MATH11188 Statistical Research Skills

Assignment 3: Scientific Report

Deadline: Friday 7th April 12.00 (midday)

Submission of assessments is via Learn

This assignment is separated into two distinct but related tasks. Task 1 focuses on producing a *group* scientific report relating to a statistical analysis; Task 2 relates to an associated *individual* (non-technical) executive summary to describe the main findings. The assignment contributes 40% of the final grade for the course (26% for Task 1; 14% for Task 2).

The scientific report and executive summary should be submitted in the form of pdf files (only pdfs will be marked!). You need to submit two separate submissions - one group report (to "Assignment 3: Group Report") and an individual report (to "Assignment 3: Individual Exec Summary").

The marking rubric for the assignment is available in Learn in the folder: "Assignment 3" with associated marking categories: Content = 40%; Structure = 20%; Executive Summary = 40%. (Content and Structure relate to Task 1; Executive Summary to Task 2).

The assignment focuses on the problem of multiple systems estimation (MSE).

Problem: Multiple Systems Estimation

Multiple systems estimation is concerned with estimating hidden or difficult to find populations. For example, the approach has been applied to the estimation of the number of people who inject drugs, modern day slaves, civilian casualties in wars and individuals with a given disease. Although the total population is difficult to observe, partial enumeration of the given population is often possible via different administrative data lists. Individuals are uniquely identifiable by each list using identifiers (such as initials, date of birth, medical number etc), allowing cross-classification of individuals across the lists. The data are then typically summarised by the number of individuals observed by each distinct combination of lists.

As an example, suppose that there are 3 data lists, labelled A, B and C. Each list has two levels corresponding to observing an individual (=1) or not observing an individual (=0). We let y_{ijk} denote the number of individuals observed by the given combination of lists, where $i = 0, 1; j = 0, 1; k = 0, 1$, corresponding to not being observed, or being observed by list A, B and C, respectively. The data are typically displayed as an incomplete contingency table:

	$A = 1$		$A = 0$	
	$B = 1$	$B = 0$	$B = 1$	$B = 0$
$C = 1$	y_{111}	y_{101}	y_{011}	y_{001}
$C = 0$	y_{110}	y_{100}	y_{010}	$y_{000} =?$

For example y_{100} corresponds to the number of individuals observed by list A but unobserved by lists B and C. Similarly, y_{111} denotes the number of individuals observed by all 3 lists. However, the number of individuals not observed by any list, denoted y_{000}, is unknown - as we do not observe these individuals by any of the lists! The total number of observed individuals is denoted by n (i.e. the number of individuals observed by at least 1 list), so that,

$$n = \sum_{ijk:ijk \neq (0,0,0)} y_{ijk}.$$

The true total population size is denoted by $N = \sum_{ijk} y_{ijk} = n + y_{000}$. The aim of multiple systems estimation is to fit a statistical model to the observed data, such that under certain assumptions (details omitted here), allows us to estimate the number of unobserved individuals, y_{000} and hence the total population size N (by combining the observed number with the observed number, n).

Statistical model

To model the data, we consider a generalised linear model (GLM), from the Poisson family with log-link function. The response variable corresponds to the number of individuals observed by each combination of sources; and (discrete) explanatory variables are the lists which have two levels (0 = unobserved; 1 = observed). In particular, we assume that the number of individuals observed by each combination of lists are (conditionally) independent, such that,

$$y_{ijk}|\mu_{ijk} \overset{ind}{\sim} Poisson(\mu_{ijk}),$$

where μ_{ijk} denotes the expected (or mean) number of individuals observed by the given combination of lists. We assume a log-link function such that the mean cell value is a linear function of the (categorical) explanatory variables, and expressed as,

$$\log \mu_{ijk} = \beta_0 + I(i=1)\beta_1 + I(j=1)\beta_2 + I(k=1)\beta_3,$$

where $I(\cdot)$ denotes the indicator function. The terms can be interpreted as follows:

β_0: intercept;

β_1: regression coefficient related to the propensity to be observed by list A;

β_2: regression coefficient related to the propensity to be observed by list B;

β_3: regression coefficient related to the propensity to be observed by list C.

Thus, for example, the expected value for the number of individuals observed by all lists is given by,

$$\log \mu_{111} = \beta_0 + \beta_1 + \beta_2 + \beta_3. \tag{1}$$

In vector notation, we can define the vector of expected values to be $\boldsymbol{\mu}$ and specify a log-link function such that,

$$\log \boldsymbol{\mu} = X\boldsymbol{\beta},$$

where X denotes the design matrix and $\boldsymbol{\beta}^T = (\beta_0, \beta_1, \beta_2, \beta_3)$ the vector of parameters. The design matrix links the expected values with the explanatory variables, and is defined via the indicator variables given in Equation (1).

Estimation of total population size

Applying the same modelling assumptions to the number of unobserved individuals, we have that,

$$y_{000}|\mu_{000} \sim Poisson(\mu_{000}),$$

where,

$$\log \mu_{000} = \beta_0.$$

Thus, the model can be fitted to the observed data and the parameters $\boldsymbol{\beta}$ estimated (either within a classical framework or a Bayesian framework). Using this fitted model we can obtain an estimate of (or predict) the number of unobserved individuals. For example, assuming a classical implementation, given the Poisson assumption, we have that the MLE for y_{000} denoted $\widehat{y}_{000}$ is equal to the the associated mean, i.e. if $\widehat{y}_{000} = \mu_{000}$. Using the invariance property of the MLE, we thus have,

$$\widehat{y}_{000} = \widehat{\mu}_{000} = \exp(\widehat{\beta}_0).$$

Finally, an estimate of the total population size is $\widehat{N} = \widehat{y}_{000} + n$.

Interactions

The above model assumes that the lists (i.e. explanatory variables) are independent of each other. For the application to multiple systems estimation this means that being observed by a given list does not change the probability of being observed by another list. However, in many situations this may not be case, and there may be list dependence, i.e. interactions between the explanatory variables. Only two-way interactions will be considered within this assignment, for example, an interaction between lists A and B, denoted $A \times B$. Interactions may be positive or negative. A positive interaction means that being observed by one list makes it more likely to be observed by the other list, and vice versa; a negative interaction means that being observed by one list makes it less likely to be observed by the other list, and vice versa. Note that the estimate of the total population size can vary substantially between different log-linear models, in terms of the interactions present in the model.

R code

The model being fitted to the data is a GLM, and so can be easily fitted using the `glm` function in `R`. The input to the `glm` function corresponds to the data and explanatory variables, with the distributional family and link function specified. For the 3 source example above, with lists A, B and C, example `R` commands for fitting the independent model (i.e. no interactions) is as follows:

```
data <- c(y100, y010, y110, y001, y101, y011, y111) # Read in the data
A <- c(1,0,1,0,1,0,1)         # Read in the values for A (i values) for given data
B <- c(0,1,1,0,0,1,1)         # Read in the values for B (j values) for given data
C <- c(0,0,0,1,1,1,1)         # Read in the values for C (k values) for given data

model <- glm(data˜A+B+C, family=poisson(link="log"))
```

Here we assume that the data are numerical values, with $y111 = y_{111}$ etc.

Interactions can be specified within the `glm` function. For example, the model with an interaction between A and B can be fitted using:

```
model <- glm(data˜A+B+C+A:B, family=poisson(link="log"))
```

As usual, the model parameters and associated useful statistics can be obtained using the `summary` function on the output from the `glm`.

Data: Modern Slavery

The data to be considered within this project relates to the number of individuals living in the UK that are potential victims living as modern day slaves in 2013. Potential victims were identified from a range of different organisations which are combined into organisation types. We consider the lists associated with four different organisation types:

S_1 - Local authority;

S_2 - Non-government organisation;

S_3 - Police force;

S_4 - Government organisation.

Collating individuals across the different sources listed above we obtain the following incomplete contingency table:

	S_4	1		0	
	S_3	1	0	1	0
S_1 =1	S_2=1	1	1	1	15
	S_2=0	0	3	19	54
S_1 =0	S_2=1	4	19	62	464
	S_2=0	76	703	1006	?

For further details see:
https://www.gov.uk/government/publications/modern-slavery-an-application-of-multiple-systems-estimation

Note that within this assignment one of the original data lists corresponding to the general public has been removed.

Task 1: Group Scientific Report - max 2 pages

You have been assigned to a group of size four (see Learn for group membership), and each group will submit a single scientific report (to be read by a statistician). The report will focus on estimating the number of potential modern day slaves in the UK in 2013, by applying a multiple systems estimation approach to the data described above. The statistical analysis needs to determine a suitable log-linear model, in terms of the interactions present within the model, and also investigate the absolute goodness-of-fit of the final selected model. You should clearly describe both the model selection process that you apply to the data as well as the associated absolute goodness-of-fit measure for the model deemed optimal in terms of the interactions present between the different lists. The results should be presented, discussed and interpreted accordingly in relation to the goodness-of-fit measures (relative and absolute) and estimation of the total population size.

Note 1: you do not need to redefine multiple systems estimation within the report - and you can consider that your scientific report is the results section following on from the description of multiple systems estimation above. This also means that you do not need to redefine the notation or describe the data - the focus is on the application of the statistical analyses, and associated interpretation of the results.

Note 2: you need only consider two-way interactions between two lists (three-way interactions between three lists is not necessary) - though of course there may be multiple such two-way interactions.

Note 3: you may consider either a classical or Bayesian approach to the problem - as long as the statistical approaches are consistent with the framework applied.

The group scientific report should be at most 2 pages. An additional appendix should be included providing the computer code used within the statistical analysis - the appendix is not included in the 2 page limit.

All members of a group need to contribute to the submitted group report. However, it is expected that members of the group may contribute in different ways and different aspects of the work, including for example, investigating relevant relative/absolute goodness-of-fit approaches; undertaking the statistical analyses in `R`; providing feedback on the analyses; interpreting the results; drafting parts of the report; proof-reading etc. Thus you may want to think about the different skills and interests of the members of the group when working together and/or if assigning individual tasks. If a member(s) of the group fails to engage and does not reasonably contribute within the group assignment, please contact the course lecturer (Ruth.King@ed.ac.uk) detailing the issues **before** 12.00 on Friday 1st April.

Task 2: Individual Executive Summary - max 1 page

An individual executive summary should be submitted that outlines the problem to be addressed, briefly describes the data and summarises the main results or findings of the statistical analyses undertaken in the context of the given problem. The summary should be concise including only the necessary information and be non-technical (to be read by, say, a member of the public). Thus, the executive summary should not simply be text "cut-and-pasted" from the scientific report, the audience of the summary is very different to that of the scientific report.

Note 1: There is an upper limit of 1 page for the executive summary, but it is quite possible that the summary may be shorter than this (1 page is the limit not necessarily an aim). Also see the marking rubric (conciseness and relevant information are explicit criteria).

Note 2: The executive summary needs to be written independently by each member of the group. Thus, whilst the material on which the summary is based will be common to individuals within the same group, there are numerous choices to be made when writing the summary, such as the material to present, structure, style etc so that the summaries should be clearly distinct from each other.