Università degli Studi di Milano - Bicocca

DIPARTIMENTO DI STATISTICA E METODI QUANTITATIVI
Corso di Dottorato di Ricerca in
Statistica e Applicazioni

Log-linear multidimensional Rasch model for Capture-Recapture

Candidato:
Elvira Pelle

Tutor:
Ch.mo Prof. Pier Francesco Perri

Relatori:
Ch.mo Prof. Peter G. M. van der Heijden
Ch.mo Prof. Dave Hessen

Anno Accademico 2012-2013

Contents

List of Tables

List of Figures

Introduction

One of the most important task for epidemiologists, biologists, ecologists and sociologists is to analyse and forecast possible changes and dynamics in a population. In order to understand and monitor these changes, accurate estimates of population characteristic are required and so sampling techniques are to be implemented.

Capture-recapture experiments may be used to obtain meaningful informations from population under study. A typical capture-recapture experiment consists in a sequence of random samples; for each sample individuals drawn are marked (or tagged) and released in the population. The rational behind this method is to account for unobserved individuals by using observed individual trapping histories. Once the data are collected, suitable statistical methodology is applied and the estimates of the population characteristic of interest are made.

Literature about capture-recapture has grown rapidly and the method was applied in many different fields with respect to the originally purpose. The first use of capture-recapture analysis can be traced to Graunt, who applied a similar technique to estimate the English population in 1625 and Laplace who estimated the population size in France in 1782. However, it is usually mentioned that the first application of capture-recapture was due by Petersen in the study of fish and wildlife populations in 1894. For this reason, in ecology the method is generally called *Petersen method*.

The first application of capture-recapture method to human populations is due to Sekar and Deming [20], who in 1949 used it to estimates birth and death rates. In this context, personal identifier such as identification numbers or names are used as marks or tags, and "being captured" is replaced by "being observed" in the registrations. Shapiro[21] applied a similar approach using birth registration in United States and census data.

Common labels for the method in human populations and record linkage are *multiple-system*, *multiple-recapture* or *multiple-records systems methods*.

The use of capture-recapture analysis to the study of epidemiologic problems came relatively late through the work of Wittes and her colleagues [24, 23] in 1968 and 1974. However, in this context some problems arise with some underlying assumptions. In 1972 Fienberg [9] approached these problems through the use of the log-linear model, as it had emerged for the analysis of multidimensional contingency tables. One of the major advantages of his solution is represented by the fact that was general and so well applicable both to animal and human populations.

A central assumption in traditional capture-recapture approach is the homogeneity of the capture probability. However, differences of character or behaviour between individuals may occur and this fact results in indirect dependence between registrations. Models that allows for varying susceptibility to capture through individuals and unequal catchability have been proposed either in the case of human populations [7] or in animal population studies [1] and psychometric models, such as the Rasch model, were successfully applied.

The Rasch model is a model for dichotomous item widely used in psychometrics. Here, the probability of a response to an item is modelled as a function of the difficulty of the item and the underlying latent ability of the individual. An extension of the dichotomous Rasch model is represented by the multidimensional Rasch model, that allows for more than one latent trait underlying the performance of a test.

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 registration and, if all registrations are supposed to be of the same kind, it is possible to treat heterogeneity in terms of constant apparent dependence between registrations (Darroch, 1993 [7], Agresti, 1994 [1], International Working Group for Disease Monitoring and Forecasting, 1995 [13]).

The basic assumptions of the Rasch model are the conditional dependence and unidimensionality. Bartolucci and Forcina, 2001 [2], shown how to relax these assumptions by adding some suitable columns to the design matrix of the model.

Contribution of this work. In the present work, we propose the use of the multidimensional Rasch model in the capture-recapture context. In particular, we assume that registrations may be divided into two or more subgroups, such that they can be view as indicators of the latent variables which account for correlations among registrations. To do so, the extension of the Dutch Identity for the multidimensional partial credit model (Hessen, 2012 [11]) can be utilized. The Dutch Identity is a tool proposed by Holland, 1990 [12] 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, 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 and to derive the parameters of the traditional log-linear model from those of the multidimensional Rasch model.

The reminder of this work is organized in the following way:

In Chapter 1 we introduce the basic concept of capture-recapture sample and the capture-recapture methodology is briefly described; then we focus on the problem of dependence among samples (or registrations).

In Chapter 2 we present the use of the log-linear models in capture-recapture framework and the principal methods utilised for model selection are briefly discussed.

In Chapter 3 is focused on the Rasch model. In particular, first the dichotomous Rash model, that is the simplest model, is described and its basic properties are discussed. Then, the attention is focused on the multidimensional Rasch model.

In Chapter 4 we propose the use of the multidimensional Rasch model in capture-recapture context. In particular, under certain assumptions discussed in the Chapter, we show how it is possible to re-express the probability of a generic capture profile in the log-linear form of the multidimensional Rasch model and the connection with the parameters of the traditional log-linear model is described.

Finally, in Chapter 5 we present an application of the methodology described in the preceding Chapter to a dataset on Neural Tube Defects (NTD's) in the Netherlands. The scope of the application is to estimate the total population size of children affected by NTD's during the period of the

study. The results of the application shows that the multidimensional Rasch model we propose presents the lowest value of AIC and BIC and thus it is the best model and selected for inference.

Chapter 1

Capture-recapture method

1.1 Introduction

A capture-recapture experiment consists of a sequence of sampling or capture occasion. On capture, previously uncaptured individuals are marked (or tagged) and marks (or tags) of previously captured individuals are noted. Once the information needed has been recorded, individuals are released back into the population. Capture occasions are usually conducted at equally spaced intervals (e.g. consecutive nights/days, breading season over a number of consecutive years, ecc.)

The resulting data from a capture-recapture experiments consist of individual capture histories, that record whether an individual has been captured or non captured at each sampling occasion. The basic capture data can be conventionally expressed in matrix form as, for example:

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & \dots & 1 \\ 1 & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 1 & \dots & 1 \end{pmatrix}$$

where

$$X_{ij} = \begin{cases} 1 & \text{if the } i^{th} \text{ individual is caught on the } j^{th} \text{ occasion} \\ 0 & \text{otherwise} \end{cases}$$

In this way, row i gives the trapping results for individual i, while column

j gives results for the j^{th} sampling occasion. Note that matrix $\mathbf{X}$ only contains capture histories for those individuals observed at least once, but does not include the capture histories for individuals that were never caught.

Once data are collected and appropriate assumptions made, the capture histories can be used to estimate population demographic characteristic of interest by fitting capture-recapture models. Assumptions depend upon the situation the researcher wants to study:

- **Closed population**: there are no changes in the population due to birth, death, emigration or immigration during the time of period when the sampling takes place[1].

- **Open population**: possible changes in the population from one sample to the next, due to birth, death, emigration or immigration, are allowed. This results in more complexity of the models used to estimate the population size.

In the remainder of the present work the attention will be restricted on closed population. The closure assumption ensures that individuals observed in a registration may be observed in other registrations. In addition, it is possible to improve population estimates taking into account individual heterogeneity, or time variation. Another assumption is that it is possible to link individuals in all registrations perfectly.

1.2 Capture-recapture method for human populations

Despite the fact that the capture-recapture method was originally developed to estimate the size of an animal population, it has been successfully applied to human populations. The earliest reference to the application of this methodology refers to Sekar and Deming, 1949 [20], who applied the method to two samples. Later Wittes and Sidel, 1968 [24], Fienberg, 1972 [9], Wittes, 1974 [25] and Wittes et al., 1974 [23] adapted it to more than two samples.

In human populations, capture-recapture techniques are applied to estimate the demographic characteristics of interest using information from

[1] Note that the demographic closure assumption is usually valid for data collected in a relatively short time.

overlapping registrations (or lists) of cases from different sources. The registrations are by their nature incomplete and the problem is to estimate the portion of population missed by all registrations. However, there are some differences between wildlife and human application of capture-recapture: in fact, in human studies there are usually two to four available registrations (or lists), while in wildlife surveys there are usually more trapping samples. There is a natural time ordering in animal experiments, but this is not true for registrations (or the order may vary with individuals). In a human population different types of ascertainment sources[2] are utilised and so the behavioural response due to the sampling scheme is not considered in the model; for animal populations all trapping samples usually use identical trapping method and this results in model that takes into account behavioural response to capture.

In the case of ascertainment data, each registration (or list) is regarded as a capture sample and identification number (or name) as tag or mark. In this way, "being captured in sample i" corresponds to "being observed in registration i" and "capture probability" corresponds to "ascertainment probability". Thus, the set up for human population is similar to capture-recapture set up for wildlife estimation.

1.3 Capture-recapture methodology

1.3.1 Two-registration model

The simplest capture-recapture model is one in which there are only two registrations (or samples). Suppose that registrations A and B are available. Let $n_{i_1 i_2}$ denotes the observed frequencies of the data. For each registration, $i_s = (0, 1)$, where $i_s = 0$ denotes "non captured" (or "non observed") and $i_s = 1$ denotes "captured" (or "observed"). Let $\pi_{i_1 i_2}$ be the corresponding probability, with $\sum \pi_{i_1 i_2} = 1$. Thus, n_{10} denotes the frequency of individuals observed only in the first registration, n_{01} is the number of individual included in the second list, but not in the first, and n_{11} is the frequency of individual observed in both registrations. Note that n_{00} and π_{00} represent, respectively, the frequency of individual "not observed" in either registration and the correspondent probability. Since they are unknown, they have to be

[2] In human populations data are usually obtained from registrations of ascertainment cases, like administrative registers, medical registrations, police registers and so on.

estimated in order to estimate the total unknown population size N.

The data can be regarded as a form of an incomplete 2^2 contingency table for which the cell corresponding to those individual unobserved in both registrations is missing:

Table 1.1: Contingency table for two registrations

	B		
A	Observed	Not Observed	Total
Observed	n_{11}	n_{10}	n_{1+}
Not Observed	n_{01}	-	
Total	n_{+1}		

where the symbol "+" denotes that the table is collapsed over the corresponding index and "−" means that the count for the corresponding cell is missing.

The rational behind capture-recapture method is to estimate the number of individuals not included in any registration using the numbers observed in only one registration and the number observed in both registrations. The assumption required for this estimate to be valid can be summarized as follows:

(i) There is no change in the population during the investigation (population is closed), that is there is no births, no deaths, no immigrations or emigrations. This results in a non-zero probability for each individual to being observed in each sample.

(ii) no misclassification of records, that is an individual can be matched without error from capture to recapture.

(iii) For at least one of the two samples, each individual has the same probability to being observed in the registration, that is homogeneity of inclusion probability.

(iv) Inclusion in registration A is independent of inclusion in registration B.

Note that if assumption (i) doesn't hold, also (iii) will not hold; in fact, individuals which stay in the population have higher probability to being ob-

served than individuals who migrate or die. Furthermore, assumption (*iv*) directly follows from assumption (*iii*), that implies that capture in the first sample does not affect capture in the second sample, so that the probability of inclusion in the first sample is independent from the probability of inclusion in the second sample (see International Working Group for Disease Monitoring and Forecasting, 1995 [13]).

Suppose that being in registration A is independent from being in registration B. The frequencies n_{10}, n_{01} and n_{11} have a multinomial distribution with probabilities π_{10}, π_{01} and π_{11} of being observed in the correspondent categories. Let m_{11}, m_{1+} and m_{+1} be the expected frequencies of n_{11}, n_{1+} and n_{+1}, respectively, and let $\hat{m}_{11}, \hat{m}_{1+}$ and $\hat{m}_{+1}$ denote the corresponding maximum-likelihood (ML) estimates. If assumption (*iii*) holds, then the proportion of the population observed in registration A is roughly the same as the proportion of individuals observed in registration A in the sub-population of those observed in registration B, that is:

$$\begin{aligned} \frac{m_{1+}}{N} &= \frac{m_{11}}{m_{+1}} \\ \frac{\hat{m}_{1+}}{N} &= \frac{\hat{m}_{11}}{\hat{m}_{+1}} \end{aligned} \tag{1.1}$$

Since the MLEs for the expected frequencies $m_{i_1 i_2}$ are just the corresponding observed frequencies $n_{i_1 i_2}$, (1.1) can be written as

$$\frac{n_{1+}}{N} = \frac{n_{11}}{n_{+1}}$$

which yields the estimator of the population size

$$\hat{N} = \frac{n_{1+} \times n_{+1}}{n_{11}} \tag{1.2}$$

This is the well-known Petersen estimator (derived independently by Petersen, 1896 [16] who was interested in the size of fish populations, and by Lincoln, 1930 [15], who was considering banding returns of wildfowl).

1.3.2 Dependence among registrations

A crucial assumption in the traditional capture-recapture approach is the homogeneity of inclusion probabilities and so independence of inclusion in registrations. However, dependence among registrations or unequal catcha-

bilities may occur. They may be caused by two different sources:

1. List dependence (or local dependence) within each individual, that is the response of an individual to one source depends on the response to the other source. In this case, conditional on any individual, the inclusion in one registration has a direct causal effect on inclusion in other registrations.

2. Heterogeneity between individuals, that is differences of character or behaviour between individuals may cause indirect dependence between registrations. In this case, even if the inclusion probabilities for the two registrations are independent within individuals, the ascertainment of the two sources may become dependent. In other words, personal behaviour has direct influence on the probability of inclusion in a registration and thus the inclusion probabilities are heterogeneous among individuals.

These two sources of dependencies are usually confounded and cannot be easily disentangled in data analysis (see International Working Group for Disease Monitoring and Forecasting, 1995 [13], and Chao et. al, 2001 [4]).

Note that in two-source capture-recapture analysis assumptions of homogeneity of inclusion probability and independence of registrations are crucial because it is impossible to check independence mathematically. In this case, in fact, there are four parameters (the total population size N, the two mean capture probabilities and a dependence measure); however, only three cells are observable, namely individual observed only in the first registration, individual present only in the second and those who are included in both. For this reason, data are insufficient for estimating dependence unless additional covariates are available.

If one or more covariates are available, instead of independence, one can make a less restrictive assumption, that is the independence conditional on covariates. In this case, estimation of the population size is improved due to the possibility to consider heterogeneity of inclusion probabilities over the level of the covariates. In addition, it is possible to make the estimates of the subpopulation size for each level of the covariates and then add up to arrive to the estimation of the size of the whole population. The traditional approach for including covariates in capture-recapture context consists in choosing only the covariates that are available in all registrations, but recently also

the use of covariates that are available for only one registration was studied (for more details see Zwane and Van der Heijden, 2007 [26] and Van der Heijden et al., 2012 [22]).

Note that, if independence assumption is made, its violation may lead to biased estimates of the population size. Gerritse et al., 2013 [10] investigated the robustness of these estimates, pointing out that, if the assumption of independence does not hold, then the estimates could be seriously biased. Bias may also occur in the case in which one or more covariates are taken into account.

If there are more than two registrations, violation of the homogeneity assumption has been successfully handled by the use of log-linear models [9, 3, 5, 1], as shown in the next chapter.

Chapter 2

Log-linear models for Capture-recapture

2.1 Introduction

The traditional capture-recapture method assumes independence between samples (or registrations). If more than two registrations are available, to handle possible dependence among samples log-linear models have been proposed. In presence of two registrations, the independence assumption is always made, as the number of observed counts is equal to the number of parameters in the independence model, that represents the saturated model. The availability of more than two registrations allows for inclusion of dependence parameters. In particular, dependencies between registrations correspond to two factor or higher-order terms in the model. Data are arranged in a 2^S contingency table (where S is the number of registrations taken into account), with one missing cell corresponding to absence in all registrations. The empty cell is treated as a "structural zero", i.e. is known a priori to have a zero value. This implies that the cell corresponding to a structural zero must remain empty under any fitted model.

Since the object of the analysis is the estimation of the missing cell, that is the number of individuals in the population who are not observed, the approach utilised is conditional: at first various log-linear models, allowing for dependencies among registrations, are fitted to the 2^S contingency table which excludes the missing cell. A model is selected, taking into account parsimony as well as fit to the data, and finally used to estimate the number

of individuals missed by all registrations.

Model selection is usually performed on the basis of likelihood function. In particular, one can use Akaike Information Criterion (AIC) or Bayes' Information Criterion (BIC).

Also the deviance can be used as criterion for the model selection, as it can be interpreted as a measure of the lack of fit of the model: in particular, the smaller the deviance, the better the fit to the data.

Using a log-linear model it is possible to take into account dependencies that can arise in several ways. In fact, positive interaction terms mean that individuals are selectively included in several samples (or registrations), and this may be due to "trap fascination" in animal capture-recapture experiments (or "social visibility" in other context, like social situations); on the other hand, negative interaction terms might be due to "trap avoidance" (or "social invisibility"), or might reflect the stratification of the population according to some latent variables (for more details, see Bishop et. al, 1975 [3]).

The remainder of the Chapter is organized as follow: in Section 2.2 a factorization of the multinomial likelihood function is described; this represents a theoretical justification of the two-stage approach used to fit a log-linear model in capture-recapture context. In Section 2.3 the two-registration problem and the model of quasi-independence are treated. It is shown that the estimate of the unknown total population size is the same as the one obtained in Chapter 1 and an example is proposed. In Section 2.4 the use of the log-linear model approach to a general situation is described. In Section 2.5 the results of Sections 2.4 and 2.2 are combined to obtain the estimation of the unknown population size and some examples are given. Finally, the problem of model selection and a brief description of principal methods utilised is given in Section 2.6.

2.2 Maximum likelihood estimation of N

The two-stage approach for the estimation of the total population size described above finds a theoretical justification in the conditional maximum likelihood estimation of N. In fact, the maximum likelihood function which involves the unknown total population size N as an unknown parameter can be factored into a product of two terms, such that one factor is a binomial

likelihood function involving N and the individuals missed by all registrations, and the other factor is a multinomial likelihood giving the conditional distribution of the observed frequencies (see Bishop et. al, 1975 [3]).

Consider a general situation in which S registrations are available. The resulting 2^S contingency table has one missing cell, corresponding to absence in all registrations. Let t be the number of observed cells in the contingency table and consider a $(t+1)$-cell multinomial random variable with cell probabilities $\pi_i, i = 1, \ldots, t$ for the first t cells and $\pi^* = 1 - \sum_{i=1}^{t} \pi_i$ for the $(t+1)$st cell (the cell containing the missing value). Let N be the unknown total number for the $(t+1)$-cell multinomial and $n_i, i = 1, \ldots, t$ be the observed counts for the first t cells. Let $n = \sum_{i=1}^{t} n_i$ the total amount of observed individuals, then the missing count for the cell not observed is $N - n$.

The multinomial likelihood function can be written as:

$$L(N;\theta) = \frac{N!}{(N-n)! \prod_{i=1}^{t} n_i!} \pi^{*N-n} \prod_{i=1}^{t} \pi_i^{n_i} \tag{2.1}$$

where the capture probabilities are rewrite as some known function of parameters (see Sanathanan, 1972 [19]) $\pi_i = \pi_i(\theta)$ and $\pi^* = \pi^*(\theta)$ and the dimension of the vector θ is at most t.

The likelihood in (2.1) can be rewritten as the product of two factors in the following way:

$$L(N;\theta) = L_1(N; \pi^*(\theta))\, L_2(\theta) \tag{2.2}$$

where

$$L_1(N; \pi^*(\theta)) = \frac{N!}{n!\,(N-n)!} \pi^{*N-n} (1 - \pi^*)^n \tag{2.3}$$

$$L_2(\theta) = n! \prod_{i=1}^{t} \frac{Q_i(\theta)^{n_i}}{n_i!} \tag{2.4}$$

with

$$Q_i(\theta) = \frac{\pi_i(\theta)}{1 - \pi^*(\theta)} \tag{2.5}$$

Although it is possible to carry out the maximum likelihood estimates of N and θ simultaneously from (2.1), it can be quite difficult due to the algebraic manipulations required. A simpler approach consists in the estimation

of θ maximizing L_2; then, it is possible to compute the maximum likelihood estimates for N using L_1.

Let $\hat{\theta}_C$ the maximum likelihood estimates for θ obtained maximizing L_2 (where C reminds the fact that θ is estimated using the conditional likelihood (2.4)). The estimator of the total population size N is

$$\hat{N}_C = \left[\frac{n}{\sum_{i=1}^{t} \pi_i \left(\hat{\theta}_C \right)} \right] \tag{2.6}$$

where the notation $[x]$ denotes the greatest integer $\leq x$.

2.3 Two-registration problem

Consider the situation with two registrations available described in Table 1.1. Let n_{1+}, n_{+1} and n_{11} be the number of individuals observed in the first registration, in the second registration and in both respectively, and let π_{1+}, π_{+1} and π_{11} denote the corresponding probabilities. Suppose that the two registrations are independent. Let $m_{i_A i_B}$ be the expected counts of frequencies $n_{i_A i_B}$ under the independent-registrations model

$$E\left(n_{i_A i_B}\right) = m_{i_A i_B} \tag{2.7}$$

where $i_A = 0$ denotes "not observed" while $i_A = 1$ indicates "observed" (i_B is defined in a similar manner).

Let K be the set of cells not containing structural zeros. Since it is known a priori that structural zeros have zero values, we have $m_{i_A i_B} = 0$ for $(i_A i_B) \notin K$.

In addition, it is possible to restrict the attention to the three cells that contain observed values. The resulting model is known in the literature as "quasi-independence" model, because is a form of independence conditional on the restriction of attention to an incomplete portion of the original contingency table.

If we have a 2×2 table with observations for all four cells, the model for the expected cell counts can be written in the natural logarithmic scale as follows:

$$\ln m_{ij} = \lambda + i_A \lambda_A + i_B \lambda_B + i_A i_B \lambda_{AB} \tag{2.8}$$

$\forall(i,j) \in K$, with constraints:

$$\sum_{i=1}^{2} \alpha_{i_A i_B}(i_A i_B)\lambda_{AB} = \sum_{j=1}^{2} \alpha_{ij}(ij)\lambda_{AB} = 0 \tag{2.9}$$

$$\sum_{i=1}^{2} \alpha_i^{(B)} i\lambda_A = \sum_{j=1}^{2} \alpha_j^{(A)} j\lambda_B \tag{2.10}$$

where

$$\alpha_{ij} = \begin{cases} 1 & \text{if } (i,j) \in K \\ 0 & \text{otherwise} \end{cases}$$

$$\alpha_j^{(A)} = \begin{cases} 1 & \text{if the } \alpha_{ij} = 1 \text{ for some } j \\ 0 & \text{otherwise} \end{cases}$$

and $\alpha_i^{(B)}$ is defined in a similar manner.

If we put no further restrictions on the λ-terms, the m_{ij} remain unrestricted. The model of quasi independence is defined by setting $(ij)\lambda_{AB} = 0$, that is

$$\ln m_{ij} = \lambda + i\lambda_A + j\lambda_B \tag{2.11}$$

If we restrict the attention to the conditional likelihood function for the cells with observed frequencies, the conditional maximum likelihood estimates of the expected counts are just the corresponding n_{ij}. We know that, under independence, the cross-product ratio for the expected values is

$$\frac{m_{11}m_{00}}{m_{10}m_{01}} = 1,$$

so that

$$m_{00} = \frac{m_{10} \times m_{01}}{m_{11}}.$$

Thus, the maximum likelihood estimate for the missing cell $(i,j) = (0,0)$ is

$$\hat{m}_{00} = \frac{\hat{m}_{10} \times \hat{m}_{01}}{\hat{m}_{11}} = \frac{n_{10} \times n_{01}}{n_{11}}.$$

and the estimate of the unknown population size $\hat{N}$ is:

$$\hat{N} = n_{11} + n_{10} + n_{01} + \frac{n_{10} \times n_{01}}{n_{11}} = \frac{n_{1+} \times n_{+1}}{n_{11}} \tag{2.12}$$

that is the same obtained in the previous chapter.

EXAMPLE. - ESTIMATION OF BIRTHS AND DEATHS IN INDIA
Sekar and Deming (1949)[20] used capture-recapture method to estimate the number of births and deaths for the residents in the Singur Health Center, an area near Calcutta in India. In their work, they described the results of an inquiry conducted during February 1947, reporting the births and deaths in years 1945 and 1946 separately. At the time of the study the total population of the area was about 64,000 people, living in about 8,300 houses. The Chowkidar, the headman in each village, periodically submit to the registrar of the area a list of births and deaths. The registrar coordinates this informations with a second report from each village and a list from the Maternity and Child Welfare Department. This list is named "registrar's list of births and deaths" (R-list). During an eleven-week period beginning on February 1947 interviewers from the All-India Institute of Hygiene and Public Health visited every house within the Singur Health Centre to prepare a list of all of the births and deaths that occurred during 1945 and 1946 (the I-list). After deleting the non-verifiable, illegible, incomplete, and incorrect items Sekar and Deming applied the two-registrations technique described above.

Data are displayed in Table 2.1

Table 2.1: Investigators' Report on Comparison of Lists of Singur Health Centre

			R-List		I-List
	Year	Total	Observed in I-List	Not Observed in I-List	Extra
Births	1945	1504	794	710	741
	1946	2242	1506	736	1009
Deaths	1945	1083	350	733	372
	1946	866	439	427	421

Source: Sekar and Deming 1949

To illustrate the capture-recapture methodology, consider the deaths occurring during 1945. The data of interest are summarized in a 2^2 contingency table as follows:

Table 2.2: Number of Deaths Occurring During 1945

	R-List		
I-List	Observed	Not Observed	Total
Observed	350	372	722
Not Observed	733	-	
Total	1083		

The estimate of the total number of deaths in 1945 is

$$\hat{N} = \frac{722 \times 1083}{350} = 2234$$

where the result is rounded to the nearest integer.

□

2.4 Generalization

Consider a situation in which S registrations are available. Let $n_{i_1 \ldots i_s}, s = 1, \ldots, S$ denotes the observed frequencies of the data, where $i_s = (0, 1)$ and $i_s = 0$ denotes "not observed" while $i_s = 1$ denotes "observed". Data can be arranged in a 2^S incomplete contingency table with one missing cell, which is unobserved by definition, corresponding to absence in all registrations.

Similarly to the two-registrations problem, let K be the set of cells obtained excluding the $(0 \ldots 0)$ cell from the 2^S contingency table. Suppose that the observed frequencies $n_{i_1 \ldots i_s}, s = 1, \ldots, S$ for the cells in K have a multinomial distribution and that the total sample size is $n = \sum n_{i_1 \ldots i_s}$, with $s = 1, \ldots, S$, where the summation is over all cells contained in K.

Let $m_{i_1 \ldots i_s}$ denote the expected count of the frequency in the $(i_1 \ldots i_s)$ cell and assume that all the expected frequencies are positive. The probability associated to a generic capture profile $(i_1 \ldots i_s)$ can be written as $m_{i_1 \ldots i_s}/n$. Let N be the unknown total population size. Thus, $n_{0 \ldots 0} = N - n$ is the number of individuals missed by all S registrations.

The most general log-linear model for the cells in K can be written as

$$\ln m_{i_1 \ldots i_S} = \lambda + i_1\lambda_1 + \cdots + i_S\lambda_S + \sum i_\omega i_\nu \lambda_{\omega\nu} + \cdots + i_1 i_2 \ldots i_S \lambda_{12 \ldots S} \quad (2.13)$$

where the sum of any individual λ term in (2.13) over any of its subscript is

zero; for example

$$\sum_{i_\omega=0}^{1} i_\omega \lambda_\omega = \sum_{i_\omega=0}^{1} i_\omega i_\nu \lambda_{\omega\nu} = \sum_{i_\nu=0}^{1} i_\omega i_\nu \lambda_{\omega\nu} = 0$$

$$\sum_{i_\omega=0}^{1} i_\omega i_\nu i_\eta \lambda_{\omega\nu\eta} = \sum_{i_\nu=0}^{1} i_\omega i_\nu i_\eta \lambda_{\omega\nu\eta} = \sum_{i_\eta=0}^{1} i_\omega i_\nu i_\eta \lambda_{\omega\nu\eta} = 0$$

Since the frequency corresponding to capture profile $(0 \ldots 0)$ is not in the set K, it is necessary to identify the model to set $\lambda_{12\ldots S(i_1 i_2 \ldots i_S)} = 0, \forall (i_1 \ldots i_S) \in K$.

By setting λ-terms in (2.13) equal to zero, it is possible to define various unsaturated log-linear models. There, the attention is restricted to the family of hierarchical models, defined as the family such that if any λ-term is set equal to zero, all its higher-order relatives must also be set to zero; conversely, if any λ-term is not zero, all its lower-order relatives must be present in the log-linear model[1]. Thus, for example in a three-registrations model if $\lambda_{12} = 0$ we must have $\lambda_{123} = 0$; on the other hand if λ_{12} is present in the model, then λ_1 and λ_2 must be also present.

The problem where the S registrations are independent corresponds to the unsaturated log-linear model given by

$$\ln m_{i_1 \ldots i_S} = \lambda + \sum_{s=1}^{S} i_s \lambda_s \tag{2.14}$$

(see Darroch, 1958 [6]).

For any unsaturated log-linear model the maximum likelihood estimates for $n_{i_1 \ldots i_s}$ can be obtained by setting the expected values of the marginal totals corresponding to the highest-order λ-term in the model equal to the

[1] The restriction to hierarchical models is due to the fact that in non-hierarchical models all the λ-terms cannot be thought of in terms of ratios of cross-product ratios (for more details, see Fienberg, 1972, p. 187).

observed frequencies. For the model in (2.14) the MLEs are given by

$$\begin{aligned} \hat{m}_{i_1++\cdots++} &= n_{i_1++\cdots++}(i_1 = 0, 1) \\ \hat{m}_{+i_2+\cdots++} &= n_{+i_2+\cdots++}(i_2 = 0, 1) \\ \hat{m}_{+++\cdots+i_S} &= n_{+++\cdots+i_S}(i_S = 0, 1) \end{aligned} \tag{2.15}$$

(2.16)

where the summation is over the cells in the set K and the symbol "+" denotes the sum over the corresponding subscript.

Except for $S = 2$ there is no closed form solution for equations in (2.15), but in general they can be obtained using numerical methods (like the Deming-Stephan iterative proportional fitting, or Newton-Raphson, or iteratively reweighted least squares). Once the MLEs are obtained it is possible to assess the goodness of fit of the model to the observed data using either

$$\begin{aligned} \chi^2 &= \sum_K \frac{(n_{i_1\ldots i_S} - \hat{m}_{i_1\ldots i_S})^2}{\hat{m}_{i_1\ldots i_S}} \\ G^2 &= 2\sum_K n_{i_1\ldots i_S} \ln\left(\frac{n_{i_1\ldots i_S}}{\hat{m}_{i_1\ldots i_S}}\right) \end{aligned} \tag{2.17}$$

The degrees of freedom of the model are determined by subtracting the number of independent parameters used in the model from the total number of cells to which the model is being fitted (that is the number of cells in the set K).

2.5 Estimation of N

Once various log-linear models are fitted to the incomplete data and the model with the best fit is chosen, it is possible to extend that model to cover the unobserved cell, whose expected value is $m_{0\ldots0}$. The maximum likelihood estimates for $m_{0\ldots0}$ can be written as

$$\hat{m}_{0\ldots0} = \frac{\hat{M}_{odd}}{\hat{M}_{even}} \tag{2.18}$$

where $\hat{M}_{odd}$ is the product of all $\hat{m}_{i_1 \ldots i_S}$ from the incomplete contingency table with $\sum i_s$ equal to an odd number and $\hat{M}_{even}$ is the product of all $\hat{m}_{i_1 \ldots i_S}$ from the incomplete contingency table with $\sum i_s$ equal to an even number. Then, the estimation of the unknown total population size is

$$\hat{N} = n + \hat{m}_{0 \ldots 0} \tag{2.19}$$

The expression in (2.18) yields the estimation of N no matter which log-linear model is selected. However, if we work with a large number of registrations and the log-linear model includes only a few interaction terms, expression in (2.18) may be numerically inefficient, even though algebraically correct (this is due to the cancellation of terms in the numerator and denominator).

In addition, it is possible to define S classes of hierarchical log-linear models having closed form MLEs for the expected frequencies of observed cells and thus for $\hat{m}_{0 \ldots 0}$. Let the i-th class be the class of models defined by setting equal to zero exactly i two-factor λ-terms ($i = 0, l \ldots, S-1$) with one dimension or variable in common.

If $i = 0$ the model is unrestricted and the expected frequencies are equal to the observed frequencies; thus, (2.18) applies directly. If two-factor terms involving a common dimension are set to zero, for example $\lambda_{12} = \lambda_{13} = 0$, then

$$\hat{m}_{0 \ldots 0} = \frac{n_{1000 \ldots 0} \times n_{0{+}{+}0 \ldots 0}}{n_{1{+}{+}0 \ldots 0} - n_{1000 \ldots 0}} \tag{2.20}$$

In general, if exactly $L < S$ two-factor terms involving a common dimension equals zero, for example $\lambda_{12} = \lambda_{13} = \cdots = \lambda_{1L+1}$, then

$$\hat{m}_{0 \ldots 0} = \frac{n_{10 \ldots 0} \times n_{0{+}{+}\cdots{+}0 \ldots 0}}{n_{1{+}{+}\cdots{+}0 \ldots 0} - n_{10 \ldots 0}} \tag{2.21}$$

where there are L consecutive subscript "+" and the remaining subscripts are equal to zero. Note that, due to the hierarchy of the model, if a two-factor order term equals zero, then all its high-order relatives must also be equal to zero.

2.5.1 The three-registrations model

In order to better understand the rational behind the use of log-linear models in capture-recapture problem, consider a situation in which three

registrations $R1, R2$ and $R3$ are available. Data can be arranged in a 2^3 contingency table with one missing cell as shown in Table 2.3

Table 2.3: Contingency table for three registrations

		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

In this case, there are 8 different hierarchical log-linear models for the incomplete 2^3 table that include parameters for the margins:

1. the saturated model (all pairwise relationship are present)

$$\ln m_{i_1 i_2 i_3} = \lambda + i_1\lambda_1 + i_2\lambda_2 + i_3\lambda_3 + i_1 i_2\lambda_{12} + i_1 i_3\lambda_{13} + i_2 i_3\lambda_{23} \quad (2.22)$$

2. three models with two two-factor terms (two pairs of registrations are related)

$$\begin{aligned} \ln m_{i_1 i_2 i_3} &= \lambda + i_1\lambda_1 + i_2\lambda_2 + i_3\lambda_3 + i_1 i_2\lambda_{12} + i_1 i_3\lambda_{13} \\ \ln m_{i_1 i_2 i_3} &= \lambda + i_1\lambda_1 + i_2\lambda_2 + i_3\lambda_3 + i_1 i_2\lambda_{12} + i_2 i_3\lambda_{23} \\ \ln m_{i_1 i_2 i_3} &= \lambda + i_1\lambda_1 + i_2\lambda_2 + i_3\lambda_3 + i_1 i_3\lambda_{13} + i_2 i_3\lambda_{23} \end{aligned} \quad (2.23)$$

3. three models with one two-factor term

$$\begin{aligned} \ln m_{i_1 i_2 i_3} &= \lambda + i_1\lambda_1 + i_2\lambda_2 + i_3\lambda_3 + i_1 i_2\lambda_{12} \\ \ln m_{i_1 i_2 i_3} &= \lambda + i_1\lambda_1 + i_2\lambda_2 + i_3\lambda_3 + i_1 i_3\lambda_{13} \\ \ln m_{i_1 i_2 i_3} &= \lambda + i_1\lambda_1 + i_2\lambda_2 + i_3\lambda_3 + i_2 i_3\lambda_{23} \end{aligned} \quad (2.24)$$

4. the "independence" model

$$\ln m_{i_1 i_2 i_3} = \lambda + i_1\lambda_1 + i_2\lambda_2 + i_3\lambda_3 \quad (2.25)$$

Since there are 7 observed cells, the degrees of freedom of the models are 0 for the saturated model, 1 for the models with two two-factor terms , 2 for

the models with one two-factor term and 3 for the independence model.

Consider the saturated model in (2.22). In this case the MLEs for the expected frequencies $m_{i_1 i_2 i_3}$ correspond to the observed frequencies; thus, applying (2.18) we have:

$$\hat{m}_{000} = \frac{n_{001} n_{010} n_{001} n_{111}}{n_{110} n_{101} n_{011}} \tag{2.26}$$

and using (2.19)

$$\hat{N} = n + \hat{m}_{000} \tag{2.27}$$

For any of the three models in (2.23) the MLEs have a closed form; for example, if $\lambda_{13} = 0$, then:

$$\begin{array}{ccc} \hat{m}_{001} = n_{001} & \hat{m}_{100} = n_{100} & \hat{m}_{101} = n_{101} \\ \hat{m}_{010} = \dfrac{n_{01+} n_{+10}}{n_{+1+}} & \hat{m}_{110} = \dfrac{n_{11+} n_{+10}}{n_{+1+}} & \hat{m}_{011} = \dfrac{n_{01+} n_{+11}}{n_{+1+}} \\ & \hat{m}_{111} = \dfrac{n_{11+} n_{+11}}{n_{+1+}} & \end{array} \tag{2.28}$$

then, after some simple manipulation, we obtain

$$\hat{m}_{000} = \frac{n_{001} n_{100}}{n_{101}} \tag{2.29}$$

Also for the three models of the form (2.24) exists a closed form for the MLEs; for the model which has only one two-factor term, corresponding to λ_{12} the MLEs are

$$\begin{array}{ccc} \hat{m}_{111} = \dfrac{n_{11+} n'_{++1}}{n'} & \hat{m}_{101} = \dfrac{n_{10+} n'_{++1}}{n'} & \hat{m}_{011} = \dfrac{n_{01+} n'_{++1}}{n'} \\ \hat{m}_{110} = \dfrac{n_{11+} n'_{++0}}{n'} & \hat{m}_{100} = \dfrac{n_{10+} n'_{++0}}{n'} & \hat{m}_{010} = \dfrac{n_{01+} n'_{++0}}{n'} \\ & \hat{m}_{001} = n_{001} & \end{array} \tag{2.30}$$

where

$$n'_{++1} = n_{++1} - n_{001} \qquad n' = n - n_{001} \tag{2.31}$$

Here (2.18) reduces to

$$\hat{m}_{000} = \frac{n_{001} n_{++0}}{n'_{++1}} = \frac{n_{001} n_{++0}}{n_{++1} - n_{001}} \tag{2.32}$$

For the independence model there is no a closed form solution for the MLEs and an iterative procedure to get the maximum likelihood estimates must be used. Then, it is possible to get the estimate $\hat{m}_{000}$ of the individuals missed by all registrations by applying (2.18) and thus the estimates of the total population size $\hat{N}$.

Note that for each of these models a different re-parametrization of the expected frequencies can be written, so that

$$\ln m_{000} = \lambda$$

and the expected number of individuals missed by all registrations may be estimated by

$$\hat{m}_{000} = \exp(\lambda) \tag{2.33}$$

In particular, for the saturated model in (2.22) a re-parametrization of the expected frequencies can be written as follows:

$$\begin{aligned}
\ln m_{100} &= \lambda + \lambda_1 \quad \ln m_{010} = \lambda + \lambda_2 \quad \ln m_{001} = \lambda + \lambda_3 \\
\ln m_{110} &= \lambda + \lambda_1 + \lambda_2 + \lambda_{12} \quad \ln m_{101} = \lambda + \lambda_1 + \lambda_3 + \lambda_{13} \\
\ln m_{011} &= \lambda + \lambda_2 + \lambda_3 + \lambda_{23} \\
\ln m_{111} &= \lambda + \lambda_1 + \lambda_2 + \lambda_3 + \lambda_{12} + \lambda_{13} + \lambda_{23}
\end{aligned}$$

For the model in (2.23) in which $\lambda_{13} = 0$ a re-parametrization is given by

$$\begin{aligned}
\ln m_{100} &= \lambda + \lambda_1 \quad \ln m_{010} = \lambda + \lambda_2 \quad \ln m_{001} = \lambda + \lambda_3 \\
\ln m_{110} &= \lambda + \lambda_1 + \lambda_2 + \lambda_{12} \quad \ln m_{101} = \lambda + \lambda_1 + \lambda_3 \\
\ln m_{011} &= \lambda + \lambda_2 + \lambda_3 + \lambda_{23} \\
\ln m_{111} &= \lambda + \lambda_1 + \lambda_2 + \lambda_3 + \lambda_{12} + \lambda_{23}
\end{aligned}$$

For the model which has only one two-factor term, corresponding to λ_{12} a re-parametrization is:

$$\begin{aligned}
\ln m_{100} &= \lambda + \lambda_1 \quad \ln m_{010} = \lambda + \lambda_2 \quad \ln m_{001} = \lambda + \lambda_3 \\
\ln m_{110} &= \lambda + \lambda_1 + \lambda_2 + \lambda_{12} \quad \ln m_{101} = \lambda + \lambda_1 + \lambda_3 \\
\ln m_{011} &= \lambda + \lambda_2 + \lambda_3 \quad \ln m_{111} = \lambda + \lambda_1 + \lambda_2 + \lambda_3 + \lambda_{12}
\end{aligned}$$

Finally, for the independence model in (2.25) we have:

$$\begin{aligned}
\ln m_{100} &= \lambda + \lambda_1 \quad \ln m_{010} = \lambda + \lambda_2 \quad \ln m_{001} = \lambda + \lambda_3 \\
\ln m_{110} &= \lambda + \lambda_1 + \lambda_2 \quad \ln m_{101} = \lambda + \lambda_1 + \lambda_3 \\
\ln m_{011} &= \lambda + \lambda_2 + \lambda_3 \quad \ln m_{111} = \lambda + \lambda_1 + \lambda_2 + \lambda_3
\end{aligned}$$

EXAMPLE. - ESTIMATION OF DEMENTIA IN SOUTH CAROLINA
Sanderson et al. (2003) applied capture-recapture methodology to evaluate the prevalence of dementia in individuals 65 years of age and older in the state of South Carolina. To do so, they used three different registrations:

- the Department of Mental Health Admissions (R1)
- the Impatient Admissions from Hospital Discharge Data (R2)
- the Emergency Room Visits from Hospital Discharge Data (R23)

Data are summarized in Table 2.4

Table 2.4: Data on dementia in South Carolina

		R3			
		Obs		Not Obs	
		R2		R2	
		Obs	Not Obs	Obs	Not Obs
R1	Obs	105	104	298	1,350
	Not Obs	1,285	2,197	9,430	–

Consider the model that allows dependence of inclusion probabilities between $R1$ and $R2$ and $R1$ and $R3$, that is

$$\ln m_{i_1 i_2 i_3} = \lambda + i_1\lambda_1 + i_2\lambda_2 + i_3\lambda_3 + i_1 i_2\lambda_{12} + i_1 i_3\lambda_{13} \qquad (2.34)$$

The estimated parameters of this model are

$$\begin{aligned}
\hat{\lambda} = 9.688 \quad \hat{\lambda}_1 = -2.525 \quad \hat{\lambda}_2 = -0.536 \\
\hat{\lambda}_3 = -1.993 \quad \hat{\lambda}_{12} = -0.747 \quad \hat{\lambda}_{13} = -0.072
\end{aligned}$$

who give the estimates of expected frequencies

$$\begin{aligned}
\hat{m}_{100} = 1,290 \quad \hat{m}_{010} = 9,430 \quad \hat{m}_{110} = 358 \quad \hat{m}_{001} = 2,197 \\
\hat{m}_{101} = 164 \quad \hat{m}_{011} = 1,285 \quad \hat{m}_{111} = 45
\end{aligned}$$

By applying (2.29) we have

$$\hat{m}_{000} = \frac{n_{010} \times n_{001}}{n_{011}} = 16,122.73 \simeq 16,123$$

or also by applying (2.33):

$$\hat{m}_{000} = \exp(9.688) \simeq 16,123$$

and

$$\hat{N} = 14,769 + 16,123 = 30,892$$

Suppose now that of inclusion probabilities of registrations $R1$ and $R3$ are independent; the model is:

$$\ln m_{i_1 i_2 i_3} = \lambda + i_1\lambda_1 + i_2\lambda_2 + i_3\lambda_3 + i_1 i_2 \lambda_{12} \qquad (2.35)$$

In this case the estimated parameters are

$$\hat{\lambda} = 9.698 \quad \hat{\lambda}_1 = -2.543 \quad \hat{\lambda}_2 = -0.545$$
$$\hat{\lambda}_3 = -2 \quad \hat{\lambda}_{12} = -0.738$$

and the corresponding estimates of expected frequencies are

$$\hat{m}_{100} = 1,281.213 \quad \hat{m}_{010} = 9,441.678 \quad \hat{m}_{110} = 355.109 \quad \hat{m}_{001} = 2,197$$
$$\hat{m}_{101} = 172.787 \quad \hat{m}_{011} = 1,273.322 \quad \hat{m}_{111} = 47.89$$

By applying (2.32) (or (2.33)) and (2.19) we have

$$\hat{m}_{000} = 16,291 \quad \text{and} \quad \hat{N} = 31,060$$

□

2.6 Model selection

In multiple-registration problems several competing models can be used to estimate the number of individuals missed by all registrations. Thus, model selection is an important part of the estimation procedure.

The aim of model selection is to find, among all the models available, a parsimonious model that fits the data well. Parsimony represents a trade-off between too few parameters and too little model structure versus too many parameters and too much model structure.

Several methods can be applied to identify the most appropriate model. For example, the likelihood ratio test (LR) can be utilised to select between two hierarchical nested log-linear models. The philosophy behind the LR test is to determine if the difference in deviance ($-2\times$log-likelihood) among the two models is statistically significant. The LR statistic approximately follows a chi-squared distribution and, using information about the degrees of freedom given by the difference in number of parameters, it is possible to determine the critical value of the test statistic and thus discriminate between the two models. In addition one can also test a model against to the data.

The likelihood can also be used to select between non-nested models: the Akaike's information criterion (AIC) or the Bayes' information criterion (BIC) can be adopted for this purpose.

Akaike defined the information criterion as

$$AIC = -2 \times \text{log-likelihood} + 2k \tag{2.36}$$

where k is the number of the parameters of the model. The first term may be interpreted as a measure of lack of fit of the model (how well the model fits the data), while the second term is a penalty for estimating k parameters. In fact, when fitting models, it is possible to increase the likelihood by adding parameters, but doing so may result in overfitting: the penalty enforces parsimony. The model with the minimum AIC value is chosen for inference.

On the other hand, the BIC can also be used for model selection. It is defined as:

$$BIC = -2 \times \text{log-likelihood} + k \ln(n) \tag{2.37}$$

where k is defined as above and n is the number of individuals observed in all registrations. Here, penalty about the number of parameters of the model is higher than in AIC. Also in this case, model with the lowest value of BIC is selected.

EXAMPLE. - SELECTION OF THE MODEL

Consider the example on Dementia in South Caroline discussed in the preceding section. The model that allows dependence of inclusion probabilities between $R1$ and $R2$ and $R1$ and $R3$ (Model 1) has 1 degree of freedom and deviance equal to 95. On the other hand, the model that allows dependence of inclusion probabilities between $R1$ and $R2$

(Model 2) has 2 degrees of freedom and deviance equal to 96.
Table 2.5 reports a summary of the two models.

Table 2.5: Summary of the models with deviance, AIC and BIC

Model	Number of parameters	Degrees of freedom	Deviance	AIC	BIC	$\hat{N}$
Model 1	6	1	95	107	153	30,892
Model 2	5	2	96	106	144	31,060

Thus, Model 2 is preferred over Model 1, since it has the lowest value of AIC and BIC.

□

Chapter 3

Rasch model as a log-linear model

3.1 Introduction

Item Response Theory (IRT) refers to a set of latent trait models widely used by psychometricians to explain the characteristics and performance of a test. The basic idea of an IRT is that the performance of a test can be described by a set of latent variables (usually called latent traits); in addition, the relationship between these latent traits and the item performance can be represented by a monotonically increasing function (known as Item Characteristic Curve), that describes relationship between the probability of a correct response to an item and the latent trait.

There are several latent trait models which differ in the number of parameters involved and the mathematical representation of the item characteristic curve. Among all the possible models, in the present work the attention is focused on the Rasch model.

The Rasch model, also known as one-parameter logistic model, postulates that the response to an item can be explained by the underlying latent ability of the individual and the difficulty of the item. Here, a set of survey items are utilised in order to produce an interval scale that describe both item difficulties and individual abilities. One of the fundamental assumptions of the Rasch model is that the comparison between two individuals is independent of the item used, so that the model is able to point out the structure of the responses.

The dichotomous Rasch model is the simplest model, since it has just one parameter describing the individual ability and just one parameter for the difficulty of an item. If it is assumed that more than one latent trait underlies the performance of a test then, instead of the dichotomous Rasch model, the multidimensional Rasch model has to be utilised. This is an extension of the original formulation of the model that allows for a multidimensional setting.

Log-linear models have been successfully applied to the Rasch model. In fact, the flexibility of log-linear model is helpful to simplify the estimation of parameters. In this context, an useful tool to re-express the Rasch equation in a log-linear form is represented by the Dutch Identity, proposed by Holland in 1990 [12], where the probability function is written as a second-order log-linear model. An extension of the Dutch Identity that allows for more than one latent trait was proposed by Hessen in 2012 [11].

The remainder of the Chapter is organised as follows: in Section 3.2 the dichotomous Rasch model and its properties are presented. Even if jointly estimation of the parameters of the model is possible, it yields inconsistent estimates; for this reason, the marginal likelihood estimation procedure, which yields consistent estimates, is discussed. Then, a situation in which the performance of a test depends of more that one latent variable is considered and an extension of the Rasch model for the multidimensional framework is described in Section 3.3. Finally, the application of the log-linear approach to the Rasch measurement context is treated in Section 3.4 and an example is given.

3.2 The Rasch model

The Rasch model was first developed by the Danish mathematician Rasch in 1960 [17]. Here, the probability of a specified response is modelled as a function of both individual and item parameters.

Consider a situation in which S items are administered to a sample of n individuals; suppose that the individual's responses to the $s-$th item can take values 0 (that denotes a wrong answer or disagree) or 1 (denoting right answer or agree).

Data can be summarised as in Table 3.1

Table 3.1: Data matrix of responses

		Items				
		1	$\cdots$	s	$\cdots$	k
Individuals	1	x_{11}		x_{1s}		x_{1S}
	$\vdots$					
	i	x_{i1}		x_{is}		x_{iS}
	$\vdots$					
	n	x_{n1}		x_{ns}		x_{nS}

where x_{is} takes the value 0 if the $i-$th individual gives an incorrect answer to the $s-$th item and x_{is} is equal to 1 if the $i-$th individual gives a right answer to the $s-$th item. In addition, note that rows correspond to individuals, while columns correspond to items.

The simplest form of the model is the dichotomous Rasch model that assumes that the probability that individual i gives a response score x_{is} to item s depends on one latent individual parameter. This probability can be written as

$$\begin{aligned} P\left(X_{is}=x_{is}\right) &= \left(\frac{e^{\theta_i-\delta_s}}{1+e^{\theta_i-\delta_s}}\right)^{x_{is}}\left(\frac{1}{1+e^{\theta_i-\delta_s}}\right)^{1-x_{is}} \\ &= \frac{e^{x_{is}(\theta_i-\delta_s)}}{1+e^{\theta_i-\delta_s}} \end{aligned} \tag{3.1}$$

where δ_s is the item parameter describing the difficulty of item s and θ_i is the individual parameter denoting the ability of person i. Thus, the probability to give a correct answer is

$$P\left(X_{is}=1\right)=\frac{e^{\theta_i-\delta_s}}{1+e^{\theta_i-\delta_s}} \tag{3.2}$$

Note that individual ability and item difficulty are measured on the same logit scale.

The Rasch model in (3.1) is also referred to as the "log-odds" model. In fact, the odds ratio for the correct answer (that is the ratio between the probability of a correct answer to the probability of getting an incorrect answer) can be written as

$$odds\left(X_{is}=1\right)=\frac{\frac{e^{\theta_i-\delta_s}}{1+e^{\theta_i-\delta_s}}}{1-\frac{e^{\theta_i-\delta_s}}{1+e^{\theta_i-\delta_s}}}=e^{\theta_i-\delta_s} \tag{3.3}$$

and the log-odds (or logit) takes the particularly simple form

$$log-odds\left(X_{is}=1\right)=\theta_i-\delta_s \tag{3.4}$$

Note that, since the estimates of individual ability and item difficulty are set to a common logit scale, the quantity $\theta_i-\delta_s$ can be interpreted as follows: if $\theta_i-\delta_s>0$ then the most probable outcome is a correct answer, as individual ability exceeds the difficulty of the item; on the other hand, if $\theta_i-\delta_s<0$ then the most probable outcome is a wrong answer, as the ability is less than item difficulty.

An important property of the Rasch model regards the invariant comparison: both the individual parameter and the item parameter can be jointly estimated in order to produce the estimates. However, although a jointly estimation procedure for the parameters of the model is possible, it is known that this approach yields inconsistent estimates for a fixed number of items and when n tends to infinity. Estimates are consistent when the number of items is large. An alternative estimation procedure that yields consistent estimates for the parameters of the model is the marginal maximum likelihood estimation. Here, the individual and the item parameters are not estimated simultaneously, but the individual parameters are integrated out (specifying the latent variable distributions) and then the item parameters are estimated. Thus, in this case the probability of a generic response pattern $\mathbf{x}=(x_1,\ldots,x_S)$ can be written as

$$P(\mathbf{X}=\mathbf{x})=\int\prod_{s=1}^{S}P\left(X_s=1|\boldsymbol{\theta}\right)^{x_s}\left\{1-P\left(X_s=1|\boldsymbol{\theta}\right)\right\}^{1-x_s}f\left(\boldsymbol{\theta}\right)d\boldsymbol{\theta} \tag{3.5}$$

where $f\left(\boldsymbol{\theta}\right)$ is the multivariate density of $\boldsymbol{\theta}$. Then, the marginal likelihood is equal to

$$L\left(\delta_1,\ldots,\delta_S\right)=\prod_{\mathbf{x}}P\left(\mathbf{X}=\mathbf{x}\right)^{n_{\mathbf{x}}} \tag{3.6}$$

Since the attention is focused to a small number of items, in the remainder of the present work we will use the marginal maximum likelihood estimation procedure.

It should be noted that the estimates of individual parameter do not depend upon the particular item used and, on the other hand the estimates of the item parameter are independent of the individuals to which items are administered. This property is reflected in the formal structure of the Rasch model: in fact, item parameter and individual parameter can be algebraic separated and θ_i can be eliminated in the estimation of δ_s. This also means that the raw scores (i.e. the number of 1's) for both individuals and items are sufficient statistics for the corresponding parameters (sufficiency property). Thus, all information on the ability of an individual is contained in the score and all information available with respect to the item, concerning the relevant latent trait, is contained in the item's score. As consequence, individuals with the same score will obtain the same estimate of the ability, even if it does not imply that they do have the same ability. This only means that if the Rasch model is valid for the situation under study, then the same score corresponds to the same estimation of the ability and no further differentiation can be made between them with respect to the ability.

Another important characteristic of the Rasch model is the local independence property, that is that the response to an item is independent of responses to other items.

In order to illustrate the logic underlying the Rasch point of view the Item Characteristic Curve (ICC) can be useful. The ICC is a curve that describes the relationship between the probability of a correct response to an item and the ability scale. Figure 3.1 shows hypothetical ICC for the Rasch Model.

On the x-axis is reported the individual ability (in logit) while on the y-axis there is the probability of success. An ICC indicates the probability that an individual that have ability delineated along the x-axis will have a response score equal to 1. The mid-way point along the curve, where the probability of a right answer is equal to 0.5, denotes the difficulty of the item. Thus, it is possible to estimate the probability of a correct response of an individual at any ability level on each item by drawing a perpendicular line through a point on the x-axis: the corresponding intersections with the ICCs denote these probabilities (Figure 3.2).

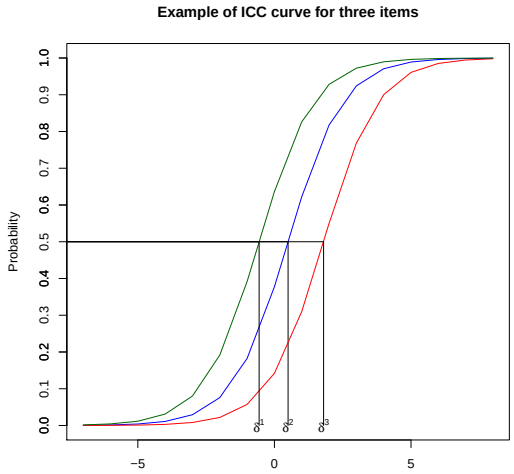

Figure 3.1: ICC curve for three items with different difficulty

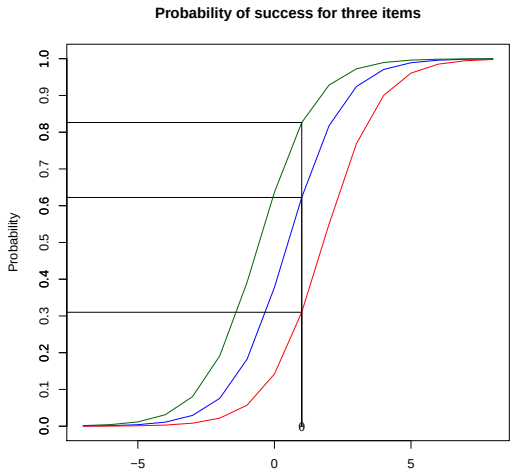

Figure 3.2: Probability of success for three items with different difficulty and fixed ability level

3.3 Multidimensional Rasch model

Suppose now that it is necessary more than one ability to give a correct answer to an item. In such case model in (3.1) may be inappropriate and an extension of the dichotomous Rasch model that allows for more than one

latent variable is needed (see Reckase, 1985 [18] and Kelderman and Rijkes, 1994 [14]).

Thus, consider a situation in which a test of S items is administered to a sample of n individuals and suppose that there are q latent variables that underlie the performance of the test. Let $\boldsymbol{\Theta} = (\Theta_1, \ldots, \Theta_q)$ denotes the vector of latent variables and let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)$ be a realization.

In other words, we are assuming that the probability to give a correct response to an item depends of a vector of latent variables, that is

$$P(X_s = 1|\boldsymbol{\theta}) = \frac{e^{\mathbf{u}_s'\boldsymbol{\theta} - \delta_s}}{1 + e^{\mathbf{u}_s'\boldsymbol{\theta} - \delta_s}} \tag{3.7}$$

where δ_s is the item parameter describing the difficulty of item s and $\mathbf{u}_s' = (u_{s1}, \ldots, u_{sq})$ is a vector of indicator variables taking the value $u_{sr} = 1$ if the response to item s depends of the $r-$th latent variable and $u_{sr} = 0$ otherwise.

To estimate the parameters of the model in (3.7) the marginal maximum likelihood estimation procedure described in the preceding section may be utilised. Thus, after specifying a multivariate distribution for the latent variables, they are integrated out and the the item parameters can be estimated.

3.4 Log-linear representation of the Rasch model

The log-linear representation of the Rasch model can be very helpful, due to the possibility to modelling and testing several hypotheses about the latent traits. In addition, log-linear models represent a general and simpler approach that allows to deal with multidimensionality models.

In this context, a useful tool is represented by the Dutch Identity proposed by Holland in 1990 [12]. Here, the probability of a response is re-expressed in a form of second-order log-linear model. Hessen in 2012[11] proposed an extension of the Dutch Identity that allows for more than one latent trait.

Let $\boldsymbol{\Theta} = (\Theta_1, \ldots, \Theta_q)$ be the vector of q latent traits that are assumed to underlie the performance of a test T administered to a sample of n individuals, and let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)$ be a realization. Suppose that the test T is composed by S items and let $\mathbf{X} = (X_1, \ldots, X_S)$ and $\mathbf{x} = (x_1, \ldots, x_S)$ denote the random vector of item scores and its realization, respectively. Let $\mathbf{u}_s' = (u_{s1}, u_{s2}, \ldots, u_{sr})$ be a vector of preassigned binary values and the

$s-$th row of the $S \times q$ full column matrix $\boldsymbol{U} = [u_{sr}]$ where $u_{sr} = 1$ if the probability distribution of X_s is assumed to depend on the latent trait θ_r, while $u_{sr} = 0$ otherwise. Following Hessen [11] the probability of a response pattern $\mathbf{x}$ can be written as

$$P(\mathbf{X} = \mathbf{x}) = P(\mathbf{X} = \mathbf{0}) \exp\left\{\sum_{s=1}^{S} x_s \delta_s + \mathbf{t}'\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}'\boldsymbol{\Gamma}\mathbf{t}\right\} \tag{3.8}$$

where $\boldsymbol{\mu}$ is the mean vector, $\boldsymbol{\Gamma}$ is a symmetric but non necessarily positive semi-definite matrix and $\mathbf{t} = (t_1, \ldots, t_q)'$ is the vector of the total scores computed using

$$\mathbf{t} = \mathbf{U}'\mathbf{x} = \begin{bmatrix} u_{11} & \ldots & u_{S1} \\ \vdots & & \vdots \\ u_{1q} & \ldots & u_{Sq} \end{bmatrix} = \begin{bmatrix} x_1 \\ \vdots \\ x_S \end{bmatrix} = \begin{bmatrix} t_1 \\ \vdots \\ t_q \end{bmatrix}.$$

If it is assumed that the population of individuals whom the test T is administered is infinite, then the expected frequencies $(e_{\mathbf{x}})$ of a response pattern $\mathbf{x}$ in a sample of size n is equal to

$$e_{\mathbf{x}} = nP\left(\mathbf{X} = \mathbf{x}\right) \tag{3.9}$$

In this case, in fact, the observed frequencies corresponding to all possible response patterns $\mathbf{x}$ have multinomial distribution. Thus, substituting (3.9) in (3.8) and taking the logarithm gives a log-linear representation of the model:

$$\ln e_{\mathbf{x}} = \delta + \sum_{s=1}^{S} x_s \delta_s + \mathbf{t}'\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}'\boldsymbol{\Gamma}\mathbf{t} \tag{3.10}$$

where $\delta = \ln\{nP(\mathbf{X} = \mathbf{0})\}$.

Without any additional constraint the model in (3.10) remains unidentified. To allow for identification it is possible to fix $\boldsymbol{\mu}$ to be equal to $\mathbf{0}$. Thus, (3.10) may be rewritten as

$$\ln e_{\mathbf{x}} = \delta + \sum_{s=1}^{S} x_s \delta_s + \frac{1}{2}\mathbf{t}'\boldsymbol{\Gamma}\mathbf{t} \tag{3.11}$$

Note that in the resulting model the observed frequency corresponding to the response pattern $\mathbf{x}$ is predicted by $r(r+1)/2$ covariates which are

functions of r sufficient statistics.

EXAMPLE. - MODEL WITH TWO LATENT TRAITS
Suppose that the performance on a test of 9 binary items is supposed to be underlie by two latent variables (named t_1 and t_2).In particular, suppose that responses on items 1 to 4 are dependent of latent trait t_1 and responses on items 5 to 9 are assumed to be dependent on latent trait t_2. The resulting 9×2 full column matrix $\mathbf{U}$ is equal to

$$\mathbf{U} = \begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \\ u_{31} & u_{32} \\ u_{41} & u_{42} \\ u_{51} & u_{52} \\ u_{61} & u_{62} \\ u_{71} & u_{72} \\ u_{81} & u_{82} \\ u_{91} & u_{92} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}.$$

Thus, the resulting model is given by

$$\begin{aligned} P(\mathbf{X} = \mathbf{x}) &= \exp\left\{\delta + \sum_{s=1}^{9} x_s \delta_s + \frac{1}{2}\boldsymbol{t}'\boldsymbol{\Gamma}\boldsymbol{t}\right\} \\ &= \exp\left\{\delta + \sum_{s=1}^{9} x_s \delta_s + \frac{1}{2}t_1^2\gamma_{11} + \frac{1}{2}t_2^2\gamma_{22} + t_1 t_2 \gamma_{22}\right\} \end{aligned}$$

where $x_s = 0$ denote a wrong response, while $x_s = 1$ denote a correct response, and $t_1 = \sum_{s=i}^{9} = u_{s1}x_s$ and $t_2 = \sum_{s=i}^{9} = \nu_{s2}x_s$ are the total scores fitted to the data.

Note that in this case there are $2(2+1)/2 = 3$ parameters of the model to be estimated to account for the two latent traits.

□

Chapter 4

Rasch model as a log-linear model for Capture-Recapture

4.1 Introduction

Modelling dependence between registrations is one of the major issue in capture-recapture framework.

As pointed out in Chapter 1 dependence among registrations may be due to registration dependence and heterogeneity between individuals. However, if both types of dependencies occur, they cannot be disentangled and interaction or common interaction terms cannot be separated (see Chao, 2001 [4]).

A way to model dependence between registrations is to include in the model two factor or higher-order interaction parameters, while to take into account dependence due to heterogeneity Rasch model can be used (see, Darroch et al., 1993 [7], and Agresti, 1994 [1]). Here, correct or incorrect answers to an item are replaced by presence or absence in a registration and heterogeneity among individuals is modelled in terms of constant apparent dependence between registration (see International Working Group for Disease Monitoring and Forecasting, 1995 [13]). With only three registrations available, the first-order heterogeneity parameter H1 is introduced in the model, taking all two-factor interaction terms to be equal and positive. If more than three registrations are available, one can include in the model the second-order heterogeneity parameter H2 (all three-factor interaction terms are supposed equal and positive), and so on.

On the other hand, if additional dependence between registrations occur, then it is possible to add in the model also two factor or higher-order interaction parameters (see Chao, 2001 [4]).

However, in this way the number of parameters to estimate may increase rapidly, as well as the complexity of the model.

Bartolucci and Forcina, 2001 [2], shown how it is possible to relax assumption of conditional dependence and unidimensionality of the Rasch model by assuming that individuals are homogeneous within a finite set of latent classes. This results in adding some suitable columns to the design matrix of the model.

The alternative approach that we propose to deal with this situation is to utilise the multidimensional Rasch model. In particular, we suppose that the registrations may be divided into subgroups which constitute the latent variables which account for correlation among registrations. This is equivalent to assume that random variables denoting the presence or absence in each registration are conditionally independent, given the latent variables.

Note that in capture-recapture method for the estimation of demographic characteristics of a human population there is usually a small number of registrations available. In this case, the jointly estimation procedure (which yields inconsistent estimates for the parameters of the model) cannot be used; thus, the marginal likelihood estimation procedure has to be adopted and the distribution of the latent variables has to be specified. Suppose that the posterior distribution of the latent variables follow a multivariate normal distribution; under this assumption the extension of the Dutch Identity for the multidimensional partial credit model (Hessen, 2012 [11]) can be applied to re-express the probability of a generic capture-profile in a log-linear form.

The remainder of the Chapter is organized as follows: after introducing of notation in Section 4.2, in Section 4.3 the attention is focused on the simpler situation in which three registrations are available. First, in Section 4.3.1 the model that allows for two latent variables is described and an example is proposed. Next, the model in presence of a stratifying variable is discussed in Section 4.3.2 and the particular case of measurement invariance is treated. The extension to a more general situation is straightforward and described in Section 4.4. Finally, the connection between the log-linear representation of the multidimensional Rasch model and the standard log-linear model is explained in Section 4.5. Here it is shown how it is possible to obtain the

parameters of the traditional log-linear model from the parameters of the multidimensional Rasch model.

4.2 Notation

Consider a situation in which S registrations, $R1, R2, \ldots, RS$ are available. Let $\pi_{0_s}, s = 1, 2, \ldots, S$ be the probability of not being observed in the $s-$th registration, let $\pi_{1_s} = 1-\pi_{0_s}$ be the probability of being observed in the $s-$th registration and let $\mathbf{i} = (i_1, i_2, \ldots, i_S)$ denote a generic capture profile for an individual. Let $I_s, s = 1, 2, \ldots, S$ be the random variables denoting the presence or absence of an individual in the corresponding registration and suppose that there are q latent variables which explain the correlation among registrations. Let $\boldsymbol{\Theta} = (\Theta_1, \Theta_2, \ldots, \Theta_q)$ denotes the vector of latent variables and $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_q)$ denotes a realization.

We assume that registrations are conditional independent given the latent variables, that is that the probability of a generic capture profile, given $\boldsymbol{\theta}$, may be written as

$$\pi_{i_1,i_2,\ldots,i_S|\boldsymbol{\theta}} = \prod_{s=1}^{S} \pi_{i_s|\boldsymbol{\theta}} \tag{4.1}$$

where $\pi_{i_s|\boldsymbol{\theta}}$ denotes the conditional probability of the $s-$th registration, given the vector of latent variables $\boldsymbol{\theta}$.

Note that, since $i_s = (0, 1), s = 1, 2, \ldots, S$ the conditional probability of the $s-$th registration may be written as

$$\pi_{i_s|\boldsymbol{\theta}} = \left(\pi_{1_s|\boldsymbol{\theta}}\right)^{i_s} \left(\pi_{0_s|\boldsymbol{\theta}}\right)^{1-i_s} \tag{4.2}$$

and thus

$$\pi_{i_1,i_2,\ldots,i_S|\boldsymbol{\theta}} = \prod_{s=1}^{S} \left(\pi_{1_s|\boldsymbol{\theta}}\right)^{i_s} \left(\pi_{0_s|\boldsymbol{\theta}}\right)^{1-i_s} \tag{4.3}$$

4.3 Three registrations problem

4.3.1 Model with three registrations and two latent variables

Suppose that registrations are not independent of each other and that correlation among registrations may be explained by some latent variables.

For the sake of simplicity, consider a situation in which three registrations are available and there are only two latent variables which explain the correlation among registrations. In other words, assume that the covariances between the random variables I_1, I_2 and I_3 can be explained by two latent variables, say θ_1 and θ_2, that is I_1, I_2 and I_3 are conditional independent given the two latent variables.

Assume, for example, that registrations R1 and R2 are indicators of the first latent variable and that R2 and R3 are indicators of the second latent variable. This situation may be illustrated as in Figure 4.1.

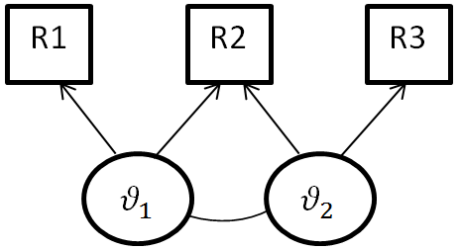

Figure 4.1: Three registrations and two latent variables

In Figure 4.1 the straight arrows between the latent variables and registrations mean that there is a direct influence of the latent variables on the connected registration, while the curved lines between the two latent variables states that there is a correlation between the two latent variables. On the other hand, since there are not direct edge between any pairs of registrations, these are conditional independent given the latent variables.

We are assuming that the capture probabilities are conditionally independent given the two latent variables, that is:

$$\begin{aligned} \pi_{i_1 i_2 i_3|\boldsymbol{\theta}} &= \prod_{s=1}^{3} \pi_{i_s|\boldsymbol{\theta}} \\ &= \prod_{s=1}^{3} \left(\pi_{1_s|\boldsymbol{\theta}}\right)^{i_s} \left(\pi_{0_s|\boldsymbol{\theta}}\right)^{1-i_s} \end{aligned} \tag{4.4}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2)$, and $\pi_{1_s|\boldsymbol{\theta}}$ denotes the probability of being observed in the $s-$th registration, conditionally on the vector of latent variables. This probability may be expressed in a logistic form in the following way:

$$\pi_{1_s|\boldsymbol{\theta}} = \frac{e^{\mathbf{u}_s'\boldsymbol{\theta}-\delta_s}}{1+e^{\mathbf{u}_s'\boldsymbol{\theta}-\delta_s}} \tag{4.5}$$

where $\mathbf{u}_s'$ is the row vector of the (3×2) full column rank matrix $\mathbf{U} = [u_{sr}]$ of weights for the latent variables, where

$$u_{sr} = \begin{cases} 1 & \text{if the registration } R_s \text{ belongs to the } r-\text{th latent variable} \\ 0 & \text{otherwise} \end{cases}$$

For the example above, the matrix $\mathbf{U}$ is then given by

$$\mathbf{U} = \begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \\ u_{31} & u_{32} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

Furthermore, to account for the two latent variables, we need the total scores $\mathrm{t}_1 = u_{11}i_1 + u_{21}i_2 + u_{31}i_3$ and $\mathrm{t}_2 = u_{12}i_1 + u_{22}i_2 + u_{32}i_3$.

Note that model in (4.5) is the multidimensional Rasch model presented in the preceding chapter. Here, δ_s is the parameter for the registration s, while θ_r is the parameter for the $r-$th latent variable.

According to the standard probability theory, the probability of a generic capture profile may be written as

$$\pi_{i_1 i_2 i_3} = \int \dots \int \pi_{i_1 i_2 i_3|\boldsymbol{\theta}} f(\boldsymbol{\theta})\, d\boldsymbol{\theta} \tag{4.6}$$

where $f(\boldsymbol{\theta})$ is the multivariate density of $\boldsymbol{\theta}$.

The Dutch Identity (Holland, 1990 [12]) represents a useful tool that allows to re-express probability in (4.6) such that integrals disappear. Hessen, 2012 [11] proposed an extension of the Dutch Identity for the multidimensional partial credit model (the multidimensional Rasch model in (4.5) is a special case of this model for dichotomous items).

We use the results of Hessen, 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.

Using the fact that

$$\pi_{000|\boldsymbol{\theta}} = \prod_{s=1}^{3} \frac{1}{1+e^{\mathbf{u}_s'\boldsymbol{\theta}-\delta_s}} \tag{4.7}$$

and

$$\pi_{000} = \int \prod_{s=1}^{3} \frac{1}{1 + e^{\mathbf{u}_s'\boldsymbol{\theta} - \delta_s}} f(\boldsymbol{\theta})\, d\boldsymbol{\theta} \tag{4.8}$$

the probability in (4.6) may be written as:

$$\pi_{i_1 i_2 i_3} = \int \prod_{s=1}^{3} \frac{e^{i_s(\mathbf{u}_s'\boldsymbol{\theta} - \delta_s)}}{1 + e^{\mathbf{u}_s'\boldsymbol{\theta} - \delta_s}} f(\boldsymbol{\theta})\, d\boldsymbol{\theta} \tag{4.9}$$

and, after some algebra

$$\pi_{i_1 i_2 i_3} = \pi_{000} e^{-\sum_s i_s \delta_s} \int e^{\mathbf{t}\boldsymbol{\theta}} g(\boldsymbol{\theta} | (i_1 i_2 i_3 = 000))\, d\boldsymbol{\theta} \tag{4.10}$$

where $g(\boldsymbol{\theta}|(i_1 i_2 i_3 = 000))$ is the posterior distribution of $\boldsymbol{\theta}$ given the capture pattern equal to zero (that is the probability of not be observed in any registration).

Note that

$$M_{\boldsymbol{\Theta}}(\mathbf{t}) = \int e^{\mathbf{t}\boldsymbol{\theta}} g(\boldsymbol{\theta} | (i_1 i_2 i_3 = 000))\, d\boldsymbol{\theta}$$

is the moment generating function conditional to the the capture pattern $(i_1 i_2 i_3 = 000)$. In order to compute the probability in (4.10), it is necessary to make an assumption about the posterior distribution of the latent variables and thus to choose a moment generating function. Assume that the posterior distribution of the latent variables follows a multivariate normal distribution, so that

$$M_{\boldsymbol{\Theta}}(\mathbf{t}) = e^{\mathbf{t}'\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}'\boldsymbol{\Gamma}\mathbf{t}} \tag{4.11}$$

This is equal to assume that the population of individuals not observed in any registration follows a normal distribution.

Then, the probability of a generic capture profile $\pi_{i_1 i_2 i_3}$ can be expressed as:

$$\begin{aligned} \pi_{i_1 i_2 i_3} &= \pi_{000} \exp\left\{ \sum_{s=1}^{3} i_s \delta_s + t_1\mu_1 + t_2\mu_2 + \frac{1}{2}t_1^2\gamma_{11} + \frac{1}{2}t_2^2\gamma_{22} + t_1 t_2 \gamma_{12} \right\} \\ &= \pi_{000} \exp\left\{ \sum_{s=1}^{3} i_s \delta_s + \mathbf{t}'\mu + \frac{1}{2}\mathbf{t}'\boldsymbol{\Gamma}\mathbf{t} \right\} \end{aligned} \tag{4.12}$$

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

Let n the number of individuals observed in all registrations. Since the probability of a generic capture pattern $i_1 i_2 i_3$ has multinomial distribution,

we can express the expected frequencies $m_{i_1 i_2 i_3}$ of the observed frequencies $n_{i_1 i_2 i_3}$ as

$$m_{i_1 i_2 i_3} = n\pi_{i_1 i_2 i_3} \tag{4.13}$$

Substituting (4.13) in (4.12) and taking the logarithm is possible to re-express (4.12) in a log-linear representation

$$\ln m_{i_1 i_2 i_3} = \delta + \sum_{s=1}^{3} i_s \delta_s + \mathbf{t}'\mu + \frac{1}{2}\mathbf{t}'\mathbf{\Gamma}\mathbf{t} \tag{4.14}$$

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

Without any additional constraint, the model in equation (4.14) cannot be identified. To go around this problem we can fix $\boldsymbol{\mu}$ to be equal to $\mathbf{0}$. Then, the model can be rewritten as:

$$\begin{aligned} \ln m_{i_1 i_2 i_3} &= \delta + \sum_{s=1}^{3} i_s \delta_s + \frac{1}{2}\mathbf{t}'\mathbf{\Gamma}\mathbf{t} \\ &= \delta + i_1\delta_1 + i_2\delta_2 + i_3\delta_3 + \frac{1}{2}t_1^2\gamma_{11} + \frac{1}{2}t_2^2\gamma_{22} + t_1 t_2 \gamma_{12} \end{aligned} \tag{4.15}$$

The resulting model is denoted as R1+R2+R3+$\theta_1 + \theta_2$.

Note that, as pointed out in the previous chapter, 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} represent the covariance between the two latent variables, given the total scores t_1 and t_2.

EXAMPLE 4.1. - CONSTRUCTING THE TWO LATENT VARIABLES
In order to better understand how to account for the two latent variables and fit the model, matrix approach may be useful.

Let $\mathbf{m}$ be the vector of expected counts

$$\mathbf{m} = \begin{pmatrix} m_{000} & m_{001} & \dots & m_{111} \end{pmatrix}'.$$

In matrix term the model in (4.15) may be written as

$$\ln \mathbf{m} = \mathbf{X}\boldsymbol{\theta}$$

where $\boldsymbol{\theta} = \begin{pmatrix} \delta & \delta_1 & \delta_2 & \delta_3 & \gamma_{11} & \gamma_{22} & \gamma_{12} \end{pmatrix}'$ is the vector of parame-

ters to be estimated and $\mathbf{X}$ is the design matrix whose columns are the vectors of coefficients of each parameter, that is

$$\mathbf{X} = \begin{pmatrix} \mathbf{1} & \mathbf{i}_1 & \mathbf{i}_2 & \mathbf{i}_3 & \mathbf{t}_1^2 & \mathbf{t}_2^2 & \mathbf{t_1 t_2} \end{pmatrix}.$$

Suppose that three registrations R1, R2 and R3 are available and that R2 is indicator of the first latent variable and that R1 and R3 are indicators of the second latent variable. The matrix $\mathbf{U}$ is given by

$$\mathbf{U} = \begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \\ u_{31} & u_{32} \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

and the total scores t_1 and t_2 for each capture profile are computed using

$$\mathbf{t} = \mathbf{U}'\mathbf{i} = \begin{bmatrix} u_{11} & u_{21} & u_{31} \\ u_{12} & u_{22} & u_{32} \end{bmatrix} \begin{bmatrix} i_1 \\ i_2 \\ i_3 \end{bmatrix} = \begin{bmatrix} t_1 \\ t_2 \end{bmatrix}.$$

Then matrix $\mathbf{X}$ may be written as

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 4 & 0 \\ 1 & 1 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 4 & 2 \end{pmatrix}$$

and the model may be fitted as a traditional log-linear model.

□

4.3.2 Model with three registrations, two strata and two latent variables

The model presented above can be applied also in the case in which a stratifying variable is available. For convenience, consider the simple situation in which three registrations are recorded in two strata (or time periods, for example two years). Here, year is a stratified variable with two categories denoted by the index j and $n_{i_1 i_2 i_3 j}$ and $\pi_{i_1 i_2 i_3 j}$ denote the observed frequencies and the probabilities for year j, respectively.

The resulting contingency table has two missing cells, one corresponding to individuals not observed in either registration for the first year, and one corresponding to individuals missed by all registrations in the second year. The corresponding contingency table is shown in Table 4.1.

Table 4.1: Contingency table for three lists and two strata

		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

Suppose that, also in this situation, we have two latent variables.

The probability of a generic capture profile may be written as

$$\pi_{i_1 i_2 i_3 j} = \int \dots \int \pi_{i_1 i_2 i_3 j|\boldsymbol{\theta}} f(\boldsymbol{\theta})\, d\boldsymbol{\theta} \tag{4.16}$$

where $\pi_{i_1 i_2 i_3 j|\boldsymbol{\theta}}$ is the probability of capture profile $i_1 i_2 i_3$ for year j and $f(\boldsymbol{\theta})$ is the multivariate density of $\boldsymbol{\theta}$.

Similarly to the previous case, assuming that the posterior distribution (given the capture pattern equal to zero) of the latent variables follows a multivariate normal distribution, model in equation (4.15) can be written as

$$\pi_{i_1 i_2 i_3 j} = \pi_{000 j} \exp\left\{ \sum_{s=1}^{3} i_s \delta_{sj} + \mathbf{t}' \mu_j + \frac{1}{2} \mathbf{t}' \boldsymbol{\Gamma}_j \mathbf{t} \right\} \tag{4.17}$$

where $\boldsymbol{\mu}_j$ is the mean vector for the j-th strata and $\boldsymbol{\Gamma}_j$ is a symmetric matrix.

Let $m_{i_1 i_2 i_3 j}$ denotes the expected frequency corresponding to the observed frequency $n_{i_1 i_2 i_3 j}$, that is

$$m_{i_1 i_2 i_3 j} = n \pi_{i_1 i_2 i_3 j} \tag{4.18}$$

Thus, substituting (4.18) in (4.17) we obtain

$$\ln m_{i_1 i_2 i_3 j} = \delta_j + \sum_{s=1}^{3} i_s \delta_{sj} + \mathbf{t}'\mu_j + \frac{1}{2}\mathbf{t}'\boldsymbol{\Gamma_j}\mathbf{t} \tag{4.19}$$

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

Without any additional constraints model in (4.19) cannot be identified; setting $\boldsymbol{\mu}_j$ equal to zero for identification we have

$$\ln m_{i_1 i_2 i_3 j} = \delta_j + \sum_{s=1}^{3} i_s \delta_{sj} + \frac{1}{2}\mathbf{t}'\boldsymbol{\Gamma_j}\mathbf{t} \tag{4.20}$$

4.3.3 Model of measurement invariance

Assume now that parameters are equal across the years. This means that the model has measurement invariance across strata (that is, the model applies across years). Under assumption of measurement invariance we have

$$\delta_{sj} = \delta_s, \qquad \forall j \tag{4.21}$$

Thus, model in equation (4.14) is equal to:

$$\ln m_{i_1 i_2 i_3 j} = \delta_j + \sum_{s=1}^{3} i_s \delta_s + \mathbf{t}'\boldsymbol{\mu}_j + \frac{1}{2}\mathbf{t}'\boldsymbol{\Gamma_j}\mathbf{t} \tag{4.22}$$

Without any additional constraint this model cannot be estimated. To identify the model we can set $\boldsymbol{\mu}_j$ to $\mathbf{0}$ for one j.

In the case of measurement invariance, it is possible to test whether $\boldsymbol{\mu}_j = \boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Gamma}_j = \boldsymbol{\Gamma}$ for all j.

If the simultaneous hypothesis holds, then model in (4.22) becomes

$$\ln m_{i_1 i_2 i_3 j} = \delta_j + \sum_{s=1}^{3} i_s \delta_s + \frac{1}{2}\mathbf{t}'\boldsymbol{\Gamma}\mathbf{t}. \tag{4.23}$$

4.4 General case

The extension of the method described in the preceding sections to a more general situation is straightforward.

Assume that we have S registrations and J strata. 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 registrations and $j = (1, 2, \ldots, J)$ is the index denoting the strata. Note that the resulting contingency table has J structural zeros (one for each strata).

Suppose now that 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 registration RS belongs to the rth 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 $t_r = \sum_{s=1}^{S} u_{sr} i_s$.

Similarly to the simpler situations, under assumption of multivariate normal distribution of the posterior distribution of the latent variables (conditional to the capture pattern of individuals not observed in any registration), the probability of a generic capture profile $\pi_{i_1 \ldots i_s j}$ is equal to

$$\pi_{i_1 \ldots i_s j} = \pi_{0 \ldots 0 j} \exp\left\{ \sum_{s=1}^{S} i_s \delta_{sj} + \mathbf{t}' \mu_j + \frac{1}{2} \mathbf{t}' \boldsymbol{\Gamma}_j \mathbf{t} \right\} \tag{4.24}$$

where $\boldsymbol{\mu}_j$ is the mean vector for the j-th strata and $\boldsymbol{\Gamma}_j$ is a symmetric matrix.

Let $m_{i_1 \ldots i_s j} = n \pi_{i_1 \ldots i_s j}$ denotes the expected counts of observed frequencies $n_{i_1 \ldots i_s j}$. Then we have the log-linear representation

$$\ln m_{i_1 \ldots i_s j} = \delta_j + \sum_{s=1}^{S} i_s \delta_{sj} + \mathbf{t}' \boldsymbol{\mu}_j + \frac{1}{2} \mathbf{t}' \boldsymbol{\Gamma_j} \mathbf{t} \tag{4.25}$$

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

$$\ln m_{i_1 \ldots i_s j} = \delta_j + \sum_{s=1}^{S} i_s \delta_{sj} + \frac{1}{2} \mathbf{t}' \boldsymbol{\Gamma_j} \mathbf{t} \tag{4.26}$$

Thus, the model in (4.26) can be treated as a traditional log-linear model and, once the parameters have been estimated can be used to obtain the estimate of the portion of population missed by all registrations and thus the total unknown population size N.

Also in this case, if assumption of measurement invariance holds, then the model in (4.26) can be written in the following way:

$$\ln m_{i_1 \ldots i_s j} = \delta_j + \sum_{s=1}^{S} i_s \delta_s + \mathbf{t}'\boldsymbol{\mu} + \frac{1}{2}\mathbf{t}'\boldsymbol{\Gamma}\mathbf{t} \tag{4.27}$$

4.5 Connection between log-linear and multidimensional Rasch models

Consider the general situation described in the preceding section with S registrations and a stratifying variable available. Consider the log-linear model in which all the two-factor interaction parameters are present and suppose that parameters differ among strata. This model can be written as:

$$\ln m_{i_1 \ldots i_S j} = \lambda_j + \sum_{s=1}^{S} i_s \lambda_{sj} + \sum_{s=1}^{S-1} \sum_{c=s+1}^{S} i_s i_c \lambda_{scj} \tag{4.28}$$

Consider now the log-linear representation of the multidimensional Rasch model in (4.25); it is equal to

$$\ln m_{i_1 \ldots i_s j} = \delta_j + \sum_{s=1}^{S} i_s \delta_{sj} + \frac{1}{2} \sum_{r=1}^{q} t_r^2 \gamma_{rrj} + \sum_{r=1}^{q-1} \sum_{\nu=r+1}^{q} t_r t_\nu \gamma_{r\nu j} \tag{4.29}$$

where $\boldsymbol{\mu}_j$ is set to zero for identification.

Using the fact that $t_r = \sum_{s=1}^{S} u_{sr} i_s$ it is possible to obtain a re-parametrization of the model that allows for a connection between the multidimensional Rasch model and the standard log-linear model. In fact, writing out t_r^2 and $t_r t_\nu$ we have:

$$t_r^2 = \sum_{s=1}^{S} u_{sr}^2 i_s^2 + 2 \sum_{s=1}^{S-1} \sum_{c=s+1}^{S} u_{sr} u_{cr} i_s i_c \tag{4.30}$$

and

$$t_r t_\nu = \sum_{s=1}^{S} u_{sr} u_{s\nu} i_s^2 + \sum_{s=1}^{S-1} \sum_{c=s+1}^{S} \left(u_{sr} u_{c\nu} + u_{s\nu} u_{cr} \right) i_s i_c \tag{4.31}$$

Substituting these expressions in (4.29) and noting that $i_s^2 = i_s$ and $u_{sr}^2 = u_{sr}$

we obtain

$$\begin{aligned}\ln m_{i_1\ldots i_S j} &= \delta_j + \sum_{s=1}^{S} i_s\delta_{sj} + \frac{1}{2}\sum_{r=1}^{q}\left[\sum_{s=1}^{S} u_{sr}i_s + 2\sum_{s=1}^{S-1}\sum_{c=s+1}^{S} u_{sr}u_{cr}i_s i_c\right]\gamma_{rrj} \\ &+ \sum_{r=1}^{q-1}\sum_{\nu=r+1}^{q}\left[\sum_{s=1}^{S} u_{sr}u_{s\nu}i_s + \sum_{s=1}^{S-1}\sum_{c=s+1}^{S}\left(u_{sr}u_{c\nu}+u_{s\nu}u_{cr}\right)i_s i_c\right]\gamma_{r\nu j}\end{aligned}$$

so that

$$\begin{aligned}\ln m_{i_1\ldots i_S j} &= \delta_j + \sum_{s=1}^{S} i_s\left[\delta_{sj} + \frac{1}{2}\sum_{r=1}^{q} u_{sr}\gamma_{rrj} + \sum_{r=1}^{q-1}\sum_{\nu=r+1}^{q} u_{sr}u_{s\nu}\gamma_{r\nu j}\right] \\ &+ \sum_{s=1}^{S-1}\sum_{c=s+1}^{S} i_s i_c\left[\sum_{r=1}^{q} u_{sr}u_{cr}\gamma_{rrj} + \sum_{r=1}^{q-1}\sum_{\nu=r+1}^{q}\left(u_{sr}u_{c\nu}+u_{s\nu}u_{cr}\right)\gamma_{r\nu j}\right]\end{aligned} \tag{4.32}$$

Note that model in (4.32) is equal to the model in (4.28), in which

$$\lambda_j = \delta_j \tag{4.33}$$

$$\lambda_{sj} = \delta_{sj} + \frac{1}{2}\sum_{r=1}^{q} u_{sr}\gamma_{rrj} + \sum_{r=1}^{q-1}\sum_{\nu=r+1}^{q} u_{sr}u_{s\nu}\gamma_{r\nu j} \tag{4.34}$$

and

$$\lambda_{scj} = \sum_{r=1}^{q} u_{sr}u_{cr}\gamma_{rrj} + \sum_{r=1}^{q-1}\sum_{\nu=r+1}^{q}\left(u_{sr}u_{c\nu}+u_{s\nu}u_{cr}\right)\gamma_{r\nu j} \tag{4.35}$$

The expressions (4.33)-(4.35) are useful to compute the parameters of the log-linear model in (4.28), using the parameters of the multidimensional Rasch model in (4.29).

Suppose now that assumption of measurement invariance holds. Under this assumption the traditional log-linear model in (4.28) takes the form

$$\ln m_{i_1\ldots i_S j} = \lambda_j + \sum_{s=1}^{S} i_s\lambda_s + \sum_{s=1}^{S-1}\sum_{c=s+1}^{S} i_s i_c\lambda_{sc} \tag{4.36}$$

while the multidimensional Rasch model is equal to

$$\ln m_{i_1\ldots i_s j} = \delta_j + \sum_{s=1}^{S} i_s\delta_s + \frac{1}{2}\sum_{r=1}^{q} t_r^2\gamma_{rr} + \sum_{r=1}^{q-1}\sum_{\nu=r+1}^{q} t_r t_\nu\gamma_{r\nu} \tag{4.37}$$

where $\boldsymbol{\mu}$ is set to zero for identification.

Similarly to the previous case, it is possible to obtain a re-parametrization of the model where

$$\lambda_s = \delta_s + \frac{1}{2}\sum_{r=1}^{q} u_{sr}\gamma_{rr} + \sum_{r=1}^{q-1}\sum_{\nu=r+1}^{q} u_{sr}u_{s\nu}\gamma_{r\nu} \tag{4.38}$$

and

$$\lambda_{sc} = \sum_{r=1}^{q} u_{sr}u_{cr}\gamma_{rr} + \sum_{r=1}^{q-1}\sum_{\nu=r+1}^{q} \left(u_{sr}u_{c\nu} + u_{s\nu}u_{cr}\right)\gamma_{r\nu} \tag{4.39}$$

and thus, it is easy to obtain the parameter for the log-linear model in (4.36) from those of the multidimensional Rasch model in (4.37).

Example 4.2.
Consider the Example 4.1. In this case the expressions (4.34)-4.35 become:

$$\lambda_s = \delta_s + \frac{1}{2}\sum_{r=1}^{2} u_{sr}\gamma_{rr} + u_{s1}u_{s2}\gamma_{12}$$

and

$$\lambda_{sc} = \sum_{r=1}^{2} u_{sr}u_{cr}\gamma_{rr} + \left(u_{s1}u_{c2} + u_{s2}u_{c1}\right)\gamma_{12}$$

Applying these formulae to the example we obtain the parameters of the traditional log-linear model in the following way:

$$\begin{aligned}
\lambda_1 = &\quad \delta_1 + \frac{1}{2}\left(u_{11}\gamma_{11} + u_{12}\gamma_{22}\right) + u_{11}u_{12}\gamma_{12} = &\quad \delta_1 + \frac{1}{2}\gamma_{22}\\
\lambda_2 = &\quad \delta_2 + \frac{1}{2}\left(u_{21}\gamma_{11} + u_{22}\gamma_{22}\right) + u_{21}u_{22}\gamma_{12} = &\quad \delta_2 + \frac{1}{2}\gamma_{11}\\
\lambda_3 = &\quad \delta_3 + \frac{1}{2}\left(u_{31}\gamma_{11} + u_{32}\gamma_{22}\right) + u_{31}u_{32}\gamma_{12} = &\quad \delta_3 + \frac{1}{2}\gamma_{22}
\end{aligned}$$

$$\begin{aligned}
\lambda_{12} = &\quad u_{11}u_{21}\gamma_{11} + u_{12}u_{22}\gamma_{22} + \left(u_{11}u_{22} + u_{12}u_{21}\right)\gamma_{12} = &\quad \gamma_{12}\\
\lambda_{13} = &\quad u_{11}u_{31}\gamma_{11} + u_{12}u_{32}\gamma_{22} + \left(u_{11}u_{32} + u_{12}u_{31}\right)\gamma_{12} = &\quad \gamma_{22}\\
\lambda_{23} = &\quad u_{21}u_{31}\gamma_{11} + u_{22}u_{32}\gamma_{22} + \left(u_{21}u_{32} + u_{22}u_{31}\right)\gamma_{12} = &\quad \gamma_{12}
\end{aligned}$$

Note that in this case there are not registrations in common between the two latent variables and, to obtain the main-effect parameters (λ_s) for the registration s, we add to the main-effect parameter in the multidimensional Rasch model half of the variance (given the total scores)

of the latent variable to which registration belongs. On the other hand, the two-factor interaction parameters (λ_{sc}) of the log-linear model correspond to the variance (given the total scores) for those two-factor interaction parameters which involve registrations which are indicator of the same latent variable; while the two-factor interaction parameters for registrations which belong to different latent variables are equal to the covariance (given the total scores) between the two latent variables.

However, if we construct the two latent variables differently, then we obtain a different parametrization for both main-effect parameters and two-factor interaction parameters of the standard log-linear model. In particular, suppose that the two latent variables have a registration in common, i.e. registrations R1 and R2 are indicator of the first latent variable and registration R1 and R3 are indicators of the second latent variable. Now, the matrix $\mathbf{U}$ is given by

$$\mathbf{U} = \begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \\ u_{31} & u_{32} \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Applying formulae (4.34)-(4.35) to this situation, we obtain the following expressions for the parameters of the standard log-linear model:

$$\lambda_1 = \delta_1 + \frac{1}{2}(\gamma_{11} + \gamma_{22}) + \gamma_{12} \qquad \lambda_2 = \delta_2 + \frac{1}{2}\gamma_{11} \qquad \lambda_3 = \delta_3 + \frac{1}{2}\gamma_{22}$$

$$\lambda_{12} = \gamma_{11} + \gamma_{12} \qquad \lambda_{13} = \gamma_{22} + \gamma_{12} \qquad \lambda_{23} = \gamma_{12}$$

Here, to obtain the main-effect parameter for the registration that belongs to both the latent variables, we have to add to the main-effect parameter in the multidimensional Rasch model half of the two variances (given the total scores) of the two latent variables and the covariance (given the total scores) between the two latent variables. The other main-effect parameters are the same to the previous situation. Concerning the two-factor interaction parameters, for the parameter involving registrations which belong to different latent variables it is equal to the covariance (given the total scores) between the two latent variables; on the other hand, the two two-factor interaction parameters which involve the registration R1 (that is the registration in common for the two latent variables) era equal to the the variance (given the total scores) of the latent variable to which the other registration belongs, plus the covariance (given the total scores) between the two latent variables.

□

Chapter 5

Application

5.1 Introduction

To illustrate the procedure presented in Chapter 4, the present application is adapted from the data set described by Zwane et al. (2004, [27]) on neural tube defects (NTD's) in the Netherlands.

Neural tube defects are serious congenital defects contributing to infant mortality and serious disability. They can occur in the first month of pregnancy and result from failure of the neural tube to close during the fetal development; consequently, the spinal cord, brain, and related structures do not form properly.

The target population of interest includes children born with an NTD's in the Netherlands during the years 1988 through 1998. In the Netherlands, several national databases record cases of neural tube defects. In the remainder of this chapter we will consider five registrations: the *Dutch Perinatal Database I*, the *Dutch Perinatal Database II*, the *National Neonate Database*, the *Dutch Monitoring System of Child Health Care* and the *Dutch Association of Patients with a NTD*.

None of these registrations record all cases of neural tube defects in the Netherlands, but children with NTD's may be included in more than one of the registrations. So, capture-recapture method can be utilised to estimate the total unknown number of children affected by NTD's in the Netherlands. The scope of this application is to estimate the total unknown number of children affected by NTD's in the Netherlands and to demonstrate that the use of the methodology presented in Chapter 4 improves the accuracy of

estimation.

The remainder of the Chapter is organised as follow: first of all a brief description of five registrations and of the structure of data set on NTD's is given in Section 5.2. Since not all registrations are operational in the same period of time, there are several unobservable cells in the data set. Zwane et al. approached the absence of observations as an incomplete data problem and proposed the E-M algorithm to estimate the unobservable cells resulting from not operational registrations. In Section 5.3 a brief description of their procedure is presented. Then, in Section 5.4, the multidimensional Rasch model discussed in Chapter 4 is applied to the NTD's data set and results are discussed in Section 5.5.

5.2 Dataset

The five registrations used in this application are described below:

The Dutch Perinatal Database I

The Dutch Perinatal Database I is an anonymous register of pregnancy and birth which record low risk pregnancies and births. Such cases are referred as primary care, that is the health care services provided by health care professionals who act as a first point of consultation for all patients within the health care system (in the Netherlands, midwives are responsible for care in such cases). Data also include cases for which care only relates to a part of pregnancy or delivery.

For this registration, data from 1988 through 1998 are utilised.

In the following we refer to the Dutch Perinatal Database I as R1.

The Dutch Perinatal Database II

The Dutch Perinatal Database II is an anonymous register which record information about birth of children in secondary care, that is that care services are provided by medical specialists and other health professionals who generally do not have first contact with patients (for example, paediatricians or gynaecologists).

Data for this registration are recorded from 1988 through 1998.

In the following we refer to the Dutch Perinatal Database II as R2.

The National Neonatal Database
The National Neonatal Database is an anonymous database recording informations about all admissions and re-admissions of newborns to paediatric departments within the first 28 days of life.

Data from 1992 through 1998 are used.

In the following we refer to the National Neonatal Database as R3.

The Dutch Monitoring System of Child Health Care
The Dutch Monitoring System of Child Health Care is a register which record information about children born alive with a NTD's who visit paediatrician for the first time. Thus, paediatric departments participated in the registration.

For this registration, data referred to period 1993-1998.

In the following we refer to the Dutch Monitoring System of Child Health Care as R4.

The Dutch Association of Patients with a NTD
This is a short questionnaire that was sent to every member of Dutch Association of Patients with a NTD with a child affected by NTD in the period 1988-1998.

In the following we refer to the Dutch Association of Patients with a NTD as R5.

Data are summarised in Table 5.1. Here, capture profiles are denoted as $i_{R1}i_{R2}i_{R3}i_{R4}i_{R5}$, so that profile 10000 indicates the frequency of children observed in registration R1, but not in the other registrations; profile 11000 denotes the frequency of observations included in registrations R1 and R2 but not included in R3, R4 and R5, and so on.

Note that for the first two registrations R1 and R2 abortions are possible, while they cannot be present in the other registrations. For this reason, attention is restricted to children with a pregnancy duration grater than 24 weeks (that represents the legal limit for abortion in the Netherlands).

Table 5.1: Observed frequencies on NTD's in the Netherlands

Year	00000	10000	01000	11000	00100	10100	01100	11100	00010	10010	01010	11010	00110	10110	01110	11110	00001	10001	01001	11001	00101	10101	01101	11101	00011	10011	01011	11011	00111	10111	01111	11111
1988	0*	4	101	24	0*				0*				0*				9	0	5	2												
1989	0*	3	114	30	0*				0*				0*				3	1	8	4												
1990	0*	3	105	43	0*				0*				0*				7	3	5	4												
1991	0*	4	95	32	0*				0*								3	1	7	8												
1992	0*	9	80	27	15	0	12	7	0*								10	1	3	3	0	0	2	3								
1993	0*	5	61	24	4	1	1	0	12	0	18	7	2	0	4	3	3	0	0	1	0	0	0	0	1	1	7	4	0	1	0	0
1994	0*	24	34	13	6	1	1	1	15	7	18	9	5	4	5	2	3	0	1	1	0	1	0	1	1	3	1	0	0	1	0	4
1995	0*	29	27	15	5	1	2	1	16	15	18	4	4	0	9	5	2	1	2	0	0	1	3	0	2	2	5	1	3	0	1	0
1996	0*	26	26	11	10	1	1	1	9	6	11	9	5	0	4	5	5	0	0	0	0	0	0	0	5	6	2	1	1	2	2	4
1997	0*	41	26	18	13	2	0	1	12	11	11	7	3	4	6	3	4	2	0	1	2	0	0	0	1	1	1	3	1	2	0	4
1998	0*	27	25	20	13	0	2	1	8	7	14	3	6	11	4	7	1	0	0	0	0	0	1	0	0	0	1	2	0	0	1	1

0* denotes structural zeros cells

Then, it should be noted that the five registrations do not refer to the same years, since only three registrations (R1, R2 and R5) cover the same period of time of 11 years (from 1988 through 1998), while registration R3 refers to the period 1992-1998 and registration R4 is available from 1993 to 1998. Thus, there are 24 structural zeros cells in the resulting contingency table; in particular, there are 11 structural zeros cells corresponding to capture profile 00000 (children missed by all registrations) for years 1988 through 1998; for each years from 1988 to 1991 there are 3 more structural zeros cells corresponding to capture profiles 00100, 00010 and 00110, that correspond to children observed in registration R3 only, in registration R4 only, and in both R3 and R4, respectively. Finally, for year 1992 there is 1 more structural zero cell corresponding to capture profile 00010, that is children observed only in registration R4.

In addition, for years with incomplete registrations, observed frequencies for some capture profiles may also include observations that could have been a different capture profile if registrations R3 and/or R4 had been active. In particular, for years 1988 to 1991 registrations R3 and R4 are not available and observed frequencies corresponding to capture profile "not observed" for R3 and R4 may be distributed also to profiles corresponding to the cross-classification in registrations R3 and R4. Thus, for example, capture profile 01000 denoting the frequency of NTD's cases recorded only in registration R2 may also include observations for capture profiles 01100 (cases observed in registrations R2 and R3), 01010 (cases observed in registrations R2 and R4), 01110 (cases observed in registrations R2, R3 and R4). Similarly, for 1992 only registration R4 is not available and observed frequencies corresponding to capture profile "not observed" for R4 may be distributed also to profiles corresponding to capture profile "observed" for R4 in this year; thus, for example, capture profile 10000, denoting cases observed only in R1 may also include observations for capture profile 10010 (cases observed in R1 and R4). In Table 5.1(a) are reported all possible configurations for incomplete years 1988 to 1991, while in Table 5.1(b) are reported all possible configurations for 1992.

(a) Years 1988 – 1991

Observed capture profiles (with only R1, R2 and R5 active)	Possible capture profiles (if all registrations had been active)
10000	10000 10100 10010 10110
01000	01000 01100 01010 01110
00001	00001 00101 00011 00111
11000	11000 11100 11010 11110
10001	10001 10101 10011 10111
01001	01001 01101 01011 01111
11001	11001 11101 11011 11111

(b) Year 1992

Observed capture profiles (with R4 non-active)	Possible capture profiles (if R4had been active)
10000	10000 10010
01000	01000 01010
00100	00100 00110
00001	00001 00011
11000	11000 11010
10100	10100 10110
10001	10001 10011
01100	01100 01110
01001	01001 01011
00101	00101 00111
11100	11100 11110
11001	11001 11011
10101	10101 10111
01101	01101 01111
11101	11101 11111

Table 5.2: Possible capture profiles

5.3 EM Algorithm

As pointed out in the preceding section, the five registrations on NTD's in the Netherlands cover different but overlapping periods of time. Zwane et al. (2004, [27]) showed that if the fact that registrations refer to different but overlapping populations is ignored, then the resulting estimates of the total population size may be biased. They approached this situation as a missing data problem and presented a version of the EM algorithm to estimate the missing entry (see Table 5.2) resulting from registrations that are non operating in some strata.

The EM (Expectation Maximization) algorithm is an iterative procedure proposed by Dempster et al. (1997, [8]) useful to compute the maximum likelihood estimates when the observations can be view as incomplete data. Data are assumed to be "missing at random" (MAR) (Rubin, 1976), that is that the missing value is conditionally independent of the actual response that would have been observed given the observed responses to other questions. In capture-recapture context, this means that observations from years where all registrations are active and observations from years with non-operating registrations with the same characteristics do not differ systematically by year (see Zwane et al., 2004).

In the EM algorithm proposed by Zwane et. al data set is divided into two groups: one group, denoted by S_1, containing years where all registrations are available (completely classified observations); the other one, denoted by S_2, consists of years for which non all registrations are available. Thus, S_1 consists of 6 years, while S_2 consists of 5 years, that is $S_1 = (1993, 1994, 1995, 1996, 1997, 1998)$ and $S_2 = (1988, 1989, 1990, 1991, 1992)$.

In the $t-$th iteration of the E-step, the expected frequencies of partially classified profiles are calculated. In particular, the partially classified frequencies in S_2 are distributed to possible capture profiles (see Table 5.2) using information from S_1.

Once all expected frequencies are computed and the data set is completed, in the M-step a log-linear model is fitted to completed data and the log-likelihood is maximised in order to calculate the estimate probabilities that will be used in the $(t+1)-$th iteration of the E-step. Thus, the updates for the completed data are derived and the log-linear model is fitted in the M-step.

This procedure is repeated until the log-likelihood function converge. Then, the parameters estimated in the last step of the algorithm are used to estimate the expected frequencies for structural zero cells, and finally the estimation of the total population size is obtained.

In order to better understand how to apply this procedure to the NTD's data set, consider, for example, the capture profile 10000 for year 1992 for the NTD's data set. Observed frequency for this profile is 9, but the EM algorithm has to distribute this value to capture profiles 10000 and 10010.

Let $n_{i_1i_2i_3i_4i_5|k}$ and $\hat{n}^{(t)}_{i_1i_2i_3i_4i_5|k}$ denote the observed frequencies and the expectation of the frequencies in the $t-$th step of the algorithm, respectively, where $i_1i_2i_3i_4i_5$ denotes the capture profile and k indicates the year.

The $(t+1)-$th E-step of the EM algorithm calculates the expectations of frequencies of capture profiles 10000 and 10010 for year 1992 in the following way:

$$\hat{n}^{(t+1)}_{10000|1992} = \frac{\sum_{k\in S_1} \hat{n}^{(t)}_{10000|k}}{\sum_{k\in S_1} \hat{n}^{(t)}_{100+0|k}} \times n_{10000|1992}$$

$$\hat{n}^{(t+1)}_{10010|1992} = \frac{\sum_{k\in S_1} \hat{n}^{(t)}_{10010|k}}{\sum_{k\in S_1} \hat{n}^{(t)}_{100+0|k}} \times n_{10000|1992}$$

where $\hat{n}^{(t)}_{100+0|k} = \hat{n}^{(t)}_{10000|k} + \hat{n}^{(t)}_{10010|k}$.

5.4 The Multidimensional Rasch Model

To apply the multidimensional Rasch model to the dataset on NTD's in the Netherlands, we assume that the five registrations R1, R2, R3, R4 and R5 may be divided into two subgroups which constitute the latent variables which account for correlation among registrations.

In order to decide which registrations belong to the same latent variable, we fit in the M-step of the EM algorithm presented above the log-linear model that allows for the presence of all the two-factor interaction parameters; we denoted this model as (R1R2+...+R4R5)+Y_{cat}, where Y_{cat} denotes the Year (that is treated as a stratifying variable). Table 5.3 summarize the estimates for the two factor interaction parameters among registrations, after the convergence of the EM algorithm.

Table 5.3: Estimates of the two-factor interaction parameters

	R1	R2	R3	R4	R5
R1	-				
R2	0.718424	-			
R3	0.185740	0.024525	-		
R4	0.557406	1.055780	1.690401	-	
R5	0.633640	-0.100489	0.467334	1.725820	-

From Table 5.3, we assume that registrations R1 and R2 belong to the first latent variable (named θ_1), while registrations R3, R4 and R5 belong to the second latent variable (called θ_2). Figure 5.1 illustrate this situation.

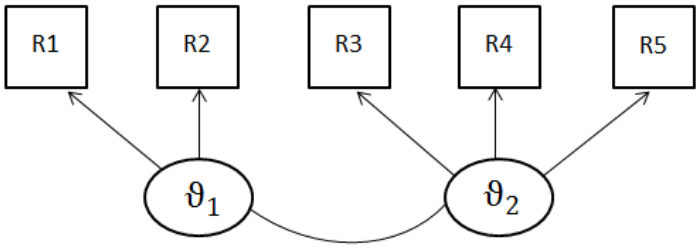

Figure 5.1: Model with five registrations and two latent variables

In this case, the matrix $\mathbf{U}$ of weights for the latent variables is given by:

$$\mathbf{U} = \begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \\ u_{31} & u_{32} \\ u_{41} & u_{42} \\ u_{51} & u_{52} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}.$$

Assuming that hypothesis of measurement invariance holds, the model considered takes the form

$$\ln m_{i_1 i_2 i_3 i_4 i_5} = \delta + \delta_j + \sum_{s=1}^{5} i_s \delta_s + \frac{1}{2} \sum_{r=1}^{2} t_r^2 \gamma_{rr} + t_1 t_2 \gamma_{12} \quad j = 1988, \ldots, 1997$$

where t_1 and t_2 are the total scores accounting for the latent variables θ_1 =R1+R2 and θ_2 =R3+R4+R5, respectively, δ is the general mean and δ_j are the main-effect parameters for years (here year 1998 is assumed as reference category).

The resulting model has 19 parameters, since there is 1 parameter for the general mean, 10 parameters(δ_j) accounting for years, 5 parameters (δ_s) accounting for registrations and 3 parameters () for the two latent variables γ_{11}, γ_{22} and γ_{12} denoting, respectively, the variance (given the total scores t_1 and t_2) of θ_1,the variance (given the total scores) of θ_2 and the covariance (given the total scores) between the latent variables. In addition, since there are 229 observed cells in the dataset (see Table 5.1) the model has 210 degrees of freedom.

From Table 5.3 it seems also reasonable to assume that registrations R1, R2 and R4 belong to the same latent variable (say θ_3), while registrations R3, R4 and R5 are indicator of the other latent variable (named θ_4). In this case, the two latent variables have registration R4 in common. Figure 5.2 shows this situation.

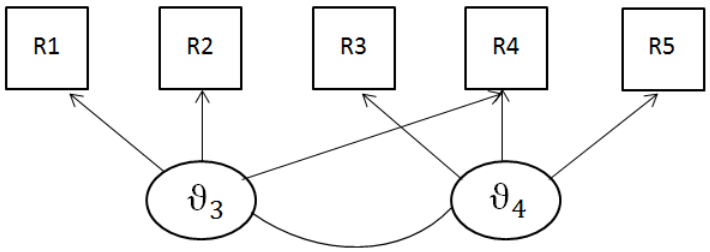

Figure 5.2: Model with five registrations and two latent variables

Now, the matrix **U** is given by

$$\mathbf{U} = \begin{bmatrix} u_{13} & u_{14} \\ u_{23} & u_{24} \\ u_{33} & u_{34} \\ u_{43} & u_{44} \\ u_{53} & u_{54} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 1 \\ 0 & 1 \end{bmatrix}.$$

and, under assumption of measurement invariance, the model considered takes the form

$$\ln m_{i_1 i_2 i_3 i_4 i_5} = \delta + \delta_j + \sum_{s=1}^{5} i_s \delta_s + \frac{1}{2} \sum_{r=3}^{4} t_r^2 \gamma_{rr} + t_3 t_4 \gamma_{34} \quad j = 1988, \ldots, 1997$$

where t_3 and t_4 are the total scores accounting for the latent variables θ_3 =R1+R2+R4 and θ_4 =R3+R4+R5, respectively.

Analogous to the preceding case, the resulting model has 19 parameters and 210 degrees of freedom.

5.5 Results

In this section we present the results obtained fitting the two multi-dimensional Rasch models discussed in the previous section to the NTD's dataset, and compare them with other log-linear models. In total, we take into account five models, which are:

- **Model 1**: the model that allows for all main-factor parameters and Year as a stratifying variable, that is

$$\ln m_{i_1 i_2 i_3 i_4 i_5} = \lambda + \lambda_j + \sum_{s=1}^{5} i_s \lambda_s \qquad j = 1988, \ldots, 1997$$

  We denote this model as 1: R1+R2+R3+R4+R5+Y_{cat}. It has 16 parameters, that are 1 parameter for the general mean, 10 parameters(λ_j) accounting for years, 5 parameters (λ_s) for registrations, and 216 degrees of freedom.

- **Model 2**: the model 1 plus all the two-factor interaction parameters among registrations, that is

$$\ln m_{i_1 i_2 i_3 i_4 i_5} = \lambda + \lambda_j + \sum_{s=1}^{5} i_s \lambda_s + \sum_{s=1}^{4} \sum_{c=2}^{5} i_s i_c \lambda_{sc} \qquad j = 1988, \ldots, 1997$$

  We denote this model as 2: 1+(R1R2+...R4R5), which has 26 parameters and 203 degrees of freedom.

- **Model 3**: the model 1 plus the first-order heterogeneity term, that is

$$\ln m_{i_1 i_2 i_3 i_4 i_5} = \lambda + \lambda_j + \sum_{s=1}^{5} i_s \lambda_s + H1 \qquad j = 1988, \ldots, 1997$$

  where H1 is computed taking all the two-factor interaction parameters among registrations to be equal, so that the resulting model has 17 parameters and 212 degrees of freedom. We denote this model as 3: 1+H1.

- **Model 4**: the first of the two models presented in the preceding section, with the two latent variables θ_1 =R1+R2 and θ_2 =R3+R4+R5. We denote this model as 4: $1+\theta_1+\theta_2$.

- **Model 5**: the second of the two models presented in the preceding section, with the two latent variables θ_3 =R1+R2+R4 and θ_4 =R3+R4+R5. We denote this model as 5: $1+\theta_3+\theta_4$.

Table 5.4 summarize the results of these models fitted to the data. In Table 5.4(a) for each model is reported the number of parameters, the degrees of freedom, the deviance, the value of AIC, the value of BIC and the estimation of the total population size $\hat{N}$, while Table 5.4(b) presents the yearly estimates $\hat{N}_j, j = 1988, \ldots, 1998$ for each model.

Table 5.4: Selected models

(a) Selected models with deviance, AIC and BIC

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

(b) Selected models with yearly estimates

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 $\theta_4 = R3 + R4 + R5$

Figure 5.3 reports the plot of the yearly estimates for each model.

Note that the model with only the main-effect parameters does not fit well the data, as it has a high deviance. The model with the first-order heterogeneity parameter improves the fit, while adding all the two-factor interaction parameters to Model 1 results in a smaller deviance, even if the number of parameters is higher (and this fact results in a higher value of

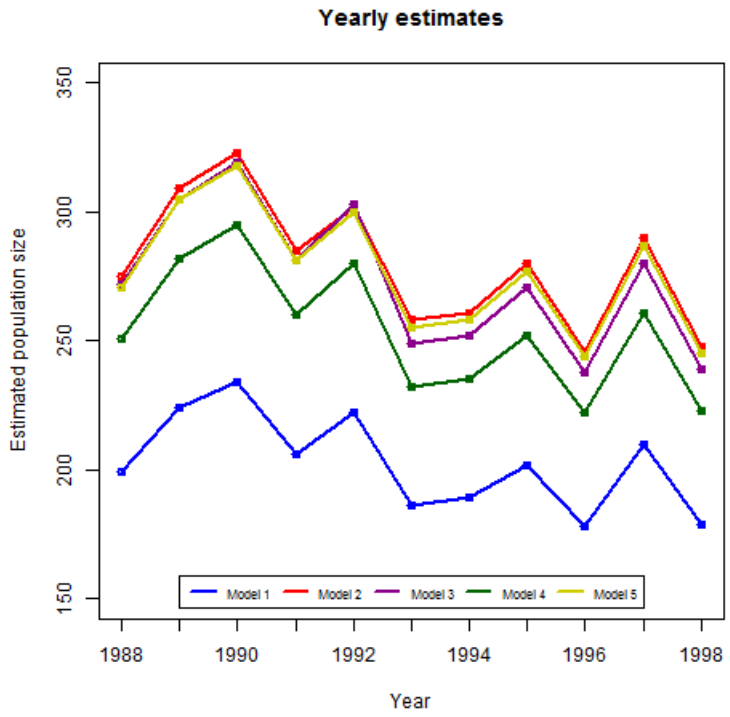

Figure 5.3: Yearly estimates for the five models

AIC and BIC).

Both of the multidimensional Rasch models fit well the data and the Model 5, with the registration R4 in common between the two latent variables is the best model, since it has the smallest value of AIC and BIC; thus, it is the selected model.

Table 5.5 reports the estimation of parameters for the selected model and the corresponding standard error.

For Model 5, the formulae (4.38)-(4.39) to obtain the estimation of parameters of the standard log-linear model, take the form:

$$\lambda_s = \delta_s + \frac{1}{2}\sum_{r=3}^{4} u_{sr}\gamma_{rr} + u_{s3}u_{s4}\gamma_{34}$$

$$\lambda_{sc} = \sum_{r=3}^{4} u_{sr}u_{cr}\gamma_{rr} + \left(u_{s3}u_{c4} + u_{s4}u_{c3}\right)\gamma_{34}$$

Table 5.5: Estimation of parameters for Model 5

Parameter	Estimate	Std. Error
δ	4.513951	0.142557
δ_{1988}	0.101292	0.116082
δ_{1989}	0.218309	0.112754
δ_{1990}	0.260357	0.111628
δ_{1991}	0.135194	0.115088
δ_{1992}	0.201906	0.111082
δ_{1993}	0.038221	0.112887
δ_{1994}	0.050644	0.112545
δ_{1995}	0.122103	0.110637
δ_{1996}	-0.00651	0.114147
δ_{1997}	0.156004	0.109768
δ_1	-2.20858	0.14922
δ_2	-1.04768	0.142911
δ_3	-3.25652	0.124767
δ_4	-2.9981	0.176131
δ_5	-4.16525	0.145811
γ_{33}	0.618927	0.082545
γ_{44}	1.108461	0.087735
γ_{34}	0.219176	0.053513

Applying these formulae, we obtain the following expressions for the parameters of the log-linear model:

$$\lambda_1 = \delta_1 + \frac{1}{2}\gamma_{33} \qquad \lambda_2 = \delta_2 + \frac{1}{2}\gamma_{33} \qquad \lambda_3 = \delta_3 + \frac{1}{2}\gamma_{44}$$

$$\lambda_4 = \delta_4 + \frac{1}{2}(\gamma_{33} + \gamma_{44}) + \gamma_{34} \qquad \lambda_5 = \delta_5 + \frac{1}{2}\gamma_{44} \qquad \lambda_{12} = \gamma_{33}$$

$$\lambda_{13} = \gamma_{34} \qquad \lambda_{14} = \gamma_{33} + \gamma_{34} \qquad \lambda_{15} = \gamma_{34}$$

$$\lambda_{23} = \gamma_{34} \qquad \lambda_{24} = \gamma_{33} + \gamma_{34} \qquad \lambda_{25} = \gamma_{34}$$

$$\lambda_{34} = \gamma_{44} + \gamma_{34} \qquad \lambda_{35} = \gamma_{44} \qquad \lambda_{45} = \gamma_{44} + \gamma_{34}$$

Thus, the main-effect parameters are equal to the main-effect parameters for the Model 5 plus half of the variance (given the total scores) of the latent variable to which the registration belongs, except for the registration R4, for which it is equal to the main-effect parameter δ_4 plus half of the variance of both the latent variables plus the covariance between θ_3 and θ_4, given the total scores. Concerning the two-factor interaction parameters, for those involving registrations which are indicator of different latent variables (that

Table 5.6: Estimation of parameters of log-linear model from Model 5

Parameter	Estimate	Std. Error
λ_1	-1.89911	0.154823
λ_2	-0.73821	0.148751
λ_3	-2.70229	0.132254
λ_4	-1.91523	0.193684
λ_5	-3.61102	0.152267
λ_{12}	0.618927	0.082545
λ_{13}	0.219176	0.053513
λ_{14}	0.838102	0.098373
λ_{15}	0.219176	0.053513
λ_{23}	0.219176	0.053513
λ_{24}	0.838102	0.098373
λ_{25}	0.219176	0.053513
λ_{34}	1.327637	0.101656
λ_{35}	1.108461	0.087735
λ_{45}	1.327637	0.101656

are $\lambda_{13}, \lambda_{15}, \lambda_{23}, \lambda_{25}$) are equal to the covariance (γ_{34}) conditional to the total scores. The two-factor interaction parameters which involve registrations belonging to the same latent variable (except those involving R4) are equal to the variance (given the total scores) of the corresponding latent variable, while other two-factor interaction parameters ($\lambda_{14}, \lambda_{24}, \lambda_{34}$ and λ_{45}) are equal to the covariance (given the total scores) plus the variance (given the total scores) of the latent variable for which the other registration is assumed to be indicator. The resulting parameters are reported in Table 5.6.

Note that the two-factor interaction parameters between registrations belonging to different latent variables (except those involving R4) are equal, as well as the two-factor interaction parameters between registration R4 and registrations belonging to θ_3 and registration R4 and registrations belonging to θ_4.

We used the parametric bootstrap with 500 replications and the percentile method in order to compute the confidence intervals for the population size estimates. Table 5.7 summarises the 95% confidence intervals for each of the five models.

Table 5.7: 95% Confidence intervals

Model	Design matrix	$\hat{N}$	95 per cent C.I.
1	R1+R2+R3+R4+R5+Y_{cat}	2229	[2164, 2297]
2	1+(R1R2+...+R4R5)	3077	[2724, 3571]
3	1+H1	3009	[2737, 3345]
4	1+$\theta_1 + \theta_2$	2793	[2559, 3104]
5	1+$\theta_3 + \theta_4$	3041	[2755, 3409]

[§]H1 is the first-order heterogeneity term

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

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

In addition, we compute the confidence intervals for the yearly estimates for Models 2 and 5. In this case, confidence intervals for the yearly estimates for Model 5 are always smaller than those of the Model 2. Results are shown in Table 5.8.

Table 5.8: 95% Confidence intervals by year

		Model 2		Model 5	
Year	Observed	$\hat{N}$	95 per cent C.I.	$\hat{N}$	95 per cent C.I.
1988	145	275	[225, 333]	271	[226, 328]
1989	163	309	[256, 385]	305	[256, 367]
1990	170	323	[272, 395]	318	[268, 382]
1991	150	285	[234, 360]	281	[235, 336]
1992	172	302	[251, 367]	300	[254, 357]
1993	160	258	[211, 311]	255	[213, 303]
1994	162	261	[216, 325]	258	[215, 305]
1995	174	280	[233, 342]	277	[235, 327]
1996	153	246	[204, 308]	244	[203, 286]
1997	180	290	[243, 355]	287	[241, 343]
1998	154	248	[200, 308]	245	[205, 297]

Chapter 6

Discussion

In the present work we proposed the use of the multidimensional Rasch model in the capture-recapture framework.

Throughout the thesis, we focused our attention on closed populations, for which we assumed that there are no births, deaths, immigrations or emigrations during the period of the study. As consequence of these assumptions, individuals in all the registrations may be perfectly linked, and if an individual is not in a registrations it is because he is simply not observed, but he must have been present in the population.

In this context, a problem widely discussed in literature concerns the way to model possible dependence among registrations. Dependence may be due to two different sources: "local dependence" (that is, the inclusion of an individual in a registration has a direct causal effect on the inclusion in other registrations), and "heterogeneity" among individuals (that is, registrations may become dependent because of the heterogeneity of inclusion probabilities among individuals).

To account for dependence among registrations log-linear models were successfully proposed; here dependence among registrations may be modelled by adding the corresponding two-factor interaction or higher-order interaction terms to the model. On the other hand, to model dependence caused by heterogeneity among individuals the dichotomous Rasch model was proposed and dependence due to heterogeneity can be modelled by adding the first-order heterogeneity or the higher-order heterogeneity parameters to the model. If extra dependence among registrations occurs, then it is possible to include the two-factor interaction or higher-order interaction terms to the

model.

To deal with this situation, the alternative approach that we proposed in this work is to use the multidimensional Rasch model. It is an extension of the dichotomous Rasch model which allows the presence of more than one latent variable underlying the performance of a test.

In the capture-recapture context, we assumed that registrations may be divided into two or more subgroups (not necessarily disjoint) which constitute the latent variables accounting for correlations among registrations. As consequence, the random variables denoting the presence or absence of an individual into a registration are assumed to be conditionally independent, given the latent variable.

In addition, we assumed that the posterior distribution of the latent variables follows a multivariate normal distribution (and this is equal to assume that the population of individuals not observed in any registration follows a normal distribution).

Under these assumptions, we applied the extension of the Dutch Identity proposed by Hessen in psychometric context to capture-recapture framework and we showed how it is possible to re-express the probability of a generic capture-profile in terms of the log-linear multidimensional Rasch model.

We also discussed the proposed model in the case in which a stratifying variable is available and in the particular situation in which the assumption of measurement invariance can be made.

Then, we presented a re-parametrization of the proposed model that allows for a connection between the multidimensional Rasch model and the standard log-linear model. Applying these formulae it is possible to compute the parameters of the standard log-linear model, starting from those of the multidimensional Rasch model.

In the last Chapter of the present work, we applied the methodology we proposed to a dataset on Neural Tube Defcts (NTD's) in the Netherlands from 1988 through 1998. The target population included children born with a NTD's and data referred to five registrations. Since these five registrations did not refer to the same years (as they covered different but overlapping periods of time), we used the E-M algorithm proposed by Zwane et. al to estimate the missing entry in the dataset.

The scope of the application was to estimate the total amount of children born with a NTD's during the period of the study.

The results showed that the selected model for inference is one of the log-linear multidimensional Rasch model obtained 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.

Finally, starting from the estimates of parameters of the selected model, we used the connecting formulae to compute the estimates of the corresponding traditional log-linear model.

Future research should be focused on the study of the multidimensional Rasch model in capture-recapture in a more general situation in which more than one latent variable is available.

Furthermore, it would be interesting to study the multidimensional Rasch model under different assumptions for the posterior distribution of the latent variables.

Appendix

```
EM_alg<-function(model){

dati<-read.csv("C:/Users/Mark/Desktop/Dataset.csv",header = TRUE, sep = ";")

NewData<-list(casenumber=dati$"Case",R1=as.numeric(dati$"R1"),R2=as.numeric(dati$"R2"),
R3=as.numeric(dati$"R3"),R4=as.numeric(dati$"R4"),R5=as.numeric(dati$"R5"),year=as.numeric(dati$"Year"))

intable<-table(NewData$R1,NewData$R2,NewData$R3,NewData$R4,NewData$R5,as.numeric(NewData$year))

incomtable<-table(NewData$R1,NewData$R2,NewData$R3,NewData$R4,NewData$R5,as.numeric(NewData$year))

options(contrasts=c("contr.treatment","contr.poly"))

options(digits=10)


####ASSEGNO I VALORI INIZIALI PER L'ALGORITMO####

#####1988-1991###
for (i in 1:4){
incomtable[1,2,1,1,1,i]<-incomtable[1,2,1,2,1,i]<-incomtable[1,2,2,1,1,i]<-
incomtable[1,2,2,2,1,i]<-intable[1,2,1,1,1,i]*0.25

incomtable[1,2,1,1,2,i]<-incomtable[1,2,1,2,2,i]<-incomtable[1,2,2,1,2,i]<-
incomtable[1,2,2,2,2,i]<-intable[1,2,1,1,2,i]*0.25

incomtable[2,1,1,1,1,i]<-incomtable[2,1,1,2,1,i]<-incomtable[2,1,2,1,1,i]<-
incomtable[2,1,2,2,1,i]<-intable[2,1,1,1,1,i]*0.25

incomtable[2,1,1,1,2,i]<-incomtable[2,1,1,2,2,i]<-incomtable[2,1,2,1,2,i]<-
incomtable[2,1,2,2,2,i]<-intable[2,1,1,1,2,i]*0.25

incomtable[2,2,1,1,1,i]<-incomtable[2,2,1,2,1,i]<-incomtable[2,2,2,1,1,i]<-
incomtable[2,2,2,2,1,i]<-intable[2,2,1,1,1,i]*0.25

incomtable[2,2,1,1,2,i]<-incomtable[2,2,1,2,2,i]<-incomtable[2,2,2,1,2,i]<-
incomtable[2,2,2,2,2,i]<-intable[2,2,1,1,2,i]*0.25
```

```
incomtable[1,1,1,1,2,i]<-incomtable[1,1,1,2,2,i]<-incomtable[1,1,2,1,2,i]<-
incomtable[1,1,2,2,2,i]<-intable[1,1,1,1,2,i]*0.25
}

###1992###

incomtable[2,1,1,1,1,5]<-incomtable[2,1,1,2,1,5]<-intable[2,1,1,1,1,5]*0.5
incomtable[1,2,1,1,1,5]<-incomtable[1,2,1,2,1,5]<-intable[1,2,1,1,1,5]*0.5
incomtable[1,1,2,1,1,5]<-incomtable[1,1,2,2,1,5]<-intable[1,1,2,1,1,5]*0.5
incomtable[1,1,1,1,2,5]<-incomtable[1,1,1,2,2,5]<-intable[1,1,1,1,2,5]*0.5
incomtable[2,2,1,1,1,5]<-incomtable[2,2,1,2,1,5]<-intable[2,2,1,1,1,5]*0.5
incomtable[2,1,2,1,1,5]<-incomtable[2,1,2,2,1,5]<-intable[2,1,2,1,1,5]*0.5
incomtable[2,1,1,1,2,5]<-incomtable[2,1,1,2,2,5]<-intable[2,1,1,1,2,5]*0.5
incomtable[1,2,2,1,1,5]<-incomtable[1,2,2,2,1,5]<-intable[1,2,2,1,1,5]*0.5
incomtable[1,2,1,1,2,5]<-incomtable[1,2,1,2,2,5]<-intable[1,2,1,1,2,5]*0.5
incomtable[1,1,2,1,2,5]<-incomtable[1,1,2,2,2,5]<-intable[1,1,2,1,2,5]*0.5
incomtable[2,2,2,1,1,5]<-incomtable[2,2,2,2,1,5]<-intable[2,2,2,1,1,5]*0.5
incomtable[2,2,1,1,2,5]<-incomtable[2,2,1,2,2,5]<-intable[2,2,1,1,2,5]*0.5
incomtable[1,2,2,1,2,5]<-incomtable[1,2,2,2,2,5]<-intable[1,2,2,1,2,5]*0.5
incomtable[2,1,2,1,2,5]<-incomtable[2,1,2,2,2,5]<-intable[2,1,2,1,2,5]*0.5
incomtable[2,2,2,1,2,5]<-incomtable[2,2,2,2,2,5]<-intable[2,2,2,1,2,5]*0.5


#####Algoritmo EM############

dati<-as.data.frame(incomtable)
H1<-(as.numeric(dati[,"Var1"])-1)*(as.numeric(dati[,"Var2"])-1)+
(as.numeric(dati[,"Var1"])-1)*(as.numeric(dati[,"Var3"])-1)+
(as.numeric(dati[,"Var1"])-1)*(as.numeric(dati[,"Var4"])-1)+
(as.numeric(dati[,"Var1"])-1)*(as.numeric(dati[,"Var5"])-1)+
(as.numeric(dati[,"Var2"])-1)*(as.numeric(dati[,"Var3"])-1)+
(as.numeric(dati[,"Var2"])-1)*(as.numeric(dati[,"Var4"])-1)+
(as.numeric(dati[,"Var2"])-1)*(as.numeric(dati[,"Var5"])-1)+
(as.numeric(dati[,"Var3"])-1)*(as.numeric(dati[,"Var4"])-1)+
(as.numeric(dati[,"Var3"])-1)*(as.numeric(dati[,"Var5"])-1)+
(as.numeric(dati[,"Var4"])-1)*(as.numeric(dati[,"Var5"])-1)

H2<-(as.numeric(dati[,"Var1"])-1)*(as.numeric(dati[,"Var2"])-1)*(as.numeric(dati[,"Var3"])-1)+
(as.numeric(dati[,"Var1"])-1)*(as.numeric(dati[,"Var2"])-1)*(as.numeric(dati[,"Var4"])-1)+
(as.numeric(dati[,"Var1"])-1)*(as.numeric(dati[,"Var2"])-1)*(as.numeric(dati[,"Var5"])-1)+
(as.numeric(dati[,"Var1"])-1)*(as.numeric(dati[,"Var3"])-1)*(as.numeric(dati[,"Var4"])-1)+
(as.numeric(dati[,"Var1"])-1)*(as.numeric(dati[,"Var3"])-1)*(as.numeric(dati[,"Var5"])-1)+
(as.numeric(dati[,"Var1"])-1)*(as.numeric(dati[,"Var4"])-1)*(as.numeric(dati[,"Var5"])-1)+
(as.numeric(dati[,"Var2"])-1)*(as.numeric(dati[,"Var3"])-1)*(as.numeric(dati[,"Var4"])-1)+
(as.numeric(dati[,"Var2"])-1)*(as.numeric(dati[,"Var3"])-1)*(as.numeric(dati[,"Var5"])-1)+
(as.numeric(dati[,"Var2"])-1)*(as.numeric(dati[,"Var4"])-1)*(as.numeric(dati[,"Var5"])-1)+
(as.numeric(dati[,"Var3"])-1)*(as.numeric(dati[,"Var4"])-1)*(as.numeric(dati[,"Var5"])-1)

t1<-(as.numeric(dati[,"Var1"])-1)+(as.numeric(dati[,"Var2"])-1)
t2<-(as.numeric(dati[,"Var3"])-1)+(as.numeric(dati[,"Var4"])-1)+(as.numeric(dati[,"Var5"])-1)
t3<-(as.numeric(dati[,"Var1"])-1)+(as.numeric(dati[,"Var2"])-1)+(as.numeric(dati[,"Var4"])-1)
```

```
t4<-(as.numeric(dati[,"Var3"])-1)+(as.numeric(dati[,"Var4"])-1)+(as.numeric(dati[,"Var5"])-1)

t1_2<-0.5*t1*t1
t2_2<-0.5*t2*t2
t12<-t1*t2
t3_2<-0.5*t3*t3
t4_2<-0.5*t4*t4
t34<-t3*t4

k<-32
iter<-0
dev<-0
devnew<-0
while(iter<10000)
{

##M-STEP####

Etable<-as.data.frame(incomtable)
Year<-relevel(Etable$Var6, ref="1998")
Etable<-cbind(Etable,H1,H2,t3,t4,t1_2,t2_2,t12,t3_2,t4_2,t34)


fit <-glm(model,family=poisson,data=Etable,
weights=c(0,1,1,1,0,1,1,1,0,1,1,1,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
0,1,1,1,0,1,1,1,0,1,1,1,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
0,1,1,1,0,1,1,1,0,1,1,1,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
0,1,1,1,0,1,1,1,0,1,1,1,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
0,1,1,1,1,1,1,1,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1))
estt<-fitted.values(fit)
est<-array(estt, dim=c(2,2,2,2,2,11))

    devnew<-deviance(fit)
   if(abs(devnew-dev) <=0.00000000000000000001) break

#####E-STEP#####
#############ANNO 1992##############

incomtable[2,1,1,1,1,5]<-(sum(est[2,1,1,1,1,6:11])/sum(est[2,1,1,,1,6:11]))*intable[2,1,1,1,1,5]

incomtable[2,1,1,2,1,5]<-(sum(est[2,1,1,2,1,6:11])/sum(est[2,1,1,,1,6:11]))*intable[2,1,1,1,1,5]

incomtable[1,2,1,1,1,5]<-(sum(est[1,2,1,1,1,6:11])/sum(est[1,2,1,,1,6:11]))*intable[1,2,1,1,1,5]

incomtable[1,2,1,2,1,5]<-(sum(est[1,2,1,2,1,6:11])/sum(est[1,2,1,,1,6:11]))*intable[1,2,1,1,1,5]
```

```
incomtable[1,1,2,1,1,5]<-(sum(est[1,1,2,1,1,6:11])/sum(est[1,1,2,,1,6:11]))*intable[1,1,2,1,1,5]

incomtable[1,1,2,2,1,5]<-(sum(est[1,1,2,2,1,6:11])/sum(est[1,1,2,,1,6:11]))*intable[1,1,2,1,1,5]

incomtable[1,1,1,1,2,5]<-(sum(est[1,1,1,1,2,6:11])/sum(est[1,1,1,,2,6:11]))*intable[1,1,1,1,2,5]

incomtable[1,1,1,2,2,5]<-(sum(est[1,1,1,2,2,6:11])/sum(est[1,1,1,,2,6:11]))*intable[1,1,1,1,2,5]

incomtable[2,2,1,1,1,5]<-(sum(est[2,2,1,1,1,6:11])/sum(est[2,2,1,,1,6:11]))*intable[2,2,1,1,1,5]

incomtable[2,2,1,2,1,5]<-(sum(est[2,2,1,2,1,6:11])/sum(est[2,2,1,,1,6:11]))*intable[2,2,1,1,1,5]

incomtable[2,1,2,1,1,5]<-(sum(est[2,1,2,1,1,6:11])/sum(est[2,1,2,,1,6:11]))*intable[2,1,2,1,1,5]

incomtable[2,1,2,2,1,5]<-(sum(est[2,1,2,2,1,6:11])/sum(est[2,1,2,,1,6:11]))*intable[2,1,2,1,1,5]

incomtable[2,1,1,1,2,5]<-(sum(est[2,1,1,1,2,6:11])/sum(est[2,1,1,,2,6:11]))*intable[2,1,1,1,2,5]

incomtable[2,1,1,2,2,5]<-(sum(est[2,1,1,2,2,6:11])/sum(est[2,1,1,,2,6:11]))*intable[2,1,1,1,2,5]

incomtable[1,2,2,1,1,5]<-(sum(est[1,2,2,1,1,6:11])/sum(est[1,2,2,,1,6:11]))*intable[1,2,2,1,1,5]

incomtable[1,2,2,2,1,5]<-(sum(est[1,2,2,2,1,6:11])/sum(est[1,2,2,,1,6:11]))*intable[1,2,2,1,1,5]

incomtable[1,2,1,1,2,5]<-(sum(est[1,2,1,1,2,6:11])/sum(est[1,2,1,,2,6:11]))*intable[1,2,1,1,2,5]

incomtable[1,2,1,2,2,5]<-(sum(est[1,2,1,2,2,6:11])/sum(est[1,2,1,,2,6:11]))*intable[1,2,1,1,2,5]

incomtable[1,1,2,1,2,5]<-(sum(est[1,1,2,1,2,6:11])/sum(est[1,1,2,,2,6:11]))*intable[1,1,2,1,2,5]

incomtable[1,1,2,2,2,5]<-(sum(est[1,1,2,2,2,6:11])/sum(est[1,1,2,,2,6:11]))*intable[1,1,2,1,2,5]

incomtable[2,2,2,1,1,5]<-(sum(est[2,2,2,1,1,6:11])/sum(est[2,2,2,,1,6:11]))*intable[2,2,2,1,1,5]

incomtable[2,2,2,2,1,5]<-(sum(est[2,2,2,2,1,6:11])/sum(est[2,2,2,,1,6:11]))*intable[2,2,2,1,1,5]

incomtable[2,2,1,1,2,5]<-(sum(est[2,2,1,1,2,6:11])/sum(est[2,2,1,,2,6:11]))*intable[2,2,1,1,2,5]

incomtable[2,2,1,2,2,5]<-(sum(est[2,2,1,2,2,6:11])/sum(est[2,2,1,,2,6:11]))*intable[2,2,1,1,2,5]

incomtable[1,2,2,1,2,5]<-(sum(est[1,2,2,1,2,6:11])/sum(est[1,2,2,,2,6:11]))*intable[1,2,2,1,2,5]

incomtable[1,2,2,2,2,5]<-(sum(est[1,2,2,2,2,6:11])/sum(est[1,2,2,,2,6:11]))*intable[1,2,2,1,2,5]

incomtable[2,1,2,1,2,5]<-(sum(est[2,1,2,1,2,6:11])/sum(est[2,1,2,,2,6:11]))*intable[2,1,2,1,2,5]

incomtable[2,1,2,2,2,5]<-(sum(est[2,1,2,2,2,6:11])/sum(est[2,1,2,,2,6:11]))*intable[2,1,2,1,2,5]

incomtable[2,2,2,1,2,5]<-(sum(est[2,2,2,1,2,6:11])/sum(est[2,2,2,,2,6:11]))*intable[2,2,2,1,2,5]
```

```
incomtable[2,2,2,2,2,5]<-(sum(est[2,2,2,2,2,6:11])/sum(est[2,2,2,,2,6:11]))*intable[2,2,2,1,2,5]


##############ANNI 1988-1991############
for(i in 1:4){
incomtable[1,2,1,1,1,i]<-((incomtable[1,2,1,1,1,5]+sum(est[1,2,1,1,1,6:11]))/
 (sum(incomtable[1,2,,,1,5])+sum(est[1,2,,,1,6:11])))*intable[1,2,1,1,1,i]
incomtable[1,2,1,2,1,i]<-((incomtable[1,2,1,2,1,5]+sum(est[1,2,1,2,1,6:11]))/
 (sum(incomtable[1,2,,,1,5])+sum(est[1,2,,,1,6:11])))*intable[1,2,1,1,1,i]
incomtable[1,2,2,1,1,i]<-((incomtable[1,2,2,1,1,5]+sum(est[1,2,2,1,1,6:11]))/
 (sum(incomtable[1,2,,,1,5])+sum(est[1,2,,,1,6:11])))*intable[1,2,1,1,1,i]
incomtable[1,2,2,2,1,i]<-((incomtable[1,2,2,2,1,5]+sum(est[1,2,2,2,1,6:11]))/
 (sum(incomtable[1,2,,,1,5])+sum(est[1,2,,,1,6:11])))*intable[1,2,1,1,1,i]
incomtable[1,2,1,1,2,i]<-((incomtable[1,2,1,1,2,5]+sum(est[1,2,1,1,2,6:11]))/
 (sum(incomtable[1,2,,,2,5])+sum(est[1,2,,,2,6:11])))*intable[1,2,1,1,2,i]
incomtable[1,2,1,2,2,i]<-((incomtable[1,2,1,2,2,5]+sum(est[1,2,1,2,2,6:11]))/
   (sum(incomtable[1,2,,,2,5])+sum(est[1,2,,,2,6:11])))*intable[1,2,1,1,2,i]
incomtable[1,2,2,1,2,i]<-((incomtable[1,2,2,1,2,5]+sum(est[1,2,2,1,2,6:11]))/
 (sum(incomtable[1,2,,,2,5])+sum(est[1,2,,,2,6:11])))*intable[1,2,1,1,2,i]
incomtable[1,2,2,2,2,i]<-((incomtable[1,2,2,2,2,5]+sum(est[1,2,2,2,2,6:11]))/
 (sum(incomtable[1,2,,,2,5])+sum(est[1,2,,,2,6:11])))*intable[1,2,1,1,2,i]
incomtable[2,1,1,1,1,i]<-((incomtable[2,1,1,1,1,5]+sum(est[2,1,1,1,1,6:11]))/
   (sum(incomtable[2,1,,,1,5])+sum(est[2,1,,,1,6:11])))*intable[2,1,1,1,1,i]
incomtable[2,1,1,2,1,i]<-((incomtable[2,1,1,2,1,5]+sum(est[2,1,1,2,1,6:11]))/
 (sum(incomtable[2,1,,,1,5])+sum(est[2,1,,,1,6:11])))*intable[2,1,1,1,1,i]
incomtable[2,1,2,1,1,i]<-((incomtable[2,1,2,1,1,5]+sum(est[2,1,2,1,1,6:11]))/
 (sum(incomtable[2,1,,,1,5])+sum(est[2,1,,,1,6:11])))*intable[2,1,1,1,1,i]
incomtable[2,1,2,2,1,i]<-((incomtable[2,1,2,2,1,5]+sum(est[2,1,2,2,1,6:11]))/
 (sum(incomtable[2,1,,,1,5])+sum(est[2,1,,,1,6:11])))*intable[2,1,1,1,1,i]
incomtable[2,1,1,1,2,i]<-((incomtable[2,1,1,1,2,5]+sum(est[2,1,1,1,2,6:11]))/
 (sum(incomtable[2,1,,,2,5])+sum(est[2,1,,,2,6:11])))*intable[2,1,1,1,2,i]
incomtable[2,1,1,2,2,i]<-((incomtable[2,1,1,2,2,5]+sum(est[2,1,1,2,2,6:11]))/
 (sum(incomtable[2,1,,,2,5])+sum(est[2,1,,,2,6:11])))*intable[2,1,1,1,2,i]
incomtable[2,1,2,1,2,i]<-((incomtable[2,1,2,1,2,5]+sum(est[2,1,2,1,2,6:11]))/
 (sum(incomtable[2,1,,,2,5])+sum(est[2,1,,,2,6:11])))*intable[2,1,1,1,2,i]
incomtable[2,1,2,2,2,i]<-((incomtable[2,1,2,2,2,5]+sum(est[2,1,2,2,2,6:11]))/
 (sum(incomtable[2,1,,,2,5])+sum(est[2,1,,,2,6:11])))*intable[2,1,1,1,2,i]
incomtable[2,2,1,1,1,i]<-((incomtable[2,2,1,1,1,5]+sum(est[2,2,1,1,1,6:11]))/
 (sum(incomtable[2,2,,,1,5])+sum(est[2,2,,,1,6:11])))*intable[2,2,1,1,1,i]
incomtable[2,2,1,2,1,i]<-((incomtable[2,2,1,2,1,5]+sum(est[2,2,1,2,1,6:11]))/
 (sum(incomtable[2,2,,,1,5])+sum(est[2,2,,,1,6:11])))*intable[2,2,1,1,1,i]
incomtable[2,2,2,1,1,i]<-((incomtable[2,2,2,1,1,5]+sum(est[2,2,2,1,1,6:11]))/
 (sum(incomtable[2,2,,,1,5])+sum(est[2,2,,,1,6:11])))*intable[2,2,1,1,1,i]
incomtable[2,2,2,2,1,i]<-((incomtable[2,2,2,2,1,5]+sum(est[2,2,2,2,1,6:11]))/
 (sum(incomtable[2,2,,,1,5])+sum(est[2,2,,,1,6:11])))*intable[2,2,1,1,1,i]
incomtable[2,2,1,1,2,i]<-((incomtable[2,2,1,1,2,5]+sum(est[2,2,1,1,2,6:11]))/
 (sum(incomtable[2,2,,,2,5])+sum(est[2,2,,,2,6:11])))*intable[2,2,1,1,2,i]
incomtable[2,2,1,2,2,i]<-((incomtable[2,2,1,2,2,5]+sum(est[2,2,1,2,2,6:11]))/
 (sum(incomtable[2,2,,,2,5])+sum(est[2,2,,,2,6:11])))*intable[2,2,1,1,2,i]
incomtable[2,2,2,1,2,i]<-((incomtable[2,2,2,1,2,5]+sum(est[2,2,2,1,2,6:11]))/
 (sum(incomtable[2,2,,,2,5])+sum(est[2,2,,,2,6:11])))*intable[2,2,1,1,2,i]
```

```
incomtable[2,2,2,2,2,i]<-((incomtable[2,2,2,2,2,5]+sum(est[2,2,2,2,2,6:11]))/
 (sum(incomtable[2,2,,,2,5])+sum(est[2,2,,,2,6:11])))*intable[2,2,1,1,2,i]
incomtable[1,1,1,1,2,i]<-((incomtable[1,1,1,1,2,5]+sum(est[1,1,1,1,2,6:11]))/
 (sum(incomtable[1,1,,,2,5])+sum(est[1,1,,,2,6:11])))*intable[1,1,1,1,2,i]
incomtable[1,1,1,2,2,i]<-((incomtable[1,1,1,2,2,5]+sum(est[1,1,1,2,2,6:11]))/
 (sum(incomtable[1,1,,,2,5])+sum(est[1,1,,,2,6:11])))*intable[1,1,1,1,2,i]
incomtable[1,1,2,1,2,i]<-((incomtable[1,1,2,1,2,5]+sum(est[1,1,2,1,2,6:11]))/
 (sum(incomtable[1,1,,,2,5])+sum(est[1,1,,,2,6:11])))*intable[1,1,1,1,2,i]
incomtable[1,1,2,2,2,i]<-((incomtable[1,1,2,2,2,5]+sum(est[1,1,2,2,2,6:11]))/
 (sum(incomtable[1,1,,,2,5])+sum(est[1,1,,,2,6:11])))*intable[1,1,1,1,2,i]

}
dev <- devnew
    iter <- iter + 1
}

dev<-round(
 sum(apply(intable[,,,,,1:4],c(1,2,5,6),sum)*log((apply(intable[,,,,,1:4],c(1,2,5,6),sum)/
 apply(est[,,,,,1:4],c(1,2,5,6),sum))) ,na.rm=T)+
 sum(apply(intable[,,,,,5],c(1,2,3,5),sum)*log((apply(intable[,,,,,5],c(1,2,3,5),sum)/
 apply(est[,,,,,5],c(1,2,3,5),sum))) ,na.rm=T)+
 sum( intable[,,,,,6:11] * log( (intable[,,,,,6:11]/est[,,,,,6:11]) ) ,na.rm=T),
 digits=3)

summary(fit)
par<-dim(summary(fit)$coefficients)

output<-cat("N_est:", round(sum(fitted.values(fit))), "\n",
"N_88:", (sum(est[,,,,,1])), "\n",
"N_89:", (sum(est[,,,,,2])), "\n",
"N_90:", (sum(est[,,,,,3])), "\n",
"N_91:", (sum(est[,,,,,4])), "\n",
"N_92:", (sum(est[,,,,,5])), "\n",
"N_93:", (sum(est[,,,,,6])), "\n",
"N_94:", (sum(est[,,,,,7])), "\n",
"N_95:", (sum(est[,,,,,8])), "\n",
"N_96:", (sum(est[,,,,,9])), "\n",
"N_97:", (sum(est[,,,,,10])), "\n",
"N_98:", (sum(est[,,,,,11])), "\n",
"Deviance:", dev, "\n")

return(list(model, output))
}
```

Bibliography

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

[2] F. Bartolucci and A. Forcina. "Analysis of Capture-Recapture Data with a Rasch-Type Model Allowing for Conditional Dependence and Multidimensionality". In: *Biometrics* 57 (2001), pp. 714–719.

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

[4] A. Chao, P.K. Tsay, S. Lin, W. Shau, and D. Chao. "Tutorial in biostatistics. The applications of capture-recapture models to epidemiological data". In: *Statistics in Medicine* 20 (2001), pp. 3123–3157.

[5] R.M. Cormack. "Loglinear model for capture-recapture". In: *Biometrics* 45 (1989), pp. 395–413.

[6] J.N. Darroch. "The multiple-recapture census I. Estimation of a closed population". In: *Biometrika* 45 (1958), pp. 343–59.

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

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin. "Maximum Likelihood from Incomplete Data via the EM Algorithm". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 39.1 (1977), pp. 1–38.

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

[10] S.C. Gerritse, P.G.M. Van der Heijden, and B.F.M. Bakker. "Robustness of population size size estimates against violation of the independence assumption". In: *59th ISI World Statistics Congress*, August 25-30, 2013. Hong Kong.

[11] D. Hessen. "Fitting and testing conditional multinormal partial credit models". In: *Psycometrika* (2012).

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

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

[14] H. Kelderman and C. P. Rijkes. "Loglinear multidimensional IRT models for polytomously scored items". In: *Psychometrika* 59.2 (1994), pp. 149–176.

[15] F.C. Lincoln. "Calculating Waterfowl Abundance on the Basis of Banding Returns". In: *United States Department of Agriculture Circular* 118 (1930), pp. 1–4.

[16] C.G.J. Petersen. "The Yearly Immigration of Young Plaice Into the Limfjord From the German Sea". In: *Report of the Danish Biological Station* 6 (1896), pp. 5–84.

[17] G. Rasch. *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danish Institute for Educational Research. (1960).

[18] M. D. Reckase. "The Difficulty of Test Items That Measure More than One Ability". In: *Applied Psychological Measurement* 9.4 (1985), pp. 401–412.

[19] L. Sanathanan. "Models and estimation methods in visual scanning experiments". In: *Technometrics* 14 (1972), pp. 813–829.

[20] C. Sekar and E.W. Deming. "On a method of estimating birth and death rates and extent of registrations". In: *Journal of the American Statistical Association* 44 (1949), pp. 101–115.

[21] S. Shapiro. "Estimating birth registration completeness". In: *Journal of the American Statistical Association* 45 (1949), pp. 261–264.

[22] P.G.M. Van der Heijden, J. Whittaker, M. Cruyff, B. Bakker, and R. Van der Vliet. "People born in the Middle East but residing in the Netherlands: Invariant Population size estimates and the role of active and passive covariates". In: *The Annals of Applied Statistics* 6.3 (2012), pp. 831–852.

[23] J. Wittes, T. Colton, and V.W. Sidel. "Capture-recapture models for assessing the completeness of case ascertainment using multiple information sources". In: *Journal of Chronic Diseases* 27 (1974), pp. 25–36.

[24] J. Wittes and V.W. Sidel. "A generalization of the simple capture-recapture model with applications to epidemiological research". In: *Journal of Chronic Diseases* 21 (1968), pp. 287–301.

[25] J.T. Wittes. "Application of a Multinomial capture-rcapture method to epidemiological data". In: *Journal od the American Statistical Association* 69 (1974), pp. 93–97.

[26] E.N. Zwane and P.G.M. Van der Heijden. "Analysing capture-recapture data when some variables of heterogeneous catchability are not collected or asked in all registrations". In: *Statistics in Medicine* 26 (2007), pp. 1069–1089.

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