A STATISTICAL TEST TO ASSESS THE NON-NORMALITY OF THE LATENT VARIABLE DISTRIBUTION

Lucia Guastadisegni[1], Irini Moustaki[2], Silvia Cagnone[1] and Vassilis Vasdekis[3]

[1] Department of Statistical Sciences ”Paolo Fortunati”, University of Bologna, (e-mail: lucia.guastadisegni2@unibo.it, silvia.cagnone@unibo.it)

[2] Department of Statistics, London School of Economics and Political Science, (e-mail: i.moustaki@lse.ac.uk)

[3] Department of Statistics, Athens University of Economics and Business, (e-mail: vasdekis@aueb.gr)

ABSTRACT: This paper presents the generalized Hausman test to detect non-normality of the latent variable distribution in unidimensional Item Response Theory (IRT) models for binary data. The test is based on the estimators resulting from the two-parameter IRT model, that assumes normality of the latent variable, and the semi-nonparametric IRT model, that assumes a more flexible latent variable distribution. The performance of the test is evaluated through a simulation study, including the cases where the latent variable is generated from a skew-normal and mixture of normals. The results highlight the good performance of the test when the latent variable is generated from a mixture of normals and from a skew-normal only with many items and large sample sizes.

KEYWORDS: generalized Hausman test, SNP-IRT model, binary data

1 Introduction

In unidimensional IRT models for binary data, the latent variable is typically assumed standard normally distributed. However, assuming normality in the model when the true latent variable distribution has a different shape than the normal one can result in large biases in parameter estimates (Ma & Genton, 2010). IRT models that assume different form of the latent variable have been proposed (for example Irincheeva *et al.*, 2012) but detecting latent variable non-normality through a statistical test remains an open issue. In this paper, we consider the generalized Hausman (GH) test (White, 1982) to detect non-normality of the latent variable distribution in unidimensional IRT models for binary data. The test is based on the maximum pairwise likelihood (PL) estimator (Lindsay, 1988) of the classical unidimensional IRT model for binary data, based on the normality assumption of the latent variable, and the quasi-maximum likelihood (ML) estimator of the unidimensional seminonparametric (SNP)-IRT model for binary data, that assumes a more flexible latent variable distribution (Irincheeva *et al.*, 2012). Some preliminary results on the performance of the GH test have been presented in Guastadisegni *et al.* (forthcoming). In details, the GH test has shown a good performance in terms of Type I error rates with many items and large sample size. The power of this test has only been evaluated when the latent variable is generated from a mixture of normals. In this paper, we evaluate the performance of the GH test also when the latent variable is generated from a skew-normal distribution.

2 The IRT models for binary data

Let $y_1,...,y_p$ denote a set of observed binary variables/items, n the number of individuals and z the latent variable with density function $h(z)$. The response probability for the i-th individual to the j-th item is modelled using a logistic model (measurement model)

$$P(y_{ij}=1|z_i)=\pi_{ij}(z_i)=\frac{\exp(\alpha_{0j}+\alpha_{1j}z_i)}{1+\exp(\alpha_{0j}+\alpha_{1j}z_i)}, \tag{1}$$

where α_{0j} is the item intercept and α_{1j} the item slope. For the classical IRT model, $h(z)=\phi(z)$, where $\phi(z)$ is the density of a standard normal. For the SNP-IRT model, the latent variable has the following SNP parametrization (Irincheeva *et al.*, 2012)

$$h(z_i)=P_L^2(z_i)\phi(z_i) \qquad P_L(z_i)=\sum_{0\leq l\leq L} a_i z_i^l. \tag{2}$$

$a_0,...,a_L$ are the real coefficients of the polynomial $P_L(z_i)$ and L is the polynomial degree. SNP_1 denotes the model for $L=1$, where $P_L(z)=a_0+a_1z$, $a_0=\sin\varphi_1$, $a_1=\cos\varphi_1$, $-\pi/2<\varphi_1\leq\pi/2$. SNP_0 denotes the model for $L=0$, where the distribution of the latent variable reduces to the normal one. To implement the GH_T test, we consider the SNP_0 and the SNP_1 model.

3 The generalized Hausman test

Consider the maximum PL estimator $\tilde{\eta}_{SNP_0}$ of the SNP_0 model, that includes the item intercepts and slopes of dimension $2p\times 1$, where p is the number of items. Under normality of the latent variable distribution, the maximum PL estimator $\tilde{\eta}_{SNP_0}$ converges in probability to the true parameter value η_0. Consider also the quasi-ML estimator $\hat{\theta}'_{SNP_1} = (\hat{\eta}'_{SNP_1}, \hat{\varphi}_1)$ of the SNP_1 model, of dimension $(2p+1) \times 1$. Under normal, multi-modal and asymmetric distributions of the latent variables and if the regularity conditions A2-A6 of White (1982) are satisfied, the quasi-ML estimator $\hat{\theta}'_{SNP_1} = (\hat{\eta}'_{SNP_1}, \hat{\varphi}_1)$ converges to $\theta'_{0*} = (\eta'_0, \varphi_1*)$, where φ_1* is the value of φ_1 that minimizes the Kullback-Leibler information criterion. The GH test is defined as

$$GH = (\hat{\eta}_{SNP_1} - \tilde{\eta}_{SNP_0})' \hat{S}(\tilde{\eta}_{SNP_0}, \hat{\theta}_{SNP_1})^{-1} (\hat{\eta}_{SNP_1} - \tilde{\eta}_{SNP_0}). \tag{3}$$

Details on the computation of the matrix $\hat{S}(\tilde{\eta}_{SNP_0}, \hat{\theta}_{SNP_1})$ can be found in Guastadisegni *at al.* (forthcoming). Under normality of the latent variable distribution, the GH test is asymptotically distributed as a χ^2_{2p}, where $2p$ are the degrees of freedom. To avoid the inversion of the matrix $\hat{S}(\tilde{\eta}_{SNP_0}, \hat{\theta}_{SNP_1})$ that is numerically unstable, we consider the following statistic

$$GH_T = (\hat{\eta}_{SNP_1} - \tilde{\eta}_{SNP_0})' (\hat{\eta}_{SNP_1} - \tilde{\eta}_{SNP_0}). \tag{4}$$

Under normality of the latent variable distribution, $GH_T \sim a\chi^2_b$, where $a = \frac{\sum_{l=1}^d \lambda_l^2}{\sum_{l=1}^d \lambda_l}$ and $b = \frac{(\sum_{l=1}^d \lambda_l)^2}{\sum_{l=1}^d \lambda_l^2}$, d is rank of $\hat{S}(\tilde{\eta}_{SNP_0}, \hat{\theta}_{SNP_1})$ and $\lambda_1, ..., \lambda_d$ are its non-zero eigenvalues.

4 Simulation study and results

The optimization of the SNP_1 model is achieved in R with direct maximization via the function "nlminb", that uses analytically computed gradient and Hessian matrix, while the SNP_0 model via the function "optim". We consider the following simulation conditions: number of items $(p = 4, 10, 20)$, sample size $(n = 500, 1000)$, 500 replications for each condition and $\alpha = 0.05$. Data are generated from a 2-PL model with the following latent variable distributions:

A $z \sim N(0,1)$
B $z \sim 0.7N(-1.5, 0.6) + 0.3N(1.5, 0.5)$, where z has an overall mean equal to -0.6 and variance equal to 2.217.
C $z \sim SN(\mu = 0, \sigma = 2.5, \lambda = 10)$, where z has mean 1.98 and variance 2.31.

Table 1 presents Type I error rates and power of the GH_T test for scenarios A, B and C. Overall, under scenario A, the GH_T test has good performance in terms

Table 1. *Type I error rates and power of the GH_T test for scenarios A, B, and C, $p = 4, 10, 20$, $n = 500, 1000$.*

		Type I error	Power	
p	n	**A**	**B**	**C**
4	500	0.016	0.796	0.03
	1000	0.086	0.92	0.234
10	500	0.018	1	0.388
	1000	0.044	1	0.59
20	500	0.056	0.986	0.744
	1000	0.06	1	0.918

of Type I error rates when the sample size is large and in general with many items. Under scenario B, the power of the GH_T test is high for most conditions. However, under scenario C, 4 and 10 items, the GH_T test has low power to detect non-normality of the latent variable distribution. It reaches a high power only with 20 items and large sample sizes. The low power of the test under scenario C can be due to the following reasons. First, the SNP_1 model does not approximate very well the skew-normal distributions (Irincheeva *et al.*, 2012). Second, the skew-normal distribution used in the simulations has a very high mean and this has a negative impact on the estimation of parameters.

References

GUASTADISEGNI, L., MOUSTAKI, I., VASDEKIS, V., & CAGNONE, S. Forthcoming. Detecting latent variable non-normality through the generalized Hausman test. *In: Quantitative Psychology: The 87th Annual Meeting of the Psychometric Society, Bologna, 2022.* Springer.

IRINCHEEVA, I., CANTONI, E., & GENTON, M. G. 2012. Generalized linear latent variable models with flexible distribution of latent variables. *Scandinavian Journal of Statistics*, **39**, 663–680.

LINDSAY, B. G. 1988. Composite likelihood methods. *Contemporary mathematics*, **80**, 221–239.

MA, Y., & GENTON, M. G. 2010. Explicit estimating equations for semiparametric generalized linear latent variable models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **72**, 475–495.

WHITE, H. 1982. Maximum likelihood estimation of misspecified models. *Econometrica*, **50**, 1–25.