Testing of Parametric Models fitted to High-dimensional Contingency Tables using Complex Survey Data

Chris Skinner and Irini Moustaki
London School of Economics, UNITED KINGDOM
Corresponding Author: Chris Skinner, email: c.j.skinner@lse.ac.uk

Abstract

Data on multiple categorical variables are often collected in surveys in order to measure a smaller number of underlying dimensions. Models of a factor analysis type may be fitted to the resulting contingency table in order to capture these dimensions. Maximum likelihood type estimation methods can be computationally burdensome and we shall consider methods which reduce this computation and can be adapted to allow for weights and other features of complex survey sampling schemes. We focus on the problem of testing the goodness-of-fit of the model or associated nested hypotheses. Because the contingency table can be sparse when the number of items is large, even for large sample surveys, we consider methods which focus on lower-order margins of the table. Such methods are sometimes called limited information tests. We consider their extension to complex survey data.

Keywords: latent variable models, limited information test statistics, pairwise estimation

1 Introduction

Data on multiple categorical variables (items) are often collected in surveys in order to measure a smaller number of underlying dimensions. Latent variable models, as in factor analysis, may then be fitted in order to capture these dimensions. In this paper we consider the problem of testing the goodness of fit of such models as well as associated nested hypotheses.

Let $y_1, \ldots, y_p$ denote p measured variables of interest. For simplicity, we just consider the case when each y_i is binary. We are interested in models which capture the association between the y_i variables in terms of latent variables $z_1, \ldots, z_q$, where $q < p$. Let $\mathbf{y}' = (y_1, \ldots, y_p)$ denote the vector of p binary observed variables. There are $R = 2^p$ possible response patterns of the form $\mathbf{y}_r' = (c_1, c_2, \ldots, c_p)$, where $c_i = 0$ or 1. A parametric model is specified by assuming that the values of $\mathbf{y}$ for units in the population are generated by a model, where $\mathbf{y} = \mathbf{y}_r$ with probability $\pi_r(\boldsymbol{\theta}) \geq 0$, $\boldsymbol{\theta}$ is a parameter vector and $\sum_{r=1}^{R} \pi_r(\boldsymbol{\theta}) = 1$.

For example, in one approach to the specification of the latent variable model, the observed binary variables y_i are taken to be manifestations of underlying continuous variables $y_i^\star$ (e.g. Muthén, 1984). It is assumed that $y_i = 0$ if $y_i^\star \leq \tau_i$ and 1 otherwise, where τ_i is the threshold associated with variable y_i^*. For convenience, the distribution of $y_i^\star$ is assumed to be standard normal. The factor model is of the form

$$\mathbf{y}^\star = \Lambda \boldsymbol{z} + \boldsymbol{\epsilon} \ , \tag{1}$$

where $\mathbf{y}^\star$ is the p-dimensional vector of the underlying variables, Λ is the $p \times q$ matrix of loadings, and $\boldsymbol{\epsilon}$ is the p-dimensional vector of unique variables. In addition, it is assumed that $\boldsymbol{z} \sim N_q(\mathbf{0}, \Phi)$ where Φ has 1's on its main diagonal, $\boldsymbol{\epsilon} \sim N_p(\mathbf{0}, \Theta)$ with Θ a diagonal matrix, $\Theta = I - \text{diag}(\Lambda\Phi\Lambda')$, and $Cov(\boldsymbol{z}, \boldsymbol{\epsilon}) = \mathbf{0}$. The parameter vector $\boldsymbol{\theta}' = (\boldsymbol{\lambda}, \boldsymbol{\varphi}, \boldsymbol{\tau})$ contains $\boldsymbol{\lambda}$ and $\boldsymbol{\varphi}$, the vectors of free non-redundant parameters in matrices Λ and Φ, respectively, and $\boldsymbol{\tau}$, the vector of free thresholds.

Under this model, the probability of response pattern r is

$$\pi_r(\boldsymbol{\theta}) = \pi\left(y_1 = c_1, \ldots, y_p = c_p; \boldsymbol{\theta}\right) = \int \ldots \int \phi_p(\mathbf{x}^\star; \Sigma_{\mathbf{y}^\star}) d\mathbf{y}^\star \ , \tag{2}$$

where $\phi_p(\mathbf{y}^\star; \Sigma_{\mathbf{y}^\star})$ is a p-dimensional normal density with zero mean, and correlation matrix $\Sigma_{\mathbf{y}^\star} = \Lambda\Phi\Lambda' + \Theta$.

In the case of complex survey designs, where the sample is drawn from a finite population of size N, the pseudo log-likelihood is given by

$$\ln L(\boldsymbol{\theta}; \mathbf{y}) = \sum_{r=1}^{R} \hat{N}_r \ln \pi_r(\boldsymbol{\theta}) \ , \tag{3}$$

where $\hat{N}_r$ is the sum of survey weights across sample units with response pattern r. The computation of the estimator which maximises this log-likelihood is very demanding, however, since it requires the evaluation of the p-dimensional integral in (2), which cannot be written in a closed form.

As a consequence, various limited information estimation methods have been proposed. The most widely used methods involve three-stage estimation (Jöreskog, 1994 ; Muthén, 1984). In this paper, we consider an alternative limited information method, which extends the pairwise likelihood (PL) approach (Katsikatsou, Moustaki, Yang-Wallentin, and Jöreskog, 2012) to complex designs. This approach is based on the bivariate marginal distributions of pairs of observed variables. Thus, $\pi_{c_i c_j}^{(y_i y_j)}(\boldsymbol{\theta})$ is the probability that

$y_i = c_i$ and $y_j = c_j$ under the model. The pairwise pseudo log-likelihood is given by:

$$pl(\boldsymbol{\theta}; \mathbf{x}) = \sum_{i<j} \sum_{c_i=0}^{1} \sum_{c_j=0}^{1} \hat{N}_{c_i c_j}^{(y_i y_j)} \ln \pi_{c_i c_j}^{(y_i y_j)}(\boldsymbol{\theta}) \ , \tag{4}$$

where $\hat{N}_{c_i c_j}$ is the sum of survey weights across sample units with $y_i = c_i$ and $y_j = c_j$. Maximizing this log-likelihood with respect to $\boldsymbol{\theta}$ gives the pairwise pseudo maximum likelihood estimator $\hat{\boldsymbol{\theta}}_{PL}$. This estimator only requires the evaluation of up to two-dimensional normal probabilities, regardless of the number of observed or latent variables. It may be argued that this estimator is consistent for $\boldsymbol{\theta}$, provided the model holds at the population level.

2 Goodness-of-fit testing

Let π_r denote the probability of response pattern r, either assumed to apply to each unit in the population, or else we simply take π_r to be a finite population proportion. We wish to test the null hypothesis $H_0 : \pi_r = \pi_r(\boldsymbol{\theta})$ against the alternative H_1 that π_r is unrestricted, subject to $\sum \pi_r = 1$.

Even if maximization of the pseudo likelihood were feasible, there are reasons against constructing goodness-of-fit tests based upon standard likelihood ratio or Pearson tests for all 2^p possible response patterns, since the resulting contingency table may be sparsE with many zero and small frequencies which will distort the approximation to the chi-square distribution (Reiser and VandenBerg, 1994). As a result, limited information goodness-of-fit tests have been proposed by (Reiser, 1996, 2008 ; Bartholomew and Leung, 2002 ; Maydeu-Olivares and Joe, 2005, 2006 ; Cai, Maydeu-Olivares, Coffman, and Thissen, 2006 ; Cagnone and Mignani, 2007), at least for the classical case of simple random sampling. These tests are based on marginal distributions rather than on all 2^p response patterns.

Vectors of marginal probabilities are defined as follows. Let $\dot{\boldsymbol{\pi}}_1 = P(y_i = 1),\ i = 1, \ldots, p$ be the $p \times 1$ vector that contains all univariate probabilities of a positive response to the ith item. Let $\dot{\boldsymbol{\pi}}_2'$ be the $\binom{p}{2} \times 1$ vector of bivariate probabilities with elements, $\dot{\pi}_{ij} = P(y_i = 1, y_j = 1), j < i$. Vectors of higher order marginal probabilities may be constructed similarly. Let $\boldsymbol{\pi}_k$ be the vector formed by stacking these vectors up to order k, so that it has dimension $s = s(k) = \sum_i^k \binom{p}{i}$. One can find an indicator matrix T of dimension $s \times 2^p$ and full row rank such that $\boldsymbol{\pi}_k = T_k \boldsymbol{\pi}$, where $\boldsymbol{\pi}$ is the vector containing the probabilities π_r for all 2^p possible response patterns r.

Let $\mathbf{p}$ and $\mathbf{p_k}$ denote the vector of weighted sample proportions corresponding to $\boldsymbol{\pi}$ and $\boldsymbol{\pi}_k$, respectively. Then standard central limit theorems for complex sampling give

$$\sqrt{n}(\mathbf{p} - \boldsymbol{\pi}) \xrightarrow{d} N(0, \Sigma), \tag{5}$$

where n is the sample size. The covariance matrix Σ takes the multinomial form, given by $\Sigma = D(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}'$, in the case of simple random sampling, but not in general. It follows that:

$$\sqrt{n}(\mathbf{p}_k - \boldsymbol{\pi}_k) \xrightarrow{d} N(0, \Sigma_k), \tag{6}$$

where $\Sigma_k = T_k \Sigma T_k$. Because T_k is of full rank s, Σ_k is also of full rank s.

In the case of a simple null hypothesis $H_0 : \boldsymbol{\pi} = \boldsymbol{\pi}_0$, a Wald test statistic, following Maydeu-Olivares and Joe (2005) for the classical case, is given by:

$$L_k = n(\boldsymbol{p}_k - \boldsymbol{\pi}_{k0})' \Sigma_k^{-1} (\boldsymbol{p}_k - \boldsymbol{\pi}_{k0}), \tag{7}$$

where $\boldsymbol{\pi}_{k0} = T_k \boldsymbol{\pi}_0$ and L_k is distributed as χ^2 with $s(k)$ degrees of freedom as n goes to infinity.

The value k should be chosen to be sufficiently large for the model to be identified from the joint moments up to k. We can derive results analogous to Theorem 14.8-3 in Bishop, Fienberg, and Holland (1975):

$$\hat{\boldsymbol{\theta}}_{PL} - \boldsymbol{\theta} = B(\boldsymbol{p} - \boldsymbol{\pi}(\boldsymbol{\theta})) + O_p(n^{-1/2}), \tag{8}$$

where $B = J^{-1} \Delta D_\pi^{-1}$, $D_\pi = diag(\boldsymbol{\pi})$, $J = \Delta' D_\pi^{-1} \Delta$ is of dimension $s \times s$ and $\Delta = \frac{\partial \boldsymbol{\pi}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ is of dimension $2^p \times dim(\theta)$.

It follows that

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_{PL} - \boldsymbol{\theta}) \xrightarrow{d} N(0, J^{-1}). \tag{9}$$

Let $\hat{\mathbf{e}} = \boldsymbol{p} - \boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_{PL})$ be the vector of unstandardized residuals. A Taylor series expansion gives:

$$\boldsymbol{\pi}(\hat{\boldsymbol{\theta}}_{PL}) = \boldsymbol{\pi}(\boldsymbol{\theta}) + \Delta(\hat{\boldsymbol{\theta}}_{PL} - \boldsymbol{\theta}) + O_p(n^{-1/2}) \tag{10}$$

So

$$\hat{\mathbf{e}} = \boldsymbol{p} - \boldsymbol{\pi}(\boldsymbol{\theta}) - \Delta B(\boldsymbol{p} - \boldsymbol{\pi}(\boldsymbol{\theta})) + O_p(n^{-1/2}) = (I - \Delta B)(\boldsymbol{p} - \boldsymbol{\pi}(\boldsymbol{\theta})).$$

So from (5):

$$\sqrt{n}\hat{\mathbf{e}} \xrightarrow{d} N(0, (I - \Delta B)\Sigma(I - \Delta B)'). \tag{11}$$

where $(I-\Delta B)\Sigma(I-\Delta B)'$ can be reduced to $D_\pi - \boldsymbol{\pi}\boldsymbol{\pi}' - \Delta(A'A)^{-1}\Delta'$ under simple random sampling.

Residuals of lower order are given by $\hat{\mathbf{e}}_k = \boldsymbol{p}_k - \boldsymbol{\pi}_k(\hat{\boldsymbol{\theta}}_{PL}) = T_k\hat{\mathbf{e}}$. From (11), we have that

$$\sqrt{n}\hat{\mathbf{e}}_k \xrightarrow{d} N(0, T_k(I-\Delta B)\Sigma(I-\Delta B)'T_k'), \tag{12}$$

where $T_k(I-\Delta B)\Sigma(I-\Delta B)'T_k' = T_k(D_\pi - \boldsymbol{\pi}\boldsymbol{\pi}' - \Delta(A'A)^{-1}\Delta')T_k'$. To estimate the asymptotic covariance matrix of $\hat{\mathbf{e}}_k$ we estimate Δ and A at $\hat{\boldsymbol{\theta}}_{PL}$ giving:

$$T_k(I-\hat{A}\hat{B})\Sigma(I-\hat{A}\hat{B})'T_k'$$

where $\hat{B} = (\hat{A}'\hat{A})^{-1}\hat{A}'\hat{D}_\pi^{-1/2}$.

Finally, the test statistic is written as:

$$L_k = n(\boldsymbol{p}_k - \boldsymbol{\pi}_k(\hat{\boldsymbol{\theta}}_{PL}))'\hat{\Sigma}_k^+(\boldsymbol{p}_k - \boldsymbol{\pi}_k(\hat{\boldsymbol{\theta}}_{PL})), \quad k = 1, \ldots, n \tag{13}$$

where $\hat{\Sigma}_k^+$ is the Moore-Penrose inverse of $\Sigma_k(\hat{\boldsymbol{\theta}}_{PL})$. Under the H_0, this test statistic is asymptotically distributed as χ^2 with degrees of freedom equal to the rank of Σ_k.

Références

Bartholomew, D. J. and S. O. Leung (2002). A goodness of fit test for sparse 2^p contigency tables. *British Journal of Mathematical and Statistical Psychology 55*, 1–15.

Bishop, Y. M., S. Fienberg, and P. Holland (1975). *Discrete Multivariate Analysis*. Cambridge, Mass.: MIT Press.

Cagnone, S. and S. Mignani (2007). Assessing the goodness of fit of a latent variable model for binary data. *Metron LXV*, 337–361.

Cai, L., A. Maydeu-Olivares, D. L. Coffman, and D. Thissen (2006). Limited information goodness-of-fit testing of item response theory models for sparse 2^p tables. *British Journal of Mathematical and Statistical Psychology 59*, 173–194.

Jöreskog, K. G. (1994). On the estimation of polychoric correlations and their asymptotic covariance matrix. *Psychometrika 59*, 381–389.

Katsikatsou, M., I. Moustaki, F. Yang-Wallentin, and K. G. Jöreskog (2012). Pairwise likelihood estimation for factor analysis models with ordinal data. *Computational Statistics and Data Analysis 56*, 4243–4258.

Maydeu-Olivares, A. and H. Joe (2005). Limited and Full-information estimation and Goodness-of-Fit testing in 2^n contingency tables: A unified framework. *Journal of the American Statistical Association 100*, 1009–1020.

Maydeu-Olivares, A. and H. Joe (2006). Limited information goodness-of-git testing in multidimensional contigency tables. *Psychometrika 71*, 713–732.

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

Reiser, M. (1996). Analysis of residuals for the multinomial item response model. *Psychometrika 61*, 509–528.

Reiser, M. (2008). Goodness-of-fit testing using components based on marginal frequencies of multinomial data. *British Journal of Mathematical and Statistical Psychology 61*, 331–360.

Reiser, M. and M. VandenBerg (1994). Validity of the chi-square test in dichotomous variable factor analysis when expected frequencies are small. *British Journal of Mathematical and Statistical Psychology 47*, 85–107.