrap sampling the pseudo-replicates are obtained by generating a Monte Carlo sample from the model described in Section 2, where the parameter values are replaced by their maximum likelihood estimates from the observed data. We used 100 pseudo-replicates for parametric bootstrapping too.

By choosing 100 pseudo-replicates one has an approximation to the sampling behaviour of the estimators. It may be that for some applications more than 100 samples should be used, but we have found for the applications in this paper there is stability in the conclusions from bootstrapping when one compares the results of 60 pseudo-replications with those of 100.

5. Applications

We shall use four sets of data to illustrate our approach. These are old and well-known sets of data, and have been chosen to show a variety of possible behaviour.

Attitudes towards the U.S.Army

The first data set used here is on attitudes to the US Army; it was presented by Stouffer et al. (1950). The four items measure attitudes towards the U.S. Army held by 1000 noncommissioned officers in 1945. A simple χ^2 test for goodness-of-fit is not significant ($\chi^2 = 7.39$ with 7 degrees of freedom), so one may say that these data are well scaled by the single latent variable logit-probit model. The maximum likelihood estimates of the parameters for the model and their estimated standard deviations as given by TWOMISS are in Table 1 together with statistics from the bootstrap replications.

Table 1: Comparison between the original ML (in brackets), empirical and parametric bootstrap estimates $\hat{\alpha}_{1i}$ for a logit-probit model for the Army data.

item	$\hat{\alpha}_{1i}$			SD($\hat{\alpha}_{1i}$)			Q3-Q1			Median	
1	(1.64)	1.68	1.68	(0.24)	0.25	0.24	(0.32)	0.34	0.32	1.67	1.63
2	(1.12)	1.13	1.11	(0.14)	0.15	0.15	(0.19)	0.20	0.16	1.13	1.10
3	(1.41)	1.45	1.41	(0.19)	0.20	0.18	(0.26)	0.31	0.21	1.44	1.39
4	(1.60)	1.63	1.60	(0.22)	0.20	0.22	(0.30)	0.27	0.28	1.64	1.58

These bootstrap values are all close to those for the original maximum likelihood estimators which suggests there is very little bias in the latter. The estimated asymptotic standard deviation seems to be an adequate

guide to the variation in $\hat{\alpha}_{1i}$ because it is in all cases close to to the standard deviation of the bootstrap estimates.

The results for the reparametrised discrimination parameters appear in Table 2, where one can see that the asymptotic results are also close to the bootstrap results.

Table 2: Comparison of the ML estimates (in brackets) and the bootstrap estimates of the reparametrisation $\hat{\alpha}^*_{1i}$ for a logit-probit model for the Army data.

item	$\hat{\alpha}^*_{1i}$			SD($\hat{\alpha}^*_{1i}$)		
1	(0.85)	0.84	0.85	(0.02)	0.02	0.03
2	(0.74)	0.73	0.74	(0.03)	0.04	0.04
3	(0.82)	0.81	0.81	(0.02)	0.03	0.03
4	(0.85)	0.85	0.84	(0.02)	0.03	0.03

It is interesting to see from the normal plots in Figure 1 that the distribution of the estimators $\hat{\alpha}_{1i}$ over the parametric bootstrap samples is approximately normal. We could have used the empirical bootstrap results

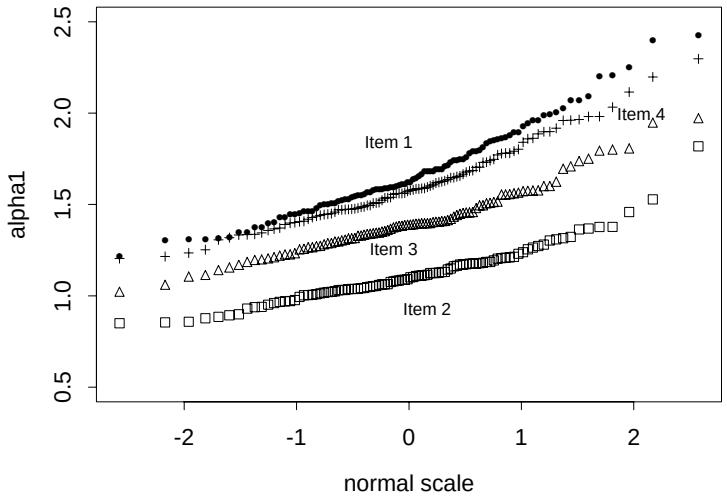

Figure 1: Normal plots of $\hat{\alpha}_{1i}$ for army attitudes data

in Figure 1 without changing the conclusions.

Attitudes towards situations of conflict

Stouffer and Toby (1951) report the answers of 216 respondents in 4 situations of conflict, for each of which the respondent can react with a universal-

istic attitude or particularistic attitude. The goodness of fit test provided a χ^2 equal to 5.85 with 3 degrees of freedom, which indicates that these data are fitted well by a single latent variable logit-probit model.

For these data the ML estimate of the discrimination parameter for item 4 is large. This causes a difference between the results from ML and bootstrap. From Table 3 one can see that the bootstrap distribution for item 4 is skewed to the right, and has a larger mean than the ML estimate. The implication is that the standard error for $\hat{\alpha}_{14}$ is underestimated by the

Table 3: Comparison between the original ML (in brackets), empirical and parametric bootstrap estimates $\hat{\alpha}_{1i}$ for logit-probit model for the Stouffer and Toby data.

item	$\hat{\alpha}_{1i}$			SD($\hat{\alpha}_{1i}$)			Q3-Q1			Median	
1	(1.15)	1.19	1.24	(0.36)	0.38	0.37	(0.48)	0.49	0.54	1.12	1.25
2	(1.58)	1.82	1.69	(0.44)	0.83	0.64	(0.59)	0.62	0.62	1.61	1.56
3	(1.35)	1.44	1.34	(0.36)	0.46	0.36	(0.48)	0.52	0.48	1.36	1.33
4	(2.10)	2.72	2.90	(0.66)	2.99	2.23	(0.89)	0.99	1.37	2.12	2.32

asymptotic results. (The ML estimate $\hat{\alpha}_{14}$ is, in fact, closer to the median of the bootstrap distribution.) The sample size is small here, and it seems that this allows large $\hat{\alpha}_{14}$ which invalidate the asymptotic theory. The problem diminishes with the reparametrisation, which eliminates the large values, as can be seen from Table 4. There is agreement between the asymptotic

Table 4: Comparison between the original ML (in brackets), empirical and parametric bootstrap estimates $\hat{\alpha}^*_{1i}$ for logit-probit model for the Stouffer and Toby data.

item	$\hat{\alpha}^*_{1i}$			SD($\hat{\alpha}^*_{1i}$)		
1	(0.75)	0.74	0.75	(0.01)	0.10	0.10
2	(0.84)	0.85	0.83	(0.03)	0.06	0.07
3	(0.80)	0.80	0.78	(0.05)	0.08	0.09
4	(0.90)	0.89	0.91	(0.02)	0.07	0.05

theory and the bootstrap.

Cancer knowledge

The third set of data comes from a study on knowledge about cancer by Lombard and Doering (1947). A total of 1729 individuals were asked about whether or not they used the following sources of general information: radio, newspaper, solid reading and lectures. These data are fitted reasonably well

by a logit-probit model (χ^2=11.68 with 6 degrees of freedom) with one latent variable as a measure of how well-informed a person is. The results here, see Table 5 and Table 6 are similar to Stouffer and Toby (1951) reported above but the sample size here is large, and so the large values for $\hat{\alpha}_{12}$ are most likely to be due to the threshold model applying to the population.

Table 5: Comparison between the original ML (in brackets), empirical and parametric bootstrap estimates $\hat{\alpha}_{1i}$ for the logit-probit model for the Lombard and Doering data.

item	$\hat{\alpha}_{1i}$			SD($\hat{\alpha}_{1i}$)			Q3-Q1			Median	
1	(0.72)	0.73	0.73	(0.09)	0.09	0.10	(0.12)	0.13	0.11	0.71	0.72
2	(3.40)	4.14	3.79	(1.14)	2.71	1.92	(1.54)	2.07	1.28	3.29	3.40
3	(1.34)	1.39	1.38	(0.19)	0.19	0.18	(0.23)	0.25	0.24	1.37	1.38
4	(0.77)	0.82	0.78	(0.22)	0.14	0.13	(0.19)	0.18	0.18	0.81	0.77

Table 6: Comparison between the original ML (in brackets), empirical and parametric bootstrap estimates $\hat{\alpha}^*_{1i}$ for the logit-probit model for Lombard and Doering data.

item	$\hat{\alpha}^*_{1i}$			SD($\hat{\alpha}^*_{1i}$)		
1	(0.58)	0.59	0.59	(0.04)	0.05	0.05
2	(0.96)	0.95	0.95	(0.21)	0.03	0.02
3	(0.80)	0.81	0.80	(0.02)	0.04	0.04
4	(0.61)	0.63	0.61	(0.09)	0.07	0.06

Figure 2 shows the normal probability plots of the empirical bootstrap distributions of $\hat{\alpha}_{1i}$. The distributions of the $\hat{\alpha}_{1i}, i = 1, 3$and 4 are approximately normal, while the values of $\hat{\alpha}_{12}$ can best be described as a skew distribution, though one might perhaps say that it is a mixture of two normal distributions with some outlying high values.

Arithmetic Reasoning test (ART) on black women

This example is a sample of 145 responses to the Arithmetic Reasoning Test (ART) from young black American women on the Armed Services Vocational Aptitude Battery, given by Mislevy (1985). The statistics of the goodness-of-fit test χ^2 is equal to 6.42 with 3 degrees of freedom, which indicates a reasonable fit. Here there is a large value for $\hat{\alpha}_{11}$, so that the response function for item 1 is effectively estimated as a threshold response function. The full details of the maximum likelihood estimation and the summary bootstrap results are in Table 7 and Table 8. Notice that in this

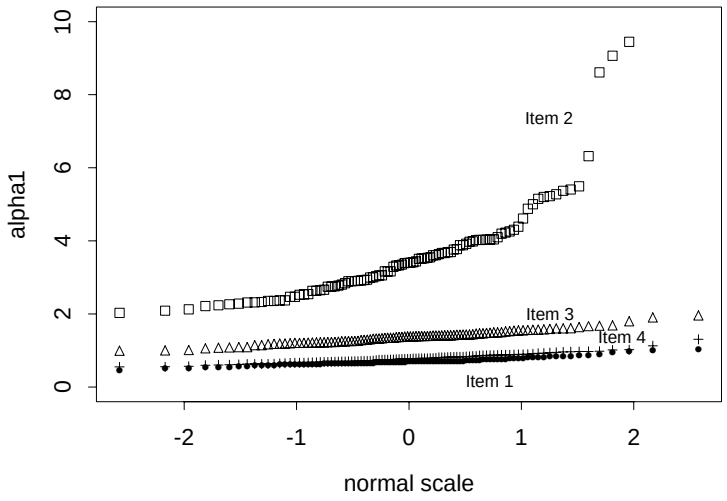

Figure 2: Normal plots of the empirical bootstrap $\hat{\alpha}_{1i}$ for cancer knowledge data

case the large value for $\hat{\alpha}_{11}$ seems to have destroyed the applicability of the asymptotic theory not only for that item, but for items 2 and 3 as well. The ML method does not seem very reliable here. The fact that the results for the parametric bootstrap are so far away from those of ML shows that this is not just a question of the model not fitting the data, but rather that ML is not working satisfactorily. The reparametrisation fails to remove all

Table 7: Comparison between the original ML (in brackets), empirical and parametric bootstrap estimates $\hat{\alpha}_{1i}$ for the logit-probit model for the ART on black women.

item	$\hat{\alpha}_{1i}$			SD($\hat{\alpha}_{1i}$)			Q3-Q1			Median	
1	(14.39)	6.79	5.70	(67.78)	7.16	5.33	(91.43)	12.37	10.68	2.96	2.53
2	(0.38)	1.63	0.71	(0.14)	3.32	0.15	(0.19)	0.20	0.16	1.13	1.10
3	(0.37)	1.56	1.17	(0.19)	0.20	0.18	(0.26)	0.31	0.21	1.44	1.39
4	(0.19)	0.14	0.24	(0.22)	0.20	0.22	(0.30)	0.27	0.28	1.64	1.58

the problems in this case, though Table 8 shows that there is perhaps less disagreement.

Conclusions on estimation of α_i

We have investigated many other data sets in the same way, and the results above are typical. The bootstrap results sometimes suggest there is

Table 8: Comparison between the original ML (in brackets), empirical and parametric bootstrap estimates $\hat{\alpha}^*_{1i}$ for the logit-probit model for the ART on black women.

item	$\hat{\alpha}^*_{1i}$			SD($\hat{\alpha}^*_{1i}$)		
1	(0.99)	0.73	0.80	(0.04)	0.34	0.25
2	(0.35)	0.45	0.40	(0.14)	0.33	0.32
3	(0.35)	0.37	0.48	(0.15)	0.39	0.27
4	(0.19)	0.04	0.17	(0.20)	0.44	0.33

bias in the ML estimates and although the bootstrap distributions must underestimate the variation which would be present in the true sampling distribution for the estimators, we believe that they give a better guide to sampling variation 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 can be inadequate. The four data sets presented above indicate broad conclusions that can be be drawn about the estimation of the discrimination parameters.

1. When the parameter estimates $\hat{\alpha}_{1i}$'s are nearly equal and not large, the asymptotic variance matrix probably can be trusted, since the bootstrap standard deviations are very close to the asymptotic ones. This similarity increases as the sample size become larger.
2. Large values for $\hat{\alpha}_{1i}$ are associated with skewed distributions or a mixture of two distributions, one Normal and another with α_{1i} 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 $\hat{\alpha}_{1i}$ to be considered large depends on the sample size.
3. When one of the $\hat{\alpha}_{1i}$'s is very large compared with the remaining $\hat{\alpha}_{1i}$'s and the sample size is small, we should probably not trust any estimates, since this item affects all the others.
4. The better the bootstrap distribution of the parameter estimates $\hat{\alpha}_{1i}$ and $\hat{\alpha}^*_{1i}$ is fitted by a Normal distribution the better is the agreement between the bootstrap and the asymptotic standard deviations.
5. As the bootstrap distribution of the parameter estimates $\hat{\alpha}^*_{1i}$ is fitted very well by a Normal distribution, most of the bootstrap results confirm those from the standard asymptotic theory. This shows that $\hat{\alpha}^*_{1i}$ are not affected by the skewness of the bootstrap $\hat{\alpha}_{1i}$.
6. In general, when there is some difference between the bootstrap (empirical and parametric) and the ML results for $\hat{\alpha}_{1i}$, bootstrap estimates

are closer to each other than to the ML estimates; and the medians of the bootstrap samples are closer to the ML estimates than the means. The strongest similarity among them is related to the interquartile difference Q3-Q1, which could be expected since most or all of the estimates responsible for the skewness of the distribution are not considered.

6. Scoring of the individuals

Very often the purpose of fitting a logit-probit model is to obtain a scoring of the subjects through their placing on the latent scale. We have checked on the stability of such scoring using the bootstrap approach. Bartholomew (1984) has shown, that an equivalent ranking is obtained from scoring either with the component scores or with the mean of the conditional distribution of Z given the observed response pattern, but we believe that it is better to use the latter, see Knott and Albanese (1993). We used the estimated means of the conditional distributions to score the response patterns of the original sample and then looked at the variation from that scoring over the bootstrap samples. We used the Pearson product moment correlation between the bootstrap sample scorings and the original one to measure the degree of agreement in the scorings.

The scorings show a high degree of stability, as can be seen from the histograms of the values of the correlation coefficients over the empirical bootstrap samples. Figure 3 below shows the results for the cancer knowledge data, for which most of the correlations were over 0.98 .

The corresponding histogram for the ART on black women is shown as Figure 4. In spite of the poor performance of ML estimation for the discrimination parameters, here also the correlation coefficients are high. It might be thought that the high degree of uncertainty about the discrimination parameters would be reflected in poor stability of the scorings, but this is not the case.

References

Albanese, M. T. (1990). *Latent Variable Models for Binary Response*. Ph. D. thesis, University of London.

Albanese, M. T. and M. Knott (1992). TWOMISS: a computer program for fitting a one- or two- factor logit-probit latent variable model to binary data when observations may be missing. Technical report, Statistics Department, London School of Economics and Political Science, England & Universidade Federal do Rio Grande do Sul, Brazil. The software will be available over EMAIL. Send inquiries to the authors. (EMAIL to M.Knott@lse.ac.uk).

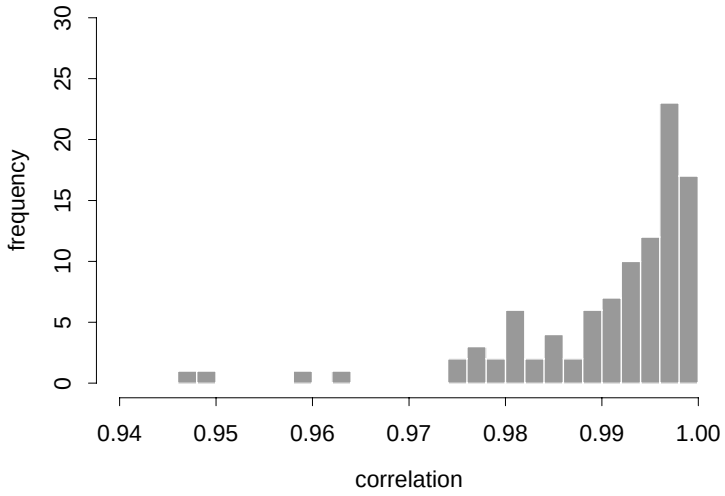

Figure 3: Histogram of empirical bootstrap correlation coefficients for conditional mean scores for cancer knowledge data

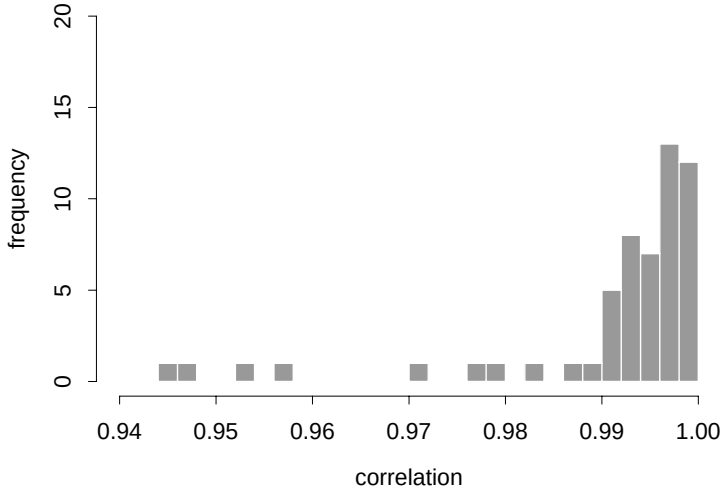

Figure 4: Histogram of empirical bootstrap correlation coefficients for conditional mean scores for the ART on black women

Bartholomew, D. J. (1984). Scaling binary data using a factor model. *Journal of the Royal Statistical Society, Series B 46*, 120–123.

Bock, R. D. and M. Aitkin (1981). Marginal maximum likelihood estimation of item parameters: application of EM algorithm. *Psychometrika 46*, 443–459.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia, Pennsylvania: Society for Industrial and Applied Mathematics.

Hall, P. J. (1992). *The Bootstrap and Edgeworth Expansion*. New York: Springer-Verlag. Springer Series in Statistics.

Knott, M. and M. T. Albanese (1993). Conditional distributions of a latent variable and scoring for binary data. *Revista Brasileira de Probabilidade e Estatística 6*, 171–188.

Lombard, H. L. and C. R. Doering (1947). Treatment of the fourfold table by partial association and partial correlation as it relates to public health problems. *Biometrics 3*, 123–128.

Mislevy, R. J. (1985). Estimation of latent group effects. *Journal of the American Statistical Association 80*, 993–997.

Stouffer, S., L. Guttman, E. Suchman, P. Lazarsfeld, S. Star, and J. Clausen (1950). *Measurement and Prediction*. Princeton, N. J.: Princeton University Press. Volume 4 of Studies in Social Psychology during World War.

Stouffer, S. and J. Toby (1951). Role conflict and personality. *Journal of Sociology 56*, 395–406.