MODELLING MISSINGNESS WITH A RASCH-TYPE MODEL

Bertoli-Barsotti, L.*
Dipartimento di Matematica, Statistica, Informatica e Applicazioni
Università di Bergamo
Via dei Canina 2 – 24127 Bergamo (Italy)
Mail: lucio.bertoli-barsotti@unibg.it; Phone&Fax: +390352052563

Punzo, A.
Dipartimento di Impresa, Culture e Società
Università di Catania
antonio.punzo@unict.it

Corresponding author: Bertoli-Barsotti, L.

Keywords: Multidimensional Rasch Models, non-ignorable missing data

Topic area of the submission: Rasch Models, Multidimensional IRT Models

Introduction

A very common problem in applications of item response theory is the presence of non-responses. Omitted responses are often treated by practitioners as not administered -e.g. when the amount of missing data is small- but this method is adequate only if data are missing at random. Recently a new approach to overcome this problem has been developed by several authors ([1], [2]); the main idea is to incorporate the non-response mechanism into the analysis, under the assumption of a bidimensional latent trait; the first dimension of the latent trait represents a general person's tendency to respond (response propensity), while the second dimension represents the usual person's attitude, or "ability". In this study, we shall refer to the case of dichotomously scored data. Let consider the following "codes" -as characters that are to be regarded as valid data for the dataset: A, representing a non-response (as a missing value code)), B if the subject responds negatively to the item and C if the subject responds positively to the item. It is worth noting that A, B and C are related in a hierarchical manner. The probability of giving a response will depend on person's position on both the first and the second dimension of the latent trait, respectively represented by θ_1 and θ_2. Finally, to combine these two dimensions in the analysis it may be assumed that θ_1 and θ_2 have a common distribution –e.g. a bivariate normal with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. More specifically, [2] introduced and studied the model $P_A \propto 1+\exp(\theta_2-\delta_2)$, $P_B \propto \exp(\theta_1-\delta_1)$, $P_C \propto \exp(\theta_1+\theta_2-\delta_1-\delta_2)$, that is equivalent to the *continuation-ratio logits* model: $\log\left(\left(P_B+P_C\right)/P_A\right)=\theta_1-\delta_1$; $\log\left(P_C/P_B\right)=\theta_2-\delta_2$. The major drawback of this approach is that it yields a *non-Rasch* model.

Modelling missingness with a Rasch-type model

We propose a new model (or class of models) that we will call Rasch-Rasch model (RRM), by adopting –instead of the above mentioned- the following *adjacent-categories logits* approach: (i) $\log\left(P_B/P_A\right)=\boldsymbol{a}'\boldsymbol{\theta}-\delta_1$, (ii) $\log\left(P_C/P_B\right)=\left(\boldsymbol{b}'-\boldsymbol{a}'\right)\boldsymbol{\theta}-\delta_2$, where $\boldsymbol{a}$ and $\boldsymbol{b}$ are know vectors of constants, and $\boldsymbol{\theta}'=\left(\theta_1,\theta_2\right)$. Indeed, both the steps A-B and B-C are modeled by a multidimensional Rasch model and, more importantly, the obtained whole model belongs to the family of Rasch models. This means that RRM is a member of the exponential family of distributions. Moreover, the RRM can be viewed as a special case of the most general structure of multidimensional Rasch models: the Multidimensional Random Coefficients Multinomial Logit Model (MRCMLM, [3]). In a more special way, the RRM can be seen as an instance of multidimensional *within-item* model [4], where the within-item multidimensionality is specified by the fact that -for every item- the different response "levels" (A, B and C) require different *linear combinations* of the latent traits θ_1 and θ_2 (note: not merely different latent traits). In order to properly interpret θ_2 as the sole dimension of "ability", one should take $a_1=b_1$ and $a_2<b_2$. By reasons of symmetry, among several possible choices of the constants $\boldsymbol{a}$ and $\boldsymbol{b}$, the most effective –avoiding possible effects of structural spurious correlation between the components of the latent trait- is the following: $\boldsymbol{a}'=\left(1,-1\right)$ and $\boldsymbol{b}'=\left(1,1\right)$. Then, in a more explicit form, the proposed model is

$$P_A\propto 1;\; P_B\propto\exp(\theta_1-\theta_2-\delta_1);\; P_C\propto\exp(\theta_1+\theta_2-\delta_1-\delta_2)\,. \tag{1}$$

Now, let $\boldsymbol{y}_{vi}=\left(y_{vi0},y_{vi1},y_{vi2}\right)$ be a selection vector, with values (1,0,0), (0,1,0) and (0,0,1), respectively, for A, B and C, with reference to respondent v and item i. By considering the RRM in its form (1), the sum $y_{v\bullet 1}+y_{v\bullet 2}$ (i.e. the total number of responses of person v) is a sufficient statistic for parameter θ_1, and the difference $y_{v\bullet 2}-y_{v\bullet 1}$ is a sufficient statistic for parameter θ_2. Then, θ_1 may be interpreted as the person's response propensity and θ_2 as the person's ability. Analogously, for item parameters: δ_1 (for which $y_{\bullet i1}+y_{\bullet i2}$ is a sufficient statistic) represents the item's tendency to provoke reluctance to respond, while δ_2 (for which $y_{\cdot i2}$ is a sufficient statistic) represents the usual item's difficulty.

Estimation of the model

As for all the other Rasch-type models, for RRM both conditional and Marginal Maximum Likelihood (MML) approaches to the estimation of item parameters are feasible. In this study we shall consider the latter, by assuming that the latent trait is distributed as a bivariate normal with a mean $\boldsymbol{\mu}$ that is set at 0 -as a constraint for identification- and a covariance matrix $\boldsymbol{\Sigma}$. The coefficient of correlation ρ can be used to measure the extent to which ability and response propensity are related. The most extreme case of $\rho = 1$ leads to the Partial Credit Model (PCM); now, since the RRM is hierarchically related to the unidimensional parameterization, the change in deviance value may be used to compare statistically the fit of these two models. The software package Conquest ([4]; Conquest is implemented with MML estimation) can be used to calibrate the parameters of the RRM. For a test containing k dichotomous items, the software produce estimates for a total of $2k+3$ parameters: $2k$ item parameters and 3 elements for the covariance matrix.

Statistical simulation

Data patterns of dimension 250×12 (i.e. nearly the same as that of the dataset used as a case study in this paper) were randomly generated from the model proposed by [2], for the following 11 equidistant values of ρ: $-1, -0.8, -0.6, \ldots, 0.6, 0.8, 1$. In detail, for each value of ρ, 100 matrices were randomly generated and a RRM was fitted; for each replication, the MML-estimated correlation coefficient ρ_{RRM} was registered and then the average $\bar{\rho}_{RRM}$ over the 100 replications was computed. The following sequence of values for $\bar{\rho}_{RRM}$ illustrates the simulation results: $-.903, -.761, -.595, -.383, -.211, -.011, .185, .396, .594, .767, .921$. In spite of the small sample, Conquest provides a satisfactory correlation parameter recovery under the RRM in its form (1).

Illustration with real data

The Voluntary Counseling and Testing Efficacy Study was a randomized clinical trial conducted to test the efficacy of voluntary HIV-1 counseling and testing in reducing sexual risk behaviour [5]. A selection of 12 dichotomously scored items –concerning opinions about condom use- from this dataset were considered for this application. Due to the negative wording of six out of the 12 items, the scores (disagree=0, agree=1) have been reversed for these questions (see Table 1). A total of 258 male subjects were considered – in such a way that the obtained dataset presented at least 1 missing for each row and each column. The total amount of missing data was about 20%. A short description of the content of the 12 items of the calibrated test, together with the estimates of the item parameters of the RRM and their standard errors are given in Table 1. The estimated correlation between dimensions was $\rho_{RRM} = 0.338$. This indicates –as expected- that the amount of missing responses went down with the increment of the level of agreement. The difference between the deviances of RRM and the PCM was also tested, using a chi-square test. The PCM fitted the data significantly less well than the RRM: the difference in deviance (38.48) was highly significant when compared to a chi square distribution with 2 (2=27-25) degrees of freedom.

Table 1: The items in the calibrated test *Attitudes towards condom*. The number of responsents in the categories A, B and C are given. The difficulty parameters are given along with their standard errors in parentheses.

	Item wording (abbreviated)	reverse scoring	A	B	C	δ_1	δ_2
I1	using condoms good protection from stds		1	25	232	-3.421 (1.00)	-2.452 (0.21)
I2	sex not as good when you use a condom	yes	66	117	75	-0.660 (0.15)	0.539 (0.14)
I3	embarrassing to buy condoms	yes	6	77	175	-2.757 (0.42)	-0.912 (0.14)
I4	using condoms good pregnancy prevention		10	31	217	-1.305 (0.33)	-2.136 (0.18)
I5	embarrassing put on condom/ or on a man	yes	20	63	175	-1.307 (0.24)	-1.119 (0.14)
I6	frnds think use condoms incldng w/spouse		45	115	98	-1.064 (0.17)	0.215 (0.14)
I7	condoms often break or slip	yes	103	81	74	0.229 (0.13)	0.170 (0.15)
I8	if sex partner wants condom I suspect	yes	17	129	112	-2.218 (0.26)	0.177 (0.13)
I9	friends use condoms w/new partner		118	47	93	0.951 (0.13)	-0.653 (0.14)
I10	friends think that condoms uncomfortable	yes	79	131	48	-0.577 (0.14)	1.151 (0.17)
I11	friends thnk alwys use condom new person		69	48	141	0.314 (0.15)	-1.125 (0.13)
I12	most people your age using condoms now		77	64	117	0.141 (0.14)	-0.603 (0.13)

References

[1] Moustaki, I.; O'Muircheartaigh, C. A one dimension latent trait model to infer attitude from nonresponse for nominal data., Statistica, **2000**, pp. 259-276.

[2] Holman, R.; Glass, A.W. Modelling non-ignorable missing-data mechanism with item response theory models, British Math. Stat. Psych, **2005**, pp. 1-17.

[3] Adams, R.J., Wilson, M.R., and Wang, W.C. The multidimensional random coefficients multinomial logit. Applied Psychological Measurement, **1997**, *21*, 1–24.

[4] Wu, M. L., Adams, R. J., Wilson, M. R. & Haldane, S., *ACER ConQuest 2.0:General item response modelling software* [computer program manual]. Camberwell, VIC: ACER Press, **2007**.

[5] The Voluntary HIV-1 Counseling and Testing Efficacy Study Group. Efficacy of voluntary HIV-1 counselling and testing in individual and couples in Kenya, Tanzania, and Trinidad: a randomised trial. Lancet**, 2000**, *256*, 103-112.