Log-linear multidimensional Rasch model for capture-recapture

Elvira Pelle, *University of Milano-Bicocca*, e.pelle@campus.unimib.it
David J. Hessen, *Utrecht University*, D.J.Hessen@uu.nl
Peter G.M. Van der Heijden, *Utrecht University, University of Southampton*, P.G.M.vanderheijden@uu.nl

Abstract. The traditional capture-recapture method assumes homogeneity of the capture probabilities. However, differences of character or behaviour between individuals may occur and models that allow for varying susceptibility to capture over individuals and unequal catchability have been proposed and psychometric models, such as the Rasch model, were successfully applied. In the present work, we propose the use of the multidimensional Rasch model in the capture-recapture context. We assume that lists may be divided into two or more subgroups, such that they can be viewed as indicators of the latent variables which account for correlations among lists. We show how to express the probability of a generic capture profile in terms of log-linear multidimensional Rasch model and apply the methodology to a real data set.

Keywords. Rasch model, capture-recapture, heterogeneity, log-linear model, EM algorithm

1 Introduction

The capture-recapture method is a statistical method originally used to estimate the size of wildlife populations [1] based on a sequence of trapping experiments where individual trapping histories are used to estimate the population size.
The capture-recapture method has been successful applied to other contexts, like human populations, where the common labels are *multiple-recapture*, *multiple-records systems* and *multiple-records systems method* [2]. In general, it can be applied to any situation in which two or more lists are available. Here, the estimation of the population size uses two or more incomplete but overlapping lists. Each list is regarded as a capture sample and data are usually arranged in an incomplete 2^s contingency table where the missing cell corresponds to absence in all s lists; then log-linear models are used to analyse the data [3].
The traditional capture-recapture method assumes that the S registrations are independent. If we allow possible dependencies between registrations, this results in interaction terms in log-linear models [4].

Another central assumption in the traditional capture-recapture approach is the homogeneity of the capture probability. However, differences of character or behaviour between individuals may cause indirect dependence between lists. Models that allow for varying susceptibility to capture through individuals and unequal catchability have been proposed either in the case of human populations [5] or in animal population studies [6] and psychometric models, such as the Rasch model, were successfully applied.

In applying the dichotomous Rasch model to the capture-recapture context, correct or incorrect answers to an item are replaced by "being observed" or "not being observed" in a list and, if all lists are supposed to be of the same kind, it is possible to treat heterogeneity in terms of constant apparent dependence between lists (Darroch [5], Agresti [6], International Working Group for Disease Monitoring and Forecasting [2]). Bartolucci and Forcina [7], shown how to relax the basic assumptions of the Rasch model (conditional independence and unidimensionality) by adding some suitable columns to the design matrix of the model. Bartolucci and Pennoni [8] proposed an extension of the latent class model for behaviour effects in which the latent class of a subject follows a Markov chain with transition probabilities depending on the previous capture history.

In the present work, we propose the use of the multidimensional Rasch model in the capture-recapture context. In particular, we assume that lists may be divided into two or more subgroups, such that they can be viewed as indicators of the latent variables which account for correlations among lists. To do so, the extension of the Dutch Identity for the multidimensional partial credit model (Hessen [9]) can be utilized. The Dutch Identity is a tool proposed by Holland [10], useful in the study of the structure of item response models, used by psychometricians to explain the characteristics and performance of a test. We use the results of Hessen [9], typically used in psychometric context, in the capture-recapture framework to express the probability of a generic capture profile in terms of log-linear multidimensional Rasch model.

We proceed as follows: in Section 2 we discuss the situation with three lists and two latent variables. In Section 3 we present the model that allows for the presence of a stratifying variable. In Section 4 we describe the method for a more general situation with S lists and J strata. In Section 5 we apply the methodology proposed to a real data set on children born with a neural tube defect (NTD's) in the Netherlands.

2 Three lists

Consider a situation in which three lists R1, R2 and R3 are available. Let $n_{i_1 i_2 i_3}$ denote the observed frequencies of the data, where $i_s = (0, 1), i = 1, 2, 3$ and $i_s = 0$ denotes "not observed" and $i_s = 1$ denotes "observed". n_{000} is the number of individuals "not observed" in any list that has to be estimated in order to estimate the total unknown population size N. Data can be arranged in a 2^3 contingency table with a missing cell corresponding to absence in all the three lists (see Table 1).

Let $I_s, s = 1, 2, 3$ be the random variables denoting the presence or absence of an individual in the corresponding list. Assume that there are two latent variables which explain the correlation among lists, and suppose that I_1, I_2 and I_3 are conditionally independent given the two latent variables. Let $\boldsymbol{\Theta} = (\Theta_1, \Theta_2)$ denotes the vector of latent variables and $\boldsymbol{\theta} = (\theta_1, \theta_2)$ denotes a realization.

		R3			
		Observed		Not Observed	
		R2		R2	
		Observed	Not Observed	Observed	Not Observed
R1	Observed	n_{111}	n_{101}	n_{110}	n_{100}
	Not Observed	n_{011}	n_{001}	n_{010}	0^*

* Missing cell is treated as structurally zero cell

Table 1. Contingency table for three lists

We are interested in analysing a log-linear model that allows the presence of the two latent variables.

Let $\pi_{0_s}, s = 1, 2, 3$ be the probability of not being observed in the $s-$th list and let $\pi_{1_s} = 1 - \pi_{0_s}$ be the probability of being observed in the $s-$th list. The probability of inclusion in list s given the vector of latent variables may be expressed in a logistic form in the following way:

$$\pi_{1_s|\boldsymbol{\theta}} = \frac{e^{\mathbf{u}_s'\boldsymbol{\theta}-\delta_s}}{1 + e^{\mathbf{u}_s'\boldsymbol{\theta}-\delta_s}} \tag{1}$$

where δ_s is the parameter for the list s, θ_r is the parameter for the $r-$th latent variable and $\mathbf{u}_s'$ is the row vector of the (3×2) full column rank matrix $\mathbf{U} = [u_{sr}]$ of weights for the latent variables, where

$$u_{sr} = \begin{cases} 1 & \text{if the list } R_s \text{ is indicator of the } r-\text{th latent variable} \\ 0 & \text{otherwise} \end{cases}$$

Let $\mathbf{t} = (t_1, t_2)$ be the vector of total scores, where the total scores are given by $\mathrm{t}_1 = u_{11}i_1 + u_{21}i_2 + u_{31}i_3$ and $\mathrm{t}_2 = u_{12}i_1 + u_{22}i_2 + u_{32}i_3$.

Let $\mathbf{i} = (i_1, i_2, i_3)$ denotes a generic capture profile for an individual. According to standard probability theory, the probability of a generic capture profile $(\pi_{i_1i_2i_3})$ may be written as

$$\pi_{i_1i_2i_3} = \int \dots \int \pi_{i_1i_2i_3|\boldsymbol{\theta}} f(\boldsymbol{\theta})\, d\boldsymbol{\theta} \tag{2}$$

where $\pi_{i_1i_2i_3|\boldsymbol{\theta}}$ is the probability of a generic capture profile conditional to $\boldsymbol{\theta}$ and $f(\boldsymbol{\theta})$ is the multivariate density of $\boldsymbol{\theta}$. Under the assumption that the posterior distribution of the vector of latent variables conditional to the capture profile $i_1i_2i_3 = 000$ follows a multivariate normal distribution, we have

$$\begin{aligned} \pi_{i_1i_2i_3} &= \pi_{000} \exp\left\{ \sum_{s=1}^{3} i_s\delta_s + t_1\mu_1 + t_2\mu_2 + \frac{1}{2}t_1^2\gamma_{11} + \frac{1}{2}t_2^2\gamma_{22} + t_1t_2\gamma_{12} \right\} \\ &= \pi_{000} \exp\left\{ \sum_{s=1}^{3} i_s\delta_s + \mathbf{t}'\mu + \frac{1}{2}\mathbf{t}'\boldsymbol{\Gamma}\mathbf{t} \right\} \end{aligned} \tag{3}$$

where $\mathbf{t} = (t_1, t_2)' = \mathbf{i}'\mathbf{U}$, $\boldsymbol{\mu}$ is the mean vector and $\boldsymbol{\Gamma} = [\gamma_{ir}]$ is symmetric.

Let n the number of individuals observed in all lists. Since the probability of a generic capture profile $i_1i_2i_3$ has multinomial distribution, we can express the expected frequencies of

$n_{i_1 i_2 i_3}$ as

$$m_{i_1 i_2 i_3} = n\pi_{i_1 i_2 i_3} \tag{4}$$

Substituting (4) in (3) and taking the logarithm we obtain:

$$\ln m_{i_1 i_2 i_3} = \delta + \sum_{s=1}^{3} i_s \delta_s + \mathbf{t}'\mu + \frac{1}{2}\mathbf{t}'\mathbf{\Gamma}\mathbf{t} \tag{5}$$

where $\delta = \ln(n\pi_{000})$.

The model in equation (5) is not identified. Setting $\boldsymbol{\mu} = \mathbf{0}$ for identification, the log-linear multidimensional Rasch model can be rewritten as:

$$\begin{aligned} \ln m_{i_1 i_2 i_3} &= \delta + \sum_{s=1}^{3} i_s \delta_s + \frac{1}{2}\mathbf{t}'\mathbf{\Gamma}\mathbf{t} \\ &= \delta + i_1\delta_1 + i_2\delta_2 + i_3\delta_3 + \frac{1}{2}t_1^2\gamma_{11} + \frac{1}{2}t_2^2\gamma_{22} + t_1 t_2 \gamma_{12} \end{aligned} \tag{6}$$

Note that there are $2(2+1)/2 = 3$ parameters to account for the two latent variables θ_1 and θ_2. In particular, γ_{11} and γ_{22} represent, respectively, the variance of the first latent variable and the variance of the second latent variable, given the total scores t_1 and t_2 , while γ_{12} represents the covariance between the two latent variables, given the total scores t_1 and t_2. In general, to account for q latent variables, we need $q(q+1)/2$ parameters.

3 Model with a stratifying variable

Suppose now that a stratifying variable is available. Let $n_{i_1 i_2 i_3 j}$ and $\pi_{i_1 i_2 i_3 j}$ denote the observed frequencies and the probabilities for strata j, respectively. In this case, n_{000j} indicates the frequency of individuals not observed in any lists in the j-th strata. With two strata the contingency table has two missing cells, as shown in Table 2.

		R3			
		Observed		Not Observed	
		R2		R2	
Year	R1	Observed	Not Observed	Observed	Not Observed
1	Observed	n_{1111}	n_{1011}	n_{1101}	n_{1001}
	Not Observed	n_{0111}	n_{0011}	n_{0101}	0^*
2	Observed	n_{1112}	n_{1012}	n_{1102}	n_{1002}
	Not Observed	n_{0112}	n_{0012}	n_{0102}	0^*

* Missing cell are treated as structurally zero cells

Table 2. Contingency table for three lists and two strata

We assume that lists are indicators of the latent variables which explain correlations among lists and the posterior distribution of the latent variables (given the capture profile of not observed) follows a multivariate normal distribution; similarly to the previous case, we have

$$\ln m_{i_1 i_2 i_3 j} = \delta_j + \sum_{s=1}^{3} i_s \delta_{sj} + \frac{1}{2}\mathbf{t}'\mathbf{\Gamma_j}\mathbf{t} \tag{7}$$

where $\delta_j = \ln(n\pi_{000j})$, $\mathbf{\Gamma}_j$ is a symmetric matrix, $\mathbf{t}$ is the vector of total scores and the mean vector for the j-th strata $\boldsymbol{\mu}_j$ is set equal to zero for identification.

If parameters are equal across the strata $\delta_{sj} = \delta_s, \forall j$, that is the assumption of measurement invariance holds, we can test whether $\boldsymbol{\mu}_j = \boldsymbol{\mu} = \mathbf{0}$ and $\mathbf{\Gamma}_j = \mathbf{\Gamma}$ for all j.

If the simultaneous hypothesis holds, then the model in (7) becomes

$$\ln m_{i_1 i_2 i_3 j} = \delta_j + \sum_{s=1}^{3} i_s \delta_s + \frac{1}{2}\mathbf{t}'\mathbf{\Gamma}\mathbf{t}. \tag{8}$$

4 Generalization

The extension to a more general situation with S registrations and J strata is straightforward.

Let $n_{i_1 \ldots i_s j}$ and $\pi_{i_1 \ldots i_s j}$ be the observed frequencies and the probabilities, respectively, where the index $i_s, s = 1, 2, \ldots, S$ denotes the cross-classification of S lists and $j = (1, 2, \ldots, J)$ is the index denoting the strata. The resulting contingency table has J structural zeros (one for each stratum).

Suppose now that the covariances between the random variables $I_1, \ldots, I_S$ can be explained by q latent variables. Let $\mathbf{u}'_s$ denotes the $s-$th row of the $SJ \times q$ full column rank matrix $\boldsymbol{U} = [u_{sr}]$, where $u_{sr} = 1$ if list RS is indicator of the $r-$th latent variable and 0 otherwise, and let $\mathbf{t} = (t_1, \ldots, t_q)$ be the vector of the total scores of the latent variables, that is $t_r = \sum_{s=1}^{S} u_{sr} i_s$.

Under the assumption of a multivariate normal distribution of the posterior distribution of the latent variables (conditional to the capture profile of individuals not observed in any list) the log-linear multidimensional Rasch model takes the form:

$$\ln m_{i_1 \ldots i_s j} = \delta_j + \sum_{s=1}^{S} i_s \delta_{sj} + \mathbf{t}'\boldsymbol{\mu}_j + \frac{1}{2}\mathbf{t}'\mathbf{\Gamma}_{\boldsymbol{j}}\mathbf{t} \tag{9}$$

where $\boldsymbol{\mu}_j$ is the mean vector for the j-th strata, $\mathbf{t}$ is the vector of total scores and $\mathbf{\Gamma}_j$ is a symmetric matrix.

Without any additional constraints the model is not identified. If we set $\boldsymbol{\mu}_j$ equal to $\mathbf{0}$ for identification we have:

$$\ln m_{i_1 \ldots i_s j} = \delta_j + \sum_{s=1}^{S} i_s \delta_{sj} + \frac{1}{2}\mathbf{t}'\mathbf{\Gamma}_{\boldsymbol{j}}\mathbf{t} \tag{10}$$

The model in (10) can be treated as a traditional log-linear model. Thus, to estimate the parameters of the model it is possible to follow the approach proposed by Sanathanan [11] which consists of maximizing the conditional likelihood given the distribution of the observed frequencies (for more details see Sanathanan [11] and Bishop et al. [4]). Once the parameters have been estimated they can be used to obtain the estimate of the portion of population missed by all lists and thus the total unknown population size N.

5 Application

We apply the methodology described in the preceding Sections to the data set of five lists described by Zwane et al. [12] on children born with a NTD's in the Netherlands. Data cover a period of 11 years (from 1988 through 1998) and Year (denoted by Y_{cat}) is treated as a

stratifying variable. Since the five lists cover different but overlapping periods of time, we use the EM algorithm proposed by Zwane et al. [12] to estimate the missing cells resulting from the fact that lists partially overlap. None of these lists record all cases of NTD's in the Netherlands, and the scope of this application is to estimate the total unknown number of children affected by NTD's.

We assume that the five lists R1, R2, R3, R4 and R5 may be divided into two subgroups such that they can be view as indicators of the latent variables which account for correlations among lists. In particular, we consider two multidimensional Rasch models obtained using the methodology described above and compare them with other log-linear models. In total, we take into account five models.

Table 3 summarizes the results of these models fitted to the data; for each model we report the number of parameters, the degrees of freedom, the deviance, the value of AIC, the value of BIC and the estimated total population size $\hat{N}$. Table 4 presents the yearly estimates $\hat{N}_j, j = 1988, \ldots, 1998$ for each model.

Model	Design matrix	Par	df*	Dev	AIC	BIC	$\hat{N}$
1	R1+R2+R3+R4+R5+Y_{cat}	16	213	400	432	487	2229
2	1+(R1R2+...+R4R5)	26	203	298	350	439	3077
3	1+H1	17	212	349	383	441	3009
4	$1+\theta_1+\theta_2$	19	210	324	362	427	2793
5	$1+\theta_3+\theta_4$	19	210	311	349	414	3041

Table 3. Selected models with deviance, AIC and BIC

Model	$\hat{N}_{88}$	$\hat{N}_{89}$	$\hat{N}_{90}$	$\hat{N}_{91}$	$\hat{N}_{92}$	$\hat{N}_{93}$	$\hat{N}_{94}$	$\hat{N}_{95}$	$\hat{N}_{96}$	$\hat{N}_{97}$	$\hat{N}_{98}$
1	199	224	234	206	222	186	189	202	178	210	179
2	275	309	323	285	302	258	261	280	246	290	248
3	272	305	319	281	303	249	252	271	238	280	239
4	251	282	295	260	280	232	235	252	222	261	223
5	271	305	318	281	300	255	258	277	244	287	245

* There are 229 observed cells

§H1 is the first-order heterogeneity term

†$\theta_1 = R1 + R2$ and $\theta_2 = R3 + R4 + R5$

‡$\theta_3 = R1 + R2 + R4$ and $\theta_4 = R3 + R4 + R5$

Table 4. Selected models with yearly estimates

Note that the model with only the main-effect parameters does not fit the data well, as it has a high deviance. The model with the first-order heterogeneity parameter improves the fit, while adding all two-factor interaction parameters to Model 1 results in a smaller deviance.

Both of the multidimensional Rasch models fit the data well and Model 5, in which lists R1, R2 and R4 are assumed to be indicators of the first latent variable (θ_3) and lists R3, R4 and R5 are supposed to be indicators of the second latent variable (θ_4), is the best model, since it has the smallest value of AIC and BIC; thus, it is the selected model.

6 Conclusion

In this manuscript we proposed the use of the multidimensional Rasch model in the capture-recapture framework.

We assumed that lists may be divided into two or more subgroups (not necessarily disjoint) which constitute the latent variables accounting for correlations among lists. As consequence, the random variables denoting the presence or absence of an individual into a list are assumed to be conditionally independent, given the latent variable.

Under the assumption that the posterior distribution of the latent variables follows a multivariate normal distribution, we used the extension of the Dutch Identity proposed by Hessen [9] in a psychometric context to the capture-recapture framework and we showed that it is possible to re-express the probability of a generic capture-profile in terms of the log-linear multidimensional Rasch model.

Finally, we applied the methodology we proposed to a dataset on NTD's in the Netherlands from 1988 through 1998. Since lists did not cover the same time periods, we used the EM algorithm proposed by Zwane et. al [12] to estimate the missing entry in the data set. The results showed that the selected model for inference is one of the log-linear multidimensional Rasch models obtained by applying the methodology proposed. In fact, it was preferable among the other log-linear model, as it presented the smallest value of both AIC and BIC.

Bibliography

[1] Seber, G. A. F. (1982) *The Estimation of Animal abundance and Related Parameters.* London: Griffin

[2] International Working Group for Disease Monitoring and Forecasting, (IWGDMF). (1995) *Capture-Recapture and Multiple-Record Systems Estimation I: History and Theoretical Development.* American Journal of Epidemiology, **142**, 1059–1068.

[3] Fienberg, S. E. (1972) *The multiple-recapture census for closed populations and the 2^k incomplete contingency tables.* Biometrika, **59**, 591–603.

[4] Bishop, Y. M. M., Fienberg, S. E. and Holland, P. W. (1975) *Discrete Multivariate Analysis: Theory and Practice.* MIT Press: Cambridge.

[5] Darroch, J. N., Fienberg, S. E., Glonek, G. F. V. and Junker, B. W. (1993) *A three-sample multiple-capture approach to census population estimation with heterogeneous catchability.* Journal of the American Statistical Association, **88**, 1137–1148.

[6] Agresti, A. (1994) *Simple capture-recapture models permitting unequal catchability and variable sampling effort.* Biometrics, **50**, 494–500.

[7] Bartolucci, F. and Forcina, A. (2001) *Analysis of Capture-Recapture data with a Rasch-Type Model allowing for Conditional Dependence and Multidimensionality.* Biometrics, **57**, 714–719.

[8] Bartolucci, F. and Pennoni, F. (2007) *A Class of Latent Markov Models for Capture-Recapture data Allowing for Time, Heterogeneity, and Behavior Effects.* Biometrics, **63**, 568–578.

[9] Hessen, D. J. (2012) *Fitting and testing conditional multinormal partial credit models.* Psychometrika, **77**, 693–709.

[10] Holland, P. W. (1990) *The Dutch Identity: a new tool for the study of item response model.* Psychometrika, **55**, 5–18.

[11] Sanathanan, L. (1972) *Models and estimation methods in visual scanning experiments.* Technometrics, **14**, 813–829.

[12] Zwane, E. N., van der Pal, K. and van der Heijden, P. G. M. (2004) *The multiple- record systems estimator when registrations refer to different but overlapping populations".* Statistics in Medicine, **23**, 2267–2281.