British Journal of Mathematical and Statistical Psychology (1989) **42**, 241–250 *Printed in Great Britain*

Tobit factor analysis†

Bengt O. Muthén‡

Graduate School of Education, University of California, Los Angeles, CA 90024, USA

A new approach is proposed for data which are skewed and have a sizeable proportion of observation at variable end points. Using a covariance structure modelling framework, the new approach assumes censored multivariate normal variables. Using bivariate information, this leads to the use of 'tobit' correlations in weighted least squares estimation. The behaviour of the tobit approach is compared to that of normal theory estimation and ADF estimation.

1. Introduction

Methods of confirmatory factor analysis and structural equation modelling have traditionally used the maximum likelihood estimator and the generalized least squares estimator under the assumption of multivariate normality. Recently, alternative methods have been developed to handle non-normal continuous data using generalized least square (GLS) estimation, for example, the 'asymptotically distribution-free' (ADF) estimator proposed by Browne (1982, 1984). These developments are important in that the common occurrence of non-normality in data need no longer impede the correct application of latent variable modelling.

Several recent studies have also examined the robustness of normal theory estimators in factor analysis and structural equation modelling, e.g. Boomsma (1983), Harlow (1985), Muthén & Kaplan (1985), and Tanaka (1984). It is frequently found that mild deviations from normality do not distort the normal theory analysis to any important degree. The need for the alternative ADF estimator is mostly in situations with severe non-normality. Frequently, however, severe non-normality occurs in the context of censored variables, i.e. variables that have limited variability with a large percentage of cases at an end-point of the scale. Such situations are obtained, for example, when a Likert scale, or sums of such scales, are used to measure extreme phenomena.

†Presented at the Fourth European Meeting of the Psychometric Society, Cambridge, England, 2–5 July 1985. This research was supported by grant number SES-8312583 from the National Science Foundation. I wish to thank David Kaplan for helpful research assistance.

‡Requests for reprints.

With censored variables, the application of the ADF approach with the aim of avoiding effects of non-normality would be inappropriate. This is because the assumptions of the standard linear measurement model will be violated. At issue here is the fact that the ordinary assumptions on the residuals do not hold, since at the end-point of the scale where the censoring takes place, residuals cannot have zero expectation. Only positive or negative residuals are possible, depending on whether the censoring is from below or above. In econometrics such response variables have been modelled under the rubric of tobit regression analysis (Amemyia, 1984; Tobin, 1958).

In this paper, the tobit approach is generalized to multivariate responses obeying latent variable model structures. A generalized least squares estimator is proposed that fits the model to a sample covariance or correlation matrix which differs from the ordinary ones. Framed in terms of correlations, the advantage of the tobit approach can be seen as avoiding the use of Pearson correlations, which are inappropriate for censored variables, and instead using correlations estimated under the assumption of a censored multivariate normal distribution. Hereby, the notion of 'latent' correlation coefficients, such as tetrachorics, polychorics, and polyserials, is generalized to 'tobit correlations'. A Monte Carlo simulation study compares the tobit estimator with normal theory generalized least squares (NTGLS) and ADF in a confirmatory factor analysis model. While both ADF and NTGLS use Pearson correlations, ADF differs from NTGLS in that it uses a weight matrix for Pearson correlations that takes into account the non-normality of the variables.

2. The tobit approach

Consider the measurement model for a set of multivariate normal latent response variables $\mathbf{y}^*$,

$$\mathbf{y}^* = \boldsymbol{\nu} + \boldsymbol{\Lambda}\boldsymbol{\eta} + \boldsymbol{\varepsilon}. \tag{1}$$

For each y^* variable, censoring may take place so that the actual measurement y_j for variable j is

$$y_j = \begin{bmatrix} c_{2j}, & \text{if} \quad y_j^* \geqq c_{2j} \\ y_j^*, & \text{if} \quad c_{1j} < y_j^* < c_{2j} \\ c_{1j}, & \text{if} \quad y_j^* \leqq c_{1j}. \end{bmatrix} \tag{2}$$

The censoring points c are known constants and not parameters to be estimated. Censoring from above or below only, and no censoring is obtained by letting cs go to $\pm\infty$.

With ordinary assumptions, we have the covariance structure of a confirmatory factor analysis model,

$$V(\mathbf{y}^*) = \boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Psi}\boldsymbol{\Lambda}' + \boldsymbol{\Theta}, \tag{3}$$

say. Let

$$\boldsymbol{\sigma} = \text{vec}^*[\boldsymbol{\Sigma}] \tag{4}$$

be a vector containing the distinct elements of the $\boldsymbol{\Sigma}$ matrix. The tobit GLS fitting function may then be written as

$$F = (\mathbf{s} - \boldsymbol{\sigma})'\mathbf{W}^{-1}(\mathbf{s} - \boldsymbol{\sigma}), \tag{5}$$

where $\mathbf{s}$ is a vector of sample statistics estimating $\boldsymbol{\sigma}$, $\boldsymbol{\sigma}$ being a function of the model parameters. Analogous to Muthén (1984), $\mathbf{W}$ is a consistent estimator of the asymptotic covariance matrix of $\mathbf{s}$, obtained by summing first-order derivative products from the univariate and bivariate log-likelihood equations of the maximum likelihood steps described below over the sample units. Alternatively, we may use $\mathbf{W} = \mathbf{I}$, giving the unweighted (ULS) estimator. Using a consistent estimator of the large sample covariance matrix of $\mathbf{s}$, a large sample chi-square measure of model fit and large sample of standard errors of parameter estimates are obtained.

As a preliminary to the minimization of F with respect to the model parameters, the elements of $\mathbf{s}$ need to be calculated. This may be done by a very rapid, two-stage approach. In the first step, univariate response information is used to estimate the mean and variance of the censored variable by maximum likelihood (see e.g. Cohen, 1950, 1955; Des Raj, 1953; Gupta, 1952). In the second step, the covariance (correlation) is estimated by maximum likelihood from bivariate information, holding the mean and variance parameters fixed at the estimated values from the univariate step.

3. A Monte Carlo study

Simulated data were analysed by the new tobit GLS estimator to study the behaviour of its chi-square measure of model fit, parameter estimates, and standard errors of estimates. Tobit GLS was compared to ordinary NTGLS and ADF. A single factor, six variable model was chosen, where the last three variables exhibited varying degrees of censoring from below: 25, 50, 75 per cent. The population variance of the factor was set at one, the loadings were all 0.7, while the error variances were all 0.51, yielding y^* variances of unity and y^* correlations of 0.49. The sample size was 500. To get reliable results a large number of replications was chosen, 500. Both the Monte Carlo study and the estimation were carried out by the LISCOMP computer program (Muthén, 1987). For all three estimators, correlation matrices were analysed, not covariance matrices. For NTGLS the weight matrix computations of Jennrich (1970) are used, while for ADF the computational formulas of Mooijaart (1985) are used.

We may first consider the attenuation of the population y^* correlations due to censoring. Muthén (1990) gives the necessary moments to be able to calculate the population correlation in the censored distribution. With 25 per cent censoring, the original correlation of 0.490 is attenuated to 0.460 when both variables are censored and to 0.464 when only one of the variables is censored. For censoring of 50 per cent, the corresponding values are 0.417 and 0.420. For censoring of 75 per cent, we have 0.344 and 0.344.

It should be noted that the differential attenuation of the population correlations destroys the single factor covariance structure for the ys. Hence, in the case of the NTGLS and ADF estimators, the assumed model will not hold and the estimators will not be consistent. We expect ADF to bring down the size of the chi-square values obtained by NTGLS, which are inflated due to non-normality. However, since the single factor model does not hold as assumed, the ADF chi-squares are also expected to be inflated. The tobit GLS approach, on the other hand, fits the model to the associations among the y^*s for which the single factor model does hold.

As a first step, the quality of the Monte Carlo generation procedure was investigated. The sample Pearson product moment correlations for the three cases of censoring are given in Table 1 for a single sample of 500. Univariate skewness and kurtosis are also given. The correlation coefficients appear reasonably close to the expected values.

A second check is obtained by considering estimation in the case of no censoring. For simplicity, only loading estimates will be reported. Parameter estimates, bias, and mean squared error are given in Table 2. The results for NTGLS and ADF showed no significant bias as expected at this sample size. However, the tobit GLS estimator (TOBIT) exhibited a small but significant overestimation of loadings. The sampling variability in the case of no censoring is given in Table 3. We note that the mean estimated standard errors are reasonably close to the empirically determined variation in the estimates. The top part of Table 4 shows the behaviour of the chi-square test statistic with no censoring. The model has nine degrees of freedom. We note that both the mean chi-square and the proportion of replications for which the model is rejected at the 5 per cent level are not significantly different from expectations. Hence, the data generation seems to work as it should.

In Table 4 the chi-square values for all three cases of censoring are given. The NTGLS and ADF estimators give appreciably too high values already at 25 per cent censoring, with a severe overestimation at 500 per cent. The tobit GLS estimator performs correctly in all cases. The skewness/kurtosis values of Table 1 for 50 per cent censoring are not at all uncommon.

In Tables 5, 6 and 7 are given the results for parameter estimates. For NTGLS and ADF the estimates of the last three, censored variables are strongly biased already at 25 per cent censoring. We note that the small tobit GLS bias observed in the case of no censoring does not increase noticeably with increased censoring.

In Tables 8, 9 and 10 are given the results for parameter estimate variability. While the tobit estimator performs reasonably well throughout, the NTGLS and ADF estimators both exhibit an underestimation of the standard errors, although for ADF it is not that marked for less than 75 per cent censoring. For 75 per cent censoring,

Table 1. Correlation, skewness, and kurtosis for all cases and variables

Case	Correlations					Skewness	Kurtosis
25%	1	2	3	4	5		
						−0.006	−0.021
2	0.491					0.007	0.007
3	0.492	0.491				0.001	−0.004
4	0.467	0.464	0.466			−0.839	0.099
5	0.466	0.464	0.467	0.463		−0.841	0.104
6	0.464	0.465	0.461	0.463	0.463	−0.841	0.116
50%						−0.006	−0.021
2	0.491					0.007	0.007
3	0.492	0.491				0.001	−0.004
4	0.422	0.420	0.421			−1.630	2.340
5	0.422	0.420	0.422	0.420		−1.632	2.357
6	0.420	0.420	0.420	0.416	0.420	−1.640	2.395
75%						−0.006	−0.021
2	0.491					0.007	0.007
3	0.492	0.491				0.001	−0.004
4	0.346	0.344	0.344			−2.995	9.992
5	0.347	0.344	0.346	0.349		−3.005	10.065
6	0.344	0.344	0.344	0.344	0.349	−3.022	10.169

Table 2. Parameter estimates, bias, and mean squared error: 0% censoring

Parameter	True value	Estimator		
		NTGLS	ADF	Tobit
λ_1	0.700	0.706** (0.9)[a] (0.001)[b]	0.707** (1.0) (0.001)	0.707** (1.1) (0.001)
λ_2	0.700	0.706** (0.9) (0.001)	0.706** (0.9) (0.001)	0.707** (0.9) (0.001)
λ_3	0.700	0.703** (0.5) (0.001)	0.704** (0.5) (0.001)	0.708** (1.1) (0.001)
λ_4	0.700	0.707** (1.0) (0.001)	0.707** (1.0) (0.001)	0.709** (1.2) (0.001)
λ_5	0.700	0.705** (0.7) (0.001)	0.704** (0.6) (0.001)	0.707** (1.0) (0.001)
λ_6	0.700	0.708** (1.1) (0.001)	0.708** (0.6) (0.001)	0.707** (1.0) (0.001)

*$P<0.05$; **$P<0.01$.
[a]Per cent bias.
[b]Mean squared error.

Table 3. Sampling variability for 0% censoring

Parameter	NTGLS	ADF	Tobit
λ_1	0.027[a]	0.027	0.027
	0.027[b]	0.027	0.028
λ_2	0.027	0.027	0.027
	0.028	0.028	0.030
λ_3	0.027	0.027	0.027
	0.028	0.028	0.027
λ_4	0.027	0.026	0.026
	0.027	0.027	0.027
λ_5	0.027	0.027	0.027
	0.027	0.027	0.029
λ_6	0.027	0.026	0.027
	0.027	0.028	0.027

[a]Mean of estimated standard errors.
[b]Empirical standard deviation of estimates.

Table 4. Chi-square and reject proportions for all cases of censoring (d.f. = 9)

Case	NTGLS	ADF	Tobit
0%			
Mean	8.800	9.181	9.167
Variance	15.635	17.769	19.878
Reject prop.	0.03	0.05	0.05
25%			
Mean	9.977**	10.213**	9.236
Variance	20.729	22.704	20.071
Reject prop.	0.10**	0.10**	0.06
Mean	14.687**	14.453**	9.235
Variance	40.767	43.977	21.537
Reject prop.	0.32**	0.30**	0.06
75%			
Mean	25.647**	20.598**	9.158
Variance	122.649	100.560	20.740
Reject prop.	0.76**	0.57**	0.06

*$P < 0.05$; **$P < 0.01$.

the tobit GLS estimation has higher variability for the last three variables than NTGLS and ADF. However, as is seen from the mean squared errors of Table 7, this is offset by the smaller bias of the tobit GLS.

As an example of the type of incorrect conclusions that may be drawn when ignoring censoring, we may finally consider the ADF estimator at 50 per cent censoring. With inflated chi-square values, the single factor model may be rejected in

Table 5. Parameter estimates, bias, and mean squared error: 25% censoring

Parameter	True value	Estimator		
		NTGLS	ADF	Tobit
λ_1	0.700	0.703* (0.4)[a] (0.001)[b]	0.702 (0.2) (0.001)	0.708** (1.1) (0.001)
λ_2	0.700	0.702 (0.3) (0.001)	0.701* (0.2) (0.001)	0.706** (0.9) (0.001)
λ_3	0.700	0.699 (−0.1) (0.002)	0.698* (−0.2) (0.001)	0.708** (1.1) (0.001)
λ_4	0.700	0.682** (−2.6) (0.001)	0.680** (−2.8) (0.001)	0.710** (1.4) (0.001)
λ_5	0.700	0.680** (−2.8) (0.001)	0.678** (−3.1) (0.001)	0.708** (1.1) (0.001)
λ_6	0.700	0.683** (−2.5) (0.002)	0.680** (−2.8) (0.001)	0.708** (1.1) (0.001)

*$P<0.05$; **$P<0.01$.
[a]Per cent bias.
[b]Mean squared error.

Table 6. Parameter estimates, bias, and mean squared error: 50% censoring

Parameter	True value	Estimator		
		NTGLS	ADF	Tobit
λ_1	0.700	0.698** (−0.1)[a] (0.001)[b]	0.694** (−0.8) (0.001)	0.707** (1.0) (0.001)
λ_2	0.700	0.698** (−0.2) (0.001)	0.695** (−0.7) (0.001)	0.707** (1.0) (0.001)
λ_3	0.700	0.695** (−0.6) (0.001)	0.691** (−1.3) (0.001)	0.708** (1.1) (0.001)
λ_4	0.700	0.640** (−8.6) (0.005)	0.624** (−10.8) (0.007)	0.710** (1.4) (0.001)
λ_5	0.700	0.638** (−8.8) (0.005)	0.623** (−10.9) (0.007)	0.709** (1.4) (0.001)
λ_6	0.700	0.642** (−8.3) (0.005)	0.626** (−10.6) (0.007)	0.708** (1.1) (0.001)

*$P<0.05$; **$P<0.01$.
[a]Per cent bias.
[b]Mean squared error.

favour of a two-factor model. From exploratory analyses, the two-factor structure of Table 11 may be chosen. As is seen from Table 11, the average chi-square value is close to the expected one of eight and the reject proportion is not significantly different from 0.05. This could lead to the erroneous conclusion of a simple, two-factor structure, whereas in reality the two factors merely correspond to the distinction between censored and non-censored variables. The high factor correlation would decrease with increasing censoring.

Table 7. Parameter estimates, bias, and mean squared error: 75% censoring

Parameter	True value	Estimator		
		NTGLS	ADF	Tobit
λ_1	0.700	0.696** (−0.6)[a] (0.001)[b]	0.692** (−1.1) (0.001)	0.707** (1.0) (0.001)
λ_2	0.700	0.696** (−0.5) (0.001)	0.693** (−1.0) (0.001)	0.706** (0.9) (0.001)
λ_3	0.700	0.696** (−0.9) (0.001)	0.689** (−1.5) (0.001)	0.707** (1.0) (0.001)
λ_4	0.700	0.561** (−19.8) (0.022)	0.494** (−29.4) (0.045)	0.713** (1.8) (0.002)
λ_5	0.700	0.560** (−20.0) (0.023)	0.495** (−29.3) (0.045)	0.714** (2.0) (0.002)
λ_6	0.700	0.565** (−19.3) (0.021)	0.498** (−28.8) (0.043)	0.712** (1.7) (0.002)

*$P<0.05$; **$P<0.01$.
[a]Per cent bias.
[b]Mean squared error.

Table 8. Sampling variability for 25% censoring

Parameter	NTGLS	ADF	Tobit
λ_1	0.028[a]	0.027	0.027
	0.028[b]	0.028	0.028
λ_2	0.028	0.027	0.027
	0.028	0.029	0.031
λ_3	0.028	0.027	0.027
	0.029	0.030	0.027
λ_4	0.029	0.029	0.028
	0.030	0.030	0.029
λ_5	0.029	0.029	0.028
	0.030	0.031	0.031
λ_6	0.029	0.029	0.028
	0.030	0.030	0.028

[a]Mean of estimated standard errors.
[b]Empirical standard deviation of estimates.

4. Summary

A new approach to factor analysis of non-normal continuous variables that are strongly skewed and censored has been proposed. This approach avoids the problems of using Pearson product moment correlations (or covariances) which assume linear relations among the variables. With censored variables linearity is questionable. While the new covariance structure estimator of ADF allows for non-normal variables by the use of a more general matrix, the linearity assumption is maintained in that Pearson correlations are still used. The new tobit factor analysis approach

Table 9. Sampling variability for 50% censoring

Parameter	NTGLS	ADF	Tobit
λ_1	0.029[a]	0.028	0.028
	0.029[b]	0.030	0.029
λ_2	0.029	0.028	0.028
	0.030	0.031	0.031
λ_3	0.029	0.028	0.028
	0.030	0.031	0.028
λ_4	0.032	0.033	0.032
	0.037	0.037	0.033
λ_5	0.032	0.033	0.032
	0.038	0.038	0.034
λ_6	0.032	0.033	0.032
	0.036	0.037	0.032

[a]Mean of estimated standard errors.
[b]Empirical standard deviation of estimates.

Table 10. Sampling variability for 75% censoring

Parameter	NTGLS	ADF	Tobit
λ_1	0.300[a]	0.030	0.029
	0.032[b]	0.034	0.031
λ_2	0.030	0.030	0.029
	0.032	0.035	0.032
λ_3	0.030	0.030	0.029
	0.032	0.035	0.029
λ_4	0.036	0.038	0.040
	0.057	0.051	0.043
λ_5	0.036	0.038	0.040
	0.056	0.051	0.044
λ_6	0.036	0.038	0.040
	0.054	0.051	0.042

[a]Mean of estimated standard errors.
[b]Empirical standard deviation of estimates.

Table 11. 50% censoring: ADF two-factor solution (d.f. = 8)

χ^2			
	Mean		8.290
	Variance		18.044
	Reject prop.		0.068
Loadings			
		0.703	0.000*
		0.705	0.000*
		0.704	0.000*
		0.000*	0.649
		0.000*	0.649
		0.000*	0.651
Factor correlation			
			0.928

uses new correlations that take the censoring into account and applies a weight matrix suitable to these new statistics. Obviously, this tobit factor analysis can be extended to structural equation modelling in general, as has been done in the LISCOMP computer program (Muthén, 1987).

A Monte Carlo study was performed where half of the indicators of a factor model were censored to various degrees. Three estimators were compared, normal theory GLS, ADF, and tobit GLS. Chi-square, estimates, and standard errors were studied. While ADF improved the behavior of chi-square and standard errors by taking non-normality into account, only the tobit GLS estimator provided satisfactory results.

References

Amemiya, T. (1984). Tobit models: A survey. *Journal of Econometrics*, **24,** 3–61.

Boomsma, A. (1983). On the robustness of LISREL (maximum likelihood estimation) against small sample size and nonnormality. PhD dissertation, University of Groningen, Groningen, The Netherlands.

Browne, M. W. (1982). Covariance structures. In D. M. Hawkins (Ed.), *Topics in Applied Multivariate Analysis*. Cambridge, MA: Cambridge University Press.

Browne, M. W. (1984). Asymptotically distribution free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology*, **37,** 62–83.

Cohen, A. C. (1950). Estimating the mean and variable of normal populations from singly and doubly truncated samples. *Annals of Mathematical Statistics*, **21,** 557–569.

Cohen, A. C. (1955). Restriction and selection in samples from bivariate normal distributions. *Journal of the American Statistical Association*, **50,** 884–893.

Des Raj (1953). On estimating the parameters of bivariate normal populations from double and singly linearly truncated samples. *Sankhya*, **12,** 277–290.

Gupta, A. K. (1952). Estimation of the mean and standard deviation of a normal population from a censored sample. *Biometrika*, **39,** 260–273.

Harlow, L. (1985). Behavior of some elliptical theory estimators with non-normal data in a covariance stuctures framework: A Monte Carlo study. Unpublished doctoral dissertation, University of California, Los Angeles.

Jennrich, R. I. (1970). An asymptotic chi-square test for the equality of two correlation matrices. *Journal of the American Statistical Association*, **65,** 904–912.

Mooijaart, A. (1985). A note on computational efficiency in asymptotically distribution-free correlational models. *British Journal of Mathematical and Statistical Psychology*, **38,** 112–115.

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika*, **49,** 115–132.

Muthén, B. (1990). Moments of the censored and truncated bivariate normal distribution. *British Journal of Mathematical and Statistical Psychology* (in press).

Muthén, B. & Kaplan, D. (1985). A comparison of some methodologies for the factor analysis of non-normal Likert variables. *British Journal of Mathematical and Statistical Psychology*, **38,** 171–189.

Muthén, B. (1987). *LISCOMP. Analysis of Linear Structural Equations with a Comprehensive Measurement Model. User's guide.* Mooresville, IN: Scientific Software.

Tanaka, J. S. (1984). Some results on the estimation of covariance structure models. PhD dissertation, University of California, Los Angeles.

Tobin, J. (1958). Estimation of relationships for limited dependent variables. *Econometrica*, **26,** 24–36.

Received 13 May 1987; revised version received 20 August 1988