Multilevel logistic modelling: Issues when working with large datasets

Dr.E. M. Y. Cheng*
Faculty of Medicine, University of Southampton, Southampton, United Kingdom
m.y.cheng@southampton.ac.uk

Scott Harris
Faculty of Medicine, University of Southampton, Southampton, United Kingdom
sharris@southampton.ac.uk

Abstract

In public health research it is becoming increasingly common for studies to combine data that is collected at the individual level with higher-level aggregated data which could come from hospitals, specialist treatment centres or GP practices. Multilevel models are often used to analyse such datasets as they allow the hierarchical structure of the data to be taken into account. With large datasets this can result in some computational issues depending on the software package being used, the computing environment and the complexity of model being fitted. For this comparison we will focus on a two-level model, which is the most commonly seen in practice. An example dataset containing in excess of 8 million cases and a higher level term with over 32,000 levels will be used as a basis for the statistical modelling. Random subsets of varying size will be taken from this dataset and models with differing levels of complexity will be applied to them all. Variations on the number of classes in the higher-level variable will also be examined and the impacts on model fitting will be noted.

Key Words: Hierarchical models, Computer intensive methods, Software comparison.

1. Introduction

In clinical practice, handling and analysing large datasets is commonplace and accessibility to ever increasing amounts of data continues to increase. Clinical data may be collected through hospitals, specialist treatment centres, various National Screening Programmes or General Practices (GP) etc. Thus, data are routinely collected from different healthcare systems and settings. Some of these settings will collect sensitive information and in some situations patient level data is aggregated at different levels to protect patients' confidentiality. Some clinical applications have shown the difficulty in handling this type of data and the associated problems that can come with their collection and analysis e.g. maintaining patients' privacy (Kamm *et al* 2013). Jackson (2006) demonstrated that data may be collected at several different levels and that when this is the case it is important that this structure in the data is retained in the statistical modelling.

When data are collected or aggregated at different levels then multilevel modeling is a suitable method for the analysis. If the researchers only analyse the observed health outcomes at the individual level then they are ignoring the nested structure of the data and as such the results can be very misleading (Centre for Multilevel Modelling, 2008). In this paper we will focus on the analysis of multilevel data. In particular we will look at the most common setting of having two levels, data collected at the individual level and data collected at a higher level such as a clinic, specialist treatment centre or GP practice. The potential issues of handling large clinical data can be separated into two main

limitations (i) computational and (ii) software. In this paper we will investigate the impact on computation and model run times of a range of different study sizes and a range of different numbers of levels contained within the higher level term. We will also consider two different model complexities by looking at both random intercept and random slope models. We will not be assessing the suitability of the models to the data but focusing purely on the computational aspects of model fit times.

National Cancer Screening Programme Application

In this study we will utilize data from the National Bowel Cancer Screening Programme as an example of a typical dataset of this type. The National Health Service (NHS) introduced Bowel Cancer Screening (NHSBCSP) in 2006 in the UK. The aims of the screening programme are to (i) detect bowel cancer at an early stage, (ii) detect polyps, which may eventually develop into cancer, (iii) to reduce bowel cancer incidence and (iv) to reduce bowel cancer related deaths. Everyone between the ages of 60 and 69 are eligible for entry into the NHSBCSP. Eligible patients will receive an invitation letter from the screening programme and will be invited to attend a local clinic.

2. Methods

Logistic and Binomial regression models are non-linear regression models, within the generalised linear regression family. These models are used when the observed outcome variable (i.e. the response variable) can only take one of two possible states such as being true or false.

The observed variable Y_{ij} is a Bernoulli random variable; this follows the Binomial distribution with two parameters (sample size n_{ij} and probability π_{ij}):

$$Y_{ij} \sim Binomial(n_{ij}, \pi_{ij})$$

The regression model can be defined as:

$$\hat{Y}_{ij} = \hat{\beta}_0 + \sum \hat{\beta} x_i + \sum \hat{\alpha} x_j + \varepsilon_{ij}$$

Where i is the individual patient i = 1, 2, …, N, j is the second level (e.g. clinic) j = 1,2,,...,M, β_0 is the intercept, β is the coefficient of the explanatory variable x_i, α is the coefficient of the explanatory variable x_j and ε_{ij} is the error term.

The GLIMMIX procedure within SAS version 9.3 has been used to fit the models. All models were run on an Intel powered Core i7-3770K @4.4GHz with 16 GB of system memory that was running SAS 9.3 under a 64 bit Windows 7 Professional environment.

Data

The full dataset contained a total of 8,274,940 observations. A couple of demographic variables (age group and gender) are included in the dataset alongside a number of socioeconomic variables. However the socioeconomic variables were only available at the higher clinic level rather than at the individual patient level. This is due to the socioeconomic variables being based on postcode information which was only available for the clinics. This example dataset will be used to demonstrate the analysis methods and

investigate the contribution to computation time and full model run time of altering certain aspects of the dataset. Various subsets of this dataset will be considered to assess the impact that increasing numbers of cases will have on the computation times. We will also merge together the levels of the higher level term so that we can assess the impact that the number of levels has on the computation time. Each scenario has been run 3 times, with the mean runtimes reported. Models with random intercepts for each clinic and more complicated models with random slopes have been investigated to see how this added complexity also affects model run times. No assessment of the suitability of models has been conducted and the only stipulation for the usability of a result was that the model was able to converge successfully.

An example section of the SAS code used for one of the random intercept models is included below:

```
PROC GLIMMIX DATA=NHSBCSPDATA;
CLASS GENDER CLINIC AGEGRP;
MODEL ADEQUATELY_SCREENED = GENDER AGEGRP SOCEC / DIST=BINARY
DDFM=KR ODDSRATIO;
RANDOM INTERCEPT / SUBJECT=CLINIC SOLUTION;
LSMEANS GENDER AGEGRP / DIFF PDIFF OR CL;
ODS EXCLUDE SOLUTIONR;
RUN;
```

3. Results

The SAS computation times for the random intercepts model are included in table 1. A variety of dataset sizes from 100,00 through to the full 8,274,940 have been considered as well as differing numbers of levels to the higher level term (from 5 to 3,000). The number of levels in the higher level term was capped at 3000 as when the number of levels is increased beyond this limit the model run times for the random slopes model (Table 2) are dramatically increased.

Table 1: Average SAS computation times (seconds) for a random intercepts model at various dataset sizes and number of levels in the higher level variable.

	Number of levels in Higher level term							
Dataset size	**5**	**100**	**500**	**1000**	**1500**	**2000**	**2500**	**3000**
100,000	1.61	1.66	2.18	22.6	83.9	212.5	446.7	827.3
250,000	3.66	4.08	4.83	24.4	87.8	214.7	449.0	831.5
500,000	8.20	7.45	9.06	29.4	92.8	218.4	452.6	835.3
1,000,000	21.8	16.4	17.3	37.6	97.3	226.1	460.1	853.7
2,500,000	37.0	41.3	38.3	59.0	120.1	263.2	483.3	865.7
5,000,000	82.0	74.5	76.6	104.6	158.2	297.8	520.8	916.6
8,274,940	133.7	135.7	125.2	146.8	223.0	338.5	571.3	976.9

Table 1 shows that increasing both the dataset size and the number of levels in the higher level term leads to an increased computation time, when looking at random intercept models. A much larger increase in computation time is seen when the number of levels in the higher level term is increased, as opposed to the raw number of cases which only has

a marginal effect. Figure 1 shows a graphical representation of these data. This shows a smooth curve as the number of levels in the higher level term is increased and a linear increase in computation time as the number of cases is increased.

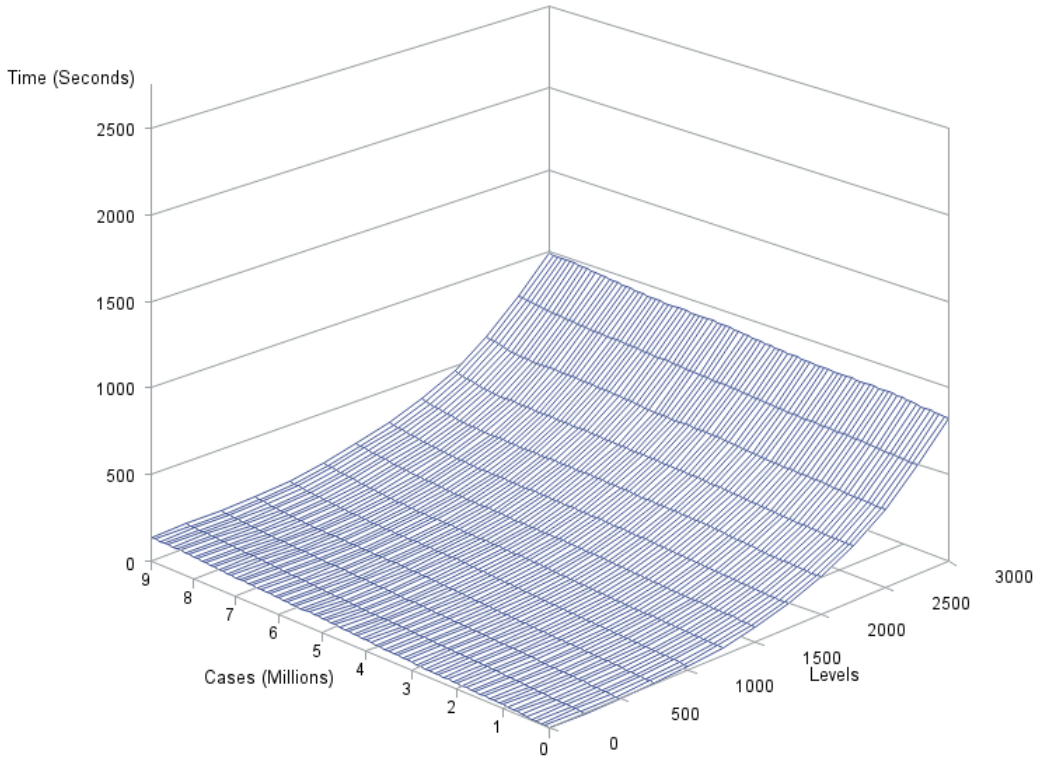

Figure 1: A surface plot showing the relationship between computation time, number of cases and the number of levels in the higher level term when fitting a random intercepts model.

In addition to looking at the computation times for a random intercept model we also looked at the more complicated random slopes model (Table 2).

Table 2: Average SAS computation times (seconds) for a random slopes model at various dataset sizes and number of levels in the higher level variable.

	Number of levels in Higher level term							
Dataset size	**5**	**100**	**500**	**1000**	**1500**	**2000**	**2500**	**3000**
100,000	1.21	1.59	3.81	31.2	107.6	279.0	587.9	1066.1
250,000	2.69	3.87	8.44	34.9	126.2	286.1	597.9	1081.5
500,000	5.90	7.06	15.5	48.6	144.0	312.6	639.4	1164.6
1,000,000	11.8	15.6	29.5	70.4	153.5	340.1	654.2	1287.1
2,500,000	27.1	38.9	64.0	122.0	231.7	492.6	784.9	1391.4
5,000,000	59.7	71.1	128.8	246.0	362.8	677.4	999.7	1867.0
8,274,940	96.5	173.9	205.8	344.9	530.2	822.1	1251.4	2072.4

Table 2 shows that increasing both the dataset size and the number of levels in the higher level term also leads to an increased computation time, when looking at random slopes models. The computation times for the random slopes models are generally longer than the comparative random intercepts model. Generally there is a larger increase in computation time when the number of levels in the higher level term is increased, as opposed to the raw number of cases, when looking at models with random slopes.

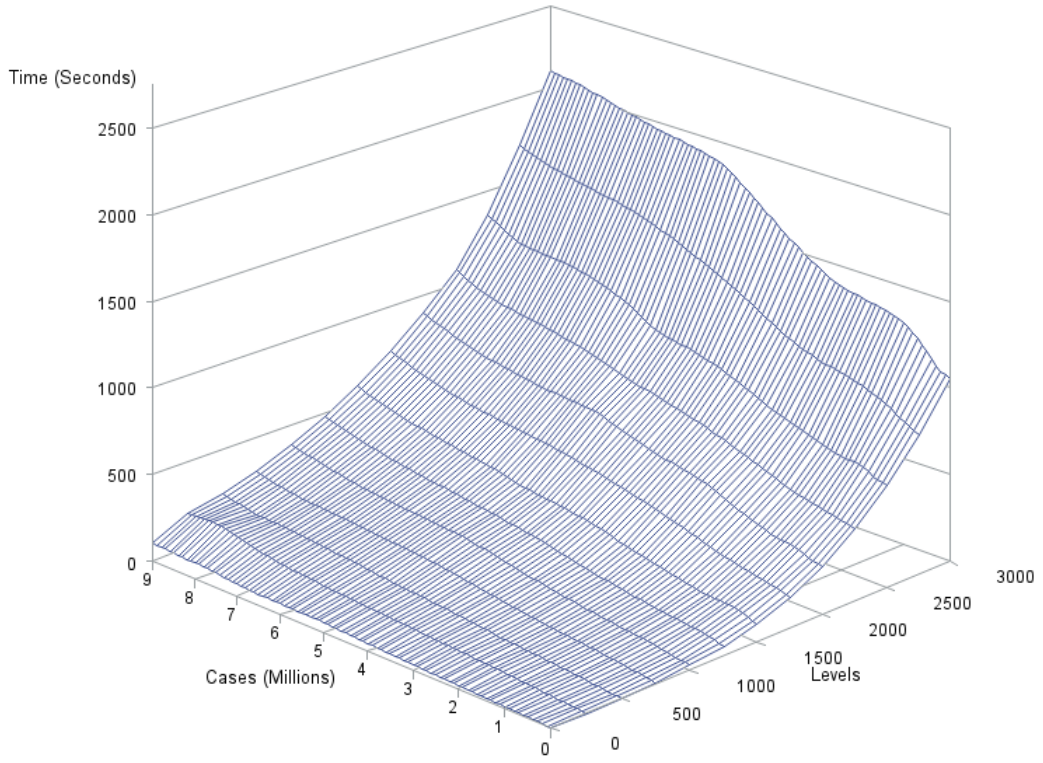

Figure 2: A surface plot showing the relationship between computation time, number of cases and the number of levels in the higher level term when fitting a random slopes model.

Figure 2 shows a graphical representation of the data from the random slope models. This figure helps to show an interaction between the raw number of cases and the number of levels in the higher level term via the "twist" in the response profile. For larger numbers of levels in the higher level term there is a greater increase in computation time due to having a larger number of cases.

4. Discussion

We have been able to show that both raw dataset size (number of cases) and the number of levels in the higher level term have an impact on the computation time for SAS PROC GLIMMIX. This is the case for both random intercept and random slope models. Generally it is the increase in the number of levels in the higher level term that results in the largest increase in computation time, although for the random slope models there appears to be an interaction between the number of cases and the number of levels.

There are many statistical packages available for multilevel modeling in addition to SAS, such as MLwiN, WinBUGS (or OperBUGS), R, Stata, SPSS etc. However, each of these software packages will come with their own limitations. With a large majority of these packages the limitation is related to reaching, or exceeding, available system memory. We initially tested our full models with MLwiN, R and SAS before deciding to stick to using SAS exclusively for this investigation. We experienced several issues with both MLwiN and R with regards to running out of or short on system memory, even when using 16GB. This occurred routinely when the number of records was larger than 2,000,000 and was independent of the number of levels being considered in the higher level term.

All of the models considered with various levels of higher level term and numbers of cases were able to be fitted successfully in SAS. The most complicated model that was considered here still took less than 35 minutes to fit. With this information and the issues experienced with both MLwiN and R then SAS seems to be the most suitable software package to use for this application, when dealing with large datasets.

5. Conclusions

The number of levels in the higher level term is the strongest driver to computation time in SAS PROC GLIMMIX, although there appears to be an interaction between the number of cases and the number of levels when looking at random slope models.

The NHSBCSP study dataset exposed some software limitations with both MLwiN and R that may be related to the amount of available system memory. SAS on the other hand had no problems with any of the models that we considered here and so seems to be the more suitable statistical software package for this application when dealing with large datasets. Austin (2010) provided more discussion on handling large data with different software packages and discussed their various limitations in more detail.

There is some potential here for future work. Stata has yet to be considered and it is commonly used in this field. The various models could also be extended into a Bayesian framework, although the computation times for such models would be expected to be noticeably longer and would be expected to stretch into hours or even days for some models.

References

Centre for Multilevel Modelling (2008), multilevel modelling.
http://www.cmm.bristol.ac.uk/MLwiN/index.shrml

Kamm, L., Bogdanov, D., Laur, S. and Vilo, J. (2013) "A new way to protect privacy in large-scale genome-wide association studies" Boinformarics, 29, 886-893.

Jackson, C., Best, N. and Richardson, S. (2006) "Impoving ecological inference using individual-level data" Statistics in medicine, 25, 2136-2159.

Austin, P.C. (2010) "Estimating Multilevel Logistic Regression Models When the Number of Cluster is Low: A Comparison of Different Statistical Software Procedure. The International Journal of Biostatistics, Volume 6, Issue 1, Article16.