A Bayes Linear Analysis of Multilevel Models

NASHAD AUCHOYBUR

A Bayes Linear Analysis of Multilevel Models

Nashad Auchoybur

A Thesis presented for the degree of
Doctor of Philosophy

Statistics Group
Department of Mathematical Sciences
University of Durham
England

August 2022

Dedicated to

Nasim, Zahraa and Zakiyyah,

Mariam and Tyyub

A Bayes Linear Analysis of Multilevel Models

Nashad Auchoybur

Submitted for the degree of Doctor of Philosophy
August 2022

Abstract

In this thesis, Bayes Linear methods for modeling multilevel data are presented and discussed. Second-order exchangeability judgements are exploited to formulate subjectivist versions of multilevel models. Bayes linear methods are applied to estimate model parameters and for diagnostic checks. Closed-form expressions of estimators are derived, allowing insight into relationships between the quantities thereof. The canonical analysis and resolution transforms are used to guide sample design and sample size determination under cost constraints. A finite version of a multilevel model is formulated, analysed and compared to infinite versions, giving further insight into sample design issues via the finite resolution transform.

A new Bayes Linear Minimum Variance Estimation (BLIMVE) approach is developed to estimate variances. Estimated variances are used to perform two-stage Bayes linear analysis of more complex multilevel models. The methods developed are shown to be applicable in cases of small level-2 samples. The Bayes linear analyses of multilevel models are applied to an educational data set using special-purpose codes written in the R Statistical Language.

Declaration

The work in this thesis is based on research carried out at the Statistics and Probability Group, the Department of Mathematical Sciences, Durham University, England. No part of this thesis has been submitted elsewhere for any other degree or qualification and it is all my own work unless referenced to the contrary in the text.

Acknowledgements

I would like to thank my supervisor Michael Goldstein for first suggesting that I do a PhD while we were peacefully walking among the trees on Réduit Campus in my beloved Island of Mauritius; little did I know then that a storm was brewing, and that soon, it would take over a sizeable portion of my life. I am very grateful to Michael for his support, expert guidance , patience and good humour over all these years. Thanks also to my earlier Statistics teachers, Brayen, Bob and late Raouf; you have been an inspiration to me. I am also grateful to all my friends, specially Kurshid and Eshan, and my colleagues at the university of Mauritius for their support. Thanks also to Muzzammil for his precious help freeing me from many administrative tasks.

Finally, a very special thanks to my wife Nasim, my daughters Zahraa and Zakiyyah for their unflinching support and love during all these years: nothing, neither this PhD nor my life, would have any meaning without you.

Contents

List of Figures

List of Tables

Chapter 1

Introduction - Multilevel modeling for multilevel data

The world we live in is complex. This complexity is rife as can be seen in natural phenomena that surround us. Complexity also permeates the biological, psychological, social and economic dimensions of our lives. In our quest to understand this complexity, we come to collect data and, unsurprisingly, this data turns out to be both complex and richly structured. A common type of richly structured data, called multilevel data, and the special class of models required to analyse such data, called multilevel models are the subject of this thesis.

Multilevel data occur in most fields of study. The most frequently cited example is in educational research, where students are grouped in classes and classes are grouped in schools forming a multilevel or hierarchical structure. The main issue in modeling multilevel data is that the usual assumption of independence is no longer valid; students in the same class are more likely to be similar as compared to students in another class. As such, commonly used models that assume independence, such as linear regression for example, are no longer valid and may lead to erroneous inferences. Hence, classical multilevel or Bayesian hierarchical modeling are better suited for accounting for the dependencies between the units at the different levels of a hierarchy. Notwithstanding these modeling approaches, there are still issues that could benefit from an alternative approach. Classical estimation methods suffer from the possibility of negative variance estimates especially with a low number

of groups. At the other end, when there are large amounts of data, Bayesian hierarchical modeling using Markov Chain Monte Carlo or dynamic Hamiltonian Monte Carlo can be very slow or difficult to tune.

In this thesis we propose to use Bayes linear methods as an alternative to existing methodologies in formulating, estimating and diagnostic checking of multilevel models. The Bayes linear approach requires only limited beliefs specifications as opposed to complete prior probability distribution specifications in the fully Bayesian approach. The Bayes linear approach is subjectivist and uses expectation rather than probability as a primitive. The principal features of the method are discussed in Goldstein and Wooff (2007).

1.1 Thesis outline

In Chapter 2 we review the concepts underlying multilevel data structures and the need for multilevel modeling in the context of some important applications. A multilevel dataset on an introductory course in Statistics, the STAT1010 dataset, is introduced and is used throughout this thesis to motivate and illustrate the analyses. We discuss a number of classical estimation methods including least squares, maximum likelihood and iterative generalized least squares methods that are all relevant to multilevel modeling. We also discuss the fully Bayesian hierarchical modeling approach and consider the difficulties in making full prior specifications, as well as in computing posterior densities. Finally, we present the concepts and methods involved within the Bayes linear approach. Using a collection of second-order exchangeable quantities, we explain the principles of adjustment of means and variances, as well as some important diagnostic checks. The notations used in this thesis are compiled in Appendix A.

In Chapter 3 we use second-order exchangeability (SOE) judgements to formulate our versions of multilevel models. We present the SOE random effects (SOEREF) model, our version of the simplest multilevel model, i.e. the random effects model. The assumptions and notations of the SOEREF model are discussed. We then extend the SOEREF model to a SOE regression (SOEREG) model. We also discuss

and illustrate suitable methods that can be used to specify priors for our models using the STAT1010 data.

In Chapter 4 we apply Bayes linear methods to adjust the population overall mean and population group means in the SOEREF model. We derive closed form expressions for the adjusted mean for balanced and unbalanced data and use these to understand how the adjusted quantities relate to prior specifications and data. We compute adjustments and diagnostics using specially written codes in the R statistical programming language and apply these to the STAT1010 data (Appendix B gives the R and also [B/D] codes as well as the STAT1010 data). We also apply a partial Bayes linear analysis and demonstrate its importance as a diagnostic tool in multilevel modeling. We exploit the canonical structure and resolution transform underlying our exchangeable adjustments to address sample design issues and sample size determination with cost constraints for both level-1 and level-2 units in the SOEREF model. We then relax the assumption of infinite exchangeability and formulate a finite version of the SOEREF model. We are interested in comparing the adjustments of the finite and infinite versions of the SOEREF model via the canonical analysis and apply these to the STAT1010 data.

In Chapter 5 we discuss the difficulties in learning about population variances and develop Bayes linear methods to estimate the level-1 variance in both the balanced and unbalanced cases. We apply these methods to the STAT1010 data and illustrate the choice of priors for fourth order quantities. The sensitivity of our adjusted variance to a higher kurtosis is also investigated.

We develop a Bayes Linear Minimum Variance Estimator (BLIMVE) to estimate the level-2 variance of the SOEREF model in Chapter 6. The method is applicable to two or more groups and we validate it using simulation. We apply BLIMVE and estimate the level-2 variance in the STAT1010 data. Having learned about both level-1 and 2 variances, we perform a two-stage analysis by substituting the estimated variances in the adjustment of the mean components. We then consider estimation of the level-1 scalar variance and level-2 variance matrix in the more complex SOEREG model. Specifications of priors, in particular for the residual variance matrix, are discussed and applied to the STAT1010 data. We describe how

to use Bayes linear methods to learn about population variances based on unbiased OLS estimators and apply these to the STAT1010 data. Finally, we apply a two-stage analysis to update intercepts and slopes in our SOEREG model and compare shrinkage in these regression coefficients for the STAT1010 data.

We conclude with a discussion of our results and some promising areas for future work in Chapter 7.

Chapter 2

Modeling data with complex structures using multilevel models

2.1 Introduction

In this chapter we review the concepts underlying a multilevel model and its application to hierarchical data. We present examples of various types of richly structured data that are commonly viewed as hierarchical or multilevel data and discuss the substantive research questions that are of interest in these multilevel data. We provide notations for multilevel models and explain the conceptual difference between multilevel and regression models. Classical and Bayesian estimation methods are presented and discussed, pointing out some of the shortcomings of these methods. We then present Bayes linear methods including adjustments of means, variances and diagnostics checks.

2.2 The pervasive multilevel data

Multilevel or hierarchical data occur frequently in most fields of study. In education for example, students are "naturally" grouped in classes and classes are in turn grouped in schools, hence forming a three-level hierarchy; the multilevel data here will then comprise of variables measured on students, classes and schools. Another example is in economics where we are interested in employment status of individuals

(employed or unemployed) grouped in regions (urban or rural) hence forming a two-level hierarchy. It is easy to find examples of multilevel data in almost all areas of the sciences. Indeed as Kreft and de Leeuw (1998) wrote "once you know that hierarchies exist, you see them everywhere".

In some situations though, the hierarchical structures may not be as explicit as in the above-mentioned examples. Consider longitudinal studies for instance, where repeated measurements are made on a sample of individuals over time. The repeated measurements taken at different points in time for one specific individual, may be viewed as grouped within that individual, thus forming a two-level hierarchy, where the repeated measurements are at the lower level and individuals are at the higher level of the hierarchy.

Acknowledging the multilevel structure in a dataset is an important step towards proper modeling of multilevel data as the clustering of individuals induces dependence. To illustrate this dependence we note in our earlier example that students in the same class share the same teacher and class environment; they are thus more likely to have similar exam scores than students in different classes. A consequence of such dependence is that it invalidates the straightforward application of traditional statistical methods such as linear regression modeling. Therefore a more flexible type of model is called for, namely a multilevel model, that properly accounts for the dependencies in multilevel data.

Over the last thirty years or so, multilevel modeling has emerged as an important modeling technique, prompting numerous textbooks (see for example, Goldstein (2010), Bryk and Raudenbush (1992), Snijders and Bosker (1999) and more recently, Gelman and Hill (2007) and softwares. A list of multilevel modeling softwares is available at the Centre for Multilevel Modelling (CMM) at Bristol University (http://www.bristol.ac.uk/cmm/). The CMM " *collaborate with a range of researchers working with multilevel models to develop new statistical methodology, implemented in software to address unsolved issues in quantitative modelling of social processes.*"

The above on-going developments have encouraged researchers in general to use multilevel modeling in the analysis of richly structured data. Below we explain the

facets of multilevel modeling that have made it gain such popularity and importance among researchers, especially in the social sciences.

2.3 The need for multilevel modeling

As we saw above, one reason for using a multilevel model is that observations in multilevel data are dependent and, unlike more familiar statistical techniques which assume independence, a multilevel model can account for this dependence. Here we consider three commonly cited advantages for using a multilevel modeling approach when analysing multilevel data. They are:

- Improved estimation of effects
- Modeling cross-level interactions
- Inference for groups with sparse data: Small area estimation

For each of the above three advantages, we begin by explaining a research issue of interest using simple examples. We then explain the benefits of using a multilevel modeling approach to address the research issue in the context of a more detailed example from the research literature.

2.3.1 Improved estimation of effects

The effect of clustering on the individuals forming the clusters is of considerable interest to researchers in applied social sciences. Thus, in school effectiveness studies, educational researchers ask to what extent school characteristics (e.g. public/private, school management) impact on pupils' performances, while in social policy research, sociologists are interested in the impact of neighbourhoods (poor/rich) on teenagers' behaviour (teenage pregnancy, school dropout). Estimating an effect at a group level (school or neighbourhood characteristic) based on an outcome at an individual level (pupil's performance, teenager's behaviour) is not straightforward and can have major pitfalls as our chosen applied example next shows.

Bennett (1976) published a report claiming that primary school children taught by teachers using a 'formal' teaching method were likely to make greater progress in learning than those pupils taught by other teaching methods. This finding , which had important policy implications, gave rise to many controversies. The statistical issue here is that in his analysis Bennett ignored the multilevel structure of the data; namely the grouping of pupils in classes and teachers. Hence, each pupil was treated as providing independent information to assess the teaching method whereas, in fact, the clustering of pupils in classes and teachers meant that pupils in the same group had correlated exam scores. Ignoring clustering and treating observations as independent led to underestimated uncertainties (small standard errors) in teaching style effects. Therefore, estimated confidence intervals were also quite narrow, leading to apparently significant differences in teaching style effects.

(Aitkin et al., 1981) analysed the same data as in the teaching styles study, but they used a multilevel model to account for the multilevel structure in the data. The outcome variable in Aitkin's model was achievement test score (Y_{pqr}) for pupil r, grouped in class q, and teaching method p, a three-level model. The result was "*greatly reduced significance of any differences in teaching style*" Aitkin et al. (1981).

It is now well known that ordinary least squares may underestimate the standard errors of regression coefficients for multilevel data while multilevel modeling provides more efficient estimates for these standard errors.

2.3.2 Modeling cross-level interactions

Are differences in mathematics achievements between boys and girls the same in private and public schools? Are differences in fertility rates between urban and rural regions the same in rich and poor countries? Each of these questions concerns the impact of variables measured at different levels of a hierarchy on a response variable. The first question, for example, concerns three variables: a response variable (achievement in mathematics) measured at the pupil level, a pupil level predictor (gender) and a school level predictor (school type). We are interested in the effects of gender and type of schools on pupils' achievement in mathematics. Suppose, for

the sake of argument, that gender differences in mathematics achievements are larger in public schools than in private schools. More specifically, suppose girls obtain on average far better grades than boys in public schools. While in private schools, because of better teachers and smaller classes, girls and boys have on average the same grades. Since the effect of gender on achievement in mathematics depends on school type, we say that there is an interaction between gender and school type. And because such interactions occur between variables measured at different levels of a hierarchy, they are termed cross-level interactions.

The identification of cross-level interactions in richly structured data are of prime importance to researchers in all fields of study. As an example we consider Shouls et al. (1996) who studied variation in an individual's chance of being long term ill based on variables such as age, low skill job, non-white ethnicity, being married and individual deprivation (unemployed, does not own a house and so on). Individuals were considered nested in local authorities and the result of Shouls et al. (1996) modeling showed significant variation in long term illness rates between local authorities. In an attempt to explain this variation they introduced a cluster level (group level) variable namely, North versus South England, in their model. While the effect of the North/South divide was significant, they found that the cross-level interaction between the North/South divide and individual deprivation was not. The Shouls et al. (1996) analysis shows how modeling cross-level interactions can seamlessly be achieved within the framework of multilevel modeling when data have a hierarchical structure.

2.3.3 Inference for groups with sparse data: small area estimation

An important survey research problem that has many applications, is that of obtaining reliable estimates of quantities such as averages, totals and rates for groups having little or even no data. This research problem frequently arises when the results of surveys, which are conducted on part of a population but with a view to

learning about the whole population, need also to be used to learn about groups or categories of the target population. Examples of these groups include people living in specific regions or belonging to specific socio-economic categories. Because surveys are costly and time consuming, sample sizes tend to be limited in practice. Hence, the groups in question may have little data in the sample or no data at all (unsampled groups), rendering estimates of the quantities for these groups difficult and inaccurate.

To understand the need for estimates in small groups, consider the economic problem of unemployment. The unemployment rate is an important indicator of the economic health of a nation and therefore national surveys are conducted on a regular basis in order to assess the level of unemployment. But unemployment affects different regions and sub-regions of a nation differently; some regions have more unemployed people than others. Hence, reliable estimates of unemployment at regional and sub-regional levels are vital for efficient policy decisions aiming at reducing unemployment.

More importantly perhaps, is the impact of unemployment on the very individual who has lost his or her job. Indeed the loss of a job to an individual not only means the loss of livelihood and the financial hardships that it entails, but also a loss of status in society, with its accompanying psychological distresses. It seems therefore natural to measure the unemployment rate among specfic groups of individuals based on their socio-economic profile, such as lone mothers aged between 18 and 25 years and having only primary education.

Obtaining reliable estimates of unemployment for regions, sub-regions and socio-economic groups such as mentioned above, is therefore important albeit difficult when such groups have little or no data in the sample. Direct estimates, based on number of unemployed and total sampled in the small group for example, are either imprecise due to the small sample size or impossible when a group is not sampled. However, many techniques have been developed over the years by survey researchers to provide reliable small area estimates. A thorough in-depth survey of these small area estimation techniques is provided by Rao (2003) in his book "Small Area Estimation". Of particular interest to us here, are the model-based techniques,

especially those that take advantage of multilevel modeling.

Multilevel modeling seems an intuitive approach to use in small area estimation, since small area estimates are required for groups based on observations that are made at the level of the individuals forming the groups. At the individual level, we construct regression-type models that relate the quantity of interest (proportion unemployed) to individual-level covariates (highest educational level attained, for example). At the group level, we include random area-specific effects to account for differences (in unemployment rates) between areas. In addition, contextual variables, available from census and administrative sources on all individuals in a small area, are used in the area-level model to explain between-area variability beyond that explained by individual-level variables. An example of such a contextual variable is the percentage of the population in each area having at least completed primary education, a known area-level predictor of unemployment.

Fitting the multilevel model then involves 'borrowing of strength' from all areas in the sample to estimate the random effect of a specific area. Such borrowing of strength increases the effective sample sizes for small areas, hence increasing the precision of the small area estimate as well as providing a measure of precision for each specific area. Hence, multilevel modeling is a powerful tool to use in small area estimation.

Following Rao (2003), we give the following advantages for using a model-based approach in small area estimation, including the specific advantages of using a multilevel modeling approach:

1. Model fitting diagnostics can be used to find a suitable model for the data as well as to investigate major discrepancies between model and data.

2. Appropriate summaries (means, totals and rates) for the small area estimates can be obtained via the model. In addition, measures of precision for the small area estimates are easily obtained from the models.

3. Multilevel models are very flexible and can model all types of response variables including continuous, binary, count and multivariate.

4. More importantly, a multilevel model can account for complex variance structures, including spatial effects and spatio-temporal effects.

The above modeling advantages are of particular relevance to this thesis. Indeed, our overarching aim is to apply Bayes linear methods as a unifying framework for model formulation, estimation and diagnostic checking in the context of multilevel modeling, including learning about variances in complex multilevel models.

2.4 A frog in a pond or the basic advantage of an expert's belief in hierarchies

As we saw above there are important gains for a researcher in recognising hierarchies in complex data and taking advantage of multilevel modeling. If a researcher's beliefs do follow a multilevel structure, then it would be a mistake to ignore it. We now consider what motivates belief in multilevel structures, why at times researchers may decide to ignore such structures and what the resulting consequences are.

The primary reason why an expert may hold belief in multilevel structures is because the expert's belief is rooted in theories about the relationship between the individual and the context (group) to which the individual belongs. Studies of contextual effects are important in all the social, economic and behavioural sciences. Educational researchers for instance, are interested in studying the impact of various contexts, such as schools or classrooms, on students. The theory underpinning these contextual effects has come to be known as 'frog pond' theory (after a classic research article 'The campus as a Frog Pond', Davis(1966). The metaphor 'a small frog in a large pond or a large frog in a small pond' is used to define the relationship between the student and the class context. For example, a rather weak student in a class of highly intelligent students may perform poorly while the same student in a class of weaker students, may gain confidence and perform better. This kind of belief naturally leads to hypotheses about cross-level interactions between variables defined at the individual and at the group levels. Our point here is that experts do hold beliefs based on theoretical considerations, and possibly irrespective of data,

and that multilevel models provide a framework to state and test these hypotheses.

Before the advent of multilevel modeling in the 1980's, researchers often had to ignore the multilevel structures when analysing complex data such as, for example, in the modeling of change using longitudinal data (which have an inherent multilevel structure). Singer and Willett (1993) report that, 'methodologists advised that researchers should not even attempt to measure change because it could not be done well' while Cronbach and Furby (1970) advised researchers studying change to 'frame their questions in other ways' because these questions pertained to cross-level hypotheses necessitating multilevel modeling.

When multilevel structures are ignored, often researchers perform single level analyses by aggregating data (at the higher group level) or disaggregating data (by using indicator variables to attach group level variables at the individual level). Unfortunately, such single-level analyses of multilevel data are inappropriate as they tend to ignore dependencies in the data.

More importantly, single-level analyses of multilevel data may result in fallacious interpretations such as the ecological fallacy which, in the multilevel modeling literature, means observing a high correlation between variables at the higher (group) level of a hierarchy and using this observation to infer a similar correlation at the individual level. The reverse, making inferences at the higher level of a hierarchy based on observations at the individual level, can also be misleading and is termed the atomistic fallacy and is related to Simpson's paradox (Hox 2002). Therefore if an expert's beliefs do follow a multilevel structure then it would be advantageous to use multilevel models to avoid the above pitfalls.

2.5 Names, notations and equations for multilevel models

So far we have seen that most real world data has a complex hierarchical structure and that there are substantial benefits in adopting a multilevel modeling approach to analyse such richly structured data. In this section we use an example to explain

some important notations and model equations for various types of multilevel models, ranging from the simplest to the more complex. Our aim is to use these common multilevel notations and equations to review multilevel model estimation methods in Section 2.6. Appendix A gives a list of notations used in this thesis. First we briefly look at the different names used for multilevel models in the literature.

2.5.1 A multitude of names for a multilevel model

Multilevel modeling is not a new statistical method; it has been around for quite a long time, albeit under various names. Indeed, rarely has a statistical method had such a plethora of names, the most common being: mixed-effects models, random-effects models, random-coefficient models, hierarchical models, covariance components models and exchangeable regressions. The reasons for such a variety of names are, in our opinion, two-fold. Firstly, the concepts and estimation methods underlying multilevel models have evolved according to specific and varied application areas. For example, animal genetics studies have given rise to variance components methods while educational and social sciences have given rise to multilevel and hierarchical modeling. Secondly, the longstanding divide in approaches to statistical inference, namely between classical and Bayesian, has also led to names such as Bayesian hierarchical models and exchangeable regressions to mark the underlying Bayesian methods used in formulating and estimating these models.

In this thesis we shall develop our own approach to analysing multilevel (or hierarchical) data based on the Bayes linear methodology. We shall then provide suitable nomenclature to reflect the specificities of our approach (see Chapter 3). When we wish to refer to classical or full Bayesian hierarchical models however, we shall use 'multilevel model' or 'Bayesian hierarchical model' as appropriate.

2.5.2 An example: The STAT1010 dataset

We shall illustrate the notations and model equations for multilevel models using a data set we collected in 2004 at the University of Mauritius, the STAT1010 data. STAT1010 is the module code of an introductory course in statistics. The module is

compulsory in several degree programmes at the university and is delivered mainly by distance education (DE). At the start of the semester, students are given a DE manual to work through and once a week they have a one-hour face-to-face session with a lecturer who may be a part-time or full-time staff member. The semester lasts for fifteen weeks. At the end of the eighth week there is a formal mid-term class test that is compulsory for all the students. The final examination is held at the end of the semester. To ensure uniformity, guidelines for marking of class test and examination questions are provided to all lecturers. Further, all marked scripts are moderated by full time lecturers.

The data comprises examination marks of 306 students grouped in each of eight classes. Table 2.1 below gives a summary of the main variables including some cases in the dataset. In five of the classes, STAT1010 is taught by part-time lecturers and the remaining three by full-time lecturers. Hence, we have a class-level variable (Part-time/Full-time). Another class-level variable is whether students study management or engineering sciences, the latter have better mathematical abilities. The response variable is the final examination marks of the students, while the explanatory variables are the mid-term class test marks and prior achievements at A level (A level scores).

Class	Student	Lecturer	Faculty	Sex	A level	Test	Exams
1	1	full-time	management	male	20	39	28
2	42	part-time	management	female	20	93	55
3	66	full-time	engineering	female	24	75	64
4	107	full-time	management	female	22	44	31
5	137	part-time	management	male	12	54	18
6	170	part-time	engineering	male	30	60	97
7	218	part-time	engineering	female	18	78	77
8	261	part-time	engineering	male	22	45	33
8	306	part-time	engineering	male	30	63	91

Table 2.1: Structure of the STAT1010 data. The first case in each of the eight classes is shown as well as the last case in class no.8.

Apart from its role in illustrating our methods, the data is also of substantive interest. The quality assurance sub-committee, of which I was a member, had expressed the need for information on issues such as the relationship between achievement at university level and performance at A level. Another question of interest that explicitly calls for multilevel modeling is whether some of the differences in students' achievement in the STAT1010 examination can be attributed to class type (part-time against full-time lecturer).

2.5.3 Notations

Along with the plethora of names for multilevel modeling that we mentioned in section 2.5.1 above, there is also a profusion of notations for basically equivalent multilevel models, but arising from different disciplines. The multilevel models of interest to us in this thesis may be considered as extensions of the general linear model (GLM) class that includes linear regression and analysis of variance models. Within this GLM class of multilevel models though, the current notations for response and predictor variables, regression coeffcients, mean and variance parameters, as well as subscripts indicating individuals and groups are varied and potentially confusing. Our choice of notations below follows somewhat those of Gelman and Hill (2007) and is suitable for fully Bayesian multilevel modeling as well as our own Bayes linear method. We introduce the notation in the context of simple linear regression which naturally leads to multilevel models.

2.5.4 Linear regression of the STAT1010 data

Regression in a single class

We are interested in the relationship between the examination and class test marks in the STAT1010 data. We first consider the marks in a single class only, say the first class, and we write the regression equation:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \tag{2.1}$$

where:

- Students are denoted by i, for $i = 1, 2, \ldots, n$. In this case $n = 41$.
- y_i is the response variable, here the examination mark of student i.
- x_i is a predictor variable, here the class test mark of student i.
- β_0 is an unknown intercept term, here the predicted examination mark of a student who scored a zero mark in the class test.
- β_1 is an unknown slope term, here representing the predicted difference between examination marks of students whose class test marks differ by 1.
- ϵ_i is an error or residual term, assumed to be independently and identically distributed.

In addition to the above, the errors are assumed to follow a normal distribution with mean zero and constant variance σ_y^2. Estimates for β_0 and β_1 are obtained by the method of ordinary least squares (OLS) that minimises the sum of squared errors, $\sum_i \epsilon_i^2$.

Fitting the above regression using, for example, the R function `lm()`, gives estimates $\hat{\beta}_0 = 21.3$ and $\hat{\beta}_1 = 0.4$. We can therefore write:

$$\hat{y}_i = 21.3 + 0.4 x_i \tag{2.2}$$

The term $\hat{y}_i$ is the fitted value, that is the estimated examination mark of student i based on (2.2). For example, the first student in class one has a class test mark of $x_1 = 39$ (see table 2.1) based on which his estimated examination mark is $\hat{y}_1 = 37$ while his actual mark is $y_1 = 28$. Therefore, based on the regression model, the student obtained 9 marks less than is expected on average from students having a

class test mark of 39. In general, the difference $y_i - \hat{y}_i$ is the estimated ith residual. An estimate of the error variance σ_y^2 is $\sum_i \hat{\epsilon}_i^2/(n-2)$.

Extending simple linear regression in a single class

One way to extend the regression of examination marks on class test marks is to add further explanatory variables, such as sex and A level scores, hence giving a multiple regression. In general, we may have p explanatory variables on each of n students. We then write the multiple regression in matrix notation as $y = X\beta + \epsilon$, where the response variable y and the residuals ϵ are n x 1 vectors each. X is a n x p matrix of predictor variables and β is the p x 1 vector of regression coefficients. As in the case of simple linear regression we derive estimates $\hat{\beta}$ of β by minimizing the sum of squared residuals:

$$\epsilon^T \epsilon = (y - X\beta)^T (y - X\beta) \tag{2.3}$$

giving

$$\hat{\beta} = (X^T X)^{-1} (X^T y) \tag{2.4}$$

and

$$var(\hat{\beta}) = (X^T X)^{-1} \sigma_y^2 \tag{2.5}$$

where the superscript T denotes matrix transpose. The fitted values are $\hat{y} = X\hat{\beta}$ and the estimate of the error variance σ_y^2 is $\hat{\epsilon}^T \hat{\epsilon}/(n-p)$ where, as before $\hat{\epsilon} = y - \hat{y}$.

Extending Ordinary Least Squares: Generalized Least Squares

In both the simple and multiple linear regressions described above, the residual error terms ϵ are assumed uncorrelated. Hence, var$(\epsilon) = \sigma_\epsilon^2 I$, where I is the identity matrix. A further generalisation of multiple linear regression is to drop this assumption and consider correlated responses. The error variance-covariance matrix is then Σ_ϵ, which is symmetric and positive definite. A simple example of correlated responses

is in longitudinal data (hence multilevel) or time series data with serially correlated errors where the variance-covariance is

$$\Sigma_y = \sigma_y^2 \begin{pmatrix} 1 & \rho_1 & \rho_2 & \dots \\ \rho_1 & 1 & \rho_1 & \dots \\ \rho_2 & \rho_1 & 1 \dots & \\ \vdots & \vdots & \ddots & \end{pmatrix}$$

where the ρ's are (auto) correlations. Generalized least squares (GLS) estimates $\hat{\beta}_{GLS}$ of β are obtained by minimizing the following generalized sum of squared residuals:

$$\epsilon^T \epsilon = (y - X\beta)^T \Sigma_y^{-1} (y - X\beta) \tag{2.6}$$

giving

$$\hat{\beta}_{GLS} = (X^T \Sigma_y^{-1} X)^{-1} (X^T \Sigma_y^{-1} y) \tag{2.7}$$

and

$$var(\hat{\beta}_{GLS}) = (X^T \Sigma_y^{-1} X)^{-1} \sigma_\epsilon^2 \tag{2.8}$$

Generalized least squares is relevant to multilevel models which can be considered as regressions with correlated errors. Such correlations are induced by the clustering of units in higher level groups as mentioned before. Indeed the multilevel estimation method used in the software MLWiN maintained at the Centre for Multilevel Modeling(see section 2.2) is based on GLS as we shall describe in section 2.8.

An important point to note here is that models with correlated errors are substantially more difficult to estimate than those with uncorrelated errors. For instance, we note that the OLS estimate $\hat{\beta}$ depends only on the data whereas $\hat{\beta}_{GLS}$ depends both on the data and the unknown variance-covariance matrix Σ_y that also needs to be estimated. We have not provided any derivation of OLS or GLS estimates as these are standard statistical methods covered in most linear models textbooks.

Having performed a linear regression in one class only, we now consider regressions in all eight classes. There are two analytical approaches that we may consider, the most obvious one being a single regression that ignores the grouping of students in classes - termed the *complete-pooling* analysis. Alternatively, we may also fit separate regressions, one for each of the eight classes - termed the *no-pooling* analysis. Figure (2.1) shows a plot of the fitted regression lines in each of the eight classes with the gray lines showing the no-pooling analysis and the single bold line, the complete-pooling analysis.

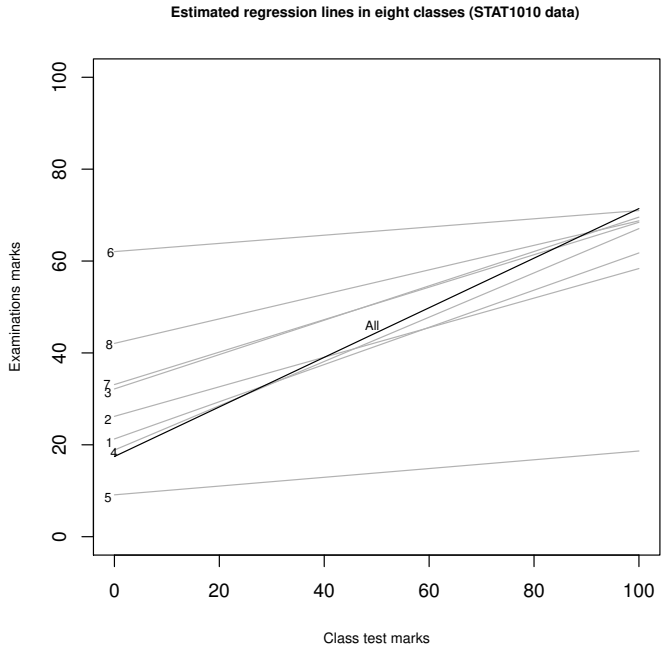


Figure 2.1: Separate regressions of classes 1 to 8 (grey lines) and a single regression for all classes (bold line labelled 'All')

It is evident from Figure (2.1) that there are some important variations in the predicted intercepts and slopes among the classes. On the one hand the complete-pooling analysis ignores these variations and therefore is not suitable if our aim is to understand variations between classes. The no-pooling analysis, on the other hand, exaggerates the variations in intercepts and slopes, more so if some classes have very few students, hence leading to inefficient estimates of the regression coefficients. Rubin (1980) calls this the *bouncing beta problem*. A compromise between no-pooling

and complete-pooling is *partial-pooling*, which is precisely what multilevel modeling does.

2.5.5 The simplest multilevel model

In the regression analyses of the eight classes above, the regression coefficients appear to vary. This variation in intercepts and slopes can be modelled via a multilevel model, the simplest of which allows the intercept only to vary by class. The model thus has only one intercept term, no predictors but two levels of variations: the usual student level variation (Level 1) and the school level variation (Level 2). Hence, the two-level random effects model is as follows.
At level 1 we have:

$$y_{ji} = \beta_{0j} + \epsilon_{ji} \tag{2.9}$$

and at level 2 we have:

$$\beta_{0j} = \mu + \alpha_j \tag{2.10}$$

where:

- Students are denoted by i, for $i = 1, 2, \ldots, n_j$ and classes, by j for $j = 1, 2, \ldots, J$. For the STAT1010 data, $J = 8$ classes and the first class has $n_1 = 41$ students.
- y_{ji} is the examinations mark of student i in class j
- μ is the underlying population average examinations mark.
- α_j is the effect of class j on examinations mark.
- ϵ_{ji} is the error or residual term of student i in class j.

The level-1 errors, ϵ_{ji}, are uncorrelated with variance σ^2_ϵ and represent variation in examinations marks between students; for example variation due to students' differing abilities in the STAT1010 examinations. The level-2 errors, α_j, are also uncorrelated with variance σ^2_α and represent variation in examinations marks between classes. The variances σ^2_ϵ and σ^2_α are jointly termed *variance components* hence the alternative name *variance components models* for multilevel models. The

probabilistic assumptions underlying the random effects model depend on the model fitting approach adopted; we discuss these assumptions when we discuss estimation in the next section.

Another way to write the two equations (2.9) and (2.10) is by combining them in the single equation form as follows:

$$y_{ji} = \mu + \alpha_j + \epsilon_{ji} \tag{2.11}$$

The single equation is more compact and applies to more complex multilevel models with several levels of variations and predictors at all these levels. In (2.11) the term in μ is called the *fixed effect* and the term in α_j, the *random effect*, hence the alternative name *linear mixed-effects model* for multilevel models.

There are four types of parameters that are of direct interest for estimation and interpretation of the model. These are namely the fixed effect μ, the level-2 residuals α_j, the level-2 variance σ^2_α, and the level-1 residual variance σ^2_ϵ. The level-1 residuals ϵ_{ji} are mostly estimated for diagnostic assessments of the fitted model, just as in the analysis of linear regression models. In the next sections we review estimation of the mean and variance components.

2.6 Estimation in multilevel modeling

Estimation in multilevel models is the central theme of this thesis. In the coming sections we shall therefore review some of the main approaches and methods commonly used for estimating parameters of multilevel models. There are three main approaches that are currently used to estimate mean and variance components in multilevel models: maximum likelihood estimation, Bayesian hierarchical modeling and bootstrap simulation. Our aim in reviewing these estimation methods is manifold.

We review the above three main estimation methods because they provide important concepts, illustrate difficulties in estimation, and may ultimately be used for comparison with our proposed estimation method, while some other methods we review, especially those based on least squares, are essential stages in the estimation

techniques we develop in this thesis. We also review the ANOVA estimation method as it illustrates the problem of negative variance estimates.

Finally, we also give an introduction to Bayes linear estimation as it is the basic methodology within which we develop our new formulation and estimation methods for both simple and complex multilevel models. For ease of exposure though, the methods we review here are illustrated via the two-level random effects model, but they apply equally well to more complex multilevel regression models.

2.7 ANOVA

One of the oldest and most popular variance estimation method, attributed to R.A. Fisher (1918), is the analysis of variance (ANOVA) method. The ANOVA estimators are obtained by equating quadratic functions of the observables to their expected values; the quadratic functions being the relevant sums of squares. Below we outline the ANOVA method for unbalanced data for the two-level random effects model. The corresponding estimators for the balanced situation are easily obtained by setting $n_j = n$.

2.7.1 ANOVA estimator of variance components for unbalanced data

Here we have different number (n_j) of observations in different groups (j). The appropriate sums of squares are given by:

$$\text{SSA} = \sum_j n_j(\bar{y}_{j.} - \bar{y}_{..})^2 \tag{2.12}$$

$$\text{SSE} = \sum_j \sum_i (y_{ji} - \bar{y}_{.j})^2) \tag{2.13}$$

$$\text{SST} = \sum_j \sum_i (y_{ji} - \bar{y}_{..})^2 \tag{2.14}$$

where SSA, SSE and SST are the between-groups, within-groups and total sums of squares respectively.

The ANOVA estimators are obtained by equating the sums of squares to their expected values. The expectations of the sums of squares are:

$$E(\text{SSA}) = (N - \sum_j n_j^2/N)\sigma_\alpha^2 + (J-1)\sigma_\epsilon^2$$

and

$$E(\text{SSE}) = (N-J)\sigma_\epsilon^2$$

Hence, the estimators are given by:

$$\begin{aligned} \hat{\sigma}_\epsilon^2 &= \frac{\text{SSE}}{(N-J)} \\ &= \text{MSE} \\ \hat{\sigma}_\alpha^2 &= \frac{\text{MSA} - \text{MSE}}{(N - \sum_j n_j^2/N)/(J-1)} \end{aligned} \tag{2.15}$$

where MSA and MSE are the between-groups and within-groups mean squares respectively. While the equation for level-1 estimator, $\hat{\sigma}_\epsilon^2$, is simple and similar for the balanced and unbalanced case, the level-2 estimator, $\hat{\sigma}_\alpha^2$, is more complex for unbalanced data when compared to balanced data.

2.7.2 Properties of ANOVA estimators

The ANOVA method does not require any underlying probability distribution and the resulting estimators are always unbiased. In the case of balanced designs, ANOVA estimators have minimum variance and they are also minimal sufficient (Searle et al., 1992). As such ANOVA has been widely used and studied. There are, however, two serious limitations of the method.

Firstly, ANOVA can yield negative estimates of variances. This is obvious from the expressions for the estimators for the level-2 variance component, $\hat{\sigma}_\alpha^2$, for either the balanced or unbalanced design. The estimator $\hat{\sigma}_\alpha^2$ will be negative whenever the variability within group (MSE) is larger than that between groups (MSA).

Secondly, in the case of unbalanced data, ANOVA estimators are no longer minimum variance and the distribution theory gets much more complicated even under

the usual normality assumption (Scheffe, 1959). Further, in contrast to the balanced case, there are no unique sums of squares to use in the ANOVA method.

Searle *et al.*(1992) describe more than a hundred years of research into variance estimation using the ANOVA methodology. They conclude that "*negativity of variance estimates, lack of distributional properties and no useful way to compare different applications of ANOVA methodology*" remain the main weaknesses of the ANOVA method.

In multilevel models with covariates occurring at several levels, the data is mostly unbalanced and there is a possibility of obtaining an estimate of a (Co-)Variance matrix containing negative variances, which is clearly undesirable. Also, accurate ANOVA estimation typically requires large data sets and is mostly used for variance component estimation in the context of experimental design rather than multilevel models.

2.8 Maximum likelihood: The principle and properties

Maximum Likelihood Estimation (MLE) is such a widely used estimation method that it may be considered a cornerstone of statistical inference. MLE requires the specification of a probability density function $p(y \mid \theta)$ where θ is the parameter to be estimated. In the traditional or classical approach, the parameter θ is considered as unknown but 'true' or 'fixed', that is not a random quantity. Given $y_1, \ldots, y_n$ are a *random sample* from the density function $p(y \mid \theta)$, the *joint sampling distribution* of the y_i's viewed as a function of θ is called the *likelihood function* and is written as:

$$L(\theta \mid y_1, \ldots, y_n) = \prod_{i=1}^{n} p(y_i; \theta)$$

The observed data $y_1, \ldots, y_n$ in the likelihood function is considered as fixed and therefore the likelihood function is often written simply as $L(\theta)$, suppressing the dependence on the data. The maximum likelihood estimator $\hat{\theta}$ is the value of θ

in Θ, the parameter space, that maximises the likelihood funtion $L(\theta)$. Therefore, MLE is deemed intuitive in that it chooses that value of the estimator that makes the data most plausible.

Apart from being intuitive, MLE has some other desirable properties. The invariance property of MLE is deemed useful as to find the MLE of $f(\theta)$, we only need to find the MLE of θ and plug it in $f(.)$. MLE though, may at times produce estimators that are biased or not uniformly minimum-variance. However, the large sample properties, such as consistency and asymptotic normality of MLE estimators, explain the widespread use of maximum likelihood estimation in numerous statistical estimation problems.

2.8.1 Maximum Likelihood Estimation of Multilevel Models

As mentioned above, to find the MLE of the parameters in the random effects model, we need to specify an appropriate probability density function. For continuous response variables such as y_{ji} in the random effects model, it is common to assume the normal distribution for the level-1 and level-2 residuals, namely $\epsilon_{ji} \sim \mathcal{N}(0, \sigma_\epsilon^2)$ and $\alpha_j \sim \mathcal{N}(0, \sigma_\alpha^2)$. It is then easy to write down the log-likelihood. For example, consider the more general multilevel regression model $y = X\beta + Zu + \epsilon$, where y is an n x 1 vector of outcomes, β is a p x 1 vector of fixed effects, X and Z are known matrices of explanatory variables, u and ϵ are level-2 and level-1 random effects respectively, with variances and covariances collected in the (Co-)Variance matrix Σ_y. The log-likelihood is

$$l = -\frac{1}{2}N - \frac{1}{2}log|\Sigma_y| - \frac{1}{2}(y - X\beta)^T\Sigma_y^{-1}(y - X\beta)$$

To find the ML estimators, we take partial derivatives of the log-likelihood with respect to the fixed effects β, and then with respect to the random-effects variances Σ_y. We thus obtain the respective ML equations, namely $\partial l/\partial\beta = 0$ and $\partial l/\partial\Sigma_y = 0$, which can be solved to obtain the desired estimators.

Solving $\partial l/\partial\beta = 0$ for the fixed effects poses no particular problem, as the elements of the vector β are unconstrained. However, solving $\partial l/\partial\Sigma_y = 0$ for the random-effects variances in Σ_y is not as straightforward, as variances are required to

be non-negative. In case one or more of the estimated random-effects variances are negative, a modification due to Herbach(1959) is applied which effectively replaces any negative estimate of variance by zero, see Searle et al. (1992) for details.

Explicit solutions to the ML equations only exist for balanced data. In the general multilevel model with covariates at one or more levels, the requirement for balance implies the same number of observations in each group and also the same number of observations for each predictor X (e.g., Bryk and Raudenbush, 1992). Such conditions however, are unlikely to be met in practice. Given unbalanced data, the solutions to the ML equations are not tractable analytically, necessitating iterative solutions. One such iterative procedure is explained in Section 2.9 below.

Restricted Maximum Likelihood (REML)

It is well known that Maximum Likelihood estimation may produce biased estimates, specially in small samples. In multilevel model estimation, ML estimates the random-effects variances conditional upon the fixed effects β, and as the ML procedure treats the estimated $\hat{\beta}$ as fixed, it thus does not account for its sampling variation. Random-effects variances will be more severely underestimated when sample sizes are small, and fewer degrees of freedom are available for estimation.

The solution to this problem is to apply ML to a linear transformation of y that is free of β, rather than to y directly. One such transformation is to use OLS to obtained estimates of the residuals, which are then used to estimate the variance components. This modification of ML is called restricted maximum likelihood or REML see Patterson and Thompson (1971).

2.9 Iterative Generalized Least Squares

Goldstein (1986) proposed an iterative generalized least squares (IGLS) procedure for estimating parameters in complex multilevel models, including models for longitudinal and multivariate data structures. The method is general enough, and can therefore fit a wide variety of multilevel models as implemented in the general purpose multilevel modeling package MLwiN. Goldstein (1986) showed that IGLS is

equivalent to MLE under the Gaussian assumption. Below we outline the general principles of the IGLS method following Goldstein (1995).

As its name says, IGLS involves the repeated application of generalized least squares (GLS)(see Sub-section 2.4). Consider the general linear model $y = X\beta + \epsilon$, where ϵ is a vector of random effects, with elements $(\alpha_j, \epsilon_{ji})$ for the random effects model for example, and (Co-)Variance matrix Σ_y, with corresponding elements $(\sigma^2_\alpha, \sigma^2_\epsilon)$ for the same random effects model. IGLS then proceeds as follows:

1. Step 1: Estimation of fixed effects.

   If Σ_y were known, then GLS of $y = X\beta + \epsilon$ would yield the estimator $\hat{\beta}_{GLS} = (X^T\Sigma_y^{-1}X)^{-1}(X^T\Sigma_y^{-1}y)$ for the fixed effect vector β.

2. Step 2: Estimation of random effects.

   If β were known, then GLS of of a new linear model $y^* = X^*\beta^* + \epsilon^*$ would yield the estimator $\hat{\beta^*}_{GLS} = (X^{*T}(\Sigma_{y^*})^{-1})X^*)^{-1}(X^{*T}(\Sigma_{y*})^{-1})y^*)$ for the random effects Σ_y.

In step 2 above, y^* is the vector of squares and products of the residuals $\epsilon\epsilon^T = (y - X\beta)(y - X\beta)^T$, stacked as a column vector (via the *vech* matrix operator). Σ_{y^*} is the covariance matrix of the vector y^*, that is the covariance of squares and products of the residuals, hence containing fourth-order moments as required for variance estimation. X^* is the design matrix linking y^* to Σ_y in the new linear model.

The initial estimates of the fixed effects β required to start IGLS are obtained via an application of OLS giving $\hat{\beta}_{OLS}$. Raw residuals (not the correct residuals, since OLS ignores group effects) are then calculated as $(y - X\hat{\beta}_{OLS})(y - X\hat{\beta}_{OLS})^T$ and used in step 2 in the iterative procedure above. IGLS then iterates between steps 1 and 2 until convergence of the estimated fixed and random parameters.

Restricted Iterative Generalized Least Squares (RIGLS)

IGLS, being equivalent to MLE, gives biased and underestimated variance components in small samples. Goldstein (1989) showed that

$$E((y - X\hat{\beta}_{OLS})(y - X\hat{\beta}_{OLS})^T) = \Sigma_y - X(X^T\Sigma_y^{-1}X)^{-1}X^T$$

and proposed using $(y - X\hat{\beta}_{OLS})(y - X\hat{\beta}_{OLS})^T + X(X^T\hat{\Sigma}_y^{-1}X)^{-1}X^T$ in step 2 of IGLS to obtain unbiased estimates of the variance components, where $\hat{\Sigma}_y^{-1}$ is the current estimates of Σ_y^{-1}. The thus modified IGLS is termed Restricted Iterative Generalized Least Squares (RIGLS), and under multivariate normality, is equivalent to REML(Goldstein (1989)).

2.10 Bayesian hierarchical modeling

The Bayesian Approach to Estimation

The estimation methods we have considered so far are termed frequentist or classical, especially when we wish to make a distinction between these methods and a different approach to statistical inference, called the Bayesian approach. While in the classical approach probability statements are only allowed for data given parameters, in the Bayesian approach in contrast, uncertainty about all quantities (including fixed but unknown parameters) can be represented probabilistically. Suppose, for example, we are interested in the unknown parameter θ and intend to collect data to learn about it. We quantify our uncertainty about θ by assigning it the prior probability $p(\theta)$. We also assign the conditional probability $p(y|\theta)$, to express our uncertainty about the future data we intend to collect given θ. The conditional probability $p(y|\theta)$ is in fact the likelihood function of θ. Using Bayes theorem, we combine the prior and likelihood to obtain the posterior probability $p(\theta|y)$. Inferences about θ, such as its mean, median and standard deviation, are then based on $p(\theta|y)$.

In his *Philosophy of Statistics*, Lindley (2000), the eminent Bayesian statistician Dennis Lindley summarises the above description of the Bayesian approach as follows:

- Statistics is the study of uncertainty
- Uncertainty should be measured by probability
- Data uncertainty is so measured, conditional on parameters
- Parameter uncertainty is similarly measured by probability

- Inference is performed within the probability calculus, mainly via Bayes theorem

The important point here is that uncertainty is measured by probability which, in the Bayesian approach, is defined as the degree of belief of an individual in an event or proposition. Hence, probability is personal or subjective. Subjectivity is discussed in greater detail by Ramsey (1926), de Finetti (1937), and in Lad (1996). For a book-length careful treatment of the Bayesian approach see Bernardo and Smith (1994) and Robert (2001).

Exchangeability and the Representation Theorem

In the analysis of scientific data, probability models are useful in tackling the important problem of predicting a future value y_{n+1} given a sequence of *observable* outcomes $\{y_1, \ldots, y_n\}$. In order to predict y_{n+1}, it is necessary and sufficient to assess the form of the joint probability $p(y_1, \ldots, y_n)$ for any n since, $p(y_{n+1}|y_1, \ldots, y_n) = p(y_1, \ldots, y_{n+1})/p(y_1, \ldots, y_n)$. For example, one may assume the y_i's to be independent. In this case no learning takes place as $p(y_{n+1}|y_1, \ldots, y_n) = p(y_{n+1})$. As argued by Bernardo and Smith (1994), some form of dependencies must be assumed among the y_i's, and there is a large number of such dependencies. One commonly used simplifying assumption, is the judgement that the future y_{n+1} and the sequence $\{y_1, \ldots, y_n\}$ form an *exchangeable* sequence of random quantities. Exchangeability was introduced by de Finetti (1937) and is equivalent to a judgement of similarity or symmetry which implies that $p(y_1, \ldots, y_n) = p(y_{\pi(1)}, \ldots, y_{\pi(n)})$ for all permutations π defined on the set $\{1, \ldots, n\}$.

The assumption of exchangeability has an important consequence regarding the probability models we are interested in. If we judge our observations as (infinitely) exchangeable, then we may apply *de Finetti's representation theorem for exchangeable sequences*, which shows that exchangeable observations should be regarded as a random sample from some probability distribution, and there exists a prior probability distribution over the parameter of the model. de Finetti (1937) stated and proved the representation theorem for binary, 0 - 1, random variables. Bernardo and Smith (1994, Ch.4) provide good coverage of various types of exchangeability

assumptions, representation theorems and the ensuing models. For a more recent interesting and accessible discussion of the logic and importance of exchangeability and the representation theorem, see Goldstein (2012).

The representation theorem has been subsequently generalized by Hewitt and Savage (1955) for real-valued random quantities, including the outcome variables y_i of interest to us here. Bernardo and Smith (1994) also give a version of the general representation theorem (see Proposition 4.3) which we adapt as follows.

Theorem 1. *If $y_1, y_2, \ldots,$ is an infinite exchangeable sequence of real-valued random quantities with probability measure P, there exists a probability measure Q over $\Im$, the space of all distribution functions on $\Re$, such that the joint distribution function of $y_1, \ldots, y_n$ has the form*

$$P(y_1, \ldots, y_n) = \int_{\Im} \prod_{i=1}^{n} F(y_i) dQ(F)$$

where

$$Q(F) = \lim_{n \to \infty} P(F_n) \tag{2.16}$$

and F_n is the empirical distribution function defined by $y_1, \ldots, y_n$.

Bernardo and Smith (1994) argue that Theorem 1 is difficult to apply explicitly in learning about a future y_{n+1} given $y_1, \ldots, y_n$, because of the infinite-dimensional,unknown distribution function F. They thus propose a more restrictive representation theorem in terms of density functions, labelled by a finite dimensional parameter, θ.

Corollary 1. *Assuming the required densities to exist, under the conditions of Theorem 1, the joint density of $y_1, \ldots, y_n$ has the form*

$$p(y_1, \ldots, y_n) = \int_{\Theta} \prod_{i=1}^{n} p(y_i|\theta) p(\theta) d(\theta) \tag{2.17}$$

where $p(.|\theta)$ denotes the density function corresponding to the unknown parameter $\theta \in \Theta$.

It is now straightforward to apply Bayes' theorem to learn about y_{n+1}.

So far we have considered exchangeability judgements for simple sequences of observable random quantities $\{y_1, y_2, \ldots\}$, corresponding to single level, not multilevel, data. The type of exchangeability judgements considered for such single level data induces complete symmetry, in the sense that our beliefs over the observable quantities is unchanged for any permutation of the subscripts. For multilevel data, which is the main focus of this thesis, the group or context is obviously important and, hence, we cannot assume our jugements to be invariant to all permutations of the subscripts of the sequence of observations; we need to restrict our exchangeability judgements in order to fully account for the multilevel structure of the data. In other words, in order to consider the meaningful subscripts of richly structured data, judgements of partial symmetry may be more appropriate and have been termed *partial exchangeability* judgements.

Partial exchangeability judgements for several examples of complex data structures, including multilevel data, are discussed in Bernardo and Smith (1994). They argue that many possible forms of partial exchangeability judgements may be contemplated, depending on the specificities of the data structures, and therefore it is difficult to give an all-embracing definition of partial exchangeability. Indeed several authors have used particular forms of partial exchangeability judgements to arrive at models suitable to the specific data under their considerations. For example, Lauritzen (2003) defines *row-column exchangeability* of binary matrices to derive the Rasch binary logistic regression model for item analysis (Rasch,1960). While, in the important area of Bayesian Nonparametrics, Rodriguez et al.(2008) use *unrestricted exchangeability* of exchangeable sequences, to arrive at a Nested Dirichlet Process model suitable to analyse data arising from multicenter studies, hence multilevel data.

Unrestricted exchangeability is a reasonable assumption for complex data that often comprise several related sequences of random quantities; the dependence between the sequences may be induced by a hierarchical or multilevel structure. For the two-level data that we considered earlier, where individual i is nested in group j, the outcome variable typically comprises the sequence $\{y_{11}, \ldots, y_{1n_1}, \ldots, y_{J1}, \ldots, y_{Jn_J}\}$, which is made up of the J related individual sequences $\{(y_{11}, \ldots, y_{1n_1}), (y_{21}, \ldots, y_{2n_2}),$

$\ldots, (y_{J1}, \ldots, y_{Jn_J})\}$. Unrestricted exchangeability implies complete exchangeability of the random quantities *within* each individual sequence along with exchangeability *between* these exchangeable sequences.

Bernardo and Smith (1994; Section 4.6.5) also propose unrestricted exchangeability of exchangeable sequences as appropriate judgements for complex data and give a joint density representation for several sequences of random quantities which we adapt as follows. The J unrestrictedly infinitely exchangeable sequences of random quantities explained above admit a representation of the form

$$p(y_{11}, \ldots, y_{1n_1}, \ldots, y_{J1}, \ldots, y_{Jn_J}) = \int_{\Theta} \prod_{j=1}^{J} \prod_{i=1}^{n_j} p(y_{ji}|\theta_j) p(\theta_1, \ldots, \theta_J) d(\theta_1), \ldots, d(\theta_J) \tag{2.18}$$

where, for each $j = 1, 2, \ldots, J$, $\theta_j \in \Theta$ is the limit, as $n_j \to \infty$, of some function of y_{ji}. For example, θ_j could be the group means and standard deviations (μ_j, σ_j).

If we now judge that the parameters $\{\theta_1, \theta_2, \ldots, \}$ also form a sequence of infinitely exchangeable quantities, we may also write the following representation.

$$p(\theta_1, \ldots, \theta_J) = \int_{\Phi} \prod_{j=1}^{J} p(\theta_j|\phi) p(\phi) d(\phi) \tag{2.19}$$

where ϕ is termed a *hyperparameter*. The hyperparameter can be identified with appropriate (strong law) limits of observables, just as we mentioned for the θ_j in (2.23) above. For a complete example of hierarchical modelling with specifications of all prior and hyperprior distributions, see Bernardo and Smith (1994; p224). To quote Bernardo and Smith (1994; p226)

> Hierarchical modelling provides a powerful and flexible approach to the representation of beliefs about observables in extended data structures, and is being increasingly used in statistical modelling and analysis.

Several excellent book-length treatments of Bayesian hierarchical modelling of multilevel data exist, see for example Gelman et.al (2013), Congdon (2003).

2.11 Applying Bayesian hierarchical modelling: prior and posterior densities

So far we have seen that the subjective approach to modelling complex multilevel data requires the use of partial exchangeability judgements and the appropriate representation theorem to formulate a suitable Bayesian hierarchical model. In practice, Bayesian fitting of hierarchical models requires (i) specification of suitable prior (and hyperprior) distributions for the model parameters and, (ii) updating from prior to posterior via the likelihood function. We briefly discuss the challenges associated with (i) and (ii) as they raise some important issues that have motivated us to consider the alternative estimation methods we propose in this thesis.

Choice of prior

Consider the simple two-level random effects model $y_{ji} = \mu + \alpha_j + \epsilon_{ji}$. The usual Gaussian assumptions required for fitting the model are:

$$y_{ji} \sim N(\mu + \alpha_j, \sigma_\epsilon^2), \quad \text{for } i = 1, 2, \ldots, n_j \text{ and } j = 1, 2, \ldots, J \tag{2.20}$$

$$\alpha_j \sim N(0, \sigma_\alpha^2), \quad \text{for } j = 1, 2, \ldots, J \tag{2.21}$$

Prior distributions are required for the hyperparameters μ, σ_α^2 and σ_ϵ^2. In principle, estimation of μ and σ_ϵ^2 pose no great problem, as sufficient data are usually available at level-1 of a hierarchy for their estimation. Hence, a *noninformative* prior, of the form $p(\mu, \sigma_\epsilon) \propto 1$, that expresses prior ignorance or that 'will let the data speak for themselves' (see Bernardo and Smith, p.357), is often assumed.

Estimation of the level-2 variance σ_α^2 however, is fraught with difficulties, more so when there are few level-2 units, J, or when the level-2 variance σ_α^2 is close to zero . In the context of Bayesian inference of variance components of the random effects model, Hill (1965) writes 'the analysis of variance opens a Pandora's box of problems which constitute a real challenge to any and all statisticians and theories of statistical inference'.

The difficulties involved in the construction of a suitable prior distribution for σ_α^2 relate to the possibility that the estimate of this level-2 variance parameter may turn out to be negative in the classical approach (see Section 2.6). The Bayesian

approach avoids this problem of a negative estimate of variance by constructing priors for σ_α^2 that place zero probability on negative values of σ_α^2. Examples of such priors include a uniform prior on σ_α or on $log(\sigma_\alpha)$ and, the inverse-gamma (ϵ, ϵ) prior, such as $p(\sigma_\alpha^2) \propto$ inverse-gamma(0.001,0.001) as used in the Winbugs package.

Use of priors that avoid negative variance estimates however, is not without problems, some of which include: the requirement of at least three groups ($J \geq 3$) to enable inference, overestimates of σ_α^2 for small J, improper posterior distributions (not integrating to one as required for any pdf), posterior inferences overly sensitive to the parameters of the prior distribution (such as ϵ in the inverse-gamma prior). These issues are discussed more fully in Gelman(2005).

The inverse-gamma (ϵ, ϵ) prior is of interest because it belongs to the family of the *conditionally conjugate* priors. Conjugacy is an important concept in Bayesian analysis. Given a likelihood $p(y|\theta)$, then the family of prior densities $p(\theta|u)$, where u is some collection of parameters, is a conjugate prior family with respect to the likelihood, if the posterior density $p(\theta|y)$ belongs to the same conjugate family for every sample size and every set of possible sample values. For example, if y represents the number of successes in n independent Bernoulli trials with (unknown) probability of success θ, then $y \sim$ Binomial(θ, n). The Beta(α, β) prior on θ is the natural conjugate as the posterior $p(\theta|y, n)$ also has a Beta density as follows:

$$\begin{aligned} p(\theta|y,n) &\propto \theta^y(1-\theta)^{n-y}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ \text{posterior} &\propto \text{likelihood} \times \text{prior} \\ &= \theta^{y+\alpha-1}(1-\theta)^{n-y+\beta-1} \end{aligned}$$

This simple example illustrates two important advantages of conjugacy. First, a conjugate prior density allows tractable calculations of the appropriate posterior density. Equally important, conjugate analysis allows the information in the prior to be interpreted in terms of equivalent data. As an examination of the posterior density in the above example shows, it is as if the prior is contributing the data equivalent of $(\alpha - 1 + \beta - 1)$ trials to the existing n trials in the sample.

As we mentioned above the inverse-gamma density prior for σ_α^2 is conditionally conjugate in the sense that the conditional posterior $p(\sigma_\alpha^2|y, \alpha, \mu, \sigma_\epsilon)$ has an inverse-

gamma density. Conditional conjugacy is useful because it is preserved when a model is expanded hierarchically (Gelman, 2005), while conjugacy is not.

Use of conjugate prior densities for tractability in the calculation of posterior densities or for convenience, can be quite restrictive, especially for models as rich as those considered in Bayesian hierarchical modelling. Fortunately, developments in computer power, as well as in computational algorithms, such as Markov Chain Monte Carlo methods, have obviated the need to restrict modelling to conjugate priors only, and have thus enlarged the range and depth of applications of Bayesian methods to solve real-life problems. We discuss computational issues next.

Computation of posterior

Bayesian inference is based on the calculation of marginal posterior densities of parameters, often a difficult task that may involve the evaluation of multi-dimensional integrals (see Section 2.10). Prior to the 1990's, the difficulties involved in the calculation of posterior densities hampered the application of Bayesian methods (Gelfand et.al.,1990). Numerical approximation methods, such as the Laplace Approximation and Iterative Quadrature amongst others, were developed during the 1980's, but their implementation was not straightforward. For a summary of these numerical approximation methods, see Bernardo and Smith (1994).

Substantial progress in Bayesian computation was made when Gelfand and Smith (1990) popularized the Gibbs sampler. The simplest implementation of the Gibbs sampler is in the following situation. Suppose the marginal posterior density, say $p(\theta, u|y)$, is difficult to calculate but the conditional posterior densities, $p(\theta|y, u)$ and $p(u|y, \theta)$, have nice closed forms. For example, $p(\theta|y, u)$ may be Gaussian and, $p(u|y, \theta)$ may have a gamma distribution. Then, after choosing a suitable starting value, say u^0 for u, random sampling from the Gaussian distribution $p(\theta|y, u = u^0)$ yields θ^1 (where the superscript 1 denotes the first sampled value of θ). Next, sampling from the gamma distribution $p(u|y, \theta = \theta^1)$ yields u^1. If this algorithm is run t times, then (θ^t, u^t), which is a realization of a Markov Chain, tends in distribution as $t \to \infty$, to a random vector whose joint distribution is the target density sought, i.e. $p(\theta, u|y)$. Hence, if the procedure for obtaining (θ^t, u^t) is replicated a large number of times, an estimate $\hat{p}(\theta, u|y)$ for $p(\theta, u|y)$ can easily be obtained.

Applying the Gibbs sampler to multilevel models follows the same principle just explained above. Gelfand et al. (1990) showed that it is relatively straightforward to perform Gibbs sampling of multilevel models by assuming conditional independence of the parameters of interest, as well as conjugate priors for each of these parameters. Consider our random effects model, $y_{ji} = \mu + \alpha_j + \epsilon_{ji}$. Assuming independent Gaussian priors for μ and α_j, the posterior densities $p(\mu|y, \alpha, \sigma_\alpha^2, \sigma_\epsilon^2)$ and $p(\alpha|y, \mu, \sigma_\alpha^2, \sigma_\epsilon^2)$ are also Gaussian. And, assuming independent inverse gamma priors for σ_ϵ^2 and σ_α^2, the posterior densities $p(\sigma_\epsilon^2|y, \mu, \alpha, \sigma_\alpha^2)$ and $p(\sigma_\alpha^2|y, \mu, \alpha, \sigma_\epsilon^2)$ are also inverse gamma. For more complex multilevel models, such as multilevel regressions, the group effects are assumed to have a multivariate normal distribution. The mean group effects are given a multivariate normal prior while the inverse covariance matrix of the group effects is assumed to follow a Wishart distribution (see Seltzer et al., 1996).

Programming the above Gibbs samplers would be uncomplicated in, for example, the R Statistical Language. But the Gibbs sampler, along with more general algorithms such as the Metropolis-Hastings algorithm (Hastings, 1970), have been implemented in softwares like WinBUGS. Together these stochastic simulation algorithms are termed Markov Chain Monte Carlo (MCMC) methods, and their developments have facilitated Bayesian estimation of complex models, including multilevel models.

2.12 Some difficulties of a fully Bayesian approach

The fully Bayesian approach to multilevel modelling described above, has important conceptual and methodological benefits. From a conceptual viewpoint, acknowledging the hierarchical structure in complex data induces judgements of exchangeability within and between groups and thus enables the borrowing of strength (Section 2.3) to effect improved inferences. From a methodological viewpoint, the Bayesian approach accounts fully for all sources of uncertainty in complex data structures, thereby leading to the formulation of richly structured hierarchical models that can be estimated via MCMC methods. In practice though, the fully Bayesian approach, like other statistical approaches, has its own limitations regarding the specification

of prior densities, computation of posterior densities, and in the design of multilevel studies.

2.12.1 Prior specification

A major stumbling block of the approach is the specification of a prior density. It is indeed difficult to elicit genuine prior densities that could capture an expert's subjective beliefs about important aspects of even a moderately complex real-life problem. For this reason, and also for tractability reasons, Bayesian analysts have used conjugate priors, though they may not represent their true subjective beliefs.

2.12.2 Computation of posterior

On the computation side, though MCMC has been a great advancement, these methods are very computer intensive. And, for complex models with many parameters, MCMC may take an exceedingly long time to converge to the target posterior density. For example, the Stan development team reports that 'a multilevel time series regression of climate on tree-ring measurements wasn't converging after hundred of thousands of iterations' (Stan reference manual, 2013). Stan is the latest high-performance software for Bayesian inference of multilevel models, and it is based on Hamiltonian Monte Carlo (Neal, 2011). In another example, Browne and Draper (2006) report months of CPU time used in a comparison of Bayesian (MCMC) and likelihood estimation of simple multilevel models.

2.12.3 The design of multilevel studies

Sampling design, like the design of experiments, is an important but complex task. This complexity stems from the designer having to consider the potentially many sources of variation, uncertainty, cost and ethical constraints, as well as subject matter knowledge and prior expertise. Such a multifaceted enterprise makes the Bayesian approach appealing. A key component of the sample design problem is sample size determination (SSD). The fully Bayesian method applied to the SSD

problem has resulted in two main approaches : a utility-based and a performance-based approach. Both approaches are explored in a series of articles in *The Statistician* (46 2,1997). Here we focus on the utility-based approach.

Lindley (1997) adapts the decision-theoretic approach of Raiffa and Schlaifer (1961) to the SSD problem as follows. Consider a random quantity X, with density $p(X|\theta)$, where the parameter θ is unknown. To make a decision d about θ, we observe n independent, identically distributed realizations $(x_1, x_2, \ldots, x_n)$ of X, which we denote by x. A crucial component of this approach is the specification of a utility function $u(n, x, d, \theta)$ describing the merit of choosing the sample size n, obtaining the result x and taking the decision d, when the parameter has the value θ. After specification of the prior $p(\theta)$, the optimum sample size is given by Lindley(1997) as

$$\max_n \Big[\sum_x \max_d \Big\{ \int u(n, x, d, \theta) p(\theta|d, x, n) d\theta \Big\} p(x|n) \Big] \tag{2.22}$$

The above Bayesian solution seems pertinent, as choosing between sampling designs is essentially a decision problem. In addition, specifying a utility function is most appropriate, as it balances the cost and performance of sampling. Furthermore, Lindley (1997) argues that *only* solutions to (2.22) above are coherent but he also agrees there are implementation difficulties.

Maximizing the above expression, especially for a multilevel model, is likely to be a difficult computational task. Even if we resort to MCMC methods, then the potentially large space of (x, d, θ) will need to be explored to find the optimum sample size. Specifying the utility function is no simple task either. Apart from the usual cost of taking additional samples, there are ethical costs also, whether in multilevel studies in educational research (involving pupils) or medical research (involving patients), making the utility function distinct for each design situation.

There are additional complications in the SSD problem for multilevel models. There is need to determine the optimum sample size for the different levels of the hierarchy, taking into consideration explanatory variables occuring at the various levels. For example, in school effectiveness studies the researcher must not only decide whether to investigate many schools with few students per school or few schools with many students per school, but also how many private or public schools

and male or female pupils to sample. With many levels and many variables 'there may be an excessively large number of possible designs from which to choose and no clear rules to guide our search' (Farrow and Goldstein,(1992)). In the related context of experimental design, Farrow and Goldstein,(1993) write '... full Bayes designs are notoriously intractable even putting aside the difficulty of making meaningful complete prior specifications.'

In view of the above problems in prior specification, computation and design, it is pertinent to consider an alternative Bayesian approach, one that requires only limited beliefs specification, and that is less computer intensive, such as the Bayes linear approach.

2.13 Introduction to Bayes linear methods

Goldstein and Wooff (1995) give a succint description of the Bayes linear approach as follow:

> The Bayes linear approach, which is based on partial belief specification with expectation as primitive, allows the straightforward construction of models reflecting second-order exchangeability.

The approach allows not only the construction of models, but also model fitting and diagnostics. [B/D], a free software for implementing the Bayes linear approach has also been developed for implementing the approach. Hence, the Bayes linear approach is a comprehensive methodology that can be applied to complex real-world problems within the subjectivist Bayesian paradigm. For a detailed exposition of the Bayes linear approach, including a description of [B/D], see the book *Bayes Linear Statistics Theory and Methods* by Goldstein and Wooff (2007).

A distinctive feature of the Bayes linear approach is that expectation, rather than probability, is used as the primitive measure of uncertainty. This is contrary to the two major approaches of Statistics, namely classical and (full) Bayesian, where expectation of an event can only be calculated after specification of all probability statements. And this is precisely what the Bayes linear approach seeks to avoid: the specification of detailed probabilities which, in realistically complex problems, may

be difficult to achieve. Treating expectation as primitive, as opposed to probability, formed the basis of the development of the subjectivist theory in de Finetti (1974).

So, if expectation is to be the primitive expression of uncertainty, then how exactly do we formulate our prior knowledge? We are required to specify only partial beliefs, in terms of means, variances and covariances. This greatly simplifies the specification task, especially in complex multilevel problems with uncertainty occuring at several levels of a hierarchy. In addition, the other important aspects of a fully Bayesian approach, such as parameter estimation, model fitting and exchangeability, are also simplified as we shall explain shortly. But more importantly, the Bayes linear approach offers an alternative to the estimation of variance components, as well as a two-stage Bayes linear analysis of mean components, so crucial to multilevel modelling, and which is at the heart of this thesis.

2.13.1 Adjusting beliefs

We present below some of the fundamentals of the Bayes linear approach. Following Goldstein and Wooff (1995), suppose we make the partial beliefs specifications for means, variances and covariances for a collection C of random quantities. We collect our beliefs about the random quantities in $B = \{B_1, B_2, \ldots, B_r \in C$. We intend to collect data $D = \{D_0, D_1, D_2, \ldots, D_k\} \in C$ (where $D_0 = 1$) in order to learn about B. For example, we may want to learn about the regression coefficients $B = \{\beta_0, \beta_1\}$ in the single level regression in Class 1 of the STAT1010 example (Section 2.5.4), after we observe $D = \{Y_1, Y_2, \ldots, Y_{41}\}$. The *adjusted expectation* of B given D, written $E_D(B)$ is the linear combination

$$E_D(B) = \sum_{i=0}^{k} h_i D_i$$

which minimizes

$$E\Big(\Big[B - \sum_{i=0}^{k} h_i D_i\Big]^2\Big)$$

Definition 2.13.1. *The* **adjusted expectation** *of a collection of random quantities B, given observation of a collection of quantities D, written $E_D(B)$ is*

$$E_D(B) = E(B) + Cov(B, D)Var(D)^{\dagger}(D - E(D)) \tag{2.23}$$

where $Var(D)^\dagger$ is the Moore-Penrose generalized inverse, with $Var(D)^\dagger = Var(D)^{-1}$ for non-singular $Var(D)$. $E_D(B)$ is called the **Bayes linear rule for B given D**.

Definition 2.13.2. *The* **adjusted variance** *of B given D, written* $Var_D(B)$ *is*

$$Var_D(B) = Var(B) - Cov(B, D)Var(D)^\dagger Cov(D, B) \tag{2.24}$$

where the value of $Var_D(B)$ depends only on prior variances and covariances, not on the data.

Definition 2.13.3. *The* **variance of** B resolved by D, *written* $RVar_D(B)$ *is*

$$RVar_D(B) = Cov(B, D)Var(D)^\dagger Cov(D, B) \tag{2.25}$$

Definition 2.13.4. *The* **adjusted covariance** $Cov_D(B_1, B_2)$ *is*

$$Cov_D(B_1, B_2) = Cov(B_1, B_2) - Cov(B_1, D)Var(D)^\dagger Cov(D, B_2) \tag{2.26}$$

Interpretations of the above beliefs adjustments are fully discussed in Goldstein and Wooff (1995), page 58. An important point made is that adjusted expectations may be viewed as tractable approximations to a full Bayes analysis. From this perspective, it is worth investigating adjusted expectations as viable alternatives to the full Bayesian hierarchical modelling of multilevel data, specially given the highly computer intensive methods required in designing multilevel studies (see Section 2.11). Also, as we shall show in Chapter 4, Bayes linear methods yield exactly the same results as a full Bayesian approach under Gaussian assumptions.

The adjusted variance $Var_D(B)$ may be interpreted as the mean square error of $E_D(B)$. $Var_D(B)$ however, does not depend on the data D, only on the prior specifications. Given the importance of variance components estimation in multilevel modelling, the Bayes linear update of variances based on D will also be reviewed below.

2.13.2 Second-order exchangeability

In Section 2.10 we explained the importance of judgements of exchangeability in the Bayesian approach. In a similar vein, Goldstein and Wooff (2007) remark that

> ...exchangeability is the fundamental judgement which gives meaning to the kinds of assumptions and modelling which characterize the usual types of statistical analysis.

but, as they subsequently argue, in practice it is difficult to make detailed prior specifications over all observables in order to apply de Finetti's exchangeability representation. Because the Bayes linear approach requires only limited beliefs specification, the restricted form of exchangeability thereby induced is exploitable, both in principle and in practice, in formulating statistical models. This form of exchangeability is termed second-order exchangeability which we now define.

Definition 2.13.5. *The collection of vectors $Z_1, Z_2, Z_3, \ldots$ is* **second-order exchangeable** *if the first- and second-order belief specification for this collection is unaffected by any permutation of the order of the vectors, so that*

$$E(Z_i) = \mu \quad Var(Z_i) = \Sigma \quad \forall i; \quad Cov(Z_i, Z_j) = \Gamma, \quad \forall i \neq j.$$

Second-order exchangeability leads to the representation theorem for infinite second-order exchangeable quantities as stated and proved in Goldstein (1986)).

Theorem 2.13.1. *If $Z_1, Z_2, Z_3, \ldots$ is an infinite second-order exchangeable sequence of random quantities, with mean and variance structure as in the above definition, then we may introduce the further random quantity $\mathcal{M}(Z)$, termed the* **population mean vector***, and also the infinite sequence $\mathcal{R}_1, \mathcal{R}_2, \ldots,$ termed the* **individual residual vectors** *We may then write:*

$$Z_i = \mathcal{M}(Z) + \mathcal{R}_i(Z)$$

where the population mean $\mathcal{M}(Z)$ has the following moments:

$$E(\mathcal{M}(Z)) = \mu \qquad Var(\mathcal{M}(Z)) = \Gamma$$

and the residuals $\mathcal{R}_i(Z)$ are themselves second-order exchangeable with

$$E(\mathcal{R}_i(Z)) = 0 \qquad Var(\mathcal{R}_i(Z)) = \Sigma - \Gamma$$

where $\mathcal{R}_1, \mathcal{R}_2, \ldots,$ are mutually uncorrelated and $\mathcal{R}_i(Z)$ is uncorrelated with $\mathcal{M}(Z)$ for each i.

2.14 Adjusting exchangeable quantities

In the coming chapters we shall construct multilevel models using second-order exchangeability assumptions and the representation theorem. We shall subsequently adjust beliefs about underlying population mean and variance components of the exchangeable models. Adjustments of exchangeable models can be considerably simplified via Bayes linear sufficiency, as explained below.

2.14.1 Adjusting mean components: Bayes linear sufficiency

Consider once again the collection of second-order exchangeable random vectors $Z_1, Z_2, Z_3, \ldots$ in Definition 2.13.5. For any individual i, the prior mean vector is $E(Z_i) = \mu$, and variance matrix is $Var(Z_i) = \Sigma$. For any two individuals $i \neq j$, the covariance matrix is $Cov(Z_i, Z_j) = \Gamma$. By Theorem 2.13.1 we construct the representation

$$Z_i = \mathcal{M}(Z) + \mathcal{R}_i(Z)$$

where $\mathcal{R}_1, \mathcal{R}_2, \ldots,$ are mutually uncorrelated and $\mathcal{R}_i(Z)$ is uncorrelated with $\mathcal{M}(Z)$ for each i. We shall collect data $D_n = \{Z_1, Z_2, Z_3, \ldots, Z_n\}$ in order to adjust beliefs about the population means and variances. To simplify notations, the latter adjustments will be written as follows:

$$E_n(\mathcal{M}(Z)) = E_{D_n}(\mathcal{M}(Z)) \qquad Var_n(\mathcal{M}(Z)) = Var_{D_n}(\mathcal{M}(Z))$$

The following theorem, from Goldstein and Wooff(2007), simplifies the adjustment of beliefs over exchangeable models. Let $Z_1, Z_2, Z_3, \ldots$ be an infinite second-order exchangeable sequence of random vectors. Given a sample $D_n = \{Z_1, \ldots, Z_n\}$, the sample mean vector

$$\bar{Z}_n = \frac{1}{n}\sum_{i=1}^{n} Z_i$$

is Bayes linear sufficient for adjusting both $\mathcal{M}(Z)$ and any values Z_i, $i \geq n$ namely

$$E_n(\mathcal{M}(Z)) = E_{\bar{Z}_n}(\mathcal{M}(Z)) \qquad Var_n(\mathcal{M}(Z)) = Var_{\bar{Z}_n}(\mathcal{M}(Z))$$

$$Var_n(Z_i) = Var_{\bar{Z}_n}(Z_i) = Var_{\bar{Z}_n}(\mathcal{M}(Z)) + Var(\mathcal{R}_i(Z))$$

Therefore in order to adjust beliefs over the mean vector and future observations, it is sufficient to adjust $\mathcal{M}(Z)$ by the sample mean vector. Hence, the adjusted expectation of $\mathcal{M}(Z)$ based on sample of size n is

$$E_{\bar{Z}_n}(\mathcal{M}(Z)) = E(\mathcal{M}(Z)) + Cov(\mathcal{M}(Z), \bar{Z}_n)Var(\bar{Z}_n)^{\dagger}(\bar{Z}_n - E(\bar{Z}_n)) \qquad (2.27)$$

In general, in order to adjust beliefs over exchangeable quantities, it is enough to adjust beliefs over the underlying population mean $\mathcal{M}(Z)$. For such adjustments, the sample mean $\bar{Z}_n$ is Bayes linear sufficient; there is no need to use the individual values of the sample. The theorem is also applicable to the adjustment of population variances as we shall illustrate next.

2.14.2 Adjusting variance components

Estimating variances and partitioning it at various levels of a hierarchy is a basic task in multilevel modelling. Variance learning however, is more difficult than learning about mean components as we explain next.

Difficulties in estimating variances

A major obstacle in variance estimation is that, by definition, a variance cannot be negative. Unfortunately, the use of an unbiased estimator of a population variance may produce such negative estimates. A similar problem may occur in the estimation of a variance-covariance matrix which is required to be non-negative definite.

Another problem is that assessing a population variance requires computations with fourth order moments (see Searle *et al.*(1992); p407), which inherently are complicated. For example, consider the rth central moment of a random quantity X with respect to the probability measure $F(x)$

$$\mu_r(X) = \int_{-\infty}^{+\infty} (x-\mu)^r dF(x).$$

If $\bar{X}$ is the sample mean of n independent and identically distributed random quantities, then the first three moments of $\bar{X}$ have simple expressions:

$$\mu(\bar{X}) = \mu(X) \qquad \mu_2(\bar{X}) = \mu_2(X)/n \qquad \mu_3(\bar{X}) = \mu_3(X)/n^2$$

while the fourth moment is

$$\mu_4(\bar{X}) = \mu_4(X)/n^3 + 3(n-1)\mu_2^2(X)/n^3$$

The complexity of the above fourth population moment also applies to its sample equivalent. Thus, calculations with the sample variance s_n^2 and its variance $Var(s_n^2)$, as required for Bayes linear adjustment of variances, is more complicated than the corresponding quantities required for the adjustment of means. Also, a Bayes linear analysis would require an individual to make well-founded specifications about variances, and also about uncertainty in these variances. The latter specifications require fourth-order moments, are unfamiliar, and as such quite challenging for the individual to make.

A fundamental difficulty in estimating a variance resides in its own definition $E(X_i-\mu)^2$. The squared quantities $(X_i-\mu)^2$ required for the Bayes linear estimation of a variance are not observable since they depend on the unknown population mean $\mathcal{M}(X)$, where $E(\mathcal{M}(X)) = \mu$. Hence, variance estimation cannot be directly made as is the case for mean estimation. This issue will be illustrated in Section 2.14.3 below.

A final problem concerns the relationship between the Bayes linear analysis of variances and the corresponding analysis of means. More specifically, the problem is whether to modify the linear Bayes estimator of a population mean using an estimate of the population variance in a **two-stage Bayes linear analysis** (Goldstein, 1983). This topic, which is also a focus of the present thesis, is considered in Section 2.14.4.

2.14.3 Estimating the population variance of a sequence of exchangeable random quantities

The Bayes linear approach reviewed here follows closely Goldstein & Woof (2007). In particular, the proof of the variance estimation method is given below as we plan to apply similar methods to learn about population variances in exchangeable multilevel models. Suppose that $Z = \{Z_1, Z_2, \ldots\}$ is an infinite exchangeable sequence of random quantities, where $\mathrm{E}(Z_k) = \mu$, $\mathrm{Var}(Z_k) = \sigma^2$, $\mathrm{Cov}(Z_k, Z_j) = \gamma$ with $\gamma \geq 0$.

We have the second-order exchangeability representation

$$Z_k = \mathcal{M}(Z) + \mathcal{R}_k(Z) \tag{2.28}$$

where the sequence $\mathcal{R}_k(Z)$ is uncorrelated and has expectation $\mathrm{E}(\mathcal{R}_k(Z)) = 0$, and variance

$$\mathrm{Var}\ (\mathcal{R}_k(Z)) = \sigma^2 - \gamma = V_R \tag{2.29}$$

We construct a representation for the corresponding population variance as follows. We let $V_k = [\mathcal{R}_k(Z)]^2$. Assuming the sequence $V_1, V_2, \ldots$ is also exchangeable, we have the following second-order representation

$$[\mathcal{R}_k(Z)]^2 = V_k = \mathcal{M}(V) + \mathcal{R}_k(V) \tag{2.30}$$

where $\mathcal{M}(V)$ represents the population variance with mean V_R and variance V_M, and the sequence $\mathcal{R}_1(V), \mathcal{R}_2(V), \ldots$ is uncorrelated with mean zero and variance $V_{R(V)}$.

As $\mathcal{M}(Z)$ is unknown, the sequence V_k is not observable. This issue of unobservable sums of squares was mentioned earlier above as a principal problem in learning about a population variance, especially in complex models such as multilevel models. In this simple situation though, it is easy to circumvent the problem by exploiting the symmetry between the squared mean deviations of the data and the residuals. Nevertheless, the simplicity of the solution illustrates an important principle: in order to learn about population variances, whether in simple or complex models, relevant sums of squares independent of any underlying (unknown) population mean should first be calculated.

Thus, given a sample of observations $(Z_1, \ldots, Z_n)$, $n \geq 2$, from (2.28) we form the following squared deviations

$$(Z_k - \bar{Z}_n)^2 = (\mathcal{R}_k(Z) - \bar{\mathcal{R}}_n)^2$$

where $\bar{Z}_n$ and $\bar{\mathcal{R}}_n$ are the observation and residual sample means respectively. For example, $\bar{\mathcal{R}}_n = (1/n) \sum_{k=1}^{n} \mathcal{R}_k(Z)$.

The squared residuals $(\mathcal{R}_k(Z) - \bar{\mathcal{R}}_n)^2$ are informative for $\mathcal{M}(V)$. By symmetry, the Bayes linear estimate for $\mathcal{M}(V)$ given the individual squared residuals depends

only on the sum of squares, or equivalently on s_n^2. We derive a representation for s_n^2 as follows.

$$\begin{aligned} s_n^2 &= \frac{1}{n-1}\sum_k (\mathcal{R}_k(Z) - \bar{\mathcal{R}}_n)^2 \\ &= \frac{1}{n-1}\{\sum_k \mathcal{R}_k(Z)^2 - \frac{1}{n}[\sum_k \mathcal{R}_k(Z)]^2\} \\ &= \frac{1}{n}\sum_k \mathcal{R}_k(Z)^2 - \frac{2}{n(n-1)}\sum_{k<j} \mathcal{R}_k(Z)\mathcal{R}_j(Z) \\ &= \mathcal{M}(V) + T \end{aligned} \tag{2.31}$$

where

$$T = \frac{1}{n}\sum_k \mathcal{R}_k(V) - \frac{2}{n(n-1)}\sum_{k<j} \mathcal{R}_k(Z)\mathcal{R}_j(Z)$$

Assuming that the residuals $\mathcal{R}_j(Z)$ have the following fourth order uncorrelated properties.

$$Cov(\mathcal{M}(V), \mathcal{R}_k(Z)\mathcal{R}_j(Z)) = Cov(\mathcal{R}_i(V), \mathcal{R}_k(Z)\mathcal{R}_j(Z)) = 0$$

for $k \neq j \neq i$. While if $k > j, w > u$, then

$$Cov(\mathcal{R}_k(Z)\mathcal{R}_j(Z), \mathcal{R}_w(Z)\mathcal{R}_u(Z)) = 0$$

unless $k = w, j = u$. Hence, we have

$$\begin{aligned} E(T) &= 0, \quad Var(T) = V_T = \frac{1}{n}V_{R(V)} + \frac{2}{n(n-1)}[V_M + V_R^2], \\ Cov(\mathcal{M}(V), T) &= 0. \end{aligned} \tag{2.32}$$

and therefore,

$$\begin{aligned} E(s_n^2) &= V_R, \quad Var(s_n^2) = V_M + V_T, \\ Cov(s_n^2, \mathcal{M}(V)) &= V_M. \end{aligned} \tag{2.33}$$

Given the above specifications, the adjusted mean and variance for the underlying population variance $\mathcal{M}(V)$ given s_n^2, are respectively

$$E_{s_n^2}(\mathcal{M}(V)) = \frac{V_M s_n^2 + V_T V_R}{V_M + V_T} \tag{2.34}$$

$$Var_{s_n^2}(\mathcal{M}(V)) = \frac{V_M V_T}{V_M + V_T} \tag{2.35}$$

The adjusted expectation of $\mathcal{M}(V)$ is the well-known precision weighted average of prior variance and data variance.

2.15 Coherence and diagnostic checks

Model checking is a vital stage of all model building approaches, Bayesian or otherwise. Within the Bayes linear approach we perform checks on both prior specifications and adjusted expectations with the aim of ensuring that they are coherent. Below we follow Goldstein & Wooff (2007) and give the various diagnostic measures in the general context of adjusting a collection of beliefs B by the, as yet unobserved, data collection D. When data is observed, it is important to check that prior specifications do not clash with such data by performing data driven diagnostics. Some of these diagnostics, such as partial diagnostics, are of special relevance to multilevel modelling.

2.15.1 Coherence

The prior specification over (D, B) is coherent if

$$Var\begin{pmatrix} D \\ B \end{pmatrix} = \begin{pmatrix} Var(D) & Cov(D,B) \\ Cov(B,D) & Var(B) \end{pmatrix} \tag{2.36}$$

is non-negative definite. Theorem 3.12 of Goldstein & Wooff (2007) gives the conditions for the variance-covariance matrix $Var(D, B)$ to be non-negative as follows:

1. $Var(D)$ is non-negative definite;
2. $Cov(D, B) \in$ range $\{Var(D)\}$;
3. $Var(B) - Cov(B, D)Var(D)^{\dagger}Cov(D, B)$ is non-negative definite for any generalized inverse of $Var(D)$.

Condition three above ensures that the adjusted variance-covariance matrix is also non-negative definite.

2.15.2 Data diagnostics

The standardized observation $S(z)$ and the discrepancy $Dis(z)$ provide simple checks on whether the prior specification for a univariate random quantity Z is in conflict with its observed value z where

$$S(z) = \frac{z - E(Z)}{\sqrt{Var(Z)}},$$

and the discrepancy

$$Dis(z) = \frac{[z - E(Z)]^2}{Var(Z)}.$$

A very large $Dis(z)$ may indicate mis-specification of $E(Z)$ or an under-estimated $Var(Z)$, while an over-estimated $Var(Z)$ may result in small $Dis(z)$. Discrepancy is similar to *discordancy* (see Barnett and Lewis, 1994) and is useful in outlier detection.

In multilevel data, outliers may occur at any level of the hierarchy and this may cause complications in their detection. While $Dis(z)$ may be useful in detecting problems with observations at each level of the hierarchy *independently*, it may be limited in detecting outliers that occur at higher levels of the hierarchy. In a two-level dataset for example, outliers at level 1, unusually low pupil outcomes say, may mask a level 2 outlier, an unusually high school outcome. Hence, to cater for the dependencies in multilevel data structures, it may be useful to also consider a multivariate discrepancy measure in addition to the univariate measure $Dis(z)$.

2.15.3 Mahalanobis distance: multivariate data discrepancy

The Mahalanobis distance, Mahalanobis (1936), is a multivariate distance measure that takes into account the correlation among the variables of interest, hence making it a suitable discrepancy measure for multilevel data where dependencies of observations (within and between groups) are of particular relevance.

Suppose Y is a multilevel (or multivariate) data set. The Mahalanobis distance or data discrepancy, Dis(Y) is given by:

$$Dis(Y) = (Y - E(Y)^T Var(Y)^{\dagger}(Y - E(Y)) \tag{2.37}$$

where $Var(Y)^{\dagger}$ is a generalized inverse, such as the Moore-Penrose pseudo inverse. As in the univariate case, large values of $Dis(Y)$ may indicate that the specification

of $E(Y)$ is not consistent with the data Y and/or that the $Var(Y)$ has been underestimated, while an overestimated $Var(Y)$ may lead to small $Dis(Y)$.

Comparing discrepancies

When comparing discrepancies across different multilevel data sets, it may be useful to standardize the data discrepancy by calculating the discrepancy ratio, $Dr(Y)$ as follows:

$$Dr(Y) = \frac{Dis(Y)}{rank(Var(Y))}. \tag{2.38}$$

Since $E(Dis(Y)) = rank(Var(Y))$, thus $Dr(Y)$ has a prior expectation of one.

Also, when comparing summary measures, including discrepancy, consideration must be given to the variance of that summary. The variance of the discrepancy $Var(Dis(Y))$ may be specified directly based, for example, on similar multilevel data. Alternatively, $Var(Dis(Y))$ may be elicited from specifications of fourth order moments over the elements of the multilevel data Y. Such fourth-order specifications may be made by assuming that the elements of Y follow a multivariate Gaussian distribution. In this case, $Dis(Y)$ will follow a chi-square distribution with $r = rank(Var(Y))$ degrees of freedom and moments:

$$E(Dis(Y)) = r \qquad Var(Dis(Y)) = 2r \tag{2.39}$$

Using the above arguments and Chebyshev's inequality, Goldstein and Wooff (1995) derive bounds for $Dr(Y)$ as follows:

$$P\big(1 - \frac{6}{\sqrt{r}} \leq Dr(Y) \leq 1 + \frac{6}{\sqrt{r}}\big) \leq 0.9444. \tag{2.40}$$

Goldstein and Wooff (1995) further argue that if the multivariate Gaussian distribution is used to specify a probabilistic distribution for $Dis(Y)$, in addition to using it to specify a value for $Var(Dis(Y))$, then the following bounds are obtained.

$$P\big(1 - \frac{2.7}{\sqrt{r}} \leq Dr(Y) \leq 1 + \frac{2.7}{\sqrt{r}}\big) \leq 0.9444. \tag{2.41}$$

These bounds can be useful in routine and fast checking of outliers in large multilevel data sets arising, for example, in simulation studies of the properties of the two-stage Bayes linear estimators of parameters of a basic SOEREG model.

2.15.4 Adjustment discrepancy

In addition to assessing the discrepancy of our multilevel data Y, we may also check whether the adjusted expectations of the mean and variance parameters of our SOEREG models agree with our prior expectations. Suppose these parameters are collected in the vector $B = (B_1, B_2, \ldots, B_r)$, then the discrepancy of the adjustment vector $Dis(E_Y(B))$ is as follows.

$$Dis_Y(B) = (E_Y(B) - E(B))^T RVar_Y(B)^\dagger (E_Y(B) - E(B)). \tag{2.42}$$

where $RVar_Y(B)$ is the variance of B resolved by Y. The corresponding adjustment discrepancy ratio is

$$Dr_Y(B) = \frac{Dis_Y(B)}{r_T}. \tag{2.43}$$

where $r_T = rank(RVar_Y(B))$.

We note that sometimes $Dis_Y(B)$ may be a little large while $Dr_Y(B)$ is not large, hence the update may still be robust. As in (2.40), using the conditions (2.39) yield the simple conservative bounds for the adjustment discrepancy ratio:

$$P\big(1 - \frac{6}{\sqrt{r_T}} \leq \frac{Dis_Y(B)}{r_T} \leq 1 + \frac{6}{\sqrt{r_T}}\big) \leq 0.9444. \tag{2.44}$$

The adjustment discrepancy and its bounds provide simple automatic checks useful when making a large number of belief adjustments such as may occur in simulation studies of our SOEREG models.

2.15.5 Partial diagnostics

We saw earlier that the grouping of data in various hierarchical levels induces complex dependencies that must be explicitly accounted for in modelling multilevel data. It is therefore important to explore how these groups of data and the prior beliefs specifications combine to give the final adjusted parameters of SOEREF and SOEREG models fitted to multilevel data.

A partial Bayes linear analysis is suitable for this task as it separates the effects of the different groups of data on our beliefs by calculating partial adjustments. The

groups mentioned here may also refer to subsets of data grouped according to, say, the predictor variable z_{ji} in our SOEREG model, in addition to the ususal groups in our hierarchies.

Partial Bayes linear analysis is dealt with extensively in Chapter five of Goldstein & Woof (2007) and may be potentially useful for the interpretation and diagnostic checking of our multilevel models. We shall investigate this in Chapter four. Below we briefly summarize some of the key definitions and concepts relating to partial adjustment of beliefs. There are potentially many partial diagnostic measures that may be considered; for our purpose we focus on partial adjustment of expectation, partial bearing and path correlation only. We follow closely Chapter five of Goldstein & Woof (2007).

Partial adjustment of beliefs

As in the previous sub-section, we collect the parameters of our SOEREG model in the vector $B = (B_1, B_2, \ldots, B_r)$. We now consider adjusting B by the collections of quantities $D = (D_1, D_2, \ldots, D_k)$ and $F = (F_1, F_2, \ldots, F_l)$. For example, D may comprise data arising from one group and F, the remaining groups in our multilevel data.

Definition 2.15.1. *The partial adjustment of B by F given D, denoted by $E_{[F/D]}(B)$, is*

$$E_{[F/D]}(B) = E_{D\cup F}(B) - E_D(B) \tag{2.45}$$

Goldstein & Woof (2007) show that expression (2.45) implies that in making the sequential beliefs adjustments of B by D and F, we may initially adjust B and F by D yielding the adjusted forms $\mathbb{A}_D(B)$ and $\mathbb{A}_D(F)$ respectively, and then adjust all the resulting adjusted beliefs by F, giving $\mathbb{A}_{\mathbb{A}_D(F)}(\mathbb{A}_D(B))$. We shall be interested to use sequential adjusted expectations of mean components of our multilevel models as a diagnostic tool.

Size of the Partial adjustment

First let us consider the size of a (full) adjustment of a scalar random quantity

Q by D. The size of an adjustment of Q when the observed value of $D = d$ is

$$Size_d(Q) = \frac{[E_d(Q) - E(Q)]^2}{Var(Q)} \tag{2.46}$$

Thus $Size_d(Q)$ is the magnitude of the standardized difference between the prior and adjusted expectation relative to the prior variance, and may be used as a diagnostic measure. Intuitively, a large $Size_d(Q)$ reveals an unanticipated change in belief, hence indicating a potential conflict between prior and adjusted expectation.

For our multilevel models, we shall consider the adjustment of a collection B by a (multilevel) data vector D. In this case, the size of the adjustment is defined as the maximum size of the adjustment of the collection B for a linear combination $h^T B$, and is given by

$$Size_d(B) = \max_{X \in \langle B \rangle} Size_d(X) \tag{2.47}$$

where $\langle B \rangle$ represents all possible linear combinations of the elements in B. Goldstein & Woof (2007) show that the maximum in (2.47) is attained when $h = Var(B)^{\dagger}[E_d(B) - E(B)]$ and they give the resulting definition of the size as follows.

Definition 2.15.2. *The size of the adjustment of the collection B by D=d is*

$$Size_d(B) = [E_d(B) - E(B)]^T Var(B)^{\dagger}[E_d(B) - E(B)] \tag{2.48}$$

We now consider the size of the partial adjustment. Goldstein & Woof (2007) page 135 give the following definition:

Definition 2.15.3. *The size of the partial adjustment, or partial size, is defined as*

$$\begin{aligned} Size_{[f/d]}(B) &= \max_{X \in \langle B \rangle} \frac{[E_{d \cup f}(X) - E_d(X)]^2}{Var(X)} \\ &= \max_{X \in \langle B \rangle} \frac{[E_{f/d}(X)]^2}{Var(X)} \end{aligned} \tag{2.49}$$

The maximum in (2.49) is

$$Size_{[f/d]}(B) = [E_{d \cup f}(B) - E_d(B)]^T Var(B)^{\dagger}[E_{d \cup f}(B) - E_d(B)] \tag{2.50}$$

The interpretation of the partial size is similar to that of the (full) size except that adjustments are made in stages.

Bearing for the Partial adjustment

If, in addition to the magnitude of the change between prior and adjusted expectation, we are also interested in the direction of this change, then we may compute the **bearing** of the adjustment of B when we observe $D = d$ as follows

$$\mathbb{Z}_d(B) = [E_d(B) - E(B)]^T Var(B)^\dagger [B - E(B)] \tag{2.51}$$

In the multilevel modelling context, we may be interested in linear combinations of beliefs quantities $\langle B \rangle$, such as differences between group means for example. The same bearing as above may also be constructed over $\langle B \rangle$ as follows.

$$\mathbb{Z}_d(B) = \sum_{i=1}^{r_B} E_d(U_i) U_i$$

where $(U_1, \ldots, U_{r_B})$ is any collection of mutually uncorrelated elements of $\langle B \rangle$ having zero mean and unit prior variance.

Goldstein & Woof (2007) give two properties of the bearing as follows: (1) the bearing is the linear combination in $\langle B \rangle$ having the largest standardized squared change in expectation and, (2) the change in adjustment for any quantity in $\langle B \rangle$ is equal to the prior covariance between the quantity and the bearing.

The above-mentioned two properties show that the bearing summarizes the actual effects of adjustments and, in cases of more complex adjustments (as in multilevel modelling), provides a simple summary of the magnitude and direction of the changes in belief following an adjustment.

For multilevel diagnostics, we are interested in the **partial bearing** which Goldstein and Wooff (2007) define as follows:

$$\mathbb{Z}_{[f/d]}(B) = \sum_{i=1}^{r_{\mathbb{P}}} E_{[f/d]}(U_i) U_i$$

for any collection $(U_1, \ldots, U_{r_{\mathbb{P}}})$ mutually uncorrelated with unit prior variance, where $r_{\mathbb{P}}$ is the rank of the partial resolution transform.

The partial bearing is related to the partial size through the following

$$Size_{[f/d]}(B) = Var(\mathbb{Z}_{[f/d]}(B)) = \sum_{i=1}^{r_{\mathbb{P}}} [E_{[f/d]}(W_i)]^2 \tag{2.52}$$

where W_i are the partial canonical directions. The partial bearing will provide a useful diagnostic analysis of the changes in magnitude and direction of expectation when we adjust beliefs about our multilevel model parameters in stages.

Partial size ratio

A further useful partial diagnostic for multilevel data is the **partial size ratio**, the ratio of the partial size in (2.50) to its expectation, that is

$$Sr_{[f/d]}(B) = \frac{Size_{[f/d]}(B)}{E(Size_{[F/D]}(B))} \tag{2.53}$$

The formula for computing $Sr_{[f/d]}(B)$ follows from the formula of the partial size given in (2.50) and $E(Size_{[F/D]}(B)) = \sum_{i=1}^{r_{\mathbb{P}}} \zeta_i$, where ζ_i are the partial canonical resolutions.

Suppose we perform an adjustment of one parameter in our multilevel model in stages, each stage corresponding to data from a separate group in our multilevel data. Suppose further that after the adjustment of the parameter by a given group, we obtain a partial size ratio which is much larger than one. This indicates an unexpectedly large change in expectation which, in turn, may indicate a conflict between data from that group and our prior specifications for the parameter. The reverse, a small partial size ratio, indicates an unexpectedly small change in expectation which may be because we were quite restrictive in our assessment of the prior variance of the parameter. There are many such parameters and groups in multilevel models, hence the potential use of the partial size ratio in diagnostic analyses of these models.

Path correlation of the Partial adjustment

In the sequential adjustment of B by D and then by F, the path correlation is the correlation between the bearing $\mathbb{Z}_d(B)$ for the data collection $D = d$ and the partial bearing $\mathbb{Z}_{[f/d]}(B)$, that is

$$PC(d, [f/d]) = Corr(\mathbb{Z}_d(B), \mathbb{Z}_{[f/d]}(B)) \tag{2.54}$$

Path correlations, like any correlation, lie between +1 and -1, with values close to +1 indicating complementarity of D and F (say two arbitrary groups in our multilevel

data), while values close to -1 indicate conflict between D and F in the sequential adjustment of B.

Chapter 3

Exchangeable Multilevel Models

In chapter 2 we explained the concept and importance of multilevel models. We also described the main features of the Bayes linear approach with the intention of applying Bayes linear methods to analyse multilevel models. In this chapter, we make use of second order exchangeability (SOE) judgements to formulate our version of multilevel models via the representation theorem for SOE random quantities. We begin by applying exchangeability judgements to observations from a two-level hierarchy, to obtain a version of the two-level random effects model that corresponds to our underlying SOE judgements. We call this model the second order exchangeable random effects model (SOEREF). The implications of the SOEREF model are briefly discussed in section 3.2. In Section 3.3 we extend SOE judgements to multilevel regression models which we call SOEREG models. These models are more general and we show that they encompass models with spatial dependencies for example.

Finally, we consider prior specifications for the parameters of the SOEREF model as applied to the STAT1010 data.

3.1 The second-order exchangeable random effects (SOEREF) model

Recall from Chapter 2 that we have a single continuous outcome variable y_{ji} representing the measurement on an individual i (level 1 unit) nested in a group j

(level 2 unit), hence a two-level hierarchy. We now wish to develop a version of the random effects model that corresponds to our underlying SOE judgements which we term a second-order exchangeable random effects (SOEREF) model. In developing the SOEREF model, we shall consider exchangeability at each level of the two-level hierarchy in turn. Initially we focus on applying exchangeability to derive the form of the SOEREF model. We then make explicit the assumptions regarding the population means and residuals introduced via the representation theorem before giving a formal definition of the SOEREF model.

Second-order exchangeability at level 1

At level 1, we consider individuals in a specific group j to be exchangeable. We assume that the sequence of outcome variables $\{y_{j1}, y_{j2}, \ldots\}$ forms an infinite SOE sequence of continuous random quantities (we discuss finite exchangeability in Section 3.2), and for any two individuals i and i', we have

$$E(y_{ji}) = \mu, \qquad Var(y_{ji}) = \sigma_y^2, \quad \forall i, j, \tag{3.1}$$

$$Cov(y_{ji}, y_{ji'}) = \sigma_u^2, \quad \forall i \neq i', \text{ and } \forall j \tag{3.2}$$

where the data variance $\sigma_y^2 > 0$ and the variance $\sigma_u^2 \geq 0$. Then, by the representation theorem, we may write the level-1 representation

$$y_{ji} = \mathcal{M}(y_j) + \mathcal{R}_i(y_j), \qquad \forall j, i. \tag{3.3}$$

where we have the following:
For each group j,

1. $\mathcal{M}(y_j)$ is termed the **population group j mean**,
2. the sequence $\mathcal{R}_1(y_j), \mathcal{R}_2(y_j), \ldots,$ are termed the **level 1 residuals**,
3. the sequence of level 1 residuals are uncorrelated among themselves and also with the population group j mean.

Second-order exchangeability at level 2: exchangeability over population group j means

The collection of population group j means $\{\mathcal{M}(y_1), \mathcal{M}(y_2), \dots\}$, is assumed to form an infinite SOE sequence with

$$E(\mathcal{M}(y_j)) = \mu, \qquad Var(\mathcal{M}(y_j)) = \sigma_u^2, \tag{3.4}$$

$$Cov(\mathcal{M}(y_j), \mathcal{M}(y_{j'})) = \gamma \qquad \forall j \neq j'. \tag{3.5}$$

where $\gamma \geq 0$.

To derive the representation for the population group j means $\mathcal{M}(y_j)$, we assume that the sequence $\{\mathcal{M}(y_1), \mathcal{M}(y_2), \dots\}$, forms an infinite exchangeable sequence with second-order moments as in (3.4) and (3.5) above. Applying once more the representation theorem, we may write

$$\mathcal{M}(y_j) = \mathcal{M}(y) + \mathcal{R}_j(\mathcal{M}(y)), \qquad \forall j. \tag{3.6}$$

Combining (3.6) and (3.3), we obtain the full model

$$y_{ji} = \mathcal{M}(y) + \mathcal{R}_j(\mathcal{M}(y)) + \mathcal{R}_i(y_j),$$

where $\mathcal{M}(y)$ is termed the **population grand mean**, and the sequence of residuals $\mathcal{R}_1(\mathcal{M}(y)), \mathcal{R}_2(\mathcal{M}(y)), \dots,$ are the **level 2 residuals**. The level-2 residuals are mutually uncorrelated and also uncorrelated with the population grand mean $\mathcal{M}(y)$.

3.1.1 Assumptions and notations for level 1 and level 2 residuals

Theorem 2.13.1 for infinite SOE quantities also states that each of the two sequences of level 1 and level 2 residuals are themselves second-order exchangeable. To complete the description of the SOEREF model, we make explicit our assumptions and notations for these two sequences of residuals. The exchangeability assumptions will be discussed in Section 3.2.

Exchangeability of level-1 residuals

We assume the level-1 residuals $\mathcal{R}_i(y_j)$ are second-order exchangeable over individuals for each group j and write $\epsilon_{ji} = \mathcal{R}_i(y_j)$ for individual i in group j with:

$$E(\mathcal{R}_i(y_j)) = 0, \qquad Var(\mathcal{R}_i(y_j)) = \sigma^2_\epsilon = \sigma^2_y - \sigma^2_u \qquad \forall i, j, \tag{3.7}$$

$$Cov(\mathcal{R}_i(y_j), \mathcal{R}_{i'}(y_j)) = 0 \qquad \forall i \neq i'. \tag{3.8}$$

where the level 1 variance $\sigma^2_\epsilon > 0$ and constant for all individuals and groups. Also, for all i and j, the level-1 residual $\mathcal{R}_i(y_j)$ is uncorrelated with the population grand mean $\mathcal{M}(y)$.

Exchangeability of level 2 residuals

We assume the level-2 residuals $\mathcal{R}_j(\mathcal{M}(y))$ are second-order exchangeable over groups j. We write $u_j = \mathcal{R}_j(\mathcal{M}(y))$ and make the following specifications:

$$E(\mathcal{R}_j(\mathcal{M}(y))) = 0, \qquad Var(\mathcal{R}_j(\mathcal{M}(y))) = \sigma^2_u - \gamma, \qquad \forall j \tag{3.9}$$

$$Cov(\mathcal{R}_j(\mathcal{M}(y)), \mathcal{R}_{j'}(\mathcal{M}(y))) = 0 \quad \forall j \neq j'. \tag{3.10}$$

The level-2 residuals $\mathcal{R}_j(\mathcal{M}(y))$ are also uncorrelated with the level-1 residuals $\mathcal{R}_i(y_j)$ and the population grand mean $\mathcal{M}(y)$. From (3.5), Var($\mathcal{M}(y)$)=γ. The level 2 variance $\sigma^2_u - \gamma \geq 0$.

The second-order exchangeable random effects (SOEREF) model

The above exchangeability assumptions and specifications lead to the following definition:

Definition 3.1.1. *Let y_{ji} represent univariate outcome measurements on each individual i nested in group j. A* **Second-order exchangeable random effects (SOEREF)** *model is given by either of the following representation form:*
Hierarchical form:

$$\begin{aligned} \textit{Level-1}: \qquad y_{ji} &= \mathcal{M}(y_j) + \mathcal{R}_i(y_j) \\ \textit{Level-2}: \qquad \mathcal{M}(y_j) &= \mathcal{M}(y) + \mathcal{R}_j(\mathcal{M}(y)). \end{aligned} \tag{3.11}$$

Single-equation form:

$$y_{ji} = \mathcal{M}(y) + \mathcal{R}_j(\mathcal{M}(y)) + \mathcal{R}_i(y_j),$$
$$i = 1, 2, \ldots, n_j, \quad \textit{and} \quad j = 1, 2, \ldots, J. \tag{3.12}$$

where $\mathcal{M}(y)$ *is the* **population grand mean**, $\mathcal{M}(y_j)$ *is the* **population group j mean**, $\mathcal{R}_j(\mathcal{M}(y))$ *are the* **level 2 residuals**, *and* $\mathcal{R}_i(y_j)$ *are the* **level 1 residuals**.

The moments of the level 1 and level 2 residuals, and of the population grand mean and population group j mean are as specified in (3.4) to (3.10).

3.2 Discussions of Exchangeability

The judgement of exchangeability, fundamental to the subjectivist approach, is a weaker requirement than the independent, identically distributed (i.i.d.) assumption of traditional inferential statistics. Exchangeability is simple, intuitive, and is generally applicable, requiring only that one's subjective judgements remain unchanged under arbitrary permutations of a given sequence of observations.

SOE assumptions weaken this requirement even further, by only imposing this restriction for second-order structure. Thus, our SOE judgements that lead to the formulation of the SOEREF model required only that first and second-order moments be specified. Together with Bayes linear methods, the SOEREF model population grand and group j means can be updated quite easily. The SOEREF variance components can similarly be updated within the Bayes linear framework, albeit with the more difficult task of specifying fourth-order moments and imposing uncorrelated fourth-order properties on products of residuals (see Section 2.14.3); these specifications and restrictions though are still less than those required for full exchangeability. Below we briefly discuss the type of SOE assumptions we have made so far.

We have assumed that our observations are infinitely exchangeable. In practice we may have a finite number n_j of individuals, nested in a finite number of groups j.

For such finite sampling situations, Goldstein & Wooff (2007, pg 188) show that the residuals are no longer orthogonal (uncorrelated) but rather have a small negative correlation, with orthogonality of order $1/r$, where r is the finite sample size. Hence, provided our finite sample sizes of level-1 individuals and of level 2 groups are each relatively small compared to their respective population sizes, departures from the assumption of infinite exchangeability may be considered unimportant. Such is the situation in practice in multilevel studies that comprise of moderate sized samples taken from large numbers of level 1 and level 2 units. In Chapter 4 we consider a finitely exchangeable SOEREF model and its application.

We assumed the observations are SOE at level one, and the resulting population group j means are also SOE at level two. Such partial or co- exchangeability allowed us to account for the two-level structure of the data. The scope of this basic SOEREF model can further be expanded as we show next.

3.3 Extending the SOEREF model

The two-level SOEREF model is the simplest multilevel model which is useful both in understanding multilevel concepts, and also in important applications such as small area estimation (see Section 2.3.3). The classical and fully Bayesian counterparts of the SOEREF model, the random effects model, has been studied extensively over the years, see Searle et al. (2006) for a history, and Khuri and Sahai (1985) for a very extensive bibliography. In Chapters 4 and 5 of this thesis we shall study the SOEREF model further.

One principal use of the SOEREF model is to obtain an estimate of the proportion of variance between groups, called the *intra-class correlation* $\rho = \sigma_u^2/(\sigma_u^2 + \sigma_\epsilon^2)$, for a given multilevel data set. If ρ is important, then it may be worthwhile to consider more complex multilevel models, with predictors at individual and group levels, to explain the variation between groups. Viewed from this perspective, the SOEREF model is an ignorance model; it ignores the potential predictors on which data is often available, specially in complex surveys and studies. We shall now extend the two-level SOEREF model to include regression predictors at the individual

and group levels. The resulting model is a system of exchangeable regressions which we term Second-Order Exchangeable Regressions (SOEREG) model.

3.4 The Second-Order Exchangeable Regression model (SOEREG)

The SOEREG model is intuitively more appealing and suitable for multilevel data as it takes into consideration the rich hierarchical structure of the data, including predictor variables that are often available at the different levels of the hierarchy.

In formulating the SOEREG model, we shall proceed in steps as follows. We start with the two-level SOEREF model and model the underlying population group j mean using simple linear regression (i.e. a single level 1 predictor). This will give us a model where the regression coefficients, namely the intercept and slope, vary across groups. These regression coefficients are subsequently modelled using a single predictor at the group level; this will yield our SOEREG model with predictors at both level 1 and 2. Finally, we write the matrix form of the SOEREG model and show that this matrix formulation encompasses more general multilevel models.

3.4.1 A note on notations

Since, we are going to consider regressions, we make use of the usual β notations for regression coefficients except that we enclose their subscripts in brackets (e.g. $\beta_{[0]}, \beta_{[1]}, \ldots$) to distinguish them from the level 1 and level 2 subscripts i and j. In place of the usual x predictor, we use z to avoid confusion with predictors in the final matrix form of the general model. Also, as we shall be dealing with several mean, variance, and covariance parameters, we shall only discuss the necessary first- and second-order specifications when we consider the more concise matrix form of the SOEREG model; for now we focus only on deriving the form of the SOEREG model.

3.4.2 A SOEREG model with a predictor at level 1 only

In the SOEREF model only the response y_{ji} was modelled, resulting in a mean component $\mathcal{M}(y_j)$, and a residual component. When additional multilevel variables are considered, for example a level 1 predictor z_{ji}, the mean component will have more structure. Thus in each group j, we now model the response y_{ji} using a simple linear regression as follows.

$$y_{ji} = \beta_{[0]j} + \beta_{[1]j} z_{ji} + \epsilon_{ji} \qquad \forall j, i, \tag{3.13}$$

where the group j regression coefficients $\beta_{[0]j}$, $\beta_{[1]j}$ are, respectively, the intercept and slope specific to group j, and ϵ_{ji} is a level 1 residual error term. If we assume that the collection of group j regression coefficients is second-order exchangeable over groups, we may then write the following representations:

$$\beta_{[0]j} = \mathcal{M}(\beta_0) + \mathcal{R}_j(\beta_0) \tag{3.14}$$

$$\beta_{[1]j} = \mathcal{M}(\beta_1) + \mathcal{R}_j(\beta_1). \tag{3.15}$$

Note that above we use the simpler notation β_0 rather than $\beta_{[0]}$ for example, when there is no subscript i or j. Replacing (3.14) and (3.15) in (3.13) we obtain the basic SOEREG model:

$$y_{ji} = \mathcal{M}(\beta_0) + \mathcal{M}(\beta_1) z_{ji} + \mathcal{R}_j(\beta_0) + \mathcal{R}_j(\beta_1) z_{ji} + \epsilon_{ji}. \tag{3.16}$$

If we set z_{ji} to zero in (3.16), we obtain the SOEREF model. Goldstein & Wooff (1995) apply the basic SOEREG model to perform a comprehensive Bayes linear analysis of exchangeable regressions in the context of an industrial process of aluminium extraction.

3.4.3 Beliefs specifications over the basic SOEREG model

To complete the description of the basic SOEREG model, we now make our second-order prior judgments over it. We shall consider the implications of these judgements in the next sub-section.

Our beliefs about the regession model (3.13) are as follows. We assume the level 1 errors ϵ_{ji} are uncorrelated with mean zero and constant variance $\sigma^2_\epsilon > 0$ as in the

SOEREF model. The ϵ_{ji}'s are also uncorrelated with the level 2 residual errors, $\mathcal{R}_j(\beta_0)$ and $\mathcal{R}_j(\beta_1)$. Our specifications for the regression coefficients are:

$$E(\beta_{[0]j}) = \mu_0, \quad E(\beta_{[1]j}) = \mu_1, \tag{3.17}$$

$$Var(\beta_{[0]j}) = \sigma_0^2, \quad Var(\beta_{[1]j}) = \sigma_1^2, \tag{3.18}$$

$$Cov(\beta_{[0]j}, \beta_{[0]j'}) = \gamma_0, \quad Cov(\beta_{[1]j}, \beta_{[1]j'}) = \gamma_1, \tag{3.19}$$

$$Cov(\beta_{[0]j}, \beta_{[1]j}) = \rho_{01}\sigma_0\sigma_1, \tag{3.20}$$

for all j and $j \neq j'$. Note that we have specified a non-zero correlation ρ_{01} between the regression coefficients in each group j as commonly assumed in multilevel models. The above level 1 judgments have implications for the level 2 representations (3.14) and (3.15) as follows:

$$E(\mathcal{M}(\beta_0)) = \mu_0, \quad E(\mathcal{M}(\beta_1)) = \mu_1, \tag{3.21}$$

$$Var(\mathcal{M}(\beta_0)) = \gamma_0, \quad Var(\mathcal{M}(\beta_1)) = \gamma_1, \tag{3.22}$$

$$Cov(\mathcal{M}(\beta_0), \mathcal{M}(\beta_1)) = 0, \tag{3.23}$$

$$Var(\mathcal{R}_j(\beta_0)) = \sigma_0^2 - \gamma_0, \quad Var(\mathcal{R}_j(\beta_1)) = \sigma_1^2 - \gamma_1, \quad \forall j, \tag{3.24}$$

$$Cov(\mathcal{R}_j(\beta_0), \mathcal{R}_{j'}(\beta_1)) = \begin{cases} \rho_{01}\sigma_0\sigma_1 & \text{if} \quad j = j', \\ 0 & \text{if} \quad j \neq j'. \end{cases} \tag{3.25}$$

Also, the sequences $\mathcal{R}_j(\beta_0), \ldots$ and $\mathcal{R}_j(\beta_1), \ldots$ are each uncorrelated amongst themselves and with $\mathcal{M}(\beta_0)$ and $\mathcal{M}(\beta_1)$.

3.4.4 Formulating a basic SOEREG model: The STAT1010 example

Here we give some plausible reasons for formulating a SOEREG model. We show how the a priori exchangeability judgements within and between groups that led to the SOEREG model in Sub-section 3.4.2, are relevant to the specific context of the STAT1010 example of Chapter 2.

In Figure 2.1 of the STAT1010 example, there are clear variations in both intercepts and slopes of the within class regressions. It makes sense, therefore, to allow

each class to have its own regression as in (3.13). There are however, some similarities between classes as well. These are due, for example, to a common syllabus and manual being used, consultation among lecturers to ensure the same coverage of topics, and a common examination. Because of the similarity between classes we assume the regression coefficients to be exchangeable across classes as in (3.14) and (3.15).

Also, a priori we know students in the Faculty of Engineering have better A level grades in mathematics, and thus will do better in the STAT1010 examination, compared to students in the Faculty of Management. This is revealed in Figure 2.1 where the regressions intercepts for engineering classes are all above those of management classes. Thus we may consider explaining variations in the varying intercepts and slopes using faculty z_{ji} as a predictor.

Finally, low values of the intercepts appear to be associated with higher slopes in Figure 2.1, hence it is sensible to posit a non-zero correlation between the regression coefficients as in (3.20).

In making the above judgements, we have referred to the STAT1010 data for illustrative purposes only. In practice, we should use experts' opinions, meta-analytic studies or auxiliary data analyses to make such judgements *a priori*. When we come to observe data, the subsequent data analysis will provide information that may cause us to refine and modify our judgements.

3.5 Extending the basic SOEREG model

In this thesis we shall develop Bayes linear methods to estimate the mean and variance components of the basic SOEREG model. Below we indicate how the basic SOEREG model may be extended to a general SOEREG model.

The basic SOEREG model has only two hierarchical levels and a single predictor at level 1. It is therefore natural to extend it to more than two hierarchical levels with several predictors defined at all these levels, and with the variance components having a more complex dependency structure. To write this general SOEREG model, we move to matrix notation.

Consider first the matrix form of the basic SOEREG model for a dataset having two groups, each with two observations. Using (3.16), we may write the following matrix equation

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \end{bmatrix} = \begin{bmatrix} 1 & z_{11} \\ 1 & z_{12} \\ 1 & z_{21} \\ 1 & z_{22} \end{bmatrix} \begin{bmatrix} \mathcal{M}(\beta_0) \\ \mathcal{M}(\beta_1) \end{bmatrix} + \begin{bmatrix} 1 & z_{11} & 0 & 0 \\ 1 & z_{12} & 0 & 0 \\ 0 & 0 & 1 & z_{21} \\ 0 & 0 & 1 & z_{22} \end{bmatrix} \begin{bmatrix} \mathcal{R}_1(\beta_0) \\ \mathcal{R}_1(\beta_1) \\ \mathcal{R}_2(\beta_0) \\ \mathcal{R}_2(\beta_1) \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{21} \\ \epsilon_{22} \end{bmatrix}$$

The above give the matrix form of the basic SOEREG model:

$$Y = X\beta + ZU + \epsilon, \tag{3.26}$$

where Y is the response vector, X is a predictor matrix of the mean components contained in the vector β, Z is the predictor matrix of the higher level residual errors contained in the vector U, and ϵ is a vector of level 1 residual error terms. It is obvious that (3.26) also applies to the general SOEREG model as, for example, Y could be multivariate with possibly more than two levels, with X and Z containing any number of predictors.

The general SOEREG model is similar to the linear mixed-effects model (see Section 2.5.5) and the general Bayesian linear model of Lindley and Smith (1972). It also encompasses more complex models, such as the small - area model of Ghosh and Rao (1994) with spatial dependencies between the random effects.

3.6 Prior specifications

An important aspect of the Bayesian approach to consider, before proceeding with the analysis of our multilevel models, is the specification of priors. In Section 2.12.1 we discussed the difficulties involved in eliciting detailed priors required for a full

Bayesian analysis. The Bayes linear methods adopted in this thesis require specifications of only first and second-order moments (means, variances and covariances) for the adjustment of mean components, and fourth-order moments for the adjustment of variance components. Specifications of even these limited beliefs can be difficult in practice.

Goldstein & Wooff (2007) (page 41) describe a six-step iterative procedure that can guide elicitation of subjective prior beliefs. Although prior specifications will obviously depend on the application context, the procedure is quite general and sensible. We shall make use of some steps of this procedure to elicit the priors required in our Bayes linear analysis of the SOEREF model applied to the STAT1010 data below.

Alternatively, the required prior moments may be specified in accordance with suitably chosen probability densities. This approach can be particularly useful when using simulation to investigate otherwise intractable properties of statistical methods, such as the two-stage Bayes linear update of the SOEREG model, including sensitivity to the choice of prior moments.

3.7 Prior specifications for the STAT1010 Example

In general, for any prior elicitation task we must first identify the quantities for which prior specifications are needed. SOE judgements, from which the quantities in our models emanate, coupled with subject matter knowledge, will be useful in guiding our thought processes in making beliefs specifications. For more complex models with complicated correlation structures, such as multilevel models with many hierarchical levels and covariates at all levels including cross-level interactions, Bayes linear graphical models (Goldstein & Wooff (2007) (page 47)) are effective tools to visualize inter-dependencies among parameters in the model thus facilitating prior elicitation.

Based on Definition 3.1.1. of the SOEREF model, we need to specify first and second order moments in respect of the following three quantities: $\mathcal{M}(y)$, $\mathcal{R}_j(\mathcal{M}(y))$,

and $\mathcal{R}_i(y_j)$.

3.7.1 Priors for the overall mean $\mathcal{M}(y)$

To specify the prior expectation for the overall mean $\mathcal{M}(y)$, we make use of the SOE representation $\mathcal{M}(y_j) = \mathcal{M}(y)+\mathcal{R}_j(\mathcal{M}(y))$ of the population group j means in (3.6). The latter implies that we must consider the mean examination scores in each class j and from these deduce our prior expectation of the overall mean $\mathcal{M}(y)$. This task is made easier from our wide experience in teaching and examining STAT1010 in various classes. We know, for instance, that students from engineering classes have higher mathematical abilities and thus do well in the STAT1010 examinations while students from the management classes, where mathematics is not a requirement, perform less well in the STAT1010 examinations. In our experience a class average would very rarely fall below 40% or rise above 70%. Judging our uncertainty to be approximately symmetric over the interval $(40\%, 70\%)$, we take the mid-point of this interval as our prior expectation, thus we set $E(\mathcal{M}(y)) = 55\%$.

Although we know that engineering and management classes differ in their mathematical and statistical abilities, we have nevertheless treated them as exchangeable above; we suppose we do not know which type of class corresponds to each label j. Similarly, in making exchangeability judgements among students (within classes), we have ignored level 1 covariates such as A level scores or gender, which could discriminate between students' examination scores. The reason for ignoring these covariates here is for illustrating Bayes linear fitting and, more importantly, diagnostic checking of the SOEREF model. The more detailed information conveyed by level 1 and level 2 covariates will be considered in the SOEREG model.

Specifying uncertainty is a rather unfamiliar, and even difficult, task in general. To simplify our specification of the uncertainty in $\mathcal{M}(y)$, we assume a Gaussian distribution for the class means. This is not an unreasonable assumption because even if observations are not Gaussian, the distribution of their means would still be closer to a Gaussian distribution. Considering a 95% interval $(40\%, 70\%)$ implies the interval size of 30 corresponds to approximately four times the standard deviation. Hence, we specify $Var(\mathcal{M}(y)) = (\frac{30}{4})^2 = 56.3$.

One possible complication is the occurrence of outliers in the form of exceptionally strong (engineering) or weak (management) classes that may invalidate the Gaussian assumption. A well known solution to the latter case, is the assumption of a t distribution with low degrees of freedom. In addition to the above suggestions, collection of auxilliary (an introductory Mathematics module) or historical data (past STAT1010 data) may ease the difficult task of eliciting genuine subjective beliefs.

Although we have referred to the Gaussian distribution, still we are not required to make a full probabilistic specification under the Bayes linear approach; we only require first and second-order moments. Therefore, if we consider that the Gaussian form somewhat underestimates the uncertainty, we can make a direct assessment as to how much to increase uncertainty, for example 10% or 20% of our initial variance specification. And because the Bayes linear analysis is fast, it is relatively straightforward to explore the sensitivity of the analysis to the amount by which uncertainty is increased.

3.7.2 Priors for the level one residual $\mathcal{R}_i(y_j)$

The expectation of the level 1 residual is zero (Definition 3.1.1). To specify the level 1 variance $Var(\mathcal{R}_i(y_j)) = \sigma^2_\epsilon$, recall that in Section 3.1.1, we assumed that σ^2_ϵ is constant for all individuals and groups in line with our SOEREF model. If however we judge level 1 variances are not constant within groups, then the SOEREF model is no longer valid. Suppose, for example, that females are more consistent in their examination performances than male, then the level 1 variance in examination score will be lower for female than male. Thus we shall need to make more detailed specifications by considering SOE judgements separately for female and for male leading to a SOEREG model as in (3.16) with the covariate z_{ji} representing gender. In contrast to (3.16) however, we also need to model the level 1 residual ϵ_{ji} as a function of z_{ji} in order to account for the dependence of the level 1 variance on gender, that is $\epsilon_{ji} = \epsilon_{[0]ji} + \epsilon_{[1]ji} z_{ji}$. Then $Var(\epsilon_{ji}) = \sigma^2_0 + 2\sigma_{01} z_{ji} + \sigma^2_1 z^2_{ji}$, that is the level 1 variance is a quadratic function in z_{ji} with the constraint $Var(\epsilon_{ji}) > 0$. Thus we obtain a SOEREG model with *complex level 1 variation*, analogous to the

complex multilevel model in Goldstein (2010).

For now we assume the SOEREF model with constant variances is suitable. In order to assess σ_ϵ^2, we reflect on the distribution of individual examination marks we have seen during our previous marking exercises. While it is difficult to make detailed beliefs statements about the distribution of examination marks, we feel more confident to express beliefs about a few percentile marks. Thus, we believe that about five percent of students across all classes get 30% or fewer marks and the same percentage get 80% or more. Using the fifth and ninety-fifth percentiles, denoted by p_5 and p_{95} respectively , Perry and Greig (1975) as reported in Hull (1978), estimate the standard deviation as follows:

$$\sigma_\epsilon = \frac{p_{95} - p_5}{3.25},$$

Thus $\sigma_\epsilon = \frac{80-30}{3.25} = 15.4$. If we assume a Gaussian distribution with the same mean 55 as $E(\mathcal{M}(y))$ above, we obtain $\Phi((80-55)/15.4) = 0.95$ and $\Phi((30-55)/15.4) = 0.05$ which agree with our assessment. Hence we specify $Var(\mathcal{R}_i(y_j)) = 15.4^2 \approx 237$.

3.7.3 Priors for the level two residual $\mathcal{R}_j(\mathcal{M}(y))$

Similar to the level 1 residual, the expectation of the level 2 residual is also zero. To specify the level 2 variance $Var(\mathcal{R}_j(\mathcal{M}(y))) = (\sigma_u^2 - \gamma)$, we may adopt several alternative approaches depending on how confident we feel in making the required assessments as well as on any additional information that may be available to us. We shall consider two such approaches here, with a view to contrasting the different specification processes involved and, more importantly, to gauge the consistency of our methods by comparing the results of the two approaches.

Our first, more direct, approach utilizes the level 2 representation $\mathcal{M}(y_j) = \mathcal{M}(y) + \mathcal{R}_j(\mathcal{M}(y))$ as in sub-section 3.7.1 above, where we specified $Var(\mathcal{M}(y)) = \gamma = 56.3$. Hence, we only need to specify our uncertainty in the population group j means, $Var(\mathcal{M}(y_j)) = \sigma_u^2$, ensuring that $(\sigma_u^2 - \gamma) \geq 0$ for coherence. However, in our assessment of $Var(\mathcal{M}(y))$ in sub-section 3.7.1 we raised concern about the possibility of the distribution of population group j means $\mathcal{M}(y_j)$ having a bimodal distribution or outliers. As a result, we feel less confident in assessing the uncertainty in $\mathcal{M}(y_j)$

compared to $\mathcal{M}(y)$, since we now have to think about how the individual $\mathcal{M}(y_j)$ vary around their mean $\mathcal{M}(y)$. To reflect our increased uncertainty, we decide on the direct assessment $\sigma_u^2 = 2\gamma$, from which we specify $Var(\mathcal{R}_j(\mathcal{M}(y))) = (\sigma_u^2 - \gamma) = 56.3$.

Our second, indirect, approach starts with the consideration that since we have already specified σ_ϵ^2, specification of $(\sigma_u^2 - \gamma)$ will automatically fix the intra-cluster correlation ρ. We shall therefore assess the level 2 variance indirectly, that is via specification of ρ as follows.

The intra-cluster correlation ρ is given by

$$\rho = \frac{(\sigma_u^2 - \gamma)}{(\sigma_u^2 - \gamma) + \sigma_\epsilon^2}, \tag{3.27}$$

where ρ is defined as the proportion of total variation in the data that is accounted for by variation between groups. It also reflects the correlation among level 1 units within a specific group; the more similar level 1 units are, the higher ρ will be. We may therefore assess the value of ρ directly by judging the similarity of level 1 units within an average group and then use (3.27) above to obtain our prior for $(\sigma_u^2 - \gamma)$. Quantifying a correlation such as ρ may be not be an easy task though.

Alternatively, we may use values of ρ as reported in relevant research as a guide in our specification task. The intra-cluster correlation is an important quantity that plays a vital role in cluster sampling (see Chapter 4) and cluster randomized trial designs and, therefore, there are a number of research studies, especially in education, reporting the magnitude of the intra-cluster correlation. One such study that is of value to us here is Hedges and Hedberg (2007). The purpose of their study, as stated by the authors, is to provide a comprehensive collection of intraclass correlations in mathematics and reading achievements. We focus on intraclass correlations for mathematics achievement tests, that we judge relevant to our STAT1010 examinations, and where values of ρ between 0.1 and 0.3 were reported. Choosing $\rho = 0.2$ and, using (3.27) with $\sigma_\epsilon^2 = 237$, we calculate $Var(\mathcal{R}_j(\mathcal{M}(y))) \approx 59$. The latter value is consistent with the value of 56.3 that we assessed directly above, though it is slightly larger. So we specify $Var(\mathcal{R}_j(\mathcal{M}(y))) = 59$.

3.7.4 Summary of prior specifications, their implications, and some reflections

We now collect together our prior specifications:

$$E(\mathcal{M}(y)) = 55, \quad Var(\mathcal{M}(y)) = 56.3, \quad Var(\mathcal{R}_i(y_j)) = 237,$$
$$Var(\mathcal{R}_j(\mathcal{M}(y))) = 59 \quad \forall i, j. \tag{3.28}$$

The underlying population overall mean and the residuals are not observable, but they are related to the observable quantities in the STAT1010 data via the SOEREF model. Hence our specifications in (3.28) have implications for these observables and it is important that we explore these.

Using the SOEREF model $y_{ji} = \mathcal{M}(y)) + \mathcal{R}_j(\mathcal{M}(y)) + \mathcal{R}_i(y_j)$ and the values in (3.28), we obtain $E(y_{ji}) = 55$ and $Var(y_{ji}) = 352.3$ (standard deviation=18.5) for a student i in class j. These values together with an assumed Gaussian distribution imply that 95% of students are expected to have examination marks in the interval $(18.1, 91.8)$. There is nothing alarming about this interval and they represent a reasonable range of marks in our experience.

For any two students i and i' in the same class j, we calculate the correlation

$$Corr(y_{ji}, y_{ji'}) = \frac{Var(\mathcal{M}(y) + \mathcal{R}_j(\mathcal{M}(y)))}{Var(\mathcal{M}(y) + \mathcal{R}_j(\mathcal{M}(y))) + Var(\mathcal{R}_i(y_j))} \qquad = 0.336 \tag{3.29}$$

which is somewhat larger than our specified value for the intra-cluster correlation $\rho = 0.2$ which also measures the correlation between marks of students in the same class. This is expected as $Var(\mathcal{M}(y)$ induces extra correlation in $(y_{ji}, y_{ji'})$. Comparing (3.29) and (3.27), the formula for ρ, we note that the difference is due to our uncertainty in the overall mean $Var(\mathcal{M}(y)))$ which appears in the numerator and denominator of (3.29) but not in (3.27). In Sub-section 3.7.3 we mentioned values of ρ as large as 0.3 based on the study by Hedges and Hedberg (2007), therefore the value of 0.336 above is no cause for concern.

For students in different classes j and j' we calculate $Corr(y_{ji}, y_{j'i'}) = 0.164$. This correlation, as expected, is smaller than that within classes. Also both correlations are low and typical of those occurring in multilevel data sets.

Finally, let us reflect on the specification task we went through above. We have used a variety of methods such as judgement of symmetry, comparison with the Gaussian distribution, use of percentiles and auxilliary information. We have also made the necessary coherence checks while ensuring that our specifications have reasonable implications for our observables. But above all what is revealing in the above exercise is how difficult the specification task is, even for the simplest SOEREF multilevel model. It is therefore not surprising that in statistical modelling shortcuts such as conjugacy based on convenient probability distributional forms are often used. In important applications though, where the statistical modelling has crucial implications for people (their health for example) and nations (the issue of climate change for example), and where data is scarce and difficult to collect but some expertise is available, then we have no other alternative than to adopt a pragmatic subjective approach. Such an approach will depend on the specification of genuine priors. To quote Goldstein (2006): *" the subjectivist Bayes approach is the only feasible method for tackling many important practical problems."* and also *" A true subjectivist formulation should start by recognising the limited abilities of the individual to make large collections of uncertainty specifications."*, making the Bayes linear approach, that only requires limited belief specifications, worth considering in tackling important real life problems.

Chapter 4

Bayes linear adjustment of mean components in SOEREF multilevel models

The Bayes linear approach, that requires only limited beliefs specifications, is a comprehensive methodology for model formulation, adjustment (estimation), and diagnostic checking. In Chapter three we used this subjectivist approach to formulate the SOEREF and SOEREG multilevel models. In this chapter we shall consider model adjustment and diagnostic checking.

We begin by using Bayes linear methods to adjust mean components in the unbalanced SOEREF model, of which the balanced model is simply a special case. Applying Bayes linear sufficiency, closed form expressions for the adjusted means are derived. These are considered difficult to obtain analytically in the case of unbalanced data (see Searle et al.(1992)), and appear rarely in the literature. The closed form expressions of the adjusted means are useful in understanding how the adjusted quantities relate to our prior specifications and the data.

We then consider applying Bayes linear methods to analyse beliefs over the unbalanced SOEREF model using observable data. This involves beliefs specification, adjustment and interpretation, as well as diagnostic checks. Beliefs specifications are illustrated in the context of the STAT1010 data. Computations of adjusted quantities, as well as diagnostic checks, have been implemented in the Bayes linear

programming tool B/D. We thus use B/D to perform the analysis of beliefs over the SOEREF model as applied to the STAT1010 data.

An important focus of this chapter is diagnostic checks. Diagnostics for multilevel models are more complicated than for ordinary linear models (Hodges, 1998). In Chapter three we have considered the potential for using Bayes linear diagnostics and, partial diagnostics in particular, for multilevel modeling. Here we apply these diagnostics to the unbalanced SOEREF model. We apply our methods to the design and sample size determination of the SOEREF model.

Finally, we relax the assumption of infinite exchangeability and consider the formulation and adjustment of the finite SOEREF model comparing these to the infinite exchangeability versions.

4.1 Updating mean components in the unbalanced SOEREF multilevel model

Below we shall derive closed-form expressions of the updated mean components for the unbalanced SOEREF model. Our primary interest in the closed-form expressions is to understand how the adjusted quantities depend on the prior and data. Searle et al. (1992) consider these expressions difficult to derive analytically for the unbalanced situation, hence the additional motivation to present them here. The main difficulty in obtaining the closed-form expressions appears to stem from the calculation of the inverse of the variance-covariance matrix of the unbalanced data. This inverse is also key to the application of the Bayes linear rule and is therefore considered first below.

4.2 Calculation of $Var(\bar{D}_n)$ and its inverse

Consider the two-level dataset y_{ji} where in each group j we have a sample of size n_j. Let $\bar{D}_n = \{\bar{y}_{1.}, \bar{y}_{2.}, \ldots, \bar{y}_{J.}\}$ be the collection of group means, where

$$\bar{y}_{j.} = \sum_{i=1}^{n_j} \frac{y_{ji}}{n_j} \quad \forall j$$

Then $\bar{D}_n$ is Bayes linear sufficient for adjusting beliefs over the mean components. The SOEREF model for $\bar{y}_{j.}$, as well as its belief specifications, can be derived from Definition 3.1.1 in Section 3.1.1. Hence, we write

$$\bar{y}_{j.} = \mathcal{M}(y) + \mathcal{R}_j(\mathcal{M}(y)) + \bar{\mathcal{R}}_{n_j}(y_{j.})$$

where $\bar{\mathcal{R}}_{n_j}(y_{j.})$ is the mean of the n_j level 1 residuals in group j. It follows that

$$E(\bar{y}_{j.}) = \mu, \quad Var(\bar{y}_{j.}) = \sigma_u^2 + \frac{\sigma_\epsilon^2}{n_j}, \quad \forall j, \tag{4.1}$$

$$Cov(\bar{y}_{j.}, \bar{y}_{j'.}) = \gamma, \quad \forall j \neq j', \tag{4.2}$$

$$Cov(\bar{y}_{j.}, \mathcal{M}(y)) = \gamma \quad \forall j \tag{4.3}$$

Expressions (4.1) to (4.3) give the specifications over $\bar{D}_n$ as follows:

$$E(\mathcal{M}(y)) = \mu, \quad E(\bar{D}_n) = (\mu, \mu, \ldots, \mu)^T = \mathbf{1}_J\mu, \quad Cov(\mathcal{M}(y), \bar{D}_n) = \mathbf{1}_J^T\gamma.$$

$$Var(\bar{D}_n) = \begin{pmatrix} \sigma_u^2 + \frac{\sigma_\epsilon^2}{n_1} & \gamma & \cdots & \gamma \\ \gamma & \sigma_u^2 + \frac{\sigma_\epsilon^2}{n_2} & \cdots & \gamma \\ \vdots & \vdots & \ddots & \vdots \\ \gamma & \gamma & \cdots & \sigma_u^2 + \frac{\sigma_\epsilon^2}{n_J} \end{pmatrix} \tag{4.4}$$

The inverse of $Var(\bar{D}_n)$

To calculate $Var^{-1}(\bar{D}_n)$ we use the following expression for the inverse of the sum of two matrices due to Henderson et al. (1959).

$$(A + UBU')^{-1} = A^{-1} - A^{-1}U(B^{-1} + U'A^{-1}U)^{-1}U'A^{-1} \tag{4.5}$$

where A and B are both non-singular and symmetric conformable matrices, and U is a column vector. The required variance $Var(\bar{D}_n)$ may be written in the form of $(A + UBU')$, where

$$A = \begin{pmatrix} \sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n_1} & 0 & \cdots & 0 \\ 0 & \sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n_J} \end{pmatrix}$$

$$U^T = \Big(1 \dots 1\Big) \quad \text{and} \quad B = \gamma. \tag{4.6}$$

Application of (4.5) involves inverting only one matrix, namely A which is diagonal. If we write $a_j = \sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n_j}$ for the jth diagonal element of A, then A^{-1} is also diagonal with element a_j^{-1}, provided $a_j \neq 0$. In fact $a_j > 0$ for all j as it is the sum of the level 1 and level 2 residual variances (see specifications (3.7) and (3.9)). It is now easy to calculate $Var^{-1}(\bar{D}_n)$.

Theorem 4.2.1. *For $Var(\bar{D}_n)$ in (4.4), $Var^{-1}(\bar{D}_n)$ exists and has elements $d_{jj'}$ given by*

$$d_{jj'} = \begin{cases} a_j^{-1}(1 - \dfrac{a_j^{-1}}{[\gamma^{-1} + \sum_{j=1}^{2} a_j^{-1}]}) & \text{if} \quad j = j', \\ -\dfrac{a_j^{-1} a_{j'}^{-1}}{[\gamma^{-1} + \sum_{j=1}^{2} a_j^{-1}]} & \text{if} \quad j \neq j'. \end{cases} \tag{4.7}$$

Proof. We have argued above that $a_j > 0$, hence $Var^{-1}(\bar{D}_n)$ exists. Applying (4.5) using A, B and U as in (4.6) gives:

$$\begin{aligned} Var^{-1}(\bar{D}_n) &= A^{-1} - A^{-1}\mathbf{1}_J[\gamma^{-1} + \mathbf{1}_J^T A^{-1}\mathbf{1}]^{-1}\mathbf{1}_J^T A^{-1} \\ &= A^{-1} - \frac{A^{-1}\mathbf{1}_J\mathbf{1}_J^T A^{-1}}{[\gamma^{-1} + \mathbf{1}_J^T A^{-1}\mathbf{1}]} \end{aligned} \tag{4.8}$$

Since the term (j, j') of the numerator in (4.8) is $a_j^{-1}a_{j'}^{-1}$ and A^{-1} is diagonal, it is trivial to show that (4.8) has elements as in (4.7). ■

For example, for $J = 2$, (4.7) gives

$$Var^{-1}(\bar{D}_n) = \begin{pmatrix} a_1^{-1}(1 - \frac{a_1^{-1}}{[\gamma^{-1} + \sum_{j=1}^{2} a_j^{-1}]}) & -\frac{a_1^{-1}a_2^{-1}}{[\gamma^{-1} + \sum_{j=1}^{2} a_j^{-1}]} \\ -\frac{a_1^{-1}a_2^{-1}}{[\gamma^{-1} + \sum_{j=1}^{2} a_j^{-1}]} & a_2^{-1}(1 - \frac{a_2^{-1}}{[\gamma^{-1} + \sum_{j=1}^{2} a_j^{-1}]}) \end{pmatrix}$$

Direct multiplication shows that the above matrix is the correct inverse.

4.3 Adjusting the population grand mean

To adjust the population grand mean $\mathcal{M}(y)$, we make use of Definition 3.1.1 of the unbalanced SOEREF model as well as its specifications in Chapter 3, and Theorem 4.2.1 for $Var^{-1}(\bar{D}_n)$.

Theorem 4.3.1. *Let y_{ji} be an observation on individual i nested in group j in a two-level data set. Following Definition 3.1.1, the unbalanced SOEREF model is $y_{ji} = \mathcal{M}(y) + \mathcal{R}_j(\mathcal{M}(y)) + \mathcal{R}_i(y_j)$, where $i = 1, 2, \ldots, n_j$, and $j = 1, 2, \ldots, J$, with second-order specifications for the population grand mean as $E(\mathcal{M}(y)) = \mu$, and $Var(\mathcal{M}(y)) = \gamma$. The sequences of level 2 and level 1 residuals are uncorrelated, have means zero and variances $Var(\mathcal{R}_j(\mathcal{M}(y))) = \sigma_u^2 - \gamma$ and $Var(\mathcal{R}_i(y_j)) = \sigma_\epsilon^2$ respectively, $\forall i, j$. The collection of sample group means $\bar{D}_n = \{\bar{y}_{1.}, \bar{y}_{2.}, \ldots, \bar{y}_{J.}\}$ is Bayes linear sufficient for the adjustment of the population grand mean. We have*

$$E_{\bar{D}_n}(\mathcal{M}(y)) = \frac{\gamma^{-1}\mu + \sum_{j=1}^{J}(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n_j})^{-1}\bar{y}_{j.}}{\gamma^{-1} + \sum_{j=1}^{J}(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n_j})^{-1}} \tag{4.9}$$

Proof. First, we note that $(\mathbf{1}_J^T A^{-1}\mathbf{1}_J) = \sum_{j=1}^{J} a_j^{-1}$ is a scalar, which will simplify the calculations below.

$$E_{\bar{D}_n}(\mathcal{M}(y)) = E(\mathcal{M}(y)) + Cov(\mathcal{M}(y), \bar{D}_n)Var^{-1}(\bar{D}_n)(\bar{D}_n - E(\bar{D}_n))$$

$$= \mu + \mathbf{1}_J^T\gamma\Big[A^{-1} - \frac{A^{-1}\mathbf{1}_J\mathbf{1}_J^T A^{-1}}{[\gamma^{-1} + \mathbf{1}_J^T A^{-1}\mathbf{1}]}\Big](\bar{D}_n - \mathbf{1}_J\mu)$$

$$= \mu + \Big[\frac{\mathbf{1}_J^T\gamma A^{-1}[\gamma^{-1} + \mathbf{1}_J^T A^{-1}\mathbf{1}] - \mathbf{1}_J^T\gamma A^{-1}\mathbf{1}_J\mathbf{1}_J^T A^{-1}}{[\gamma^{-1} + \mathbf{1}_J^T A^{-1}\mathbf{1}]}\Big]$$
$$(\bar{D}_n - \mathbf{1}_J\mu)$$

$$= \mu + \Big[\frac{\mathbf{1}_J^T A^{-1} + \gamma\mathbf{1}_J^T A^{-1}(\mathbf{1}_J^T A^{-1}\mathbf{1}) - \gamma(\mathbf{1}_J^T A^{-1}\mathbf{1}_J)\mathbf{1}_J^T A^{-1}}{[\gamma^{-1} + \mathbf{1}_J^T A^{-1}\mathbf{1}]}\Big]$$
$$(\bar{D}_n - \mathbf{1}_J\mu)$$

$$= \mu + \Big[\frac{\mathbf{1}_J^T A^{-1}}{[\gamma^{-1} + \mathbf{1}_J^T A^{-1}\mathbf{1}]}\Big](\bar{D}_n - \mathbf{1}_J\mu)$$
$$= \Big[1 - \frac{\mathbf{1}_J^T A^{-1}\mathbf{1}}{[\gamma^{-1} + \mathbf{1}_J^T A^{-1}\mathbf{1}]}\Big]\mu + \Big[\frac{\mathbf{1}_J^T A^{-1}}{[\gamma^{-1} + \mathbf{1}_J^T A^{-1}\mathbf{1}]}\Big]\bar{D}_n$$

$$= \Big[\frac{\gamma^{-1}}{[\gamma^{-1} + \mathbf{1}_J^T A^{-1}\mathbf{1}]}\Big]\mu + \Big[\frac{\mathbf{1}_J^T A^{-1}}{[\gamma^{-1} + \mathbf{1}_J^T A^{-1}\mathbf{1}]}\Big]\bar{D}_n$$

$$= \Big[\frac{1}{[1 + \gamma\sum_{j=1}^J a_j^{-1}]}\Big]\mu + \Big[\frac{\gamma}{1 + \gamma\sum_{j=1}^J a_j^{-1}]}\Big]\sum_{j=1}^J a_j^{-1}\bar{y}_{j.}$$

$$= \frac{\gamma^{-1}\mu + \sum_{j=1}^J(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n_j})^{-1}\bar{y}_{j.}}{\gamma^{-1} + \sum_{j=1}^J(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n_j})^{-1}} \tag{4.10}$$

■

Corollary 4.3.1. *In the balanced case the adjusted grand mean is*

$$E_{\bar{D}_n}(\mathcal{M}(y)) = \frac{\gamma^{-1}\mu + J(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n})^{-1}\bar{y}_{..}}{\gamma^{-1} + J(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n})^{-1}} \tag{4.11}$$

Proof. Substituting $n_j = n$ in (4.10) gives (4.11). ■

The adjusted expectation $E_{\bar{D}_n}(\mathcal{M}(y))$ in (4.10) reveals the familiar Bayesian precision weighted average of prior mean μ and data mean $\bar{y}_j$. The weights are the prior precision $\gamma^{-1} = (var(\mathcal{M}(y)))^{-1}$ and data precision $\sum_{j=1}^J(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n_j})^{-1} = \sum_{j=1}^J Var(\mathcal{R}_j(\mathcal{M}(y)) + \bar{\mathcal{R}}_{n_j}(y_{j.}))^{-1}$, that is the sum over all J groups of the reciprocal of the level 2 and level 1 residual variances.

Note that putting $Var(\mathcal{R}_j(\mathcal{M}(y)) = (\sigma_u^2 - \gamma) = 0$, results in a single level data and, removing all terms in j in (4.11), we obtain the well-known Bayesian estimator of an unknown population mean

$$\frac{\frac{1}{\gamma}\mu + \frac{n}{\sigma^2}\bar{y}}{\frac{1}{\gamma} + \frac{n}{\sigma^2}}$$

To compare (4.10) with the result of a full Bayesian analysis, consider the linear mixed-effects model

$$y_{ji}|\theta_j \sim N(\theta_j, \sigma^2), \quad i = 1, 2, \ldots, n_j; \quad j = 1, 2, \ldots, J.$$

$$\theta_j \sim N(\mu, \tau^2), \quad j = 1, 2, \ldots, J,$$

which corresponds to our SOEREF model except that, like in most Bayesian analyses of multilevel models, a non-informative uniform prior is assumed for the "fixed-effect" μ, the argument being that typically enough data is available to estimate the overall mean, even in small studies. If normal priors are assumed for the level 1 and level 2 residuals both with mean zero and variances σ_j^2 and τ^2 respectively as in Gelman et al. (2009; p140) (see also, for example, Hill (1965) or (Lindley and Smith (1972)), then the estimated posterior mean is

$$\frac{\sum_{j=1}^{J} \frac{1}{\sigma_j^2+\tau^2}\bar{y}_j}{\sum_{j=1}^{J} \frac{1}{\sigma_j^2+\tau^2}}, \tag{4.12}$$

which is in agreement with (4.10) since for a noninformative prior on $\mathcal{M}(y)$, $\gamma^{-1} \to 0$.

In addition to the adjustment being a compromise between prior and data, (4.10) also reveals that a specific group j with greater precision due, for example, to a larger sample size n_j in that group, will contribute more weight to that $\bar{y}_j$ in the final adjustment.

The preceding argument also implies that in the balanced case all groups of data will contribute equally to the final adjustment as is clear from Corollary 4.3.1. Note that (4.11) agrees with the result for the balanced case given in (Searle et al.,1992), page 335.

4.3.1 The adjusted variance of $\mathcal{M}(y)$

Theorem 4.3.2. *The adjusted variance of $\mathcal{M}(y)$ by $\bar{D}_n$ is*

$$Var_{\bar{D}_n}(\mathcal{M}(y)) = \frac{1}{\gamma^{-1} + \sum_{j=1}^{J}(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n_j})^{-1}} \tag{4.13}$$

Proof. The Bayes linear rule for the adjusted variance is

$$\begin{aligned} Var_{\bar{D}_n}(\mathcal{M}(y)) &= Var(\mathcal{M}(y)) - Cov(\mathcal{M}(y), \bar{D}_n) Var^{-1}(\bar{D}_n) Cov(\bar{D}_n, \mathcal{M}(y)) \\ &= \gamma - \frac{\mathbf{1}_J^T A^{-1} \mathbf{1} \gamma}{[\gamma^{-1} + \mathbf{1}_J^T A^{-1} \mathbf{1}]}, \quad \text{see (4.10)} \\ &= \frac{1}{[\gamma^{-1} + \mathbf{1}_J^T A^{-1} \mathbf{1}]}, \end{aligned}$$

which, upon replacing A^{-1}, gives the required result. ■

We may also write (4.13) in terms of precision, that is inverse variance, as follows

$$\frac{1}{Var_{\bar{D}_n}(\mathcal{M}(y))} = \frac{1}{\gamma} + \frac{1}{\sum_{j=1}^{J}(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n_j})},$$

which is the familiar Bayesian form where *posterior precision equals the prior precision plus the data precision.*

4.3.2 The resolved variance of $\mathcal{M}(y)$

The variance of $\mathcal{M}(y)$ resolved by $\bar{D}_n$ (Definition 2.13.2) is

$$\begin{aligned} RVar_{\bar{D}_n}(\mathcal{M}(y)) &= Cov(\mathcal{M}(y), \bar{D}_n) Var^{-1}(\bar{D}_n) Cov(\bar{D}_n, \mathcal{M}(y)) \\ &= \frac{\mathbf{1}_J^T A^{-1} \mathbf{1} \gamma}{[\gamma^{-1} + \mathbf{1}_J^T A^{-1} \mathbf{1}]} \\ &= \frac{\gamma \sum_{j=1}^{J}(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n_j})^{-1}}{\gamma^{-1} + \sum_{j=1}^{J}(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n_j})^{-1}} \end{aligned} \tag{4.14}$$

and the resolution is

$$\begin{aligned} R_{\bar{D}_n}(\mathcal{M}(y)) &= \frac{RVar_{\bar{D}_n}(\mathcal{M}(y))}{Var(\mathcal{M}(y))} \\ &= 1 - \frac{Var_{\bar{D}_n}(\mathcal{M}(y))}{Var(\mathcal{M}(y))} \\ &= 1 - \frac{1}{1 + \gamma \sum_{j=1}^{J}(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n_j})^{-1}} \end{aligned} \tag{4.15}$$

The resolution $R_{\bar{D}_n}(\mathcal{M}(y))$, a scale-free measure of the efficiency of an adjustment, shows the proportion of prior variance explained for the population grand mean $\mathcal{M}(y)$ after adjusting $\mathcal{M}(y)$ by $\bar{D}_n$.

$R_{\bar{D}_n}(\mathcal{M}(y))$ lies between zero and one. When $Var_{\bar{D}_n}(\mathcal{M}(y)) = Var(\mathcal{M}(y))$, $R_{\bar{D}_n}(\mathcal{M}(y))$ will be zero. This implies that $\bar{D}_n$ is not informative in adjusting $\mathcal{M}(y)$. At the other end, as the number of level 2 groups J increases, $\gamma \sum_{j=1}^{J}(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n_j})^{-1}$ in (4.15) will tend towards infinity, hence $R_{\bar{D}_n}(\mathcal{M}(y))$ will approach its maximum value of one.

4.4 Adjusting the population group j mean

In multilevel modelling, learning about the effect of clustering (grouping) on an outcome defined at the individual level is of considerable interest to researchers

(Section 2.3.1). For the SOEREF model, this involves learning about population group means which we shall consider in this section. To adjust the population group means $\mathcal{M}(y_j)$, we first need to adjust the level 2 residuals, that is $E_{\bar{D}_n}(\mathcal{R}_j(\mathcal{M}(y))$. The level 2 residuals represent group effects and so are themselves of importance in multilevel modelling.

4.4.1 Adjusting the level 2 residuals

Theorem 4.4.1. *The adjustment of each level 2 residual $\mathcal{R}_j(\mathcal{M}(y)$ of the unbalanced SOEREF model by the collection of sample group means $\bar{D}_n = \{\bar{y}_{1.}, \bar{y}_{2.}, \ldots, \bar{y}_{J.}\}$ is*

$$E_{\bar{D}_n}(\mathcal{R}_j(\mathcal{M}(y)) = \frac{(\sigma_u^2 - \gamma)}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n_j})}\left[(\bar{y}_{j.} - \mu) - \frac{\sum_{j'=1}^{J}(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n_j})^{-1}(\bar{y}_{j.} - \mu)}{\gamma^{-1} + \sum_{j=1}^{J}(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n_j})^{-1}}\right] \quad (4.16)$$

Proof. Let $\mathcal{C}_{R_j}$ denote the vector of level 2 residuals $\{\mathcal{R}_1(\mathcal{M}(y)),$ $\mathcal{R}_2(\mathcal{M}(y)), \ldots \mathcal{R}_J(\mathcal{M}(y))\}$ in the unbalanced SOEREF model with specifications $\mathrm{E}(\mathcal{R}_j(\mathcal{M}(y)) = 0$ and $\mathrm{Var}(\mathcal{R}_j(\mathcal{M}(y)) = \sigma_u^2 - \gamma$, for $j = 1, 2, \ldots, J$ (see Section 3.1.1). Based on these specifications,

$$E(\mathcal{C}_{R_j}) = \mathbf{1}_J 0, \quad Cov(\mathcal{C}_{R_j}, \bar{D}_n) = \mathbf{I}_J(\sigma_u^2 - \gamma), \forall j, \quad (4.17)$$

where $\mathbf{1}_J$ is a column of J ones and $\mathbf{I}_J$ is an identity matrix of dimension J. Applying

the Bayes linear rule gives the adjusted vector of level 2 residuals as follows.

$$E_{\bar{D}_n}(\mathcal{C}_{R_j}) = E(\mathcal{C}_{R_j}) + Cov(\mathcal{C}_{R_j}, \bar{D}_n)Var^{-1}(\bar{D}_n)(\bar{D}_n - E(\bar{D}_n))$$

$$= \mathbf{1}_J 0 + (\sigma_u^2 - \gamma)\mathbf{I}_J\Big[A^{-1} - \frac{A^{-1}\mathbf{1}_J\mathbf{1}_J^T A^{-1}}{[\gamma^{-1} + \mathbf{1}_J^T A^{-1}\mathbf{1}]}\Big](\bar{D}_n - \mathbf{1}_J\mu)$$

$$= (\sigma_u^2 - \gamma)\Big[\mathbf{I}_J - \frac{A^{-1}\mathbf{1}_J\mathbf{1}_J^T}{[\gamma^{-1} + \mathbf{1}_J^T A^{-1}\mathbf{1}]}\Big]A^{-1}(\bar{D}_n - \mathbf{1}_J\mu)$$

$$= (\sigma_u^2 - \gamma)\begin{pmatrix} 1 - \frac{1}{da_1} & -\frac{1}{da_1} & \cdots & -\frac{1}{da_1} \\ -\frac{1}{da_2} & 1 - \frac{1}{da_2} & \cdots & -\frac{1}{da_2} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{1}{da_J} & -\frac{1}{da_J} & \cdots & 1 - \frac{1}{da_J} \end{pmatrix}\begin{pmatrix} \frac{1}{a_1}(\bar{y}_{1\cdot} - \mu) \\ \frac{1}{a_2}(\bar{y}_{2\cdot} - \mu) \\ \vdots \\ \frac{1}{a_J}(\bar{y}_{J\cdot} - \mu) \end{pmatrix}$$

$$= (\sigma_u^2 - \gamma)\begin{pmatrix} \frac{1}{a_1}(\bar{y}_{1\cdot} - \mu) - \frac{1}{da_1}\sum_{j=1}^J \frac{1}{a_j}(\bar{y}_{j\cdot} - \mu) \\ \frac{1}{a_2}(\bar{y}_{1\cdot} - \mu) - \frac{1}{da_2}\sum_{j=1}^J \frac{1}{a_j}(\bar{y}_{j\cdot} - \mu) \\ \vdots \\ \frac{1}{a_J}(\bar{y}_{J\cdot} - \mu) - \frac{1}{da_J}\sum_{j=1}^J \frac{1}{a_j}(\bar{y}_{j\cdot} - \mu) \end{pmatrix}, \tag{4.18}$$

where $a_j = (\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n_j})$ *and* $d = \gamma^{-1} + \sum_{j=1}^J (\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n_j})^{-1}$. *The* j*th row of (4.18) gives (4.16).* ■

Corollary 4.4.1. *In the balanced case, the adjusted level 2 residual* $\mathcal{R}_j(\mathcal{M}(y)$*, for* $j = 1, 2, \ldots, J$ *is*

$$E_{\bar{D}_n}(\mathcal{R}_j(\mathcal{M}(y)) = \frac{(\sigma_u^2 - \gamma)}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n})}\left[(\bar{y}_{j\cdot} - \mu) - \frac{J\gamma(\bar{y}_{\cdot\cdot} - \mu)}{J\gamma + (\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n})}\right] \tag{4.19}$$

Proof. Substituting $n_j = n$ in (4.16) and summing over the resulting constant terms gives

$$E_{\bar{D}_n}(\mathcal{R}_j(\mathcal{M}(y)) = \frac{(\sigma_u^2 - \gamma)}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n})}\left[(\bar{y}_{j\cdot} - \mu) - \frac{J(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n})^{-1}(\bar{y}_{\cdot\cdot} - \mu)}{\gamma^{-1} + J(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n})^{-1}}\right],$$

which, upon multiplying the numerator and denominator of the right-most term in brackets by $\gamma(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n})$, yields (4.19). ■

We note that (4.19) agrees with the result in Searle et al.(1992; page 336), hence providing a check on our calculation. Also, since (4.16) involves μ, we could combine Theorem 4.3.1 and Theorem 4.4.1 to yield the following corollary.

Corollary 4.4.2. *The adjusted jth level 2 residual can also be written as*

$$E_{\bar{D}_n}(\mathcal{R}_j(\mathcal{M}(y))) = \frac{(\sigma_u^2 - \gamma)}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n_j})}(\bar{y}_{j.} - E_{\bar{D}_n}(\mathcal{M}(y))) \tag{4.20}$$

Proof. Re-write (4.16) of Theorem 4.4.1 as

$$\begin{aligned}
E_{\bar{D}_n}(\mathcal{R}_j(\mathcal{M}(y))) &= \frac{(\sigma_u^2 - \gamma)}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n_j})}\left[\bar{y}_{j.} - \Big(\mu + \frac{\sum_{j=1}^J (\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n_j})^{-1}(\bar{y}_{j.} - \mu)}{\gamma^{-1} + \sum_{j=1}^J (\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n_j})^{-1}}\Big)\right] \\
&= \frac{(\sigma_u^2 - \gamma)}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n_j})}\left[\bar{y}_{j.} - \Big(\mu + \frac{\sum_{j=1}^J a_j^{-1}(\bar{y}_{j.} - \mu)}{\gamma^{-1} + \sum_{j=1}^J a_j^{-1}}\Big)\right] \\
&= \frac{(\sigma_u^2 - \gamma)}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n_j})}\left[\bar{y}_{j.} - \Big(\frac{\mu\gamma^{-1} + \mu\sum_{j=1}^J a_j^{-1}}{\gamma^{-1} + \sum_{j=1}^J a_j^{-1}}\right. \\
&\left.\frac{+\sum_{j=1}^J a_j^{-1}\bar{y}_{j.} - \sum_{j=1}^J a_j^{-1}\mu}{\gamma^{-1} + \sum_{j=1}^J a_j^{-1}}\Big)\right] \\
&= \frac{(\sigma_u^2 - \gamma)}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n_j})}\left[\bar{y}_{j.} - \Big(\frac{\mu\gamma^{-1} + \sum_{j=1}^J (\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n_j})^{-1}\bar{y}_{j.}}{\gamma^{-1} + \sum_{j=1}^J (\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n_j})^{-1}}\Big)\right] \\
&= \frac{(\sigma_u^2 - \gamma)}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n_j})}(\bar{y}_{j.} - E_{\bar{D}_n}(\mathcal{M}(y))),
\end{aligned}$$

where $a_j = (\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n_j})$. ■

The form of the Bayes linear adjustment in (4.20) is similar to the traditional multilevel shrinkage estimator of the level 2 residual, (see, for example, Goldstein(1995; page 10)). The deviation $(\bar{y}_{j.} - E_{\bar{D}_n}(\mathcal{M}(y)))$ is similar to the raw level 2 residual. We write

$$\eta = \frac{(\sigma_u^2 - \gamma)}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n_j})},$$

where η is the familiar shrinkage factor. The factor η, which is always less than or equal to one, shows by how much the estimated level 2 residual is shrunk towards its prior specified value of zero. We shall see that η occurs frequently in the forthcoming Bayes linear analyses in this chapter. We interpret $E_{\bar{D}_n}(\mathcal{R}_j(\mathcal{M}(y)))$ as follows.

1. *Effect of small n_j.* If n_j is small for a specific group j, then η is close to zero, hence $E_{\bar{D}_n}(\mathcal{R}_j(\mathcal{M}(y)))$ is shrunk to zero.
   Explanation. When n_j is small, the data $\bar{y}_{j.}$ is not a reliable estimator of the group j mean, and is thus not informative for the adjustment $E_{\bar{D}_n}(\mathcal{R}_j(\mathcal{M}(y)))$ which is therefore shrunk to its prior mean zero. A similar explanation holds in cases where σ_ϵ^2 is large or $(\sigma_u^2 - \gamma)$ is small, resulting in η being close to zero.

2. *Effect of large n_j.* If n_j is large for a specific group j, then η is close to one, hence $E_{\bar{D}_n}(\mathcal{R}_j(\mathcal{M}(y))) = (\bar{y}_{j.} - E_{\bar{D}_n}(\mathcal{M}(y)))$.
   Explanation. When n_j is large, the data $\bar{y}_{j.}$ is a precise estimator of the group j mean, and thus $(\bar{y}_{j.} - E_{\bar{D}_n}(\mathcal{M}(y))$ is a reliable estimator of the jth level 2 residual. Likewise, when σ_ϵ^2 is small or $(\sigma_u^2 - \gamma)$ is large η is close to one, and $E_{\bar{D}_n}(\mathcal{R}_j(\mathcal{M}(y))) = (\bar{y}_{j.} - E_{\bar{D}_n}(\mathcal{M}(y)))$.

4.4.2 Adjusting the population group j mean

Having calculated the adjusted level 2 residual, it is now straightforward to derive the adjustment of the population group j mean $\mathcal{M}(y_j)$.

Theorem 4.4.2. *The adjustment of each population mean $\mathcal{M}(y_j)$, for $j = 1, 2, \ldots, J$ in the unbalanced SOEREF model, by the collection of sample group means $\bar{D}_n = \{\bar{y}_{1.}, \bar{y}_{2.}, \ldots, \bar{y}_{J.}\}$ is*

$$E_{\bar{D}_n}(\mathcal{M}(y_j)) = \frac{(\sigma_u^2 - \gamma)}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n_j})}\bar{y}_{j.} + \frac{\frac{\sigma_\epsilon^2}{n_j}}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n_j})}E_{\bar{D}_n}(\mathcal{M}(y)) \tag{4.21}$$

Proof. By the SOE level 2 representation of Definition 3.1.1, namely

$$\mathcal{M}(y_j) = \mathcal{M}(y) + \mathcal{R}_j(\mathcal{M}(y)) \quad j = 1, 2, \ldots, J. \tag{4.22}$$

Since adjusted expectation is linear (see Property 3.2.1 of Goldstein and Wooff (2007; page 56)), we obtain the following adjustments

$$E_{\bar{D}_n}(\mathcal{M}(y_j)) = E_{\bar{D}_n}(\mathcal{M}(y)) + E_{\bar{D}_n}(\mathcal{R}_j(\mathcal{M}(y)). \tag{4.23}$$

Substituting $E_{\bar{D}_n}(\mathcal{R}_j(\mathcal{M}(y))$ from Corollary 4.4.2, we obtain

$$E_{\bar{D}_n}(\mathcal{M}(y_j)) = E_{\bar{D}_n}(\mathcal{M}(y)) + \frac{(\sigma_u^2 - \gamma)}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n_j})}(\bar{y}_{j.} - E_{\bar{D}_n}(\mathcal{M}(y))), \tag{4.24}$$

which can easily be simplified to yield (4.21). ■

To interpret (4.21), we write it as

$$E_{\bar{D}_n}(\mathcal{M}(y_j)) = \eta \bar{y}_{j.} + (1 - \eta) E_{\bar{D}_n}(\mathcal{M}(y)), \tag{4.25}$$

showing that the adjusted population group j mean is a weighted average of the sample group j mean $\bar{y}_{j.}$ and the adjusted overall population mean, $E_{\bar{D}_n}(\mathcal{M}(y))$, where the weight η is the shrinkage factor. Hence, our interpretation follows that of the previous section. For example, if n_j is small for a specific group j, then η will be closer to zero, thus less weight will be put on $\bar{y}_{j.}$, and $E_{\bar{D}_n}(\mathcal{M}(y_j))$ will be pulled towards the overall adjusted mean $E_{\bar{D}_n}(\mathcal{M}(y))$.

Also, we note the similarity between the form of our adjustment in (4.24) and the BLUP estimator (Searle et al. (1992; page 57))

$$BLUP(\mu + \alpha_i) = GLSE(\mu) + \frac{n_i \sigma_\alpha^2}{n_i \sigma_\alpha^2 + \sigma_e^2}[\bar{y}_{i.} - GLSE(\mu)], \tag{4.26}$$

where i indicates group and σ_α^2, the level 2 variance. The BLUP estimator has many desirable properties and is very versatile as it can be used to derive the Kalman filter, the method of Kriging, and Credibility theory in insurance (see Robinson (1991)). We may consider our estimator as a Bayes linear BLUP, with the important difference that within the Bayes linear approach all quantities reflect our judgements of uncertainty, while in the traditional BLUP these quantities are true but unknown population values. But similar to the comment in Searle et al. (1992), the implementation of our Bayes linear BLUP requires estimation of the variance components which is an important aspect of this thesis.

4.4.3 The adjusted variance of $\mathcal{R}_j(\mathcal{M}(y))$

In order to adjust the variances and covariances of the population group j means $\mathcal{M}(y_j)$ by $\bar{D}_n$, we first need to adjust these same quantities for the level 2 residuals $\mathcal{R}_j(\mathcal{M}(y))$.

Theorem 4.4.3. *For any level 2 residual $\mathcal{R}_j(\mathcal{M}(y))$ in group j, the adjusted variance is*

$$Var_{\bar{D}_n}(\mathcal{R}_j(\mathcal{M}(y))) = \frac{(\sigma_u^2-\gamma)\sigma_\epsilon^2}{n_j(\sigma_u^2-\gamma+\frac{\sigma_\epsilon^2}{n_j})} + \frac{(\sigma_u^2-\gamma)^2(\sigma_u^2-\gamma+\frac{\sigma_\epsilon^2}{n_j})^{-2}}{\gamma^{-1}+\sum_{j=1}^{J}(\sigma_u^2-\gamma+\frac{\sigma_\epsilon^2}{n_j})^{-1}}, \tag{4.27}$$

and for any pair of level 2 residuals $\mathcal{R}_j(\mathcal{M}(y))$ and $\mathcal{R}_{j'}(\mathcal{M}(y))$ in groups j and j' respectively, the adjusted covariance is

$$Cov_{\bar{D}_n}(\mathcal{R}_j(\mathcal{M}(y)),\mathcal{R}_{j'}(\mathcal{M}(y))) = \frac{(\sigma_u^2-\gamma)^2(\sigma_u^2-\gamma+\frac{\sigma_\epsilon^2}{n_j})^{-1}(\sigma_u^2-\gamma+\frac{\sigma_\epsilon^2}{n_{j'}})^{-1}}{\gamma^{-1}+\sum_{j=1}^{J}(\sigma_u^2-\gamma+\frac{\sigma_\epsilon^2}{n_j})^{-1}}. \tag{4.28}$$

Proof. The Bayes linear rule for the adjusted variance is

$$\begin{aligned}
Var_{\bar{D}_n}(\mathcal{C}_{R_j}) &= Var(\mathcal{C}_{R_j}) - Cov(\mathcal{C}_{R_j},\bar{D}_n)Var^{-1}(\bar{D}_n)Cov(\bar{D}_n,\mathcal{C}_{R_j}) \\
&= (\sigma_u^2-\gamma)\mathbf{I}_J - (\sigma_u^2-\gamma)\mathbf{I}_J\Big[A^{-1} - \frac{A^{-1}\mathbf{1}_J\mathbf{1}_J^TA^{-1}}{[\gamma^{-1}+\mathbf{1}_J^TA^{-1}\mathbf{1}]}\Big](\sigma_u^2-\gamma)\mathbf{I}_J \\
&= (\sigma_u^2-\gamma)\mathbf{I}_J - (\sigma_u^2-\gamma)^2\mathbf{I}_J\Big[A^{-1} - \frac{A^{-1}\mathbf{1}_J\mathbf{1}_J^TA^{-1}}{[\gamma^{-1}+\mathbf{1}_J^TA^{-1}\mathbf{1}]}\Big] \\
&= (\sigma_u^2-\gamma)(\mathbf{I}_J - (\sigma_u^2-\gamma)A^{-1}) + \frac{(\sigma_u^2-\gamma)^2}{[\gamma^{-1}+\mathbf{1}_J^TA^{-1}\mathbf{1}]}\Big[A^{-1}\mathbf{1}_J\mathbf{1}_J^TA^{-1}\Big] \\
&= (\sigma_u^2-\gamma)\mathbf{diag}\left\{1-\frac{(\sigma_u^2-\gamma)}{a_1},\cdots,1-\frac{(\sigma_u^2-\gamma)}{a_J}\right\} \\
&\quad + \frac{(\sigma_u^2-\gamma)^2}{[\gamma^{-1}+\mathbf{1}_J^TA^{-1}\mathbf{1}]}\begin{pmatrix} a_1^{-2} & a_1^{-1}a_2^{-1} & \cdots & a_1^{-1}a_J^{-1} \\ a_2^{-1}a_1^{-1} & a_2^{-2} & \cdots & a_2^{-1}a_J^{-1} \\ \vdots & \vdots & \ddots & \vdots \\ a_J^{-1}a_1^{-1} & a_J^{-1}a_2^{-1} & \cdots & a_J^{-2} \end{pmatrix}
\end{aligned}$$

$$= (\sigma_u^2 - \gamma)\mathbf{diag}\left\{\frac{\sigma_\epsilon^2}{n_1 a_1}, \cdots, \frac{\sigma_\epsilon^2}{n_J a_J}\right\}$$
$$+ \frac{(\sigma_u^2 - \gamma)^2}{[\gamma^{-1} + \mathbf{1}_J^T A^{-1}\mathbf{1}]} \begin{pmatrix} a_1^{-2} & a_1^{-1}a_2^{-1} & \cdots & a_1^{-1}a_J^{-1} \\ a_2^{-1}a_1^{-1} & a_2^{-2} & \cdots & a_2^{-1}a_J^{-1} \\ \vdots & \vdots & \ddots & \vdots \\ a_J^{-1}a_1^{-1} & a_J^{-1}a_2^{-1} & \cdots & a_J^{-2} \end{pmatrix} \tag{4.29}$$

■

After substituting $a_j = (\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n_j})$, the diagonal elements of (4.29) give the required variances, and the off-diagonal elements, the covariances. We may now calculate the above adjusted quantities for the balanced case.

Corollary 4.4.3. *Let $\mathcal{C}_{R_j}$ be the vector of level 2 residuals as defined in Theorem 4.5.1. For the balanced case, the adjusted variances and covariances are given by the diagonal and off-diagonal entries of the following variance-covariance matrix.*

$$Var_{\bar{D}_n}(\mathcal{C}_{R_j}) = \frac{(\sigma_u^2 - \gamma)\sigma_\epsilon^2}{n(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n})} I_J + \frac{(\sigma_u^2 - \gamma)^2 \gamma K_{J\times J}}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n})(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)} \tag{4.30}$$

Proof. Substituting $n_j = n$ and $a_j = a$ in (4.29) yield

$$Var_{\bar{D}_n}(\mathcal{C}_{R_j}) = (\sigma_u^2 - \gamma)\mathbf{diag}\left\{\frac{\sigma_\epsilon^2}{na}, \cdots, \frac{\sigma_\epsilon^2}{na}\right\}$$
$$+ \frac{(\sigma_u^2 - \gamma)^2}{[\gamma^{-1} + Ja^{-1}]} \begin{pmatrix} a^{-2} & a^{-2} & \cdots & a^{-2} \\ a^{-2} & a^{-2} & \cdots & a^{-2} \\ \vdots & \vdots & \ddots & \vdots \\ a^{-2} & a^{-2} & \cdots & a^{-2} \end{pmatrix}, \tag{4.31}$$

■

which is easily seen to be equal to (4.30) upon substituting $a = (\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n})$. Note that we use K rather than the traditional J as the matrix of ones as the latter is used for the number of level 2 units. Also, the adjusted variances and covariances for the balanced case (4.30) agree with those given in Searle et al. (1992; page 336).

The interpretation of the adjusted variances and covariances of the level 2 residuals is no longer straightforward. Nevertheless, we may easily re-write (4.27) as

$$\begin{aligned} Var_{\bar{D}_n}(\mathcal{R}_j(\mathcal{M}(y))) &= \frac{(\sigma_u^2-\gamma)}{(\sigma_u^2-\gamma+\frac{\sigma_\epsilon^2}{n_j})}\left[\frac{\sigma_\epsilon^2}{n_j}+\right. \\ &\left.\frac{(\sigma_u^2-\gamma)}{(\sigma_u^2-\gamma+\frac{\sigma_\epsilon^2}{n_j})}\frac{1}{\gamma^{-1}+\sum_{j=1}^{J}(\sigma_u^2-\gamma+\frac{\sigma_\epsilon^2}{n_j})^{-1}}\right] \\ &= \frac{(\sigma_u^2-\gamma)}{(\sigma_u^2-\gamma+\frac{\sigma_\epsilon^2}{n_j})}\frac{\sigma_\epsilon^2}{n_j}+\frac{(\sigma_u^2-\gamma)^2}{(\sigma_u^2-\gamma+\frac{\sigma_\epsilon^2}{n_j})^2}Var_{\bar{D}_n}(\mathcal{M}(y)) \\ &= \eta\frac{\sigma_\epsilon^2}{n_j}+\eta^2 Var_{\bar{D}_n}(\mathcal{M}(y)) \end{aligned} \tag{4.32}$$

revealing that the adjusted variance of the level 2 residual for group j is a shrunken estimator of the group j data variance $\frac{\sigma_\epsilon^2}{n_j}$ and the adjusted variance of the overall mean. Again, the shrinkage factor η plays a pivotal part in the adjusted quantities. It can similarly be shown that the adjusted covariance $Cov_{\bar{D}_n}(\mathcal{R}_j(\mathcal{M}(y)), \mathcal{R}_{j'}(\mathcal{M}(y)))$ can be interpreted in terms of the shrinkage factor and the adjusted variance of the overall population mean $\mathcal{M}(y)$.

But it must now be evident from the above that calculating and interpreting adjustments over group-level quantities *individually* may not be that straightforward, especially since these quantities are correlated. We shall therefore argue next for the adjustment of a collection of beliefs over group-level quantities in the SOEREF model.

4.5 Canonical analysis

As we saw in the previous section, analysing and interpreting beliefs individually from a collection of quantities can get increasingly complicated. Alternatively we may choose to analyse and interpret overall changes in beliefs over the whole collection of quantities using, what Goldstein and Wooff (2007) have termed, a canonical analysis. We collect the results of the canonical analysis in Theorem 4.5.1 as follows.

Theorem 4.5.1. *For the adjustment of B by D we have the following:*

- *The resolution transform matrix is*

$$\mathbb{T}_{B:D} = Var(B)^\dagger Cov(B,D) Var(D)^\dagger Cov(D,B). \tag{4.33}$$

- *The canonical resolutions are the (ordered) eigenvalues of* $\mathbb{T}_{B:D}$*, i.e.* $1 \geq \lambda_1 \geq \ldots \geq \lambda_r \geq 0$*, where* $r = \mathbf{rank}\{Var(B)\}$*.*

- *The canonical quantities are given by* $W_i = v_i^T(B - E(B))$
  where, for each i*,* $E(W_i) = 0$*,* $Var(W_i) = 1$ *and* $Cov(W_i, W_{i'}) = 0$ $(i \neq i')$ *and* $v_1, ..., v_r$ *correspond to the normed right eigenvectors of* $\mathbb{T}_{B:D}$ *scaled such that* $v_i^T Var(B) v_i = 1$*.*

- *The system resolution for* $\langle B \rangle$ *is*

$$R_D(B) = \frac{\mathbf{tr}\{\mathbb{T}_{B:D}\}}{\mathbf{rk}\{Var(B)\}}. \tag{4.34}$$

  $R_D(B)$ *gives the average resolution for the canonical directions obtained from the adjustment over the collection* $\langle B \rangle$*.*

The canonical quantities have the properties that W_1 has the largest relative reduction in variance amongst elements of $\langle B \rangle$, W_2 the next largest reduction amongst elements of $\langle B \rangle$ that are uncorrelated with W_1, and so forth. The actual reductions in variance are given as $Var_D(W_i) = 1 - \lambda_i$.

A canonical analysis is quite versatile and can be helpful in various stages of a Bayes linear analysis. At the data collection stage, if alternative sources of data are available, a canonical analysis can be used to assess the strengths and weaknesses of these competing data sets. Canonical quantities can also be useful in revealing those linear combinations for which data are expected to be informative/uninformative. A canonical analysis can also be useful in identifying problems in our belief specifications, especially in complex multilevel structures; such problems would show up as unanticipated results of our belief adjustments. Finally, for exchangeable adjustments such as in our SOE multilevel structures, a canonical analysis can simplify our sample size design calculations.

In fact, for our SOEREF model we have already calculated the adjustment over the linear combinations of the population group j means $\langle \mathcal{M}(y_j) \rangle$ since, when evaluating the adjusted expectation $E_{\bar{D}_n}(\mathcal{M}(y_j))$ for $j = 1, 2, \ldots, J$, we also implicitly evaluated the adjusted value of each element of $\langle \mathcal{M}(y_j) \rangle$, as

$$E_{\bar{D}_n}(\sum_{j=1}^{J} c_j \mathcal{M}(y_j)) = \sum_{j=1}^{J} c_j E_{\bar{D}_n}(\mathcal{M}(y_j)).$$

Thus, it remains for us to evaluate the expected effects, such as reductions in variances and diagnostics, of the Bayes linear adjustments over $\langle \mathcal{M}(y_j) \rangle$.

4.5.1 The resolution transform for the adjustment of $\mathcal{M}(y_j)$ by $\bar{D}_n$

Below we calculate the resolution transform $\mathbb{T}_{B:D}$ for the balanced case only. We shall see in Section 4.10 that a balanced design is optimal for the adjustment of the overall mean $\mathcal{M}(y)$. So from a design perspective it is consistent that we also consider the adjustment of $\mathcal{M}(y_j)$ for the balanced case. Further, for the balanced case the eigenstructure of $\mathbb{T}_{B:D}$ can be easily obtained from Lemma 11.62 of Goldstein and Wooff (2007; page 449) or more directly from (Searle et al. (1992; page 443)).

Theorem 4.5.2. *The resolution transform matrix for the adjustment of the collection of population group j means $\mathcal{M}(y_j)$, for $j = 1, 2, \ldots, J$ in the balanced SOEREF model, by the collection of sample group means $\bar{D}_n = \{\bar{y}_{1.}, \bar{y}_{2.}, \ldots, \bar{y}_{J.}\}$ is*

$$\mathbb{T}_n = \frac{(\sigma_u^2 - \gamma)}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n})} \left[I_J + \frac{\frac{\gamma \sigma_\epsilon^2}{n}}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)(\sigma_u^2 - \gamma)} K_J \right] \tag{4.35}$$

Proof. Let $\mathcal{C}_\mathcal{M}$ be the collection of the population means $\mathcal{M}(y_j)$, for $j = 1, 2, \ldots, J$. For simplicity we shall write $\mathbb{T}_n$ for $\mathbb{T}_{\mathcal{C}_\mathcal{M}:\bar{D}_n}$. Using (4.33) of Theorem 4.5.1 we obtain

$$\mathbb{T}_n = Var(\mathcal{C}_\mathcal{M})^{-1} Cov(\mathcal{C}_\mathcal{M}, \bar{D}_n) Var(\bar{D}_n)^{-1} Cov(\bar{D}_n, \mathcal{C}_\mathcal{M}) \tag{4.36}$$

To calculate the second-order quantities in (4.36), we make use of the SOE level 2 representation $\mathcal{M}(y_j) = \mathcal{M}(y) + \mathcal{R}_j(\mathcal{M}(y))$.

(i) Calculation of $Var(\mathcal{C}_\mathcal{M})$

$Var(\mathcal{M}(y_j)) = Var(\mathcal{M}(y)) + Var(\mathcal{R}_j(\mathcal{M}(y))) = \gamma + (\sigma_u^2 - \gamma) = \sigma_u^2$, while $Cov(\mathcal{M}(y_j), \mathcal{M}(y_{j'})) = Var(\mathcal{M}(y)) = \gamma$, for $j \neq j'$. Therefore,

$$Var(\mathcal{C}_{\mathcal{M}}) = \begin{pmatrix} \sigma_u^2 & \gamma & \cdots & \gamma \\ \gamma & \sigma_u^2 & \cdots & \gamma \\ \vdots & \vdots & \ddots & \vdots \\ \gamma & \gamma & \cdots & \sigma_u^2 \end{pmatrix}$$
$$= (\sigma_u^2 - \gamma)I_J + \gamma K_J \tag{4.37}$$

(ii) $Cov(\mathcal{C}_{\mathcal{M}}, \bar{D}_n) = Var(\mathcal{C}_{\mathcal{M}})$

Since the jth element of $\bar{D}_n$ is $\bar{y}_{j\cdot} = \mathcal{M}(y)+\mathcal{R}_j(\mathcal{M}(y))+\bar{\mathcal{R}}_{n_j}(y_{j\cdot}) = \mathcal{M}(y_j)+\bar{\mathcal{R}}_{n_j}(y_{j\cdot})$, and the level 1 residual is uncorrelated with both the level 2 residual and the overall population mean, therefore $Cov(\mathcal{C}_{\mathcal{M}}, \bar{D}_n) = Var(\mathcal{C}_{\mathcal{M}})$.

(iii) For the balanced case $Var(\bar{D}_n) = (\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n})I_J + \gamma K_J$, the inverse of which is a standard result.

$$Var(\bar{D}_n)^{-1} = \frac{1}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n})}\left[I_J - \frac{\gamma}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)}K_J\right]$$

Hence using (i) to (iii),

$$\begin{aligned}\mathbb{T}_n &= Var(\bar{D}_n)^{-1}Cov(\bar{D}_n, \mathcal{C}_{\mathcal{M}}) \\ &= \frac{1}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n})}\left[I_J - \frac{\gamma}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)}K_J\right]\left[(\sigma_u^2 - \gamma)I_J + \gamma K_J\right],\end{aligned}$$

which upon simplication reduces to (4.35). ■

4.5.2 The canonical resolutions

The canonical resolutions are the ordered eigenvalues of $\mathbb{T}_n$. Given the special form of $\mathbb{T}_n$ in (4.35), we use the results in Searle et al. (1992; page 443), namely that the eigenvalues of $aI_n + bK_n$ are a, with multiplicity $n - 1$ and $a + nb$. Hence, the

eigenvalues λ_j of $\mathbb{T}_n$ are as follows.

$$\begin{aligned}
\lambda_1 &= \frac{(\sigma_u^2-\gamma)}{(\sigma_u^2-\gamma+\frac{\sigma_\epsilon^2}{n})}\left[1+\frac{\frac{J\gamma\sigma_\epsilon^2}{n}}{(\sigma_u^2-\gamma+\frac{\sigma_\epsilon^2}{n}+J\gamma)(\sigma_u^2-\gamma)}\right], \\
&= \frac{n(\sigma_u^2-\gamma)}{n(\sigma_u^2-\gamma)+\sigma_\epsilon^2}\left[1+\frac{nJ\gamma\sigma_\epsilon^2}{(n(\sigma_u^2-\gamma)+\sigma_\epsilon^2+nJ\gamma)(n(\sigma_u^2-\gamma))}\right], \\
&= \frac{n(\sigma_u^2-\gamma)}{n(\sigma_u^2-\gamma)+\sigma_\epsilon^2}+\frac{nJ\gamma\sigma_\epsilon^2}{(n(\sigma_u^2-\gamma)+\sigma_\epsilon^2)(n(\sigma_u^2-\gamma)+\sigma_\epsilon^2+nJ\gamma)}, \\
&= \frac{n(\sigma_u^2-\gamma)(n(\sigma_u^2-\gamma)+\sigma_\epsilon^2+nJ\gamma)+nJ\gamma\sigma_\epsilon^2}{(n(\sigma_u^2-\gamma)+\sigma_\epsilon^2)(n(\sigma_u^2-\gamma)+\sigma_\epsilon^2+nJ\gamma)}, \\
&= \frac{n(\sigma_u^2-\gamma)(n(\sigma_u^2-\gamma)+\sigma_\epsilon^2)+nJ\gamma(n(\sigma_u^2-\gamma)+\sigma_\epsilon^2)}{(n(\sigma_u^2-\gamma)+\sigma_\epsilon^2)(n(\sigma_u^2-\gamma)+\sigma_\epsilon^2+nJ\gamma)}, \\
&= \frac{n(\sigma_u^2-\gamma)+nJ\gamma}{n(\sigma_u^2-\gamma)+\sigma_\epsilon^2+nJ\gamma}
\end{aligned}$$

$$\lambda_2 = \frac{(\sigma_u^2-\gamma)}{(\sigma_u^2-\gamma+\frac{\sigma_\epsilon^2}{n})}, \tag{4.38}$$

where λ_2 has multiplicity $(J-1)$.

It is obvious that $\lambda_1 \geq \lambda_2$ (λ_1 corresponds to $a+nb$ and λ_2 to a above), with equality if $\gamma = 0$, that is no uncertainty about the population overall mean $\mathcal{M}(y)$. We need to ensure that the largest possible reduction in variance is one, that is $\lambda_j \leq 1$, for each j. From the definition of the shrinkage factor, $0 \leq \lambda_2 \leq 1$. For λ_1 we have

$$\begin{aligned}
\lambda_1 \leq 1 &\implies \frac{(\sigma_u^2-\gamma)+J\gamma}{(\sigma_u^2-\gamma)+\frac{\sigma_\epsilon^2}{n}+J\gamma} \leq 1 \\
&\implies \frac{\sigma_\epsilon^2}{n} \geq 0,
\end{aligned} \tag{4.39}$$

which is a coherence condition, namely that the belief specification of the level 1 variance is non-negative.

We now have the following results. The largest eigenvalue λ_1 depends on both the levels 1 and 2 sample sizes n and J respectively. Further, as either n or $J \to \infty$, $\lambda_1 \to 1$. This implies that the more we increase the level 1 and/or level 2 sample sizes, the more the uncertainty we expect to resolve about the corresponding component in $\langle \mathcal{M}(y_j) \rangle$ by observing $\bar{D}_n$.

The smallest eigenvalue λ_2 however, depends only on the level 1 sample size, n, and as $n \to \infty$, $\lambda_2 \to 1$. Hence, the more we increase the level 1 sample size, the more the uncertainty we expect to resolve in the direction of the corresponding component in $\langle \mathcal{M}(y_j) \rangle$ by observing $\bar{D}_n$.

Again the above results underline the importance of the shrinkage factor, since the largest eigenvalue, λ_1, is a function of η while the smallest eigenvalue, λ_2, is equal to η.

4.5.3 The canonical quantities

While the canonical resolutions (eigenvalues of $\mathbb{T}_n$) gave the *magnitude* of the resolved uncertainty, the canonical quantities (eigenvectors of $\mathbb{T}_n$), will provide the *type* of resolution associated with each component in $\langle \mathcal{M}(y_j) \rangle$ upon observing $\bar{D}_n$. To calculate these canonical quantities, we exploit the special form of $\mathbb{T}_n$ for which the eigenvectors are proportional to the columns of the Helmert matrix of order J, denoted by $\mathbf{H}_J$ (see Lemma 11.62 of Goldstein and Wooff (2007; page 448)). For example, the first two canonical quantities are

$$W_1 = \alpha_1 \begin{pmatrix} 1 & 1 & \dots & 1 \end{pmatrix} \begin{pmatrix} \mathcal{M}(y_1) - \mu \\ \mathcal{M}(y_2) - \mu \\ \dots \\ \mathcal{M}(y_J) - \mu \end{pmatrix},$$

$$W_2 = \alpha_2 \begin{pmatrix} 1 & -1 & 0 & \dots & 0 \end{pmatrix} \begin{pmatrix} \mathcal{M}(y_1) - \mu \\ \mathcal{M}(y_2) - \mu \\ \dots \\ \mathcal{M}(y_J) - \mu \end{pmatrix},$$

where the first eigenvector W_1 is the average of the group means, while W_2, as well as all the remaining eigenvectors, are all of the contrasts, i.e. all of the vectors summing to zero. We choose α_1 and α_2 so that W_1 and W_2 have unit variances and zero means (after re-scaling $\mathcal{M}(y_j)$). Hence, the canonical quantities become

$$W_1 = \frac{1}{\sqrt{[J\sigma_u^2 + J(J-1)\gamma]}}(\mathcal{M}(y_1) - \mu + \ldots + \mathcal{M}(y_J) - \mu), \tag{4.40}$$

$$W_2 = \frac{1}{\sqrt{2[\sigma_u^2 + \gamma]}}(\mathcal{M}(y_1) - \mathcal{M}(y_2)), \tag{4.41}$$

where the denominator in (4.40) is the standard deviation of $\sum_{j=1}^{J} \mathcal{M}(y_j)$.

We learn most about W_1 since it has the largest resolution, λ_1. At the other end, we learn least in the direction of W_2 as it has the least resolution, λ_2. In addition, since λ_2 has multiplicity $(J-1)$, we learn equally about all possible contrasts, a consequence of the balanced design we chose. If however, some contrasts are more important than others, then we might choose an unbalanced design. An examination of the canonical directions reveals that they do not depend on the level 1 sample size n, and can thus be used to guide sample size choice in our multilevel design problem.

4.6 Example: Bayes linear analysis of the STAT1010 data

We shall now apply Bayes linear methods to analyze the Stat1010 data using [B/D], the freely available Bayes linear programming language (Goldstein & Wooff, 1995). [B/D] provides the required facilities for the specification and analysis of beliefs, including diagnostic checks. In addition, [B/D] also produces Bayes linear diagnostic influence diagrams. These may be particularly useful for analyzing the effects on

our beliefs of multiple sources of information, as well as of potential covariates (both at level 1 and level 2) using a partial Bayes linear analysis.

We shall update our beliefs about the SOEREF model using the prior specifications in Chapter 3. To check our [B/D] programme, we compare its output with those from an R programme (Appendix B) we have written based on our derivations for the adjusted quantities in Sections 4.3 and 4.4.

4.7 Discrepancy

Before adjusting our beliefs, we must check that our prior specifications do not conflict with the observed data. We do so by applying the discrepancy measures of Section 2.15. The corresponding R codes are in Appendix B.

To identify potential problems with individual observations, we compute the standardized observation as follows.

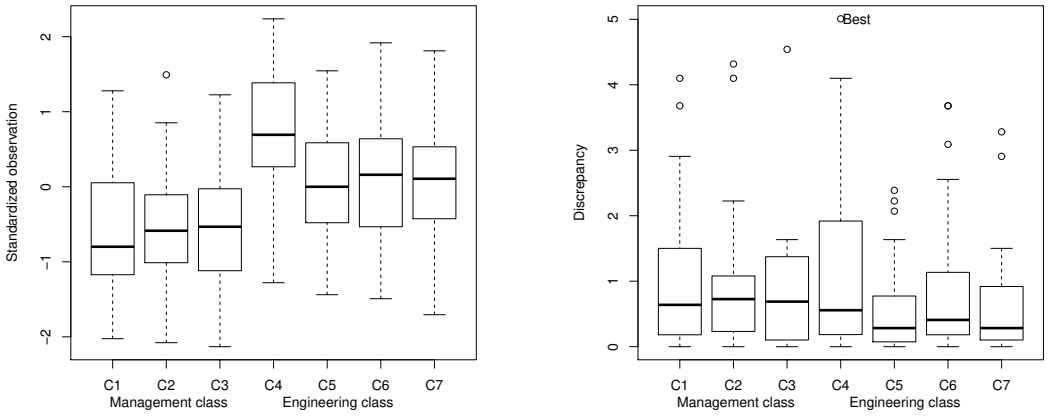
$$S(y_{ji}) = \frac{(y_{ji} - E(y_{ji}))}{\sqrt{Var(y_{ji})}} \tag{4.42}$$

The discrepancy between observation and prior assessments is given by

$$Dis(y_{ji}) = \frac{[y_{ji} - E(y_{ji})]^2}{Var(y_{ji})} \tag{4.43}$$

The interpretation of $S(y_{ji})$ and $Dis(y_{ji})$ is not straightforward, depending on context, including the experts making the prior judgements, and also on sample size (see Goldstein & Wooff (2007), page 96). As far as our judgements are concerned, we have been quite thorough in making our prior specifications. Also the data can be trusted to be correct given they are official examinations data. As for sample sizes, we have a total of 269 students (level 1) grouped in seven classes (level 2) with a minimum of 23 students in one class and a maximum of 47 students in another. Thus we have a sufficiently large sample of level 1 observations but a rather modest sample of level 2 observations.

The standardized and discrepancy measures of the individual examinations marks (y_{ji}) are shown in the box-plots in Figure 4.1 (Appendix B, R codes lines 16 to 24). We have grouped the students' marks according to their respective class. From



(a) Standardized observations (b) Discrepancy

Figure 4.1: *The distributions of standardized observations (a) and discrepancy (b) for each class in the STAT1010 data. The data marked "Best" in (b) is a student scoring 97% in the exams.*

Figure 4.1(a), there do not appear to be any outlying observation. In fact, almost all the observations lie in the interval (-2,+2), which looks quite short. Under a Gaussian distribution we would expect 5% (13 out of 269) of the students to have standardized marks outside the (-2,+2) interval but as we are sampling only seven groups, there is no cause for concern.

In Figure 4.1(b) all the median (and mean) discrepancies are less than one except that the means for C1 (1.01) and C4 (1.16) are slightly above one due to the large examinations marks of a few students. Since each discrepancy has prior expectation one, this again points to an over-estimated prior uncertainty in $Var(y_{ji})$.

A useful guide to detect outliers from any uni-modal continuous distribution Z with standard deviation σ_Z , is the **three-sigma rule** (Pukelsheim 1994) given by $P(|Z - E(Z)| \leq 3\sigma_Z) = 0.95$. Thus if many observations are three or more standard deviations away from the mean, they may be considered a possible diagnostic signal. In Figure 4.1(b), 12 out of 269 students have discrepancies above 3, representing 4.5% which is close to the expected 5%. The largest discrepancy (marked "Best" in plot (b)) is for a student in an Engineering class who scored 97% in the examinations and, although this observation is more than three standard deviations distant from

our expectation, it is clearly not an outlier.

Figure 4.1(a) also supports our earlier arguments that on average students from management classes (C1-C3) perform less well than those of engineering classes (C4-C7), thus a SOEREG model with type of class as a level 2 predictor would be a more suitable model than the present SOEREF model.

Discrepancy ratio

To compare discrepancies across classes, we calculate the discrepancy ratio $Dr(\bar{y}_{.j})$ based on the Mahalanobis distance (Chapter 2) as follows

$$Dr(\bar{y}_{.j}) = \frac{(\bar{y}_{.j} - E(\bar{y}_{.j}))^T Var(\bar{D}_n)^{-1}(\bar{y}_{.j} - E(\bar{y}_{.j}))}{rank(Var(\bar{D}_n))} \tag{4.44}$$

(Appendix B, R code lines 21 to 31). For the STAT1010 data we obtain $Dr(\bar{y}_{.j}) = 1.1347$ which is less than the upper bound for a discrepancy ratio based on Chebyshev's inequality, i.e. $1 + \frac{6}{\sqrt{7}} = 3.2678$. Thus the maximal data discrepancy across classes is in line with our prior expectations.

4.8 Adjusting beliefs about the overall and group means

We shall now adjust beliefs about $\mathcal{M}(y)$ and $\mathcal{M}(y_j)$ using the STAT1010 data. The adjustments are calculated using [B/D] (Appendix B, lines 32 to 115). As a check, we compare our [B/D] output with those of BALM (BAyes Linear Modeling), our purposely written R programme (Appendix B, lines 116 to 136). The adjustments using [B/D] shown in Table 4.1 are exactly the same as the output from BALM (Appendix B, lines 137 to 153).

Table 4.1 summarizes the key results from our adjustments. The first two columns show the mean components and their prior expectation. Column (3) reveals that the adjusted expectations for management classes are below the prior expectation while the adjustments for engineering classes are above. The direction and magnitude of the changes from prior to adjustment are shown by the standardized adjustment discrepancies between brackets in Column (3). The standardized adjust-

		Expectation			Variance	
Element	Prior	Adjusted	Observed	Prior	Adjusted	Resolution(%)
(1)	(2)	(3)	(4)	(5)	(6)	(7)
$\mathcal{M}(y)$	55.0	53.86(-0.16)	↓	56.3	8.03	85.7
$\mathcal{M}(y_1)$	55.0	44.64(-0.99)	43.73(41)	115.3	5.33	95.4
$\mathcal{M}(y_2)$	55.0	46.02(-0.87)	44.65(23)	115.3	8.95	92.2
$\mathcal{M}(y_3)$	55.0	47.61(-0.71)	46.71(28)	115.3	7.53	93.5
$\mathcal{M}(y_4)$	55.0	68.51(1.28)	69.77(47)	115.3	4.69	95.9
$\mathcal{M}(y_5)$	55.0	55.24(0.02)	55.37(43)	115.3	5.10	95.6
$\mathcal{M}(y_6)$	55.0	57.75(0.26)	58.09(46)	115.3	4.79	95.8
$\mathcal{M}(y_7)$	55.0	56.08(0.10)	56.29(41)	115.3	5.33	95.4

Table 4.1: *Adjusting overall mean $\mathcal{M}(y)$ and group j means $\mathcal{M}(y_j)$ in the SOEREF model using the STAT1010 data. Column (3) shows the adjusted expectations and the standardized adjustment discrepancy in brackets. Column (4) shows the effect of the observed data on the adjustments. Adjustment of $\mathcal{M}(y)$ (and each $\mathcal{M}(y_j)$ also) depend on all group means and sample sizes ($\bar{y}_{j.}$,n_j) in column (4) as indicated by ↓. For the adjustment of each $\mathcal{M}(y_j)$ the most influential data and sample size ($\bar{y}_{j.}$,n_j) is shown.*

ment discrepancy for $\mathcal{M}(y)$ is given in (4.47) below. The corresponding expression for $\mathcal{M}(y_j)$ can be obtained by substituting $\mathcal{M}(y_j)$ in (4.47).

$$S(\mathcal{M}(y)) = \frac{(E_{\bar{D}_n}(\mathcal{M}(y)) - E(\mathcal{M}(y)))}{\sqrt{RVar_{\bar{D}_n}(\mathcal{M}(y))}} \tag{4.45}$$

The discrepancies are in the range (-0.99,1.28) while in practice we would expect such standardized discrepancies to be in the range (-2,+2). It is also clear that the pattern in the negative and positive signs in the standardized discrepancies is associated with type of class. Column(4) shows the observed group mean $\bar{y}_{j.}$ with the corresponding group sample size n_j between brackets. Each population mean in column (1) depends on all the observed group means. $\mathcal{M}(y)$ depend in a complex way on all the observed group means, thus the arrow pointing downwards. For each $\mathcal{M}(y_j)$, the most influential data for the adjustment, namely $\bar{y}_{j.}$ is shown. A comparison of column(3) and (4) shows clearly the proximity of the adjusted and observed means. Columns (5) and (6) show that all prior variances have been

reduced considerably leading to the substantial resolutions in column (7). The sizes of these resolutions clearly depend on the sample sizes (column(4)). The latter results, as well as the proximity of the adjusted and observed means, are most probably due to the combined effects of the large group sample sizes as well as to our prior uncertainties being too big. These are further discussed next.

4.9 Sensitivity of the adjustments to the prior and sample sizes

Figure 4.2 shows the prior, data and adjusted means with their respective three standard deviations limits. It is clear that the prior judgements have only a small effect on the adjustments. As mentioned in the previous section, this is most likely due to the large sample sizes and to our prior uncertainties which are too big. As these adjustments form a key part of our analysis, we shall therefore explore the interplay between changes in prior uncertainties and sample sizes in more detail.

For illustrative purposes, we consider that the prior variances that we have specified were upper bounds for our uncertainty. We introduce a scalar multiplier, c say, and explore what happens to the adjustment when c is reduced from one towards zero. We only scale the prior variance γ of $\mathcal{M}(y)$ by c, keeping the level 1 and level 2 variances unchanged so that we can explore the effects of changes in our prior confidence in the overall mean relative to the remaining uncertainties. In addition, we also consider the effect of sample size on our adjustments by reducing all samples by a common proportion, d say. We consider the joint effects of (c,d) on the adjustments. The results are given in Table 4.2 and Figures 4.3 and 4.4.

Table 4.2 shows the effects of reducing the sample size and uncertainty in γ on the adjusted overall and group means. We show only the more significant changes, those that result from reductions of 1.0, 0.5 and 0.1. For any adjusted mean, the changes are more significant across rows (change in sample size) than down columns (change in uncertainty). This is mostly due to γ representing only about 16% of our total prior uncertainty.

As the sample size and/or the prior uncertainty γ is reduced, all the adjusted

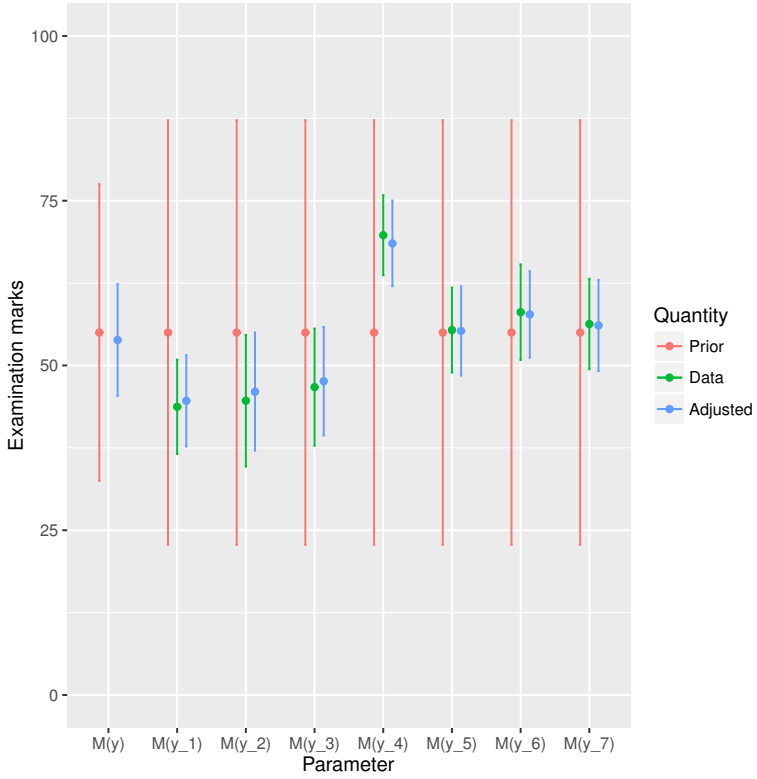


Figure 4.2: *The effect of STAT1010 data on the adjustments of overall and group means. For $\mathcal{M}(y)$ no data is shown as the adjustment depends in a complex way on all the group means, the green bars. For each $\mathcal{M}(y_j)$, the most influential data, namely $\bar{y}_{j\cdot}$ is shown. All three types of bars are $\pm$ three standard deviations.*

means are pulled progressively towards the prior expectation of 55. Even for $\mathcal{M}(y_4)$,the group with largest sample size, the adjusted mean changes quite appreciably from 68.5 to 63.1 in the direction of the prior of 55.

We illustrate the effect on the adjustments of the changes in sample size and uncertainty in Figure 4.3. The prior, data and adjusted means along with three standard deviations are as shown in Figure 4.2. We have included the adjustment of the means based on the sample and uncertainty reduced by 0.1 (in purple). All the newly adjusted means (purple dots) are closer to the prior mean of 55. For the overall mean, the adjusted standard deviations for the reduced sample and uncertainty are smaller than for the full sample and uncertainty. This is because the adjusted variance of $\mathcal{M}(y)$ depends quite strongly on γ (see Section 4.3.1) and we have reduced γ to one tenth its value. In contrast the adjusted standard deviations (purple bars) of the group means are larger than for the full data (the blue bars). As

Group Means and Sample reduction												
Resolution	$\mathcal{M}(y)$			$\mathcal{M}(y_1)$			$\mathcal{M}(y_2)$			$\mathcal{M}(y_3)$		
reduction	0.1	0.5	1.0	0.1	0.5	1.0	0.1	0.5	1.0	0.1	0.5	1.0
0.1	54.9	54.6	54.5	49.3	45.5	47.7	51.5	47.1	46.1	51.4	48.5	47.7
0.5	54.7	54.1	54.0	49.2	45.5	47.6	51.3	47.0	46.0	51.3	48.4	47.6
1.0	54.6	54.0	53.9	49.2	45.4	47.6	51.3	47.0	46.0	51.2	48.3	47.6
Group Means and Sample reduction												
Resolution	$\mathcal{M}(y_4)$			$\mathcal{M}(y_5)$			$\mathcal{M}(y_6)$			$\mathcal{M}(y_7)$		
reduction	0.1	0.5	1.0	0.1	0.5	1.0	0.1	0.5	1.0	0.1	0.5	1.0
0.1	63.1	67.6	68.6	55.1	55.2	55.3	56.7	57.6	57.8	55.6	56.0	56.1
0.5	63.0	67.5	68.5	55.0	55.2	55.3	56.6	57.5	57.8	55.5	55.9	56.1
1.0	63.0	67.5	68.5	55.0	55.2	55.2	56.5	57.5	57.7	55.4	55.9	56.1

Table 4.2: *Changes in the adjusted means $\mathcal{M}(y)$ and $\mathcal{M}(y_j)$ as the uncertainty and sample sizes are reduced by a factor of 0.1. As the sample and/or uncertainty are reduced, the adjusted means are pulled closer to the prior mean of 55 and away from the data means $\bar{y}_{j..}$ and $\bar{y}_{j.}$.*

can be seen from Section 4.4.3, two important components of the adjusted variance of $\mathcal{M}(y_j)$ are $(\sigma_u^2 - \gamma)$ and $\frac{\sigma_\epsilon^2}{n_j}$. Hence reducing n_j leads to the increase in the adjusted standard deviations of the group means, more so since we have also kept the level 2 variance $(\sigma_u^2 - \gamma)$ fixed and large relative to the other prior variances. To summarize, the reduced sample size and uncertainty in γ pull the adjusted means towards the prior mean and increase the adjusted variance of the group means due to the reduced group sample sizes.

4.9.1 Design curves and the choice of sample size for adjusting $\mathcal{M}(y)$

In Figure 4.4 we construct a spaghetti plot showing how resolutions for the adjustment of $\mathcal{M}(y)$ change when its uncertainty γ and the sample size are reduced. We shall refer to each curve in the spaghetti plot as a *design curve*. For example, the lowest design curve corresponds to $d = 0.1$, that is the smallest sample $n_j = (4, 2, 3, 5, 4, 5, 4)$ while the uppermost design curve corresponds to the full sample size n_j. All the design curves slope downwards since for any quantity Z the resolution $R_D(Z)$ decreases with decreasing prior uncertainty $Var(Z)$ as

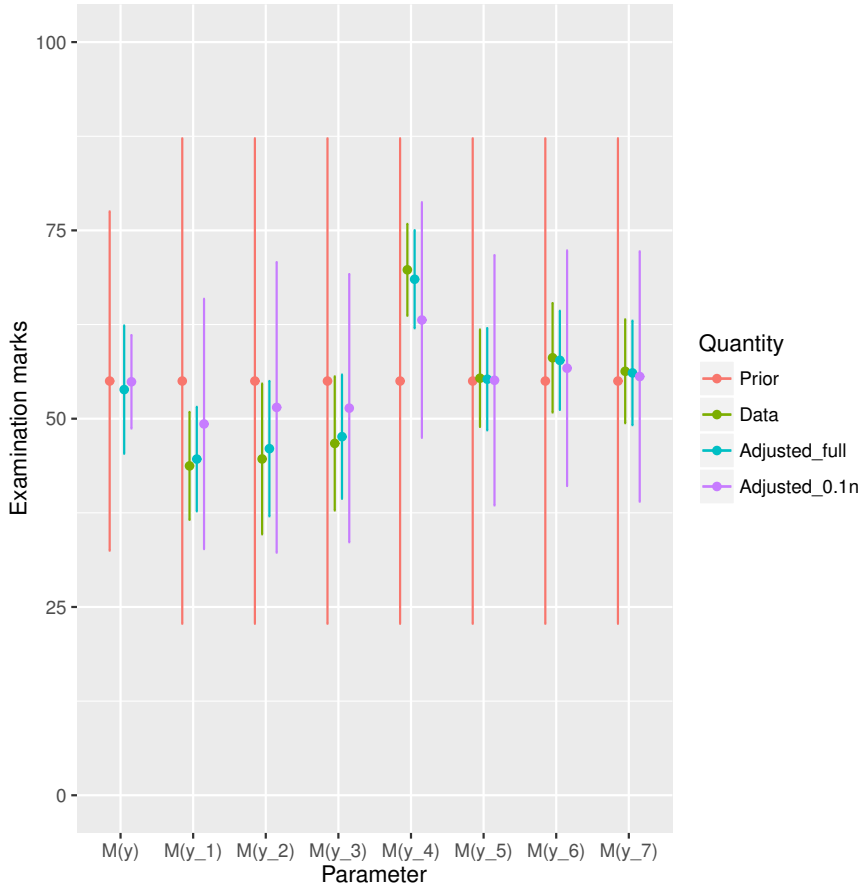


Figure 4.3: *The effects of reducing the sample size and uncertainty on the adjustments of overall and group means (purple bars). The red, green and blue bars are as defined in Figure 4.2.*

$R_D(Z) = 1 - \frac{var_D(Z)}{Var(Z)}$.

A surprising feature of Figure 4.4 is that all the uppermost design curves (d between 1.0 and 0.6) are very close together. This implies that only marginal losses in resolution will be incurred if we sample 60% rather than the full sample size for any given level of uncertainty. Hence, in planning a study similar to the design of the STAT1010 data, it might be cost-effective to sample only half the full sample size, as the corresponding design curve for $d = 0.5$ is quite close to the other uppermost design curves. This choice will result in a resolution of about 73% if the uncertainty in γ is reduced by 0.5. For the same reduction in γ, we could still achieve a decent resolution of 62% even if we choose the smallest design d=0.1, i.e. $n_j = (4, 2, 3, 5, 4, 5, 4)$. This design is quite close to a balanced one with $n_j = 4$ for all j and we would prefer it on the ground of simplicity.

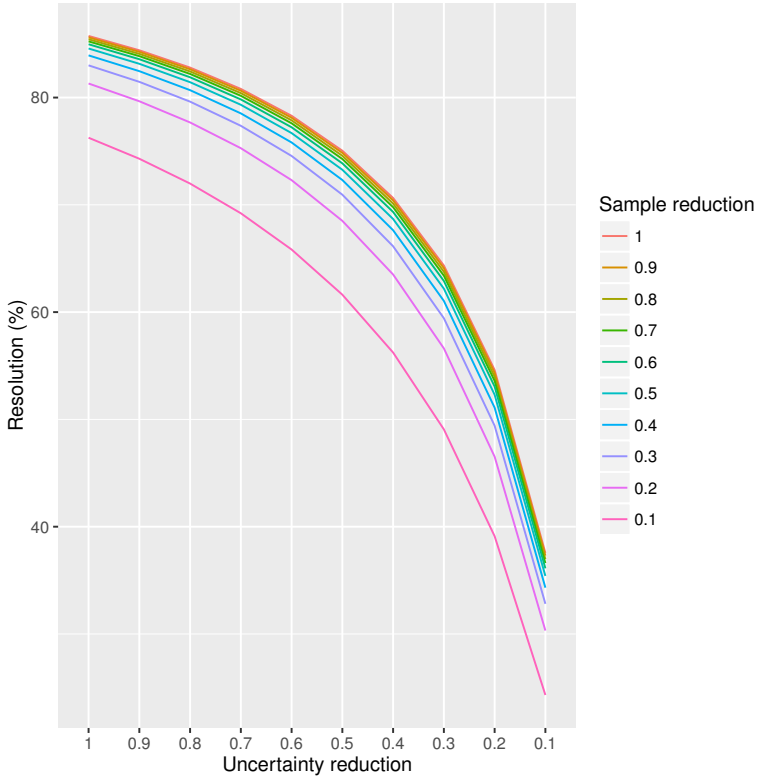


Figure 4.4: *Spaghetti plot showing changes in resolutions resulting from reductions in prior uncertainty and sample size when adjusting the overall mean* $\mathcal{M}(y)$. *The initial prior uncertainty* $\gamma = 56.3$ *and group sample sizes* $n_j = (41, 23, 28, 47, 43, 46, 41)$ *are decreased successively by 0.1. Small uncertainties and small sample sizes are associated with low resolution.*

4.9.2 Design curves and the choice of sample size for adjusting $\mathcal{M}(y_j)$

Figure 4.5 shows the design curves for adjusting each group mean $\mathcal{M}(y_j)$. In each class the design curve slopes downwards as in Figure 4.4 showing that the resolution decreases as uncertainty is reduced. Across the classes C1 to C7, classes with small sample sizes n_j have lower design curves than those with larger sample sizes since the adjustment of a specific group mean $\mathcal{M}(y_j)$ is largely influenced by its own sample size n_j. For example, the design curves for C2, the group with the smallest sample size ($n_2 = 23$), are lower than the corresponding design curves for C4, the group with the largest sample size ($n_4 = 47$). Also, for each class the upper three or four design curves are quite close, although not as close as in Figure 4.4.

We may exploit these differences in resolution patterns at different uncertainty

levels between classes to choose how much to sample in each of the seven classes. For example, suppose we now feel confident that our prior uncertainty is half the value we specified initially and we would like to achieve a resolution of at least 80% for adjusting each $\mathcal{M}(y_j)$. Then using Figure 4.4, we would sample $0.3 \times n_j$ for classes C1, C4, C5, C6, C7 and $0.4 \times n_j$ for classes C2 and C3. Such a sampling scheme is intuitive: we sample proportionately more in small classes (C2 and C3) and less in the remaining larger classes. Note that despite the small classes having roughly half the sample sizes of the larger ones, we sample only ten percent more (i.e. $0.4n_j$ vs. $0.3n_j$) in these classes to achieve the uniform resolution of 80% in all classes; the reason being that in estimating the group mean of any given class we borrow strength from the remaining classes.

The design question being addressed here and in Sub-section 4.9.1 is quite specific and should be contrasted with two-stage cluster sampling design where normally an optimal design is sought with costs constraints. Under the latter condition a more efficient sample design is a balanced one where we sample as few level 1 units as possible and use our resources to maximize the number of level 2 units (see our proof of Theorem 4.12.1 and the recommendation that follows.) The important difference with Theorem 4.12.1 is that here the number of level 2 units $J = 7$ is fixed and we are seeking how many level 1 units to sample (independently) in each group based on changes in our prior uncertainty γ while in Theorem 4.12.1 we seek the optimum design for a given design cost by allowing both J and n_j to vary freely.

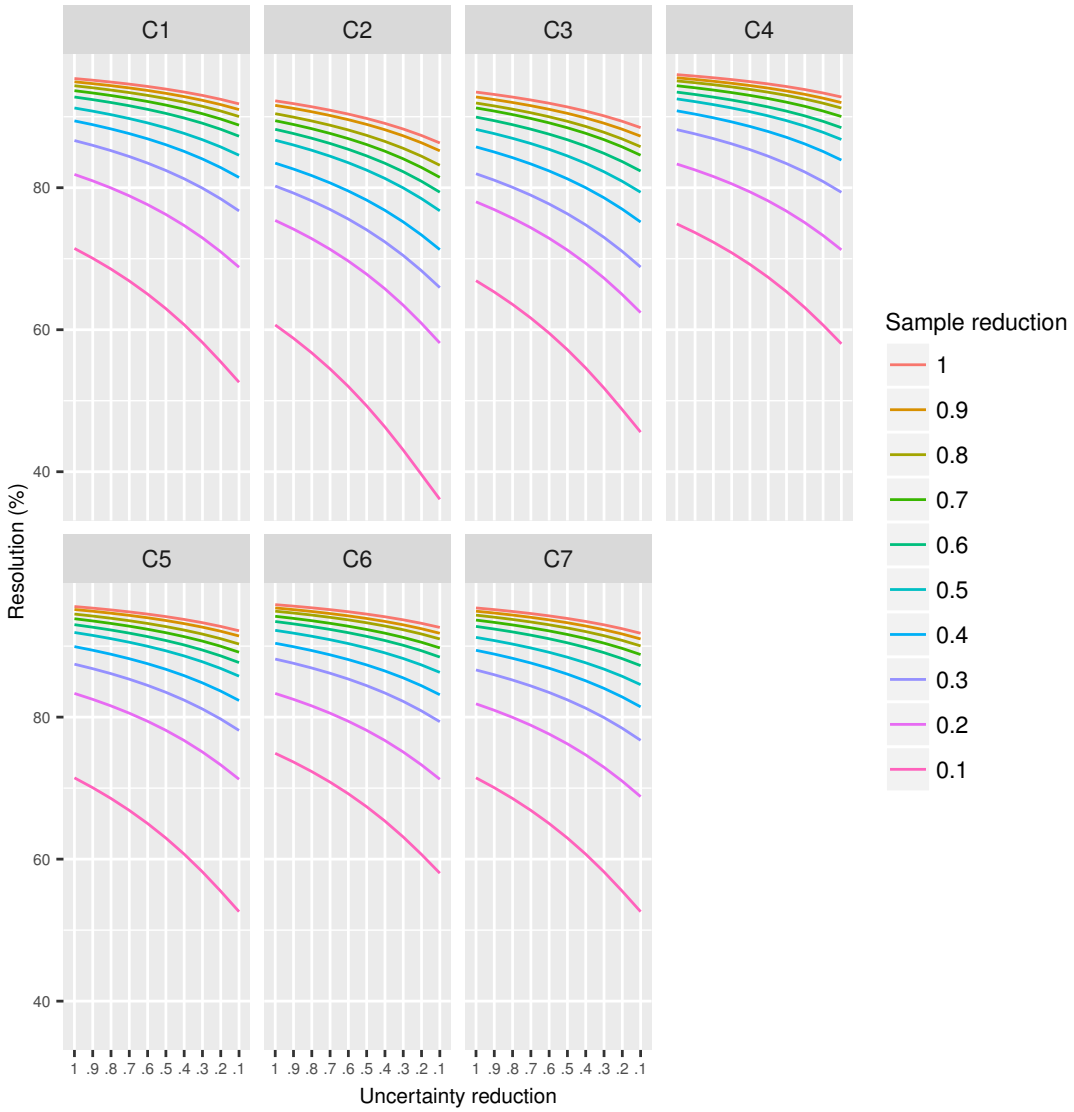


Figure 4.5: *Spaghetti plots showing changes in resolutions resulting from reductions in prior uncertainty and sample size when adjusting each group mean* $\mathcal{M}(y_j)$. *The initial prior uncertainty* $Var(y_{ji}) = 352.3$ *and group sample sizes* $n_j = (41, 23, 28, 47, 43, 46, 41)$ *are decreased by 0.1 until* $Var(y_{ji}) = 35.23$ *and* $n_j = (4, 2, 3, 5, 4, 5, 4)$. *The pattern of changes in resolution varies by class.*

4.10 The resolution $R_{\bar{D}_n}(\mathcal{M}(y))$ and the design of cluster sampling

The resolution $R_{\bar{D}_n}(\mathcal{M}(y))$ has some important implications in the design of two-stage cluster sampling as discussed below. In two-stage cluster sampling clusters (J level 2 units) are sampled at the first stage, and within each sampled cluster, indi-

viduals (n_j level 1 units) are sampled at the second stage, thus generating two-level hierarchies suitable to be analysed by our SOEREF models.

Implication of $R_{\bar{D}_n}(\mathcal{M}(y))$ for determination of sample size at level 2 and level 1

In general, ignoring cost of sampling, for a fixed overall sample size (level 1 and level 2 together), it is preferable to sample more level 2 units than level 1 units. This is because increasing level 2 units is expected to reduce the uncertainty about $\mathcal{M}(y)$ more than increasing level 1 units (and keeping level 2 units fixed). To see this, first we note that $R_{\bar{D}_n}(\mathcal{M}(y))$ is maximized when $\sum_{j=1}^{J}(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n_j})^{-1} \to \infty$ which happens when $J \to \infty$, since $\sum_{j=1}^{J}(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n_j})^{-1}$ is a sum of positive quantities. Therefore, the larger the number of level 2 units J, the higher the resolution. Whereas for fixed J, as $n_j \to \infty, \sum_{j=1}^{J}(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n_j})^{-1} \to J(\sigma_u^2 - \gamma)^{-1}$, hence $R_{\bar{D}_n}(\mathcal{M}(y)) < 1$ and cannot attain its maximum.

As a simple example, consider the following two extreme sample designs for sampling a total of ten observations. If we sample ten groups each with a single observation ($J = 10$, $n_j = 1$, for all j), then $\sum_{j=1}^{J}(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n_j})^{-1} = 10(\sigma_u^2 - \gamma + \sigma_\epsilon^2)^{-1}$ which is greater than $(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{10})^{-1}$, as is obtained from sampling only one group with ten observations ($J = 1$, $n_1 = 10$). Hence, in this case ten groups of a single observation each are expected to produce a higher resolution than a single group with ten observations.

Implication of $R_{\bar{D}_n}(\mathcal{M}(y))$ for balanced sample

Ignoring cost of sampling, it is preferable to aim for a balanced design rather than an unbalanced one for the same overall sample size. We demonstrate this using the following theorem.

Theorem 4.10.1. *For a fixed overall sample size, the resolution $R_{\bar{D}_n}(\mathcal{M}(y))$ is larger for balanced designs compared to unbalanced designs.*

Proof. We need to show that for a fixed number of level 2 units J, the sample $(n_1, \cdots, n_J)$ maximizes $\sum_{j=1}^{J}((\sigma_u^2 - \gamma) + \frac{\sigma_\epsilon^2}{n_j})^{-1}$, which in turn maximizes the resolution $R_{\bar{D}_n}(\mathcal{M}(y))$, when $n_j = n$ for $j = 1, \ldots, J$.

Using the method of Lagrange multipliers, we find

$$\max_{n_j,\ j=1,\ldots,J} \quad f(n_1,\cdots,n_J) \ = \sum_{j=1}^{J} \Big((\sigma_u^2-\gamma)+\frac{\sigma_\epsilon^2}{n_j}\Big)^{-1} \tag{4.46}$$

$$\text{subject to} \quad g(n_1,\cdots,n_J) \ = \sum_{j=1}^{J} n_j - nJ = 0 \tag{4.47}$$

$$\forall j \in \{1,\ldots,J\}$$

The Lagrangian is

$$L(n_1,\cdots,n_J;\lambda) = f - \lambda g \tag{4.48}$$

Equating the partial derivatives $\frac{\partial L}{\partial n_j}$ and $\frac{\partial L}{\lambda}$ to zero, gives

$$\frac{\partial L}{\partial n_j} = \gamma\frac{\sigma_\epsilon^2}{n_j^2}\Big(\frac{(\sigma_u^2-\gamma)n_j+\sigma_\epsilon^2}{n_j}\Big)^{-2} - \lambda = 0 \tag{4.49}$$

$$\frac{\partial L}{\partial \lambda} = -\sum_{j=1}^{J} n_j + nJ = 0 \tag{4.50}$$

Simplifying the above, we obtain n_j in terms of λ as follows.

$$n_j \ = \frac{\sqrt{\gamma}\sigma_\epsilon}{\sqrt{\lambda}(\sigma_u^2-\gamma)} - \frac{\sigma_\epsilon^2}{(\sigma_u^2-\gamma)} \tag{4.51}$$

Replacing n_j in the constraint yields

$$\sum_{j=1}^{J} n_j \ = \ J\Big(\frac{\sqrt{\gamma}\sigma_\epsilon}{\sqrt{\lambda}(\sigma_u^2-\gamma)} - \frac{\sigma_\epsilon^2}{(\sigma_u^2-\gamma)}\Big) = nJ$$

from which

$$\lambda \ = \Big(\frac{\sqrt{\gamma}\sigma_\epsilon}{n(\sigma_u^2-\gamma)+\sigma_\epsilon^2}\Big)^2$$

Substituting λ from the above and simplifying, yields

$$n_j = n, \quad \forall j \in \{1,\ldots,J\} \tag{4.52}$$

Finally, we need to check that $f(n,\cdots,n)$ is indeed a maximum; we do so by examining the effect of imbalance on (4.46). To create imbalance, suppose we swap one unit between the first two groups, giving $(n-1,n+1,n\cdots,n)$ so that the total sample remains nJ. The decrease $\Big((\sigma_u^2-\gamma)+\frac{\sigma_\epsilon^2}{n}\Big)^{-1}$-$\Big((\sigma_u^2-\gamma)+\frac{\sigma_\epsilon^2}{n-1}\Big)^{-1}$ is larger than

the increase $\left((\sigma_u^2-\gamma)+\frac{\sigma_\epsilon^2}{n+1}\right)^{-1}$ - $\left((\sigma_u^2-\gamma)+\frac{\sigma_\epsilon^2}{n}\right)^{-1}$ since $f(n)=\left((\sigma_u^2-\gamma)+\frac{\sigma_\epsilon^2}{n}\right)^{-1}$, being an increasing function with gradient $f'(n)\propto\frac{1}{n^2}$ (from (4.47) above), decreases at a faster rate than it increases. The more units are swapped to create imbalance, the greater will be the decrease in (4.46) compared to the balanced case. Hence, $f(n_1,n_2\cdots,n_J)<f(n,\cdots,n)$, when n_j are unequal. ■

To summarize, based on the resolution $R_{\bar{D}_n}(\mathcal{M}(y))$ and ignoring cost of sampling, for a fixed overall sample we would recommend the following for a two-stage cluster sample to reduce uncertainty in $\mathcal{M}(y)$:

1. Sample more level 2 units.
2. Sample less level 1 units, reduce n_j so as to maximize J to achieve 1. above.
3. Use a balanced sample.

It is interesting to note that the above recommendations agree with those of the United Nations Statistics Division (2005) for a two-stage cluster sample design in the context of designing an efficient household sample survey.

4.10.1 The canonical structure and the choice of sample size

In order to determine the desired sample size for our SOEREF model, we exploit the result that the eigenvectors of the resolution transform $\mathbb{T}_n$ are the same for each sample n in the adjustment of the collection of population group j means $\mathcal{M}(y_j)$. To this end, we re-state Theorem 6.5 of Goldstein and Wooff (2007; page 198) below, and prove it in the specific context of the adjustment of the SOEREF model.

Theorem 4.10.2. *The eigenvectors of $\mathbb{T}_n$ are the same for each n. Further, if eigenvector W has eigenvalue λ for $\mathbb{T}_1$, then the corresponding eigenvalue $\lambda_{(n)}$ for W as an eigenvector of $\mathbb{T}_n$ is*

$$\lambda_{(n)}=\frac{n\lambda}{(n-1)\lambda+1} \tag{4.53}$$

Proof. The resolution transform matrix $\mathbb{T}_n$ has an eigenvector v corresponding to the eigenvalue λ if

$$\mathbb{T}_n v = \lambda v$$

Substituting $\mathbb{T}_n = Var(\bar{D}_n)^{-1}Cov(\bar{D}_n, \mathcal{C}_\mathcal{M})$ from the proof of Theorem 4.5.2, we have

$$Var(\bar{D}_n)^{-1}Cov(\bar{D}_n, \mathcal{C}_\mathcal{M})v = \lambda v$$

Multiplying both sides of the above equation by $Var(\bar{D}_n)$

$$Cov(\bar{D}_n, \mathcal{C}_\mathcal{M})v = \lambda Var(\bar{D}_n)v$$

Substituting expressions for $Cov(\bar{D}_n, \mathcal{C}_\mathcal{M})$ and $Var(\bar{D}_n)$ (see proof of Theorem 4.5.2), we obtain

$$\begin{aligned}
[(\sigma_u^2 - \gamma)I_J + \gamma K_J]v &= \lambda[(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n})I_J + \gamma K_J]v \\
&= \lambda[(\sigma_u^2 - \gamma)I_J + \gamma K_J]v + \lambda\frac{\sigma_\epsilon^2}{n}I_J v \\
[(\sigma_u^2 - \gamma)I_J + \gamma K_J](1-\lambda)v &= \lambda\frac{\sigma_\epsilon^2}{n}I_J v \\
[(\sigma_u^2 - \gamma)I_J + \gamma K_J]v &= \frac{\lambda}{n(1-\lambda)}\sigma_\epsilon^2 I_J v \qquad (4.54)
\end{aligned}$$

Putting $n = 1$ in (4.54),

$$[(\sigma_u^2 - \gamma)I_J + \gamma K_J]v = \frac{\lambda}{(1-\lambda)}\sigma_\epsilon^2 I_J v \qquad (4.55)$$

Equating the right-hand sides of (4.54) and (4.55) gives

$$\frac{\lambda}{(1-\lambda)} = \frac{\lambda_{(n)}}{n(1-\lambda_{(n)})},$$

from which we obtain the required result

$$\lambda_{(n)} = \frac{n\lambda}{(n-1)\lambda + 1}$$

■

It is straightforward to verify that both $\lambda_{1(n)}$ and $\lambda_{2(n)}$, corresponding to λ_1 and λ_2 respectively of Sub-section 4.5.2, satisfy (4.53).

Goldstein and Wooff (2007) use relation (4.53) to simplify sample size design for general exchangeable adjustments. In this context, they also provide, via their Corollary 6.6, inequality (4.56) below.

$$n \geq \frac{\alpha}{1-\alpha}\frac{1-\lambda}{\lambda}, \tag{4.56}$$

where α is the proportionate reduction in variance. We shall consider using (4.56) to determine a suitable sample size to achieve the desired variance reduction over the elements of $\langle \mathcal{M}(y_j) \rangle$.

4.10.2 The design of sample size at Level 1

For our SOEREF model the smallest canonical resolution is $\lambda_{min} = \lambda_2$ with $n = 1$, hence

$$\lambda_{min} = \frac{(\sigma_u^2 - \gamma)}{(\sigma_u^2 - \gamma + \sigma_\epsilon^2)}, \tag{4.57}$$

that is $\lambda_{min} = \rho$, the intra-cluster correlation, which upon substitution in (4.56) yields

$$n \geq \frac{\alpha}{1-\alpha}\frac{1-\rho}{\rho}. \tag{4.58}$$

As a simple application of (4.58), consider a reduction in variance of α=90%. If level 1 units within a group are very similar, for example $\rho = 0.9$, then using (4.58) a sample of $n = 1$ in each group is enough for a 90% reduction in variance when learning about $M(Y_1) - M(Y_2)$ or any of the $J - 2$ remaining contrasts associated with λ_2. This result is intuitive since, if units in a group were all similar, then only one level 1 unit would provide all the information available from that group. Whereas, if level 1 units within a group differ substantially for example $\rho = 0.1$, then a sample of $n = 81$ is required to achieve the 90% reduction in variance.

4.10.3 Optimal design for a two-level model

In the previous section we considered the choice of level 1 sample size only. In a multilevel setting the design problem is more complex for two main reasons. Firstly,

in a simple two-level hierarchy, we need to decide on two sample sizes, each corresponding to a level in the hierarchy. Secondly, in practical survey design, we need to balance costs of sampling units at each level of the hierarchy against a total allocated cost for the survey since it is usually more costly to survey an extra level 2 rather than an extra level 1 unit.

The problem, therefore, is to choose a design (n, J), the number of level 1 and level 2 units respectively, subject to a given cost constraint. In Section 4.4, we saw that the resolution $\mathcal{R}_{\bar{D}_n}\mathcal{M}(y)$ depends on, and has important implications for, the design (n, J). We therefore aim to choose a design (n, J) that maximizes the resolution of the population grand mean $\mathcal{R}_{\bar{D}_n}\mathcal{M}(y)$ subject to a simple cost function. We have the following theorem.

Theorem 4.10.3. *In the two-level SOEREF model, suppose c_1 and c_2 are the costs associated with sampling a single level 2 and a single level 1 unit respectively. Then a simple cost function when sampling J level 2 and n level 1 units is $C = c_1 J + c_2 Jn$, where n and J each take integer values $\{1, 2, \ldots\}$. The optimal level 1 and level 2 sample sizes (n_{opt}, J_{opt}) is the sample design obtained by maximizing the resolution of the population grand mean $\mathcal{R}_{\bar{D}_n}\mathcal{M}(y)$ subject to the cost constraint C, and is given by*

$$n_{opt} = \sqrt{\frac{c_1}{c_2}\frac{\sigma_\epsilon^2}{(\sigma_u^2 - \gamma)}}, \tag{4.59}$$

and

$$J_{opt} = \frac{C}{(c_1 + c_2 n_{opt})}. \tag{4.60}$$

Proof. From Section 4.3.2, the resolution $\mathcal{R}_{\bar{D}_n}\mathcal{M}(y)$ for the balanced case may be written as

$$\mathcal{R}_{\bar{D}_n}\mathcal{M}(y) = \big(1 + \frac{1}{J\gamma}(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n})\big)^{-1} \tag{4.61}$$

Using the cost function we obtain

$$J = \frac{C}{(c_1 + c_2 n)}, \tag{4.62}$$

and substituting J in (4.59) yields

$$\mathcal{R}_{\bar{D}_n}\mathcal{M}(y) = \big(1 + \frac{1}{C\gamma}(c_1 + c_2 n)(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n})\big)^{-1}. \tag{4.63}$$

Taking the first derivative with respect to n and equating to zero gives

$$\frac{d\mathcal{R}_{\bar{D}_n}\mathcal{M}(y)}{dn} = -(1 + \frac{1}{C\gamma}(c_1 + c_2 n)(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n})\big)^{-2}$$
$$(c_2(\sigma_u^2 - \gamma) - c_1\sigma_\epsilon^2 n^{-2}) = 0 \tag{4.64}$$

Simplifying (4.64) yields

$$n^2 = \frac{c_1}{c_2}\frac{\sigma_\epsilon^2}{(\sigma_u^2 - \gamma)}, \tag{4.65}$$

from which n_{opt} is obtained. The optimal level 2 sample size J_{opt} is obtained by substituting n_{opt} in (4.62).

We compute the second derivative to verify that n_{opt} is indeed a maximum. First we note that (4.62) may be written as

$$\frac{d\mathcal{R}_{\bar{D}_n}\mathcal{M}(y)}{dn} = -(\mathcal{R}_{\bar{D}_n}\mathcal{M}(y))^2(c_2(\sigma_u^2 - \gamma) - c_1\sigma_\epsilon^2 n^{-2}) \tag{4.66}$$

from which the second derivative is

$$\frac{d^2\mathcal{R}_{\bar{D}_n}\mathcal{M}(y)}{dn^2} = -2(\mathcal{R}_{\bar{D}_n}\mathcal{M}(y))\frac{d\mathcal{R}_{\bar{D}_n}\mathcal{M}(y)}{dn}(c_2(\sigma_u^2 - \gamma) - c_1\sigma_\epsilon^2 n^{-2})$$
$$-2(\mathcal{R}_{\bar{D}_n}\mathcal{M}(y))^2 c_1\sigma_\epsilon^2 n^{-3}. \tag{4.67}$$

Putting $n = n_{opt}$ and noting that $(c_2(\sigma_u^2 - \gamma) - c_1\sigma_\epsilon^2 n_{opt}^{-2}) = 0$, (4.65) simplifies to

$$\frac{d^2\mathcal{R}_{\bar{D}_n}\mathcal{M}(y)}{dn^2} = -2(\mathcal{R}_{\bar{D}_n}\mathcal{M}(y))^2 c_1\sigma_\epsilon^2 n_{opt}^{-3}. \tag{4.68}$$

The above is clearly negative as the right-hand side is the product of -2 and terms that are all positive. Since the second derivative is negative, therefore (n_{opt}, J_{opt}) is a maximum. ■

Compare (4.59) with the classical optimal sample size (n_{opt}^*) for two-stage cluster sampling using the same cost function as above (see Cochran, 1999).

$$n_{opt}^* = \sqrt{\frac{c_1}{c_2}\frac{S_2^2}{S_u^2}}, \tag{4.69}$$

where S_2^2 and S_u^2 are the ANOVA estimates of level 1 and level 2 variances. Note that we use c1 to denotes level 2 cost and c2 level 1 cost. This is in line with the classical literature, where c1 denote cost of primary sampling units (level 2 units) and c2 cost of second-stage units (level 1 units).

In both (4.59) and (4.69) the optimal sample depends on the ratio of the level 1 and level 2 costs and variances. An important difference is that the classical n^*_{opt} depends on the ratio of the ANOVA estimates of level 1 and level 2 variances (see Section 2.7.1) and, therefore, suffers from the disadvantages associated with these estimators, especially the possibility of a negative estimate of the level 2 variance. In contrast, our Bayes Linear n_{opt} depends on the ratio of the level 1 and level 2 variances that we specify subjectively. In addition, for our specifications to be coherent, we restrict both these variances such that $\sigma_\epsilon^2 > 0$ and $(\sigma_u^2 - \gamma) \geq 0$ (see Section 3.7). In learning about variances, however, the possibility of a negative level 2 update cannot be excluded. The conditions under which a negative estimate of $(\sigma_u^2 - \gamma)$ occurs will be considered in Chapter 5.

4.10.4 Some considerations in the application of n_{opt}

Below we discuss some practical implications of the formula for n_{opt} in the choice of an optimal design.

The form of n_{opt}

Firstly we note that n_{opt} may also be written as

$$n_{opt} = \sqrt{\frac{c_1}{c_2}\frac{(1-\rho)}{\rho}} \tag{4.70}$$

showing the dependence of n_{opt} on the intracluster correlation ρ. When ρ is small, then the optimum size n_{opt} increases. In other words when there is more variation within than between groups, more level 1 units should be sampled. Similarly, n_{opt} increases when c_1 is larger compared to c_2. That is when it is more costly to sample a level 2 unit compared to sampling a level 1 unit, then more level 1 units should be sampled.

Also, as n_{opt} depends on the square root of the cost ratio c_1/c_2 and the intra-cluster correlation ratio $(1-\rho)/\rho$, the optimum size is not too sensitive to small variations in these quantities.

The cost function C

We have assumed a simple cost function C, which is adequate in sampling situations where the cost of travel between clusters is negligible. This is the case for our STAT1010 data in which the cost of travel between classrooms (clusters) is negligible because the University of Mauritius has a single small campus. When travel costs between clusters are substantial, $C = c_1 J + c_2 Jn + c_3\sqrt{J}$ is a more suitable cost function. The derivation of this cost function, along with more general cost functions are considered in detail in Hansen, Hurwitz & Madow (1953).

Because of the analogous form of the Bayes linear n_{opt} and the classical n^*_{opt}, the above implications are similar to those discussed in Hansen, Hurwitz & Madow (1953). The Bayes linear approach to the design of two-stage cluster sampling that we develop in this thesis aims to combine the strengths of both the classical and Bayes linear methods.

4.11 Example: Two-level design for the STAT1010 data

Application of n_{opt} to the STAT1010 data requires estimates of the costs c_1 and c_2, and careful elicitation of the prior level 1 and 2 variances, σ^2_ϵ and $(\sigma^2_u - \gamma)$. The latter prior variances were elicited in Section 3.7. Therefore we only need to estimate the various costs.

The costs c_1 and c_2, or their ratios, may also be estimated using historical data where available, and/or expert opinion, just as in the elicitation of prior variances. But while variances are notoriously difficult to estimate for the many reasons that we explained in Chapter 2, it is more common for an individual to think in terms of, and assess, costs, rather than variances. Besides, even in the assessment of (subjective) probability itself, one often thinks in terms of costs (prices) of lotteries and their

rewards. The assessment of various elements of costs arising in cluster sampling, as well as the construction of costs functions, are explained in detail in Chapter 6, page 270 of Hansen, Hurwitz & Madow (1953). Furthermore, small inaccuracies in the elicited costs and variances will not unduly affect the optimal sample size as the resolution $\mathcal{R}_{\bar{D}_n}\mathcal{M}(y)$ is mostly flat around n_{opt}. We illustrate this below using the STAT1010 data.

Based on our experience in conducting local surveys, we specify reasonable costs (in Mauritian Rupees) for a small project as follows. The total budget available is Rs 5000, the cost of sampling a class is Rs 500 and that of collecting data on a student is Rs 100. The variances are specified as in Section 3.7. Thus we specify

$$C = 5000 \quad c_1 = 500 \quad c_2 = 100 \quad \sigma^2_\epsilon = 237 \quad \gamma = 56.3 \quad \sigma^2_u = 115$$

Using the above specifications and $\mathcal{R}_{\bar{D}_n}\mathcal{M}(y)$ as in (4.61), we calculate the resolutions for different sample sizes as shown in Table 4.3.

Level 1 size $n =$	1	2	3	4	5	6	7	8	9	10
Resolution (%)	61.295	69.359	71.812	72.548	72.553	72.189	71.623	70.940	70.187	69.393
Relative Resolution	0.84	0.96	0.99	1.00	1.00	0.99	0.99	0.98	0.97	0.96

Table 4.3: *resolutions and relative resolutions (the proportion of resolution relative to the maximum resolution at $n = 5$, if a sample size other than the optimal is chosen) for different level 1 sample sizes of the STAT1010 data.*

As shown in Table 4.3, the resolution around n_{opt} is flat; the maximum loss in resolution relative to the optimum is 16% and corresponds to $n = 1$. Note that taking a larger level 1 sample than the optimum also reduces the resolution. This is due to the cost constraint that requires balancing level 1 and 2 sample sizes; increasing n can only be done at the expense of reducing J. But, using our result in Section 4.4, a one-unit reduction in J will result in a larger reduction in the resolution $\mathcal{R}_{\bar{D}_n}\mathcal{M}(y)$ than a one-unit reduction in n.

Hence for our STAT1010 example, the optimal size is $n_{opt} = 5$ and it is expected to resolve 72.6% of the prior uncertainty in the overall population mean $\mathcal{M}(y)$. Using (4.60), the corresponding optimal level 2 sample size is $J_{opt} = 5$. The use of the

resolution to guide sample size design (n_{opt}, J_{opt}) requires prior specifications of the level 1 and 2 variances. This contrasts with the classical design (n^*_{opt}, J^*_{opt}) that requires estimates of these variances based on historical data or data from pilot studies. Both these approaches have their own strengths. Incorporating prior information from carefully elicited experts' beliefs is an advantage of the general Bayesian approach, especially when data is scarce, while learning about model parameters from data is equally important. Thus estimating variances based on observable data is also advantageous. In this thesis, we aim to combine the strengths of these two approaches via the two-stage Bayes linear methodology that we will develop in Chapter 6.

For the simple cost function, use of the explicit formula for n_{opt} further simplifies the calculation of the optimum. For more complicated cost functions, the calculations leading to Table 4.1 demonstrate an alternative method for obtaining n_{opt}, that is by evaluating $\mathcal{R}_{\bar{D}_n}\mathcal{M}(y)$ for various n and the given costs. We illustrate this for the more complex cost function that includes travel costs between clusters.

4.11.1 Example: Application of two-level design for a more complex cost function

Consider now the cost function $C = c_1 J + c_2 Jn + c_3\sqrt{J}$, where c_3, the travel costs between clusters, tends to be proportional to $\sqrt{J}$ (see Hansen, Hurwitz & Madow, 1953). For comparison sake we make the same specifications for the costs C, c_1, c_2 and the variances as above. We assume that we are sampling classes in different parts of Mauritius and that we are travelling by taxi. The cost of travel to and from classes and the waiting time for a taxi is around Rs1,000. Thus we specify $c_3 = 1000$ which is twice c_1, the cost of sampling an additional cluster.

Our aim is to evaluate the resolution for various sample sizes just as in Table 4.1. We substitute $n = \frac{C - c_1 J - c_3\sqrt{J}}{c_2 J}$ in the formula (4.61) for $\mathcal{R}_{\bar{D}_n}\mathcal{M}(y)$ and write an R function to evaluate the latter . We also need to ensure that $n > 0$. Table 4.4 reveals that the optimal design is $(n_{opt} = 6, J_{opt} = 3)$ with a resolution of about 63% compared to $(n_{opt} = 5, J_{opt} = 5)$ with a resolution of about 73% for the simple cost function. The additional traveling cost has thus resulted in a lower overall sample

Level 2 size J	1	2	3	4	5
Level 1 size n	35	13	6	3	1
Resolution (%)	45.988	59.156	62.870	59.291	35.534

Table 4.4: *Determination of the optimal level 1 and 2 sample sizes and resolutions for the STAT1010 data for a complex cost function.*

size of 18 (a 28% reduction), fewer clusters $J = 3$ and slightly more level 1 units $n_j = 6$ being sampled. But what if we wanted to achieve the same resolution of 73% as in the simple cost situation, even at a higher sampling cost? We address this issue next.

4.11.2 Determining the optimal design and cost to achieve a desired level of resolution

So far we have determined the optimal sample size by maximizing the resolution when the total sampling cost C is given. In some situations though, it may be quite difficult to ascertain C while it may still be relatively easy to estimate the ratio of the level 2 cost c_1 and the level 1 cost c_2. We may then determine the optimal size by minimizing C for a pre-determined resolution. We show below that minimizing cost subject to a fixed resolution or maximizing resolution for a fixed cost results in optimal designs of the same form.

Corollary 4.11.1. *Under the conditions of Theorem 4.10.2, minimizing the simple cost function $C = c_1 J + c_2 Jn$ when the resolution of the population grand mean $\mathcal{R}_{\bar{D}_n}\mathcal{M}(y)$ is kept fixed gives the same optimal design (n_{opt}, J_{opt}) as in Theorem 4.10.2., namely*

$$n_{opt} = \sqrt{\frac{c_1}{c_2}\frac{\sigma_\epsilon^2}{(\sigma_u^2 - \gamma)}}, \tag{4.71}$$

and

$$J_{opt} = \frac{C}{(c_1 + c_2 n_{opt})}. \tag{4.72}$$

Proof. To minimize the cost C subject to keeping the resolution $\mathcal{R}_{\bar{D}_n}\mathcal{M}(y)$ fixed, we re-write expression (4.63) for $\mathcal{R}_{\bar{D}_n}\mathcal{M}(y)$ as follows.

$$C = \frac{\mathcal{R}_{\bar{D}_n}\mathcal{M}(y)}{\gamma(1-\mathcal{R}_{\bar{D}_n}\mathcal{M}(y))}(c_1 + c_2 n)(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n}). \tag{4.73}$$

Taking the first derivative gives

$$\frac{dC}{dn} = \frac{\mathcal{R}_{\bar{D}_n}\mathcal{M}(y)}{\gamma(1-\mathcal{R}_{\bar{D}_n}\mathcal{M}(y))}(c_2(\sigma_u^2 - \gamma) - c_1\sigma_\epsilon^2 n^{-2}). \tag{4.74}$$

Equating (4.74) to zero yields n_{opt} as in (4.71) and substituting it in $C = c_1 J + c_2 Jn$ yields J_{opt} as in (4.72). Also, from (4.74) we note that $\frac{d^2C}{dn^2} \propto 2c_1\sigma_\epsilon^2 n^{-3} > 0$, therefore (n_{opt}, J_{opt}) yields the minimum cost C as required.

■

As an application of Corollary 4.11.1, we answer the question posed at the end of Sub-section 4.11.1, namely using the complex cost function, which design will achieve a resolution of 73% (as was obtained using the simple cost function) and at what cost. We wrote a simple R function that finds the minimum cost by evaluating C for varying n using (4.73). Using this function we find that the design to achieve the given resolution of 73% is $(n_{opt} = 4, J_{opt} = 6)$ with a minimum cost of Rs7,395. We note that as the n's must be integers, we must suitably round it up or down.

We may compare the above design with the one in the previous section, where, using the same complex cost function, a resolution of 63% gave the optimal design $(n_{opt} = 6, J_{opt} = 3)$ for a given cost of Rs5,000. Thus to raise the resolution from 63% to 73%, we need to sample less level 1 units ($n_{opt} = 4$ instead of 6) and more level 2 units ($J_{opt} = 6$ instead of 3) resulting in an increase in overall sample size of 33% (from 18 to 24) and an increase in cost of almost 48% (from Rs5,000 to Rs7,395).

4.12 The finite SOEREF model

In formulating the SOEREF model in Chapter 3, we have assumed that, for each group j, the level 1 outcome variables $\{y_{j1}, y_{j2}, \ldots\}$ form a potentially infinite sequence. At level 2, we again assumed a potentially infinite sequence for the population group j means $\{\mathcal{M}(y_1), \mathcal{M}(y_2), \ldots\}$ induced by the level 1 SOE judgements.

These assumptions allowed us to make use of the representation theorem for infinite second-order exchangeable random quantities and to subsequently introduce and adjust the overall and group j means of our SOEREF model.

Although multilevel data structures tend to be large in general, so that the assumption of infinite sequences may be reasonable, there are situations however, such as in longitudinal studies, where the populations at each level of a hierarchy are more restricted. Besides, the very nature of a multilevel dataset implies that the higher the level of the hierarchy, the fewer the units of observations: there are fewer districts than schools, fewer schools than classes and fewer classes than pupils.

In such cases the assumption of infinite sequences may be viewed as a modeling simplification and we wish to consider the consequences of relaxing this assumption on the adjustments of the SOEREF model. We begin by reviewing finite exchangeability and finite second-order exchangeability as defined by Goldstein (1986). The latter will also enable us to introduce some relevant notations.

4.12.1 Finite exchangeability

In explaining the concept of exchangeability, Bernardo & Smith (2000;p169) proceed by first defining finite exchangeability as follows.

Definition 4.12.1. *(***Finite exchangeability***). The random quantities $x_1, x_2, \ldots, x_n$ are said to be judged finitely exchangeable under a probability measure P if the implied joint degree of belief distribution satisfies*

$$P(x_1, x_2, \ldots, x_n) = P(x_{\pi(1)}, x_{\pi(2)}, \ldots, x_{\pi(n)}) \tag{4.75}$$

for all permutations π defined on the set $1, \ldots, n$.

They then extend Definition 4.12.1 to sequences that are potentially infinite as follows.

Definition 4.12.2. *(***Infinite exchangeability***). The infinite sequence of random quantities $x_1, x_2, \ldots,$ is said to be judged infinitely exchangeable if every finite subsequence is judged exchangeable in the sense of Definition 4.12.1.*

It seems natural to consider finite exchangeability first since real life populations are finite and so are their associated sequences of random variables. For example, in a multilevel data set, the sequences of pupils, classes and schools are all finite and, using registration statistics, we can easily put an upper bound on the lengths of these sequences. In other situations though, it may not be that straightforward to place an upper bound on the sizes of populations being considered. In fact many surveys have as primary objective the estimation of the population sizes themselves. When it is difficult to specify an upper bound for a large population, we often resort to assuming an in principle, infinite sequence. Such an assumption, like any modelling assumption, requires careful examination.

Assuming infinite exchangeability for an arbitrary (finite) sequence in the sense of Definition 4.14.2 implies that this sequence can be deemed to form part of an infinite sequence. But as Bernardo & Smith (2000) show, not all finitely exchangeable sequences can be embedded in a larger finitely exchangeable sequence, let alone an infinite one. Using a mathematical example they show that an exchangeable sequence of three binary random variables cannot be extended to a sequence of four binary random variables. In addition, in some situations, there may be no logical basis to extend a finite sequence. We have only 200 secondary schools for example, and there is no logical basis to extend this sequence of schools.

Diaconis & Freedman (1980) show that deFinetti's representation theorem for infinite sequences does not hold exactly for a finite sequence. However, they prove that the difference in probabilities when assuming an infinite approximation to a finite sequence of length n is of the order $(1/n)$. Hence the infinite assumption may not result in any sizeable difference, especially when n is large.

4.12.2 Finite second-order exchangeability and finite population representation theorem

We have already defined second-order exchangeability and stated the infinite population representation theorem in Chapter 2. Here we consider an alternative, but equivalent definition of second-order exchangeability as given in Goldstein(1986). The notations used in this definition and in formulating the finite representation

theorem are most suitable for deriving the finite representation version of our SOEREF model.

Suppose a series of observations are made on a sample of individuals and we group these in the collection $\mathcal{C} = \{X_1, X_2, \ldots\}$. Further, we denote the collection for individual i by $\mathcal{C}_i = \{X_{1i}, X_{2i}, \ldots\}$ and the full population collection by $\mathcal{C}^*$, where $\mathcal{C}^*$ is the union of all the elements in all of the individual collections.

Definition 4.12.3. *(Goldstein (1986)). The collection of measurements $\mathcal{C}$ is second-order exchangeable over the full collection $\mathcal{C}^*$ if*

$$E(X_{vi}) = m_v \forall v, i; \tag{4.76}$$

$$Cov(X_{vi}, X_{wi}) = d_{vw} \forall v, w, i; \tag{4.77}$$

$$Cov(X_{vi}, X_{wj}) = c_{vw} \forall v, w, i \neq j; \tag{4.78}$$

The above definition applies to both finite and infinite collections. Using the specifications in Definition 4.12.3, Goldstein (1986) states the finite population representation theorem as follows.

Theorem 4.12.1. *(Goldstein (1986)). If the population collection consists of N individuals, that is $\mathcal{C}^* = \bigcup_{i=1}^{N} \mathcal{C}_i$, and $\mathcal{C}$ is second-order exchangeable over $\mathcal{C}^*$, then we may introduce the further collections of random quantities $\mathcal{M}^{[N]}(\mathcal{C}) = \{\mathcal{M}^{[N]}(X_1), \mathcal{M}^{[N]}(X_2), \ldots\}$, and, for each $i = 1, \ldots, N$, $\mathcal{R}_i^{[N]}(\mathcal{C}) = \{\mathcal{R}_i^{[N]}(X_1), \mathcal{R}_i^{[N]}(X_2) \ldots\}$, and write*

$$X_{vi} = \mathcal{M}^{[N]}(X_v) + \mathcal{R}_i^{[N]}(X_v) \tag{4.79}$$

where $\mathcal{M}^{[N]}(X_v) = (1/N) \sum_{i=1}^{N} X_{vi}$. The collections $\mathcal{M}^{[N]}(\mathcal{C})$ and $\mathcal{R}_i^{[N]}(\mathcal{C})$ satisfy the following relationships

$$E(\mathcal{M}^{[N]}(X_v)) = m_v \ \ \forall v; \tag{4.80}$$

$$E(\mathcal{R}_i^{[N]}(X_v)) = 0 \ \ \forall v, i; \tag{4.81}$$

$$Cov(\mathcal{M}^{[N]}(X_v), \mathcal{M}^{[N]}(X_w)) = c_{vw} + \frac{1}{N}(d_{vw} - c_{vw}) \ \ \forall v, w; \tag{4.82}$$

$$Cov(\mathcal{M}^{[N]}(X_v), \mathcal{R}_j^{[N]}(X_w)) = 0 \ \ \forall v, w, j; \tag{4.83}$$

$$Cov(\mathcal{R}_i^{[N]}(X_v), \mathcal{R}_j^{[N]}(X_w)) = \begin{cases} \frac{N-1}{N}(d_{vw} - c_{vw}) \ \textit{if } i{=}j \ \forall v,w; \\ -\frac{1}{N}(d_{vw} - c_{vw}) \ \textit{otherwise.} \end{cases} \tag{4.84}$$

Just as in the case of infinite exchangeability, each observation is expressed as the sum of a population mean quantity $\mathcal{M}^{[N]}(X_v)$ and a residual $\mathcal{R}_i^{[N]}(X_v)$. And, as before, the population mean collection is uncorrelated with the residual collection. In contrast to the infinite case, the residuals are correlated to the order $(1/N)$. Furthermore, the population mean and residuals are unobservable in the infinite case while they are clearly observable by virtue of their definitions in the above theorem. The residuals $\mathcal{R}_i^{[N]}(X_v)$ will be uncorrelated in the limit $N \to \infty$. Similarly, the finite population representation theorem reduces to the infinite representation theorem in the limit $N \to \infty$, the limit being in mean square.

4.13 The representation theorem for the finite SOEREF model

To derive the representation theorem for the finite two-level SOEREF model, we assume that our multilevel population has a finite number of groups G, each with a finite number of individuals N. We also assume that we sample J groups from G where $(J \leq G)$ and, in each of these J groups, we sample n individuals out of N where $(n \leq N)$, that is a balanced design. We observe a single response variable y_{ji} on each individual i in group j, where $i = 1, 2, \ldots, n$ and $j = 1, 2, \ldots, J$.

Our exchangeability judgements are similar to those of the infinite SOEREF case (see Chapter 3) as follows. For each group j we assume individuals within a group are similar, hence second-order exchangeable (Level 1 exchangeability). Further we assume that groups also are similar, that is the group j means resulting from the level 1 exchangeability are themselves second-order exchangeable (Level 2 exchangeability). We have the following finite population representation theorem for the SOEREF model.

Theorem 4.13.1. *Suppose a two-level population consists of a finite number of G groups and a finite number of N individuals in each group. We consider exchangeability judgements at each level of the hierarchy in turn and make use of the same second-order specifications as in Chapter 3 as follows.*

$E(y_{ji}) = \mu, \quad Var(y_{ji}) = \sigma_y^2 \quad \forall\ i,j, \quad Cov(y_{ji}, y_{ji'}) = \sigma_u^2 \quad i \neq i',$
$Cov(y_{ji}, y_{j'i'}) = \gamma \quad i \neq i' and \quad j \neq j'.$

Level 1 exchangeability

If at level 1 of the hierarchy we judge that individuals within group j are second-order exchangeable, then we may introduce the further collection of random quantities $\{\mathcal{M}^{[N]}(y_1), \mathcal{M}^{[N]}(y_2), \ldots, \mathcal{M}^{[N]}(y_G)\}$, and, for each $i = 1, 2, \ldots, N$, $\{\mathcal{R}_i^{[N]}(y_1), \mathcal{R}_i^{[N]}(y_2), \ldots, \mathcal{R}_i^{[N]}(y_G)\}$, and write

$$y_{ji} = \mathcal{M}^{[N]}(y_j) + \mathcal{R}_i^{[N]}(y_j) \tag{4.85}$$

where the finite group j mean $\mathcal{M}^{[N]}(y_j) = \frac{1}{N}\sum_{i=1}^{N} y_{ji}$. The collections $\mathcal{M}^{[N]}(y_j)$ and $\mathcal{R}_i^{[N]}(y_j)$ satisfy the following relationships

$$E(\mathcal{M}^{[N]}(y_j)) = \mu, \quad \forall j \tag{4.86}$$

$$Var(\mathcal{M}^{[N]}(y_j)) = \sigma_u^2 + \frac{1}{N}(\sigma_y^2 - \sigma_u^2), \quad \forall j \tag{4.87}$$

$$E(\mathcal{R}_i^{[N]}(y_j)) = 0, \quad \forall i, j \tag{4.88}$$

$$Cov(\mathcal{M}^{[N]}(y_j), \mathcal{M}^{[N]}(y_{j'})) = \gamma, \quad \forall j \neq j' \tag{4.89}$$

$$Cov(\mathcal{M}^{[N]}(y_j), \mathcal{R}_i^{[N]}(y_j)) = 0 \tag{4.90}$$

$$Cov(\mathcal{R}_i^{[N]}(y_j), \mathcal{R}_{i'}^{[N]}(y_j)) = \begin{cases} \frac{N-1}{N}(\sigma_y^2 - \sigma_u^2) \text{ if } i = i' \quad \forall j \\ -\frac{1}{N}(\sigma_y^2 - \sigma_u^2) \text{ otherwise.} \end{cases} \tag{4.91}$$

Level 2 exchangeability

If at level 2 of the hierarchy we judge that groups are second-order exchangeable, then we may introduce the further random quantity $\mathcal{M}^{[G]}(y)$, and, for each $j = 1, 2, \ldots, G$, $\{\mathcal{R}_1^{[G]}(\mathcal{M}^{[N]}(y)), \mathcal{R}_2^{[G]}(\mathcal{M}^{[N]}(y)), \ldots, \mathcal{R}_G^{[G]}(\mathcal{M}^{[N]}(y))\}$, and write

$$\mathcal{M}^{[N]}(y_j) = \mathcal{M}^{[G]}(y) + \mathcal{R}_j^{[G]}(\mathcal{M}^{[N]}(y)) \tag{4.92}$$

where finite population grand mean $\mathcal{M}^{[G]}(y) = \frac{1}{G}\sum_{j=1}^{G} \mathcal{M}^{[N]}(y_j)$. The collection

$\mathcal{R}_j^{[G]}(\mathcal{M}^{[N]}(y))$ *and the quantity* $\mathcal{M}^{[G]}(y)$ *satisfy the following relationships.*

$$E(\mathcal{M}^{[G]}(y))) = \mu, \quad \forall j \tag{4.93}$$

$$Var(\mathcal{M}^{[G]}(y)) = \gamma + \frac{1}{G}((\sigma_u^2 + \frac{1}{N}\sigma_\epsilon^2) - \gamma), \tag{4.94}$$

$$E(\mathcal{R}_j^{[G]}(\mathcal{M}^{[N]}(y))) = 0, \quad \forall j \tag{4.95}$$

$$Cov(\mathcal{M}^{[G]}(y), \mathcal{R}_j^{[G]}(\mathcal{M}^{[N]}(y))) = 0 \tag{4.96}$$

$$Cov(\mathcal{R}_j^{[G]}(\mathcal{M}^{[N]}(y)), \mathcal{R}_{j'}^{[G]}(\mathcal{M}^{[N]}(y))) = \begin{cases} \frac{G-1}{G}((\sigma_u^2 + \frac{1}{N}\sigma_\epsilon^2) - \gamma) \ \textit{if } j = j' \\ -\frac{1}{G}((\sigma_u^2 + \frac{1}{N}\sigma_\epsilon^2) - \gamma) \ \textit{otherwise.} \end{cases} \tag{4.97}$$

Proof. The proof of Theorem 4.13.1 follows from an application of the finite population representation theorem of Goldstein (1986) as follows. For the level 1 representation, we compare the specifications for y_{ji} with those of X_{vi} in Definition 4.12.3, we see that $c_{vw} = \sigma_u^2$ and $d_{vw} = \sigma_y^2$ and substituting in (4.82) and (4.84) yield the corresponding relationships in (4.87) and (4.91).

For the level 2 representation, we derive the specifications for $\bar{y}_{j.}$ from those of y_{ji}. Thus, $E(\bar{y}_{j.}) = \mu$, $Cov(\bar{y}_{j.}, \bar{y}_{j'.}) = \sigma_u^2 + \frac{1}{N}\sigma_\epsilon^2 \quad j = j'$,
$Cov(\bar{y}_{j.}, \bar{y}_{j'.}) = \gamma \quad j \neq j'$,
Again comparing the above with the specifications for X_{vi}, we see that $c_{vw} = \gamma$ and $d_{vw} = \sigma_u^2 + \frac{1}{N}\sigma_\epsilon^2$ and substituting again in (4.82) and (4.84) yield the corresponding relationships in (4.94) and (4.97). ■

4.13.1 Comparing the finite and the infinite SOEREF model

Our SOE judgements are similar for both the finite SOEREF model of Theorem 4.13.1 and the infinite version of Chapter 3. And in both cases our observations are expressed as the sum of a population mean quantity and a residual from this mean via the appropriate representation theorem. Furthermore, below we demonstrate that infinite and finite SOEREF models are similar if the level 1 and level 2 populations are large compared to the respective sample sizes. Therefore all our analyses could be carried out in terms of finite exchangeability (which would be more precise because populations are not infinite) but provided the populations are large compared to the sample sizes, then this would make little difference.

At level 1, $\mathcal{M}^{[N]}(y_j)$ is the finite population group j mean and $\mathcal{R}_i^{[N]}(y_j)$ is the collection of level 1 residuals, that is discrepancy for individual i in group j from its group j mean. As in the infinite case, (4.90) shows that each $\mathcal{M}^{[N]}(y_j)$ is uncorrelated with the collection of level 1 residuals. In (4.91) the level 1 residuals have a small negative correlation to the order of $(1/N)$. In the limit $N \to \infty$ all the quantities in (4.87) and (4.89) tend to their infinitely exchangeable counterparts, that is

$$\lim_{N\to\infty} Var(\mathcal{M}^{[N]}(y_j)) = \sigma_u^2 \quad \forall j \tag{4.98}$$

$$\lim_{N\to\infty} Cov(\mathcal{R}_i^{[N]}(y_j), \mathcal{R}_{i'}^{[N]}(y_j)) = \begin{cases} (\sigma_y^2 - \sigma_u^2) = \sigma_\epsilon^2 \text{ if } i = i' & \forall \text{j} \\ 0 \text{ otherwise.} \end{cases} \tag{4.99}$$

We note that the level 1 residuals are uncorrelated as in the infinitely exchangeable case. In the infinite SOEREF model the population group j mean, $\mathcal{M}(y_j)$ and the level 1 residual for individual i, $\mathcal{R}_i(y_j)$ are unobservable while their finite counterparts $\mathcal{M}^{[N]}(y_j)$ and $\mathcal{R}_i^{[N]}(y_j)$ are observable.

At level 2, $\mathcal{M}^{[G]}(y)$ is the finite population grand mean and $\mathcal{R}_j^{[G]}(\mathcal{M}^{[N]}(y))$ is the level 2 residual, that is the discrepancy for the group j mean from the grand mean. The grand mean is uncorrelated with the collection of level 2 residuals as in the infinite case. In (4.97) the level 2 residuals have a small negative correlation to the order of $(1/G)$. In the limit as both $N \to \infty$ and $G \to \infty$ all the quantities in (4.94) and (4.97) tend to their infinitely exchangeable counterparts as follows

$$\lim_{(N,G)\to\infty} Var(\mathcal{M}^{[G]}(y)) = \gamma, \tag{4.100}$$

$$\lim_{(N,G)\to\infty} Cov(\mathcal{R}_j^{[G]}(\mathcal{M}^{[N]}(y)), \mathcal{R}_{j'}^{[G]}(\mathcal{M}^{[N]}(y))) = \begin{cases} (\sigma_u^2 - \gamma) \text{ if } j = j' \\ 0 \text{ otherwise.} \end{cases} \tag{4.101}$$

The level 2 residuals are now uncorrelated as in the infinitely exchangeable case. Both $\mathcal{M}^{[G]}(y)$ and $\mathcal{R}_j^{[G]}(\mathcal{M}^{[N]}(y))$ are observable while the corresponding $\mathcal{M}(y)$ and $\mathcal{R}_j(\mathcal{M}(y))$ in the infinite case are not observable.

In multilevel data we may have a smaller population of level 2 units than of level 1 units - there are fewer schools than children in these schools. The reverse, a potentially infinite population of level 2 units and a finite population of level 1 units, is also possible: a population of dyads, for instance, have a larger level 2

population (married couples for example) but with only two level 1 units (husband and wife) in each dyad. It is straightforward to examine the effects of these two cases on Theorem 4.13.1 by setting $N \to \infty$ while G remains finite in the first case, while in the second one $G \to \infty$ while N remains finite.

4.13.2 Comparing the finite SOEREF and the finite exchangeable multivariate model

The proof of Theorem 4.13.1 reveals an important result: application of the finite population representation theorem of Goldstein (1986) (here Theorem 4.12.1) to each level of the hierarchy yields the finite representation for the two-level SOEREF model. Although, Theorem 4.12.1 was stated for the multivariate variable X_{vi}, it is equally applicable to the multilevel variable y_{ji} since the clustering of individuals in a group j induces a correlation structure analogous to the multivariate situation.

In the multivariate context, Goldstein (1986) emphasizes that, whatever the total number of individuals in the population, we only need to consider our SOE judgements for two individuals, with all other specifications *following from the perceived symmetries in the population*. But how does this translate to our multilevel model where not only we have a population of individuals, but we have also a population of groups? Intuitively, we need to consider only two individuals ($N = 2$) in each of only two groups ($G = 2$). We demonstrate this as follows.

Using the expressions in (4.93), $Var(\mathcal{R}_i^{[N]}(y_j)) = \frac{N-1}{N}(\sigma_y^2 - \sigma_u^2)$, hence

$$Cov(\mathcal{R}_i^{[N]}(y_j), \mathcal{R}_{i'}^{[N]}(y_j)) = -\frac{1}{N-1} Var(\mathcal{R}_i^{[N]}(y_j)). \tag{4.102}$$

Similarly, using (4.99), $Var(\mathcal{R}_i^{[G]}(y)) = \frac{G-1}{G}((\sigma_u^2 + \frac{1}{N}\sigma_\epsilon^2) - \gamma)$, hence

$$Cov(\mathcal{R}_j^{[G]}(\mathcal{M}^{[N]}(y)), \mathcal{R}_{j'}^{[G]}(\mathcal{M}^{[N]}(y))) = -\frac{1}{G-1} Var(\mathcal{R}_i^{[G]}(y)) \tag{4.103}$$

For (4.102) and (4.103) to be coherent both N and G should be at least 2, that is a minimum of two groups with a minimum of two individuals in each of these two groups. Conceptually, this means that, whatever the number of groups and the number of individuals in these groups, for our SOE judgements to be coherent, we only need to consider beliefs between two individuals in each of two groups and

beliefs between the two groups. All other specifications will then follow from the perceived symmetries in our finite multilevel data.

4.14 Comparing the adjusted population grand mean for the finite and the infinite SOEREF model

We stated earlier that the assumption of infinite exchangeability is a modelling simplification. We now explore the effect of this simplification on our adjusted beliefs by comparing the adjustments of the population grand mean for the finite and infinite balanced SOEREF model. To adjust the finite population grand mean we use the same specifications as in the infinite case.

Theorem 4.14.1. *Suppose a two-level population consists of a finite number of G groups and a finite number of N individuals in each group, that is a balanced design. Given the representation and SOE specifications for the finite SOEREF model as in Theorem 4.13.1, we adjust the population grand mean using the collection of observed group means* $\bar{D}_n = \{\bar{y}_{1.}, \bar{y}_{2.}, \ldots, \bar{y}_{J.}\}$*, where* $\bar{D}_n$ *is Bayes linear sufficient to adjust beliefs over the mean components. The adjusted grand mean is*

$$E_{\bar{D}_n}(\mathcal{M}^{[G]}(y)) = \Big(1 - \frac{J\big((\sigma_u^2 - \gamma) + \frac{\sigma_\epsilon^2}{N} + G\gamma\big)}{G(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)}\Big)\mu + \frac{J\big((\sigma_u^2 - \gamma) + \frac{\sigma_\epsilon^2}{N} + G\gamma\big)}{G(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)}\bar{y}_{..} \tag{4.104}$$

Proof. Application of the Bayes linear rule gives:

$$E_{\bar{D}_n}(\mathcal{M}^{[G]}(y)) = E(\mathcal{M}^{[G]}(y)) + Cov(\mathcal{M}^{[G]}(y), \bar{D}_n)Var^{-1}(\bar{D}_n)(\bar{D}_n - E(\bar{D}_n))$$

We make use of Theorem 4.13.1 to obtain the quantities on the right of $E_{\bar{D}_n}(\mathcal{M}^{[G]}(y))$. From (4.93) we obtain $E(\mathcal{M}^{[G]}(y)) = \mu$.

Using the representation for the finite SOEREF model we may write $\bar{y}_{j.} = \mathcal{M}^{[G]}(y) + \mathcal{R}_j^{[G]}(\mathcal{M}^{[N]}(y)) + \frac{1}{n}\bar{\mathcal{R}}_i^{[N]}(y_j)$, where $\bar{\mathcal{R}}_n^{[N]}(y_j) = \frac{1}{n}\sum_{i=1}^{n}\mathcal{R}_i^{[N]}(y_j)$. Together with the

SOE specifications and relationships in Theorem 4.15.1 we obtain:

$$\begin{aligned}
Cov(\mathcal{M}^{[G]}(y), \bar{D}_n) &= Cov\big(\mathcal{M}^{[G]}(y), \bar{y}_{1.}, \ldots, \bar{y}_{J.}\big) \\
&= Cov\big(\mathcal{M}^{[G]}(y), \bar{y}_{j.}\big)\mathbf{1}_J^T, \forall j \\
&= Cov\big(\mathcal{M}^{[G]}(y), (\mathcal{M}^{[G]}(y) + \mathcal{R}_j^{[G]}(\mathcal{M}^{[N]}(y)) + \frac{1}{n}\bar{\mathcal{R}}_i^{[N]}(y_j))\big)\mathbf{1}_J^T \\
&= Var(\mathcal{M}^{[G]}(y))\mathbf{1}_J^T, \\
&= \frac{1}{G}\Big((\sigma_u^2 + \frac{1}{N}\sigma_\epsilon^2) + (G-1)\gamma\Big)\mathbf{1}_J^T \\
&= \frac{1}{GN}\Big(N(\sigma_u^2 - \gamma) + \sigma_\epsilon^2 + GN\gamma\Big)\mathbf{1}_J^T
\end{aligned}$$

where $Var(\mathcal{M}^{[G]}(y))$ is given in (4.94).

$Var(\bar{D}_n)$ depends solely on our second-order specifications which are similar to the infinite case in Section 4.2. In Theorem 4.5.2 we obtained the inverse of $Var(\bar{D}_n)$ as follows.

$$Var(\bar{D}_n)^{-1} = \frac{1}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n})}\Big[I_J - \frac{\gamma}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)}K_J\Big]$$

From the above results, we obtain

$$\begin{aligned}
E_{\bar{D}_n}(\mathcal{M}^{[G]}(y)) &= E(\mathcal{M}^{[G]}(y)) + Cov(\mathcal{M}^{[G]}(y), \bar{D}_n)Var^{-1}(\bar{D}_n)(\bar{D}_n - E(\bar{D}_n)) \\
&= \mu + \frac{1}{GN}\Big(N(\sigma_u^2 - \gamma) + \sigma_\epsilon^2 + GN\gamma\Big)\mathbf{1}_J^T \\
&\frac{1}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n})}\Big[I_J - \frac{\gamma}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)}K_J\Big](\bar{D}_n - E(\bar{D}_n)) \\
&= \mu + \frac{1}{GN}\Big(N(\sigma_u^2 - \gamma) + \sigma_\epsilon^2 + GN\gamma\Big) \\
&\frac{1}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n})}\Big[1 - \frac{J\gamma}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)}\Big](J\bar{y}_{..} - J\mu) \\
&= \mu + \frac{1}{GN}\Big(N(\sigma_u^2 - \gamma) + \sigma_\epsilon^2 + GN\gamma\Big) \\
&\Big[\frac{J}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)}\Big](\bar{y}_{..} - \mu) \\
&= \Big(1 - \frac{J\big((\sigma_u^2 - \gamma) + \frac{\sigma_\epsilon^2}{N} + G\gamma\big)}{G(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)}\Big)\mu + \frac{J\big((\sigma_u^2 - \gamma) + \frac{\sigma_\epsilon^2}{N} + G\gamma\big)}{G(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)}\bar{y}_{..}
\end{aligned}$$

■

We now compare the adjustment of the finite grand mean with the infinite balanced case (4.13) of Corollary 4.3.1 which is re-written in the same form as the

above adjusted finite grand mean.

$$E_{\bar{D}_n}(\mathcal{M}(y)) \;=\; \Big(1 - \frac{J\gamma}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)}\Big)\mu + \frac{J\gamma}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)}\bar{y}_{..}$$

Comparing $E_{\bar{D}_n}(\mathcal{M}(y))$ above with the finite version in (4.106), we may write

$$E_{\bar{D}_n}(\mathcal{M}^{[G]}(y)) = E_{\bar{D}_n}(\mathcal{M}(y)) + \frac{J\big((\sigma_u^2 - \gamma) + \frac{\sigma_\epsilon^2}{N}\big)}{G(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)}(\bar{y}_{..} - \mu) \tag{4.105}$$

From (4.105) above, the difference between the finite and infinite adjustments depends on the level 2 sampling fraction J/G, on our specified variances for level 1 and level 2 units, and also on our uncertainty for $\mathcal{M}(y)$ and its prior expectation via the difference $(\bar{y}_{..} - \mu)$. For example, if the level 2 sampling fraction J/G is small and/or μ has been specified close to $\bar{y}_{..}$, then the finite and infinite adjustments will not differ by much.

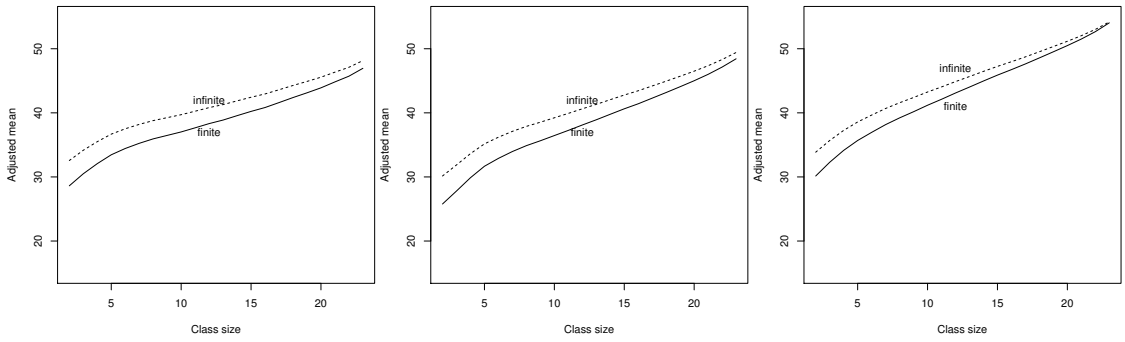
From (4.105) we also note that if we allow only the finite population group size G to increase, then $\lim_{G\to\infty} E_{\bar{D}_n}(\mathcal{M}^{[G]}(y)) = E_{\bar{D}_n}(\mathcal{M}(y))$, that is the finitely adjusted grand mean becomes similar to its infinite version in the limit, irrespective of whether the level 1 population is finite or infinite. Next, if we allow only the level 1 population size N to increase while keeping G fixed, we obtain:

$$\lim_{N\to\infty} E_{\bar{D}_n}(\mathcal{M}^{[G]}(y)) = E_{\bar{D}_n}(\mathcal{M}(y)) + \frac{J\big((\sigma_u^2 - \gamma)\big)}{G(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)}(\bar{y}_{..} - \mu) \tag{4.106}$$

The above expression shows that the difference between the adjusted finite grand mean when N is large and its infinite version is as discussed for expression (4.105) above.

4.14.1 Differences between the finite and infinite adjusted grand mean in the STAT1010 data

We now explore the effect of the level 1 and level 2 sampling fractions on the differences between the finite and infinite adjusted grand mean for the STAT1010 data. In order to obtain a balanced data set to which our finite results apply, we consider only the first $N = 23$ observations for each of $G = 7$ classes. From this finite population we have to sample a minimum of $n = 2$ students in a minimum of $J = 2$ classes



No. of classes, J=2 No. of classes, J=4 No. of classes, J=7

Figure 4.6: *Disparities between the finite and infinite adjustments of the population grand mean* $\mathcal{M}(y)$ *for two, four and seven classes. The prior mean* $\mu = 55$ *and the data mean* $\bar{y}_{..} = 54.04$.

for our judgements to be coherent (see Sub-section 4.15.2). Thus our sample designs (J, n) comprise of $(2, 2), \ldots, (2, 23), (3, 2), \ldots, (3, 23), \ldots, (7, 2), \ldots, (7, 23)$ with the corresponding sample means $\bar{y}_{j.}$ which is here denoted as $\bar{y}_{j.}(J, n)$ to highlight the associated design. For example $\bar{y}_{j.}(2, 2)$ corresponds to the mean of the four observations in the design $J = 2, n = 2$. We sort both the examination scores within each class and classes (by their group means $\bar{y}_{j.}$) in ascending order so that the sample means $\bar{y}_{j.}(J, n)$ increase with n and J from design $(2, 2)$ through to design $(23, 7)$. Using $\bar{y}_{j.}(J, n)$ for each design we compute the finite and infinite adjusted population grand mean.

The results for two, four and seven classes are shown in Figure 4.6. The observed increasing trends in the graph are a consequence of the effects of the increasing sample means $\bar{y}_{j.}$ coupled with the increasing sample sizes (J, n) on the finite and infinite adjustments. Also, as $\bar{y}_{j.}$ increases the component $(\bar{y}_{j.} - \mu)$ (see (4.105)) shrinks, explaining the convergence of the finite and infinite curves.

We also note that when $J = 7$ and $n = 23$ then $E_{\bar{D}_n}(\mathcal{M}^{[G]}(y)) = \bar{y}_{..}$, the sample grand mean. This can be verified by setting $J = G$ and $n = N$ in (4.104). That is when we sample the whole of our finite population, the prior receives zero weight and the data, the maximum weight: the data swamp the prior. Sampling the whole finite population also pushes the infinite adjustment $E_{\bar{D}_n}(\mathcal{M}(y))$ very close to $\bar{y}_{..}$ but, unlike the finite case, the prior does retain some of the weight.

In conclusion Figure 4.7 reflects mostly changes in $(\bar{y}_{j.} - \mu)$ and these mask all the other components in (4.105) that influence the disparities between the finite and infinite adjustments. A further impediment of our example data here is that it is quite small; the level 2 population is only $G = 7$, and that hinders studying the effect of the level 2 sampling fraction.

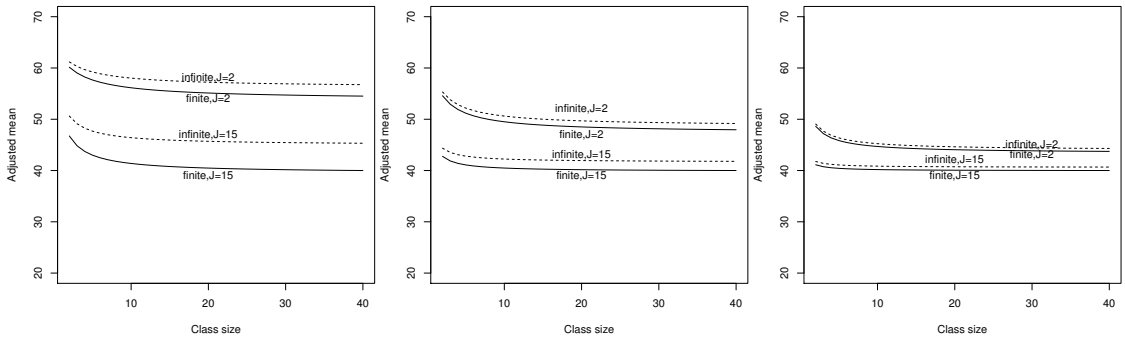
4.14.2 Conditions for ignoring the difference between the finite and infinite situations.

To overcome the above-mentioned limitations of our STAT1010 example, we shall consider a hypothetical finite population with $N = 40$ and $G = 15$, about twice the level 1 and 2 population sizes of the STAT1010 example. This gives us sufficient data to explore the conditions for which the finite/infinite issue may be ignored, that is they both hold, so that we may use the simpler infinite calculations. We put $\bar{y}_{j.} = 40$ and $\mu = 65$, thereby also fixing $(\bar{y}_{j.} - \mu)$, for all sample sizes, so that the finite and infinite curves are spaced sufficiently. This will allow us to focus on the effects on our adjustments of varying the level 1 and level 2 sampling fractions, as well as the prior uncertainty γ for the underlying population grand mean $\mathcal{M}(y)$. We set γ to 16, 56 and 156.

The resulting adjustments are shown in Figure 4.7. All the curves slope downward towards the data mean of 40 as more level 1 and 2 data are used in the adjustment. The level 2 sample size J has quite a large influence on the adjustment; in all three panels the curves with $J = 15$ are closer to the data mean of 40 than when $J = 2$. This is not the same for the level 1 sample size n where all the curves, whether for the finite or infinite adjustments, tend to remain rather flat as n increases from 2 to 40 for a given value of J. The effect of the prior uncertainty γ is also quite important on the adjustments; increasing γ reduces the difference between the finite and infinite adjustments. We summarize our findings as follows.

Conditions where the finite/infinite issue is ignorable:

- The level 2 sampling fraction J/G is small.
- The prior uncertainty γ is large.



Gamma=16 Gamma=56 Gamma=156

Figure 4.7: *Disparities between the finite and infinite adjustments of the population grand mean $\mathcal{M}(y)$ for the hypothetical data with prior mean $\mu = 65$ and data mean $\bar{y}_{..} = 40$, finite level 1 and 2 populations of $N = 40$ and $G = 15$ respectively. Adjustments are shown for level 1 samples of $n = 2$ to 40 and level 2 samples of $J = 2$ and 15 for $\gamma = 16, 56$ and 156. The dashed lines represent the adjustments of $\mathcal{M}(y)$ for the infinite SOEREF model.*

The smaller the level 2 sampling fraction, the closer will the finite and infinite adjustments be. However, a large prior uncertainty γ coupled with a small level 2 sampling fraction, will result in almost similar finite and infinite adjustments. The above two conditions can also be deduced from expression (4.105) that relates $E_{\bar{D}_n}(\mathcal{M}^{[G]}(y))$ and $E_{\bar{D}_n}(\mathcal{M}(y))$. The level 1 sampling fraction n/N has almost no bearing on the differences between the finite and infinite adjustments.

4.15 The adjusted variance of the population grand mean in the finite and the infinite SOEREF model

Below we state and prove a theorem for the adjusted variance for the finite population grand mean and we relate it to the infinite situation.

Theorem 4.15.1. *Consider the balanced two-level population consisting of a finite number G of groups and a finite number of N individuals. Using the representation and SOE specifications for the finite SOEREF model as in Theorem 4.13.1, the ad-*

justed variance of $\mathcal{M}^{[G]}(y)$ *given the collection of group means* $\bar{D}_n = \{\bar{y}_{1.}, \bar{y}_{2.}, \ldots, \bar{y}_{J.}\}$ *is*

$$Var_{\bar{D}_n}(\mathcal{M}^{[G]}(y)) = \frac{1}{G}(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{N} + G\gamma)\Big[1 - \frac{J(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{N} + G\gamma)}{G(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)}\Big] \tag{4.107}$$

Proof.

$$\begin{aligned}
Var_{\bar{D}_n}(\mathcal{M}^{[G]}(y)) &= Var(\mathcal{M}^{[G]}(y)) \\
&\quad - Cov(\mathcal{M}^{[G]}(y), \bar{D}_n) Var^{-1}(\bar{D}_n) Cov(\bar{D}_n, \mathcal{M}^{[G]}(y)) \\
&= \frac{1}{G}(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{N} + G\gamma) - \frac{1}{G^2}(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{N} + G\gamma)^2 \mathbf{1}_J^T \\
&\quad \frac{1}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n})}\Big[I_J - \frac{\gamma}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)} K_J\Big]\mathbf{1}_J \\
&= \frac{1}{G}(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{N} + G\gamma) - \frac{1}{G^2}(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{N} + G\gamma)^2 \\
&\quad \frac{1}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n})}\Big[1 - \frac{J\gamma}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)}\Big] J \\
&= \frac{1}{G}(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{N} + G\gamma) - \frac{1}{G^2}(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{N} + G\gamma)^2 \\
&\quad \frac{J}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)} \\
&= \frac{1}{G}(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{N} + G\gamma)\Big[1 - \frac{J(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{N} + G\gamma)}{G(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)}\Big],
\end{aligned}$$

where the expression for $Var(\mathcal{M}^{[G]}(y))$ comes from (4.94) and the covariance and inverse variance expressions are as in Theorem 4.14.1. ■

The adjusted finite variance depends on the level 1 and level 2 sample sizes (n, J) and finite populations sizes (N, G) as well as on our prior variance specifications. Since $n \leq N$ and $J \leq G$, and all variance quantities in (4.107) are non-negative, $J(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{N} + G\gamma)/G(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma) \leq 1$. Hence, we deduce that $Var_{\bar{D}_n}(\mathcal{M}^{[G]}(y)) \geq 0$, which is equivalent to the coherence condition that the variance-covariance matrix over our beliefs and data is non-negative definite (see Goldstein and Wooff (2007; page 67)). We note that $Var_{\bar{D}_n}(\mathcal{M}^{[G]}(y)) = 0$ when $n = N$ and $J = G$, which has the intuitive implication that when we sample all the finite level 1 and 2 populations there is no uncertainty left in $\mathcal{M}^{[G]}(y)$.

Irrespective of the size of the level 1 population N, as the finite level 2 population G becomes large, in the limit we have from (4.107)

$$\begin{aligned}\lim_{G\to\infty} Var_{\bar{D}_n}(\mathcal{M}^{[G]}(y)) &= \gamma - \frac{J\gamma^2}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)} \\ &= \frac{1}{\gamma^{-1} + J(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n})^{-1}}\end{aligned} \tag{4.108}$$

which is the same as the adjusted variance of the overall mean in the infinite balanced case.

However, if G is small and only N becomes large then we have

$$\lim_{N\to\infty} Var_{\bar{D}_n}(\mathcal{M}^{[G]}(y)) = \frac{1}{G}(\sigma_u^2 - \gamma + G\gamma)\Big[1 - \frac{J(\sigma_u^2 - \gamma + G\gamma)}{G(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)}\Big] \tag{4.109}$$

As mentioned before, it is more realistic for a hierarchy to comprise of a level 2 population with few groups (G) and a larger level 1 population N in each of these groups so that (4.109) may be more useful in practice than (4.108).

4.16 The finite and infinite resolution

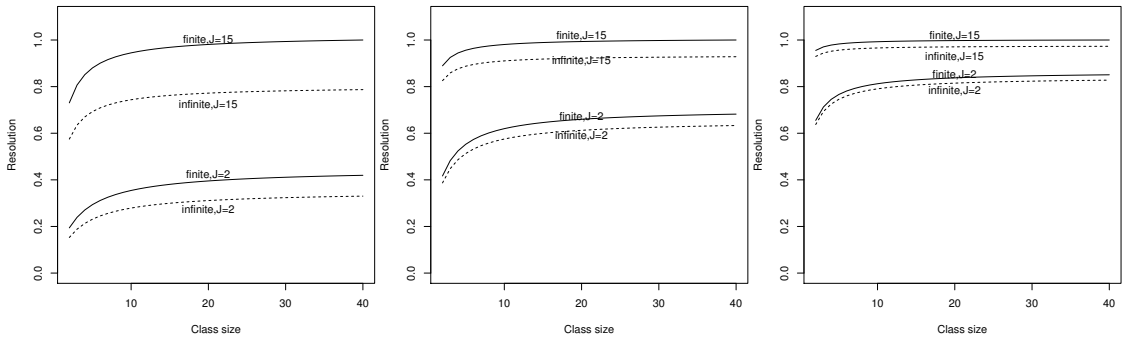
The resolution for adjusting the finite mean $\mathcal{M}^{[G]}(y)$ by $\bar{D}_n$ is given by

$$\begin{aligned}R_{\bar{D}_n}(\mathcal{M}^{[G]}(y)) &= 1 - \frac{Var_{\bar{D}_n}(\mathcal{M}^{[G]}(y))}{Var(\mathcal{M}^{[G]}(y))} \\ &= \frac{J(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{N} + G\gamma)}{G(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)}\end{aligned} \tag{4.110}$$

It is quite easily seen from (4.110) that $R_{\bar{D}_n}(\mathcal{M}^{[G]}(y))$ lies between zero and one. $R_{\bar{D}_n}(\mathcal{M}^{[G]}(y))$ will be zero when $Var_{\bar{D}_n}(\mathcal{M}^{[G]}(y)) = Var(\mathcal{M}^{[G]}(y))$. This will happen if either $\bar{D}_n$ is not informative for adjusting $\mathcal{M}^{[G]}(y)$ or our beliefs specification is not sufficiently detailed to exploit the information in $\bar{D}_n$. The resolution will be one if we sample all the level 1 and level 2 finite populations, i.e. $(n = N)$ and $(J = G)$, hence $\bar{D}_n$ will contain all the information required to adjust $\mathcal{M}^{[G]}(y)$, and all uncertainty will be resolved.

To show the connection between the finite resolution and its infinite counterpart, we rewrite (4.110) as the following sum of two terms

$$R_{\bar{D}_n}(\mathcal{M}^{[G]}(y)) = \frac{\frac{J}{G}(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{N})}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)} + \frac{J\gamma}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)} \tag{4.111}$$



Gamma=16 Gamma=56 Gamma=156

Figure 4.8: *Disparities in the proportion of uncertainty resolved in the finite and infinite adjustments of the population grand mean $\mathcal{M}(y)$ for the hypothetical data. The resolutions are shown for level 1 samples of $n = 2$ to 40 and level 2 samples of $J = 2$ and 15 for $\gamma = 16, 56$ and 156.*

The first term in (4.111) is always positive, while the second term is the resolution of the overall mean in the infinite balanced case, i.e. $R_{\bar{D}_n}(\mathcal{M}(y))$. Thus the finite resolution $R_{\bar{D}_n}(\mathcal{M}^{[G]}(y))$ is larger than $R_{\bar{D}_n}(\mathcal{M}(y))$.

This difference between the two resolutions is also evident in Figure 4.8. The latter shows the disparity between the finite and infinite resolutions for our hypothetical data set with finite level 1 population $N = 40$ students, from which successive samples of $n = 2, 3, \ldots, 40$ are taken, and level 2 population $G = 15$ classes, from which only two extreme samples $J = 2$ and $J = 15$ are considered. As before, three levels of uncertainty $\gamma = 16, 56, 156$ in $\mathcal{M}(y)$ are considered. As γ increases the difference between $R_{\bar{D}_n}(\mathcal{M}^{[G]}(y))$ and $R_{\bar{D}_n}(\mathcal{M}(y))$ decreases for both $J = 2$ and $J = 15$.

The differences between the resolutions are influenced to a large extent by the level 2 sampling fraction J/G. For example, $R_{\bar{D}_n}(\mathcal{M}^{[G]}(y))$ converges to $R_{\bar{D}_n}(\mathcal{M}(y))$ when the level 2 sampling fraction is decreased from $J/G = 15/15$ to $J/G = 2/15$ (see first panel of Figure 4.9). This convergence effect can also be demonstrated by making the finite level 2 population size G large while keeping the level 2 sample size J fixed. Irrespective of the size of the level 1 population N, as G becomes large,

in the limit we have from (4.111)

$$\lim_{G\to\infty} R_{\bar{D}_n}(\mathcal{M}^{[G]}(y)) = \frac{J\gamma}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)} \tag{4.112}$$

which is equal to $R_{\bar{D}_n}(\mathcal{M}(y))$, the resolution of the overall mean in the infinite balanced case.

Figure 4.9 reveals that for a level 1 sampling fraction $n/N = 10/40$ or more, the difference between $R_{\bar{D}_n}(\mathcal{M}^{[G]}(y))$ and $R_{\bar{D}_n}(\mathcal{M}(y))$ remains more or less the same for all values of J and γ. In fact, this discrepancy between the two resolutions persists even if we allow the finite level 1 population N in $R_{\bar{D}_n}(\mathcal{M}^{[G]}(y))$ to become large while keeping the level 1 sample size n and the level 2 population size G fixed as follows.

$$\lim_{N\to\infty} R_{\bar{D}_n}(\mathcal{M}^{[G]}(y)) = \frac{\frac{J}{G}(\sigma_u^2 - \gamma)}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)} + \frac{J\gamma}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)}. \tag{4.113}$$

The second term on the right of (4.113) is the infinite resolution $R_{\bar{D}_n}(\mathcal{M}(y))$, so that the difference between the finite resolution as $N \to \infty$ and $R_{\bar{D}_n}(\mathcal{M}(y))$ is given by the first term of the sum in (4.113).

4.17 The finite adjustments of level 2 quantities

In order to update the population group j mean $\mathcal{M}^{[N]}(y_j)$, for $j = 1, 2, \ldots, G$ in the balanced finite SOEREF model, we make use of the level 2 representation $\mathcal{M}^{[N]}(y_j) = \mathcal{M}^{[G]}(y) + \mathcal{R}_j^{[G]}(\mathcal{M}^{[N]}(y))$. Since we have already updated the population grand mean $\mathcal{M}^{[G]}(y)$, we only need to update the level 2 residual $\mathcal{R}_j^{[G]}(\mathcal{M}^{[N]}(y))$ using the sample group means $\bar{D}_n = \{\bar{y}_{1.}, \bar{y}_{2.}, \ldots, \bar{y}_{J.}\}$.

Theorem 4.17.1. *The adjustment of each level 2 residual $\mathcal{R}_j^{[G]}(\mathcal{M}^{[N]}(y))$ of the balanced finite SOEREF model by the collection of sample group means $\bar{D}_n = \{\bar{y}_{1.}, \bar{y}_{2.}, \ldots, \bar{y}_{J.}\}$ is*

$$E_{\bar{D}_n}(\mathcal{R}_j^{[G]}(\mathcal{M}^{[N]}(y))) = \frac{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{N})}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n})}\left[(\bar{y}_{j.} - \mu) - \frac{J(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + G\gamma)(\bar{y}_{..} - \mu)}{G(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)}\right] \tag{4.114}$$

Proof. Let $\mathcal{C}_{R_j}$ denote the vector of level 2 residuals $\{\mathcal{R}_1^{[G]}(\mathcal{M}^{[N]}(y)),$ $\mathcal{R}_2^{[G]}(\mathcal{M}^{[N]}(y)), \ldots \mathcal{R}_J^{[G]}(\mathcal{M}^{[N]}(y))\}$ in the finite SOEREF model. Using the SOE relationships from Theorem 4.15.1, and in particular

$$E(\mathcal{R}_j^{[G]}(\mathcal{M}^{[N]}(y))) = 0, \quad \forall j$$

$$Cov(\mathcal{R}_j^{[G]}(\mathcal{M}^{[N]}(y)), \mathcal{R}_{j'}^{[G]}(\mathcal{M}^{[N]}(y))) = \begin{cases} \frac{G-1}{G}((\sigma_u^2 + \frac{1}{N}\sigma_\epsilon^2) - \gamma) \text{ if } j = j' \\ -\frac{1}{G}((\sigma_u^2 + \frac{1}{N}\sigma_\epsilon^2) - \gamma) \text{ otherwise.} \end{cases}$$

we calculate the required second order quantities over $(\mathcal{C}_{R_j}, \bar{D}_n)$. We have $E(\mathcal{C}_{R_j}) = \mathbf{1}_J 0$ and $Cov(\mathcal{C}_{R_j}, \bar{D}_n) = (\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{N})(\mathbf{I}_J - \frac{1}{G}\mathbf{K}_J)$ where $\mathbf{1}_J$ is a column of J ones, $\mathbf{I}_J$ is an identity matrix of dimension J and $\mathbf{K}_J$ is a $J \times J$ matrix of ones. Applying the Bayes linear rule gives the adjusted vector of level 2 residuals as follows.

$$\begin{aligned}
E_{\bar{D}_n}(\mathcal{C}_{R_j}) &= E(\mathcal{C}_{R_j}) + Cov(\mathcal{C}_{R_j}, \bar{D}_n) Var^{-1}(\bar{D}_n)(\bar{D}_n - E(\bar{D}_n)) \\
&= \mathbf{1}_J 0 + (\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{N})(\mathbf{I}_J - \frac{1}{G}\mathbf{K}_J) \\
&\frac{1}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n})}\left[\mathbf{I}_J - \frac{\gamma}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)}\mathbf{K}_J\right](\bar{D}_n - \mathbf{1}_J\mu) \\
&= \frac{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{N})}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n})}(\mathbf{I}_J - \frac{1}{G}\mathbf{K}_J)\left[\mathbf{I}_J - \frac{\gamma}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)}\mathbf{K}_J\right] \\
&(\bar{D}_n - \mathbf{1}_J\mu) \\
&= \frac{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{N})}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n})}\left[\mathbf{I}_J - \frac{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + G\gamma)}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)}\mathbf{K}_J\right](\bar{D}_n - \mathbf{1}_J\mu) \\
&= \frac{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{N})}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n})}\begin{pmatrix} 1-f & -f & \cdots & -f \\ -f & 1-f & \cdots & -f \\ \vdots & \vdots & \ddots & \vdots \\ -f & -f & \cdots & 1-f \end{pmatrix}\begin{pmatrix} (\bar{y}_{1.} - \mu) \\ (\bar{y}_{2.} - \mu) \\ \vdots \\ (\bar{y}_{J.} - \mu) \end{pmatrix} \\
&= \frac{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{N})}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n})}\begin{pmatrix} (\bar{y}_{1.} - \mu) - f\sum_{j=1}^J(\bar{y}_{j.} - \mu) \\ (\bar{y}_{2.} - \mu) - f\sum_{j=1}^J(\bar{y}_{j.} - \mu) \\ \vdots \\ (\bar{y}_{J.} - \mu) - f\sum_{j=1}^J(\bar{y}_{j.} - \mu) \end{pmatrix} \qquad (4.115)
\end{aligned}$$

where $f = \frac{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + G\gamma)}{G(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)}$. Setting $\sum_{j=1}^J(\bar{y}_{j.} - \mu) = J(\bar{y}_{..} - \mu)$ in the jth row of

(4.115) gives (4.114). ■

We note that as $G, N \to \infty$, $\mathcal{R}_j^{[G]}(\mathcal{M}^{[N]}(y)) \to \mathcal{R}_j(\mathcal{M}(y))$ as in Theorem 4.4.1 for the balanced SOEREF design.

Using the finite level 2 representation, we calculate the adjusted finite population group j mean $E_{\bar{D}_n}(\mathcal{M}^{[N]}(y_j))$ by adding $E_{\bar{D}_n}(\mathcal{M}^{[G]}(y))$ (4.106) and $E_{\bar{D}_n}(\mathcal{R}_j^{[G]}(\mathcal{M}^{[N]}(y)))$ (4.116). However, unlike the finite case, it is not straightforward to interpret $E_{\bar{D}_n}(\mathcal{M}^{[N]}(y_j))$ as it is the sum of two complex adjusted quantities.

4.17.1 The finite adjusted variance of $\mathcal{R}_j^{[G]}(\mathcal{M}^{[N]}(y))$

After calculating the adjusted mean of the level 2 residual $\mathcal{R}_j^{[G]}(\mathcal{M}^{[N]}(y))$, we now consider the adjustment of the variance and covariances of the collection of level 2 residuals $\mathcal{C}_{R_j}$. We may then compare our results with the infinite adjustments derived in Corollary 4.4.3 for the balanced SOEREF model.

Theorem 4.17.2. *The adjusted variances and covariances of the collection of level 2 residuals* $\mathcal{C}_{R_j} = \{\mathcal{R}_1^{[G]}(\mathcal{M}^{[N]}(y)), \mathcal{R}_2^{[G]}(\mathcal{M}^{[N]}(y)), \ldots \mathcal{R}_J^{[G]}(\mathcal{M}^{[N]}(y))\}$ *in the balanced finite SOEREF model by the collection of sample group means* $\bar{D}_n = \{\bar{y}_{1.}, \bar{y}_{2.}, \ldots, \bar{y}_{J.}\}$ *are obtained respectively from the diagonal and off-diagonal elements of*

$$Var_{\bar{D}_n}(\mathcal{C}_{R_j}) = \frac{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{N})}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n})} \Bigg\{ \frac{(N-n)}{N} \frac{\sigma_\epsilon^2}{n} (\mathbf{I}_J - \frac{1}{G}\mathbf{K}_J) + \frac{(G-J)}{G^2} \frac{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{N})(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + G\gamma)}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)} \mathbf{K}_J \Bigg\} \tag{4.116}$$

Proof. First, we note that $(\mathbf{I}_J - a\mathbf{K}_J)(\mathbf{I}_J - b\mathbf{K}_J) = (\mathbf{I}_J - (a + b - Jab)\mathbf{K}_J)$.

$$\begin{aligned} Var_{\bar{D}_n}(\mathcal{C}_{R_j}) &= Var(\mathcal{C}_{R_j}) - Cov(\mathcal{C}_{R_j}, \bar{D}_n) Var^{-1}(\bar{D}_n) Cov(\bar{D}_n, \mathcal{C}_{R_j}) \\ &= (\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{N})(\mathbf{I}_J - \frac{1}{G}\mathbf{K}_J) - (\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{N})(\mathbf{I}_J - \frac{1}{G}\mathbf{K}_J) \\ &\quad \frac{1}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n})} \left[\mathbf{I}_J - \frac{\gamma}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)} \mathbf{K}_J \right] (\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{N}) \\ &\quad (\mathbf{I}_J - \frac{1}{G}\mathbf{K}_J), \end{aligned}$$

All the matrices above are of the form $(\mathbf{I}_J - a\mathbf{K}_J)$ and their multiplications result in the same symmetric matrix form, simplification of which results in (4.116). ■

It is straightforward to verify that as $N, G \to \infty$, the variance and covariance terms in (4.116) are the same as the corresponding infinitely adjusted quantities that we derived in (4.30) of Corollary 4.4.3. As we mentioned in Corollary 4.4.3, calculating and interpreting adjusted level 2 quantities individually is not that straightforward, more so since these quantities as in $\mathcal{C}_{R_j}$ are correlated. We shall therefore proceed to analyze and interpret overall changes in beliefs over the collection $\mathcal{C}_{R_j}$ via a canonical analysis as we did for our infinite adjustments of group level quantities.

4.18 Canonical analysis for the adjustment of the finite population group means

We shall now analyze and interpret overall changes in beliefs over the collection of finite population group means $\mathcal{M}^{[N]}(y_j)$ using a canonical analysis. The motivations for such an analysis are as explained in Section 4.5 for the adjustments of infinite populations. In addition, here we are also interested in comparing the canonical analysis of the finite and infinite cases. We begin by calculating the resolution transform matrix which has a central role in Bayes linear statistics, (Goldstein and Wooff, 2007).

4.18.1 The resolution transform matrix for the adjustment of $\mathcal{M}^{[N]}(y_j)$

First, let $\mathcal{C}_{\mathcal{M}^{[N]}} = \{\mathcal{M}^{[N]}(y_1), \mathcal{M}^{[N]}(y_2), \ldots, \mathcal{M}^{[N]}(y_J)\}$ denote the collection of finite population group means. In accordance with the notation for the resolution transform matrix in Section 4.5, viz. $\mathbb{T}_{B:D}$, we have for the finite case, $\mathbb{T}_{\mathcal{C}_{\mathcal{M}^{[N]}}:\bar{D}_n}$ which we write as $\mathbb{T}_n^{[N]}$ for simplicity.

Theorem 4.18.1. *The resolution transform matrix for the adjustment of the collection of finite population group j means $\mathcal{C}_{\mathcal{M}^{[N]}} = \{\mathcal{M}^{[N]}(y_1), \mathcal{M}^{[N]}(y_2),$ $\ldots, \mathcal{M}^{[N]}(y_J)\}$ in the balanced SOEREF model, by the collection of sample group*

means $\bar{D}_n = \{\bar{y}_{1.}, \bar{y}_{2.}, \ldots, \bar{y}_{J.}\}$ *is*

$$\mathbb{T}_n^{[N]} = \frac{1}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n})}\left[(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{N})\mathbf{I}_J + \frac{\gamma\frac{(N-n)}{N}\frac{\sigma_\epsilon^2}{n}}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)}\mathbf{K}_J\right] \tag{4.117}$$

Proof. From our second order specifications for the finite SOEREF model in Theorem 4.13.1 we obtained $Var(\mathcal{M}^{[N]}(y_j))=\sigma_u^2 + \frac{1}{N}(\sigma_y^2 - \sigma_u^2)=\sigma_u^2 + \frac{1}{N}\sigma_\epsilon^2$, $\forall j$, and $Cov(\mathcal{M}^{[N]}(y_j), \mathcal{M}^{[N]}(y_{j'}))=\gamma$, $\forall j \neq j'$. Hence,

$$Var(\mathcal{C}_{\mathcal{M}^{[N]}}) = (\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{N})\mathbf{I}_J + \gamma\mathbf{K}_J \tag{4.118}$$

Since, $Cov(\mathcal{M}^{[N]}(y_j), \mathcal{R}_i^{[N]}(y_j)) = 0$, we have

$$Cov(\mathcal{C}_{\mathcal{M}^{[N]}}, \bar{D}_n) = Var(\mathcal{C}_{\mathcal{M}^{[N]}}) \tag{4.119}$$

Using the above, the resolution transform matrix is

$$\begin{aligned}
\mathbb{T}_n^{[N]} &= Var(\mathcal{C}_{\mathcal{M}^{[N]}})^{-1}Cov(\mathcal{C}_{\mathcal{M}^{[N]}}, \bar{D}_n)Var(\bar{D}_n)^{-1}Cov(\bar{D}_n, \mathcal{C}_{\mathcal{M}^{[N]}}) \\
&= Var(\bar{D}_n)^{-1}Cov(\bar{D}_n, \mathcal{C}_{\mathcal{M}^{[N]}}) \\
&= \frac{1}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n})}\left[\mathbf{I}_J - \frac{\gamma}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)}\mathbf{K}_J\right]\left[(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{N})\mathbf{I}_J + \gamma\mathbf{K}_J\right] \\
&= \frac{1}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n})}\Bigg[(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{N})\mathbf{I}_J \\
&+ \left(\gamma - \frac{\gamma(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{N})}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)} - \frac{J\gamma^2}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)}\right)\mathbf{K}_J\Bigg].
\end{aligned}$$

On simplifying the multiplier term of $\mathbf{K}_J$ between brackets, we obtain (4.117). ■

We note that as $N \to \infty$, $\mathbb{T}_n^{[N]}$ in (4.117) above is equal to the resolution transform matrix $\mathbb{T}_n$ in (4.35) for the infinite population case. Also, if we sample the whole level 1 population, i.e., $n = N$, then $\mathbb{T}_n^{[N]} = \mathbf{I}_J$, that is the resolution transform matrix becomes the identity matrix, implying that all the uncertainties in the collection of linear combinations of the finite group means $\langle \mathcal{M}^{[N]}(y_j) \rangle$ will be resolved. This is to be expected since we have observed the whole finite level 1 populations.

Based on $\mathbb{T}_n^{[N]}$, we shall now calculate the canonical resolutions (eigenvalues) and their associated canonical directions (eigenvectors) that will be useful in achieving a better understanding of the magnitude and type of information gained by observing our multilevel data.

4.18.2 The canonical resolutions

The canonical resolutions, i.e. the ordered eigenvalues of $\mathbb{T}_n^{[N]}$, are easily obtained from $\mathbb{T}_n^{[N]}$ since it is of the form $(a\mathbf{I}_J + b\mathbf{K}_J)$ which has eigenvalues a, with multiplicity $n-1$, and $a + nb$ (see Section 4.5.2).

Corollary 4.18.1. *Let $\lambda_1^{[N]}$ and $\lambda_2^{[N]}$ be the two distinct eigenvalues of $\mathbb{T}_n^{[N]}$, where $\lambda_1^{[N]}$ is the largest eigenvalue and $\lambda_2^{[N]}$ is the smallest eigenvalue with multiplicity $(J-1)$. Then*

$$\lambda_1^{[N]} = \frac{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{N} + J\gamma)}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)} \tag{4.120}$$

$$\lambda_2^{[N]} = \frac{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{N})}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n})}, \tag{4.121}$$

where $n \leq N, J \leq G$.

We use the special form $(a\mathbf{I}_J + b\mathbf{K}_J)$ of $\mathbb{T}_n^{[N]}$ to prove Corollary 4.18.1.

Proof. The coefficient of $\mathbf{I}$ plus J times the coefficient of $\mathbf{K}_J$ for $\mathbb{T}_n^{[N]}$ in (4.119) gives

the largest eigenvalue as follows.

$$\lambda_1^{[N]} = \frac{1}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n})}\left[(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{N}) + \frac{J\gamma\frac{(N-n)}{N}\frac{\sigma_\epsilon^2}{n}}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)}\right],$$

$$= \frac{1}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n})}\left[\frac{N(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{N})(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma) + J\gamma(N-n)\frac{\sigma_\epsilon^2}{n}}{N(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)}\right],$$

$$= \frac{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{N} + J\gamma)}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n} + J\gamma)}$$

The coefficient of $\mathbf{I}$ for $\mathbb{T}_n^{[N]}$ in (4.117) gives the smallest eigenvalue as follows.

$$\lambda_2^{[N]} = \frac{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{N})}{(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n})},$$

■

First we note that as $N \to \infty$, $\lambda_1^{[N]} \to \lambda_1$ and $\lambda_2^{[N]} \to \lambda_2$, their infinite counterparts. Both $\lambda_1^{[N]}$ and $\lambda_2^{[N]}$ depend on the level 1 sample and population sizes, n and N respectively. As $n \to N$, $\lambda_j^{[N]} \to 1$, for each j. Hence, the more we increase the level 1 sample size, the more the uncertainty we expect to resolve in each of the two directions of the corresponding components in $\langle \mathcal{M}^{[N]}(y_j) \rangle$ by observing $\bar{D}_n$.

We need to verify that $\lambda_1^{[N]} \geq \lambda_2^{[N]}$. Using expressions (4.120) and (4.121), we obtain $\lambda_1^{[N]} \geq \lambda_2^{[N]}$ if $n \leq N$, which is a trivial condition. Also, $\lambda_1^{[N]} = \lambda_2^{[N]}$ when $\gamma = 0$, that is when there is no uncertainty about the population overall mean $\mathcal{M}^{[G]}(y)$.

We also need to ensure that the largest possible reduction in variance is one, that is $\lambda_j^{[N]} \leq 1$, for each j. Both $\lambda_1^{[N]}$ and $\lambda_2^{[N]}$ attain their maximum values when their denominators are minimum, and this occurs when n is maximum. Putting $n = N$ in (4.120) and (4.121) we see that, for each j, $\lambda_j^{[N]}$ attain its maximum value of one as required. In other words when we sample the whole level 1 finite population, i.e. $n = N$, we resolve all the uncertainty about $\langle \mathcal{M}^{[N]}(y_j) \rangle$ by observing $\bar{D}_n$ as we would expect.

4.18.3 The eigenstructure of $\mathbb{T}_n^{[N]}$

We now show that the SOE samples selected from finite and infinite multilevel populations have similar coherence relationships.

Theorem 4.18.2. *The eigenvectors of $\mathbb{T}_n^{[N]}$ are the same for each n. If $Y^{[N]}$ is an eigenvector of $\mathbb{T}_1^{[N]}$ with corresponding eigenvalue $\lambda^{[N]}$, then the corresponding eigenvalue $\lambda_n^{[N]}$ for $\mathbb{T}_n^{[N]}$ is*

$$\lambda_{(n)}^{[N]} = \frac{n(N-1)\lambda^{[N]}}{(n-1)N\lambda^{[N]} + (N-n)} \tag{4.122}$$

Proof. The resolution transform matrix $\mathbb{T}_n^{[N]}$ has an eigenvector v corresponding to the eigenvalue $\lambda^{[N]}$ if

$$\mathbb{T}_n^{[N]} v = \lambda^{[N]} v$$

Substituting $\mathbb{T}_n^{[N]} = Var(\bar{D}_n)^{-1} Cov(\bar{D}_n, \mathcal{C}_{\mathcal{M}^{[N]}})$ from the proof of Theorem 4.18.1, we have

$$Var(\bar{D}_n)^{-1} Cov(\bar{D}_n, \mathcal{C}_{\mathcal{M}^{[N]}}) v = \lambda^{[N]} v$$

Multiplying both sides of the above equation by $Var(\bar{D}_n)$

$$Cov(\bar{D}_n, \mathcal{C}_{\mathcal{M}^{[N]}}) v = \lambda^{[N]} Var(\bar{D}_n) v$$

Substituting expressions for $Cov(\bar{D}_n, \mathcal{C}_{\mathcal{M}^{[N}})$ and $Var(\bar{D}_n)$ (see proof of Theorem 4.18.1), we obtain

$$\begin{aligned}
[(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{N})I_J + \gamma K_J]v &= \lambda^{[N]}[(\sigma_u^2 - \gamma + \frac{\sigma_\epsilon^2}{n})I_J + \gamma K_J]v \\
&= \lambda^{[N]}[(\sigma_u^2 - \gamma)I_J + \gamma K_J]v + \lambda^{[N]}\frac{\sigma_\epsilon^2}{n} I_J v \\
[(\sigma_u^2 - \gamma)I_J + \gamma K_J]v + \frac{\sigma_\epsilon^2}{N} I_J v &= \lambda^{[N]}[(\sigma_u^2 - \gamma)I_J + \gamma K_J]v + \lambda^{[N]}\frac{\sigma_\epsilon^2}{n} I_J v \\
[(\sigma_u^2 - \gamma)I_J + \gamma K_J](1 - \lambda^{[N]})v &= (\frac{\lambda^{[N]}}{n} - \frac{1}{N})\sigma_\epsilon^2 I_J v \\
[(\sigma_u^2 - \gamma)I_J + \gamma K_J]v &= \frac{(\frac{\lambda^{[N]}}{n} - \frac{1}{N})}{(1 - \lambda^{[N]})}\sigma_\epsilon^2 I_J v
\end{aligned} \tag{4.123}$$

Putting $n = 1$ in (4.123),

$$[(\sigma_u^2 - \gamma)I_J + \gamma K_J]v = \frac{(\lambda^{[N]} - \frac{1}{N})}{(1 - \lambda^{[N]})}\sigma_\epsilon^2 I_J v \tag{4.124}$$

Equating the right-hand sides of (4.123) and (4.124) gives

$$\frac{\lambda^{[N]} - \frac{1}{N}}{(1 - \lambda^{[N]})} = \frac{\frac{\lambda_{(n)}^{[N]}}{n} - \frac{1}{N}}{(1 - \lambda_{(n)}^{[N]})},$$

which, after solving for $\lambda_{(n)}^{[N]}$, gives expression (4.124). ■

Theorem 4.18.2 shows that the canonical directions for the adjustment of $\langle \mathcal{M}^N(y_j) \rangle$ by $\bar{D}_n$ are the same for each n as was the case for our infinite adjustments in Theorem 4.10.2.

Further, dividing the numerator and denominator of (4.122) by N, we obtain

$$\lambda_{(n)}^{[N]} = \frac{(n - \frac{n}{N})\lambda^{[N]}}{(n-1)\lambda^{[N]} + (1 - \frac{n}{N})}. \tag{4.125}$$

It is clear that as N becomes large or if the sampling fraction $\frac{n}{N}$ is small, then the eigenvalue $\lambda_{(n)}^{[N]}$ is the same as the corresponding eigenvalue $\lambda_{(n)}$ for infinite sampling in Theorem 4.10.2

4.19 Finite adjustment of population group means $\mathcal{M}^{[N]}(y_j)$ for the STAT1010 data

To apply our results of the preceding sections to the STAT1010 data, we shall consider a balanced subset of the original data. We sort the classes (groups) in increasing order of group means, see column (2) of Table 4.5. We sample $n = 23$ students from each class which, for simplicity, we assume to have a finite population of N=30 students each. There are J=7 classes where the first three C1 to C3 are from the Faculty of Law and Management,and the remaining four classes are from the Faculty of Engineering. We shall use the same prior specifications as before, namely $E(\mathcal{M}(y)) = 55$, $Var(\mathcal{M}(y)) = 56.3$, $Var(\mathcal{R}_i(y_j)) = 237$, and $Var(\mathcal{R}_j(\mathcal{M}(y))) = 59 \quad \forall i, j$.

Table 4.5 shows the results of our finite and infinite adjustments for the balanced data. The prior expectation and variance are the same for all seven classes. The finite adjusted expectations (column 3) for each class are shrunk towards the prior expectation of 55%. The changes in expectation relative to the resolved variance

in the respective group means are shown in column 4. For example, for the first class, the prior and adjusted expectations are 55% and 43.598% respectively, and the change in expectation relative to the resolved variance in $\mathcal{M}(y_1)$ is -1.073 calculated as follows

$$S(E_{\bar{D}_n}(\mathcal{M}(y_1))) = \frac{E_{\bar{D}_n}(\mathcal{M}(y_1)) - E(\mathcal{M}(y_1))}{\sqrt{RVar_{\bar{D}_n}(\mathcal{M}(y_1))}} = -1.073.$$

The changes in standardized expectations for the seven classes range from -1.073 to 1.395, that is there is no surprising change as standardized expectations should have expectation zero and variance unity. The finite adjustments differ to some extent from the corresponding infinite adjustments since quite a large proportion of the finite population was sampled (i.e. n=23 out of N=30).

The change in variance for each $\mathcal{M}(y_j)$ from prior to finitely adjusted is 115.3 - 2.331 = 112.969, so that 98% of the prior variance is resolved, which is larger compared to the infinite resolution of 92.2%.

Although the changes in expectation show no cause for concern, the pattern in these changes indicate that variation in STAT1010 marks are associated with faculty. All the changes (column 4) for Faculty of Law & Management ($\mathcal{M}(y_1)$ to $\mathcal{M}(y_3)$) are negative while those for Faculty of Engineering ($\mathcal{M}(y_4)$ to $\mathcal{M}(y_7)$ are positive. Students from the Faculty of Engineering require good A level maths and thus perform better in STAT1010 compared to students from the Faculty of Law &Management. A SOEREG model will be more suitable to account for these differences.

It is of interest to compare the quality of the infinite population size approximation to the finite population. We randomly select a smaller sample of $n = 10$ students only from each of the 7 classes and compare the finite and infinite adjustments of the faculty mean score. The results are in Table 4.6 below. There are very little differences between the infinite and finite adjusted means and also in their associated changes in expectations relative to the resolved variances. Hence, when the sampling fraction is small compared to the population (here n=10 out of N=30), we may use the simpler infinite adjustments.

		Expectation			
Element	Class means	Finite Adjustment	Change	Infinite Adjustment	Change
(1)	(2)	(3)	(4)	(5)	(6)
$\mathcal{M}(y_1)$	43.217	43.598	-1.073	44.848	-0.984
$\mathcal{M}(y_2)$	44.652	44.983	-0.942	46.070	-0.866
$\mathcal{M}(y_3)$	46.087	46.368	-0.812	47.291	-0.747
$\mathcal{M}(y_5)$	56.826	56.734	0.163	56.434	0.139
$\mathcal{M}(y_7)$	58.087	57.952	0.278	57.507	0.243
$\mathcal{M}(y_6)$	59.043	58.875	0.364	58.321	0.322
$\mathcal{M}(y_4)$	70.391	69.829	1.395	67.982	1.259
Prior expectation	55.0	Prior variance	115.3		
Finite adjusted variance	2.331	Finite Resolution	98.0%		
Infinite adjusted variance	8.958	Infinite Resolution	92.2%		

Table 4.5: *Finite and infinite adjustment of group j means $\mathcal{M}(y_j)$ in the SOEREF model using a balanced sample of the STAT1010 data.*

		Expectation			
Element	Class means	Finite Adjustment	Change	Infinite Adjustment	Change
(1)	(2)	(3)	(4)	(5)	(6)
$\mathcal{M}(y_1)$	41.200	41.641	-1.257	43.089	-1.155
$\mathcal{M}(y_2)$	40.500	40.965	-1.320	42.493	-1.213
$\mathcal{M}(y_3)$	52.900	52.935	-0.194	53.050	-0.189
$\mathcal{M}(y_5)$	57.100	56.989	0.187	56.625	0.158
$\mathcal{M}(y_7)$	52.200	52.259	-0.258	52.454	-0.247
$\mathcal{M}(y_6)$	60.100	59.885	0.460	59.179	0.405
$\mathcal{M}(y_4)$	72.000	71.372	1.540	69.310	1.388
Prior expectation	55.0	Prior variance	115.3		
Finite adjusted variance	2.331	Finite Resolution	98.0%		
Infinite adjusted variance	8.958	Infinite Resolution	92.2%		

Table 4.6: *Finite and infinite adjustment of group j means $\mathcal{M}(y_j)$ in the SOEREF model using a balanced but small sample of 10 classes of the STAT1010 data.*

Chapter 5

Bayes linear estimation of the level-1 variance

In Chapter 4 we derived both infinite and finite adjustments of beliefs about the population overall and population group means in the SOEREF model. We showed that all these adjustments and the associated adjusted quantities, such as the resolutions for example, depended on the prior level 1 and 2 variances. However, we did not learn about the population variances using the available data.

Learning about population variances is somewhat more complex when compared to the estimation of mean components as explained in Section 2.14. In Section 2.14.3, we followed Goldstein & Woof (2007, pg. 265) to carry out a Bayes linear adjustment of the variance of a sequence of exchangeable random quantities. We now wish to extend Bayes linear methods to learn about population variances in multilevel models.

We start with the simplest multilevel model, namely the SOEREF model which has two variance components, one for each level of the hierarchy. Our interest centers on the level 2 variance component which can be difficult to estimate, particularly when data is scarce or when the true population variance is close to zero, in which cases the estimate of the level 2 variance could even be negative. We shall consider adjustment of level-2 variances in Chapter 6.

Below we consider adjustment of the level-1 variance for both the balanced and unbalanced cases.

5.1 Adjusting the level-1 variance - balanced situation

First, we revisit the SOEREF model. In Chapter 3 we used SOE judgements at levels 1 and 2 of the hierarchy to derive the SOEREF model:

$$y_{ji} = \mathcal{M}(y) + \mathcal{R}_j(\mathcal{M}(y)) + \mathcal{R}_i(y_j),$$
$$i = 1, 2, \ldots, n_j, \quad \text{and} \quad j = 1, 2, \ldots, J.$$

In the balanced case $n_j = n$. For assessing the level-1 variance, we reiterate our SOE judgments of the level-1 residuals.

We assume the level-1 residuals $\mathcal{R}_i(y_j)$ are second-order exchangeable over individuals for each group j and write $\epsilon_{ji} = \mathcal{R}_i(y_j)$ for individual i in group j with:

$$E(\mathcal{R}_i(y_j)) = 0, \qquad Var(\mathcal{R}_i(y_j)) = \sigma^2_\epsilon \qquad \forall i, j,$$

$$Cov(\mathcal{R}_i(y_j), \mathcal{R}_{i'}(y_j)) = 0 \qquad \forall i \neq i'.$$

where the level 1 variance $\sigma^2_\epsilon > 0$ and constant for all individuals and groups. Also, for all i and j, $\mathcal{R}_i(y_j)$ is uncorrelated with the population grand mean $\mathcal{M}(y)$.

In order to learn about the level-1 population variance, we need to construct a representation for the corresponding quantity as follows. If we assume the sequence ϵ^2_{ji} is second order exchangeable then we have the decomposition

$$\epsilon^2_{ji} = V_{\epsilon_{ji}} = \mathcal{M}(V_\epsilon) + \mathcal{R}_{ji}(V_\epsilon) \tag{5.1}$$

Since $E[\epsilon_{ji}] = E[\mathcal{R}_i(y_j)] = 0$ for all i and j, $E[\epsilon^2_{ji}]$ is the variance of $\mathcal{R}_i(y_j)$. In line with Goldstein & Woof (2007, pg. 265), we write $Var[\mathcal{R}_i(y_j)] = V_{R_\epsilon}$. Hence, E($\mathcal{M}(V_\epsilon)$)=$V_{R_\epsilon}$. The sequence $\mathcal{R}_{11}(V_\epsilon), \mathcal{R}_{12}(V_\epsilon), \ldots$ is uncorrelated with mean zero and constant variance $V_{R(V_\epsilon)}$. Also, each element $\mathcal{R}_{ji}(V_\epsilon)$ is uncorrelated with $\mathcal{M}(V_\epsilon)$. The level-1 population variance is denoted by $\mathcal{M}(V_\epsilon)$ while the variance of $\mathcal{M}(V_\epsilon)$ is denoted by V_{M_ϵ}. For simplicity, below we shall write $\mathcal{R}_{ji}(y)$ in place of $\mathcal{R}_i(y_j)$.

Based on representation (5.1), updating the level-1 population variance is equivalent to updating a mean component, here $\mathcal{M}(V_\epsilon)$, using a suitable statistic. One such statistic is the ANOVA estimator of the level-1 variance (see Section 2.7 for the

properties of this estimator), that is the mean squared error (MSE) which we denote by $\hat{\sigma}^2_\epsilon$. Thus we shall use the decomposition $\hat{\sigma}^2_\epsilon = \mathcal{M}(V_\epsilon) + T_\epsilon$ to adjust $\mathcal{M}(V_\epsilon)$. The difference between the decomposition for $\hat{\sigma}^2_\epsilon$ and the corresponding decomposition (5.1) is that $V_{\epsilon_{ji}}$ cannot be measured directly as $\mathcal{M}(y) + \mathcal{R}_j(\mathcal{M}(y))$ is unknown. However, we need to make some additional assumptions about T_ϵ which comprises of products of residuals $\mathcal{R}_{ji}(V_\epsilon)$ and $\mathcal{R}_{ji}(y)$ as follows.

Following Goldstein & Woof (2007, pg. 267), we assume the following fourth-order uncorrelated properties:

$$Cov(\mathcal{M}(V_\epsilon), \mathcal{R}_{ji}(y)\mathcal{R}_{ji'}(y)) = Cov(\mathcal{R}_{ji}(V_\epsilon), \mathcal{R}_{ji}(y)\mathcal{R}_{ji'}(y)) = 0, \tag{5.2}$$

and

$$Cov(\mathcal{R}_{ji}(y)\mathcal{R}_{ji'}(y), \mathcal{R}_{j'i}(y)\mathcal{R}_{j'i'}(y)) = 0, \text{ for } j \neq j' \tag{5.3}$$

We note that $j \neq j'$ implies that $i \neq i'$, i.e. the same student i cannot belong to two different classes j and j'.

Thus, to adjust the level 1 variance of the SOEREF model, we have the following theorem.

Theorem 5.1.1. *In the SOEREF model, we write the level-1 residuals as* $\epsilon_{ji} = \mathcal{R}(y_{ji})$ *with specifications* $E(\epsilon_{ji}) = 0$ *and* $Var(\epsilon_{ji}) = \sigma^2_\epsilon$. *To learn about the population level - 1 variance, we construct a representation for the squared level-1 residuals* $\epsilon^2_{ji} = V_{\epsilon_{ji}} = \mathcal{M}(V_\epsilon) + \mathcal{R}_{ji}(V_\epsilon)$. *We adjust* $\mathcal{M}(V_\epsilon)$ *based on the mean squared error (MSE)* $\hat{\sigma}^2_\epsilon$ *via the decomposition* $\hat{\sigma}^2_\epsilon = \mathcal{M}(V_\epsilon) + T_\epsilon$. *Using the second-order specifications over* $\mathcal{M}(V_\epsilon)$ *and* $\mathcal{R}_{ji}(V_\epsilon)$, *and assuming the following fourth order uncorrelated properties*

$$Cov(\mathcal{M}(V_\epsilon), \mathcal{R}_{ji}(y)\mathcal{R}_{ji'}(y)) = Cov(\mathcal{R}_{ji}(V_\epsilon), \mathcal{R}_{ji}(y)\mathcal{R}_{ji'}(y)) = 0,$$

and

$$Cov(\mathcal{R}_{ji}(y)\mathcal{R}_{ji'}(y), \mathcal{R}_{j'i}(y)\mathcal{R}_{j'i'}(y)) = 0, \text{ for } j \neq j',$$

we derive the following joint prior assessments

$$E(\hat{\sigma}^2_\epsilon) = V_{R_\epsilon}, \quad Var(\hat{\sigma}^2_\epsilon) = V_{M_\epsilon} + V_{T_\epsilon}, \quad Cov(\hat{\sigma}^2_\epsilon, \mathcal{M}(V_\epsilon)) = V_{M_\epsilon}. \tag{5.4}$$

The Bayes linear adjusted mean and variance of the level-1 population variance $\mathcal{M}(V_\epsilon)$ *based on the mean squared error (MSE)* $\hat{\sigma}^2_\epsilon$ *is given by*

$$E_{\hat{\sigma}^2_\epsilon}(\mathcal{M}(V_\epsilon)) = \frac{V_{M_\epsilon}\hat{\sigma}^2_\epsilon + V_{T_\epsilon}V_{R_\epsilon}}{V_{M_\epsilon} + V_{T_\epsilon}}, \tag{5.5}$$

with the corresponding adjusted variance

$$Var_{\hat{\sigma}^2_\epsilon}(\mathcal{M}(V_\epsilon)) = \frac{V_{M_\epsilon}V_{T_\epsilon}}{V_{M_\epsilon} + V_{T_\epsilon}}. \tag{5.6}$$

Proof. First, we derive (5.4). The MSE is the mean squared deviations of each observation y_{ji} from its respective group mean $\bar{y}_{j.}$, and it can be written in terms of the squared residuals as follows:

$$\begin{aligned}
\hat{\sigma}^2_\epsilon &= \frac{1}{g(n-1)}\text{SSE} \\
&= \frac{1}{g(n-1)}\sum_j\sum_i (y_{ji} - \bar{y}_{j.})^2 \\
&= \frac{1}{g(n-1)}\sum_j\sum_i (\epsilon_{ji} - \bar{\epsilon}_{j.})^2
\end{aligned}$$

We use the above to construct a representation for $\hat{\sigma}^2_\epsilon$ as follows:

$$\begin{aligned}
\hat{\sigma}^2_\epsilon &= \frac{1}{g(n-1)}\sum_j\left[\sum_i \epsilon^2_{ji} - \frac{1}{n}(\sum_i \epsilon_{ji})^2\right] \\
&= \frac{1}{g(n-1)}\sum_j\left[\sum_i \frac{(n-1)}{n}\epsilon^2_{ji} - \frac{2}{n}\sum_{i<i'}\epsilon_{ji}\epsilon_{ji'}\right] \\
&= \sum_j\sum_i \frac{1}{gn}[\mathcal{M}(V_\epsilon) + \mathcal{R}_{ji}(V_\epsilon)] - \frac{2}{gn(n-1)}\sum_j\sum_{i<i'}\mathcal{R}_{ji}(y)\mathcal{R}_{ji'}(y) \\
&= \mathcal{M}(V_\epsilon) + T_\epsilon,
\end{aligned} \tag{5.7}$$

where

$$T_\epsilon = \frac{1}{gn}\sum_j\sum_i \mathcal{R}_{ji}(V_\epsilon) - \frac{2}{gn(n-1)}\sum_j\sum_{i<i'}\mathcal{R}_{ji}(Y)\mathcal{R}_{ji'}(y) \tag{5.8}$$

From (5.8) and the fourth-order uncorrelated properties (5.2) and (5.3), we may now derive the following properties of T_ϵ:

$$\mathrm{E}(T_\epsilon) = 0 \tag{5.9}$$

$$\mathrm{Var}(T_\epsilon) = V_{T_\epsilon} = \frac{1}{gn}V_{R(V_\epsilon)} + \frac{2}{gn(n-1)}(V_{M_\epsilon} + V^2_{R_\epsilon}) \tag{5.10}$$

$$\mathrm{Cov}(\mathcal{M}(V_\epsilon), T_\epsilon) = 0 \tag{5.11}$$

Using the above we obtain the results in (5.4), that is

$$\mathrm{E}(\hat{\sigma}_\epsilon^2) = V_{R_\epsilon}, \quad \mathrm{Var}(\hat{\sigma}_\epsilon^2) = V_{M_\epsilon} + V_{T_\epsilon}, \quad \mathrm{Cov}(\hat{\sigma}_\epsilon^2, \mathcal{M}(V_\epsilon)) = V_{M_\epsilon}.$$

The proofs for the adjusted level-1 population variance (5.5) and its corresponding adjusted variance (5.6) in Theorem 5.2.1. simply follows from the application of the Bayes linear equations for updating a mean given $\hat{\sigma}_\epsilon^2$. ■

In order to perform a Bayes linear adjustment of the level-1 population variance $\mathcal{M}(V_\epsilon)$, we need to specify V_{M_ϵ}, the variance of $\mathcal{M}(V_\epsilon)$ and $V_{R(V_\epsilon)}$ (which is part of V_{T_ϵ}), the variance of $\mathcal{R}_{ji}(V_\epsilon)$.

5.2 Priors for fourth order quantities

Assessing a population variance requires computations with fourth order moments (see Searle *et al.*(1992); p407). While the first three moments have simple mathematical forms, the fourth moment involves rather more complicated expressions. For example, consider the rth central moment of a random quantity X with respect to the probability measure $F(x)$

$$\mu_r(X) = \int_{-\infty}^{+\infty} (x-\mu)^r dF(x).$$

If $\bar{X}$ is the sample mean of n *i.i.d* random quantities then the formulae for the first three moments of $\bar{X}$ are simple:

$$\mu(\bar{X}) = \mu(X), \quad \mu_2(\bar{X}) = \mu_2(X)/n, \quad \mu_3(\bar{X}) = \mu_3(X)/n^2$$

while the fourth moment is

$$\mu_4(\bar{X}) = \mu_4(X)/n^3 + 3(n-1)\mu_2^2(X)/n^3$$

Also, a Bayesian subjectivist perspective would require an individual to make well-sourced specifications about variances along with specifications of his uncertainty about these variances. Such specifications are unfamiliar and, as such, quite challenging for the individual. An important issue here is whether it is simpler, and/or

more meaningful, for the individual to specify his uncertainty through fourth order moments or whether it makes more sense to specify his uncertainties about variances directly (e.g. by thinking as to what is the variance of the specified variance). The fourth order moment is related to the kurtosis of a probability density function.

5.3 Choice of the priors for V_{M_ϵ} and $V_{R(V_\epsilon)}$

Elicitation of an expert's beliefs is quite challenging in general but more so for a fourth order quantity such as $V_{R(V_\epsilon)}$, the variance of the zero-mean uncorrelated sequence $\mathcal{R}_{11}(V_\epsilon), \mathcal{R}_{12}(V_\epsilon), \ldots$ in (5.1). Since $V_{R(V_\epsilon)}$ relates to the shape of the distribution of y_{ji}, for its specification we follow the approach in Goldstein & Wooff (2007). We assume that the level-1 population variance $\mathcal{M}(V_\epsilon)$ acts like a scale parameter and write:

$$\mathcal{R}_{ji}(y) = \sqrt{\mathcal{M}(V_\epsilon)} Z_{ji} \quad \forall i, j \tag{5.12}$$

where $\mathrm{E}(Z_{ji}) = 0$, $\mathrm{Var}(Z_{ji}) = 1$, Z_{ji} is independent of $\mathcal{M}(V_\epsilon)$ and the sequence $Z_{11}, Z_{12}, \ldots$ are also independent.

Using representation (5.1) we have $[\mathcal{R}_{ji}(y)]^2 = \mathcal{M}(V_\epsilon) + \mathcal{R}_{ji}(V_\epsilon)$ and substituting $\mathcal{R}_{ji}(y)$ from (5.12) and simplifying, we have:

$$\mathcal{R}_{ji}(V_\epsilon) = \mathcal{M}(V_\epsilon)(Z_{ji}^2 - 1) \tag{5.13}$$

We thus require the variance of the product of the two independent quantities on the right of (5.13). Now, if A and B are independent random quantities with means μ_A and μ_B respectively, then

$$\mathrm{Var}(AB) = \mu_B^2 \mathrm{Var}(A) + \mu_A^2 \mathrm{Var}(B) + \mathrm{Var}(A)\mathrm{Var}(B),$$

from which we obtain

$$\begin{aligned} \mathrm{Var}(\mathcal{R}_{ji}(V_\epsilon)) &= V_{R(V_\epsilon)} \\ &= (\mathrm{Var}(\mathcal{M}(V_\epsilon)) + [\mathrm{E}(\mathcal{M}(V_\epsilon))]^2)\mathrm{Var}(Z_{ji}^2) \\ &= (V_{M_\epsilon} + V_{R_\epsilon}^2)\mathrm{Var}(Z_{ji}^2) \\ &= (V_{M_\epsilon} + V_{R_\epsilon}^2)(\mathrm{Kur}(Z_{ji}) - 1) \end{aligned} \tag{5.14}$$

If for example, we consider Z_{ji} to be approximately Gaussian, then we would assign $\text{Kur}(Z_{ji})=\text{E}(Z_{ji}^4)=3$. Since $\text{E}(Z_{ji}^2)=1$, we have $\text{Var}(Z_{ji}^2) = \text{E}(Z_{ji}^4)-[\text{E}(Z_{ji}^2)]^2 = 2$. Else, we may choose a fat tail distribution if we consider it appropriate. For example if we choose the t distribution scaled to have variance 1, then

$$Z_{ji} = \sqrt{\frac{\nu}{\nu-2}}T_\nu,$$

where T_ν has a t distribution with ν degrees of freedom. The kurtosis for this distribution is

$$\text{Kur}(Z_{ji}) = \frac{3(\nu-2)}{\nu-4},$$

giving the following variance.

$$\text{Var}(Z_{ji}^2) = \frac{2(\nu-1)}{\nu-4}.$$

Hence, if we decide on $\nu = 5$, then we are choosing a high kurtosis leading to $\text{Var}(Z_{ji}^2)=8$, that is a high variance for the squared residuals $V_{R(V_\epsilon)}$ which in turn will lead to a higher variance for V_{T_ϵ} in (5.6). This implies that the observed $\hat{\sigma}_\epsilon^2$ will receive less weight compared to our prior V_{R_ϵ} in our adjustment (5.5). Conversely, if we judge a smaller kurtosis than the Gaussian one above is suitable, then a uniform distribution centered on zero such as $Z_{ji} \sim \text{Unif}(-1,1)$ will give $\text{Kur}(Z_{ji}) = 1.8$, leading to a smaller variance $\text{Var}(Z_{ji}) = 0.8$. This will give more weight to our data in the adjustment (5.5.).

Suppose we use representation (5.12) and we specify values for $\text{Var}(Z_{ji}^2)$ and V_{R_ϵ}. One method to specify our fourth order quantities uses the following theorem.

Theorem 5.3.1. *We assume that the level-1 population variance $\mathcal{M}(V_\epsilon)$ acts like a scale parameter so that $\mathcal{R}_{ji}(y) = \sqrt{\mathcal{M}(V_\epsilon)}Z_{ji} \quad \forall i,j$. The proportion of variance resolved $Var_{\hat{\sigma}_\epsilon^2}(\mathcal{M}(V_\epsilon))$, relative to the prior V_{M_ϵ} is*

$$\frac{Var_{\hat{\sigma}_\epsilon^2}(\mathcal{M}(V_\epsilon))}{Var(\mathcal{M}(V_\epsilon))} = \frac{1}{1+\frac{n-1}{\kappa}\frac{c}{c+1}}, \tag{5.15}$$

where, for simplicity, we write $V_{M_\epsilon} = cV_{R_\epsilon}^2$, $c>0$ and $\kappa = \frac{1}{gn}[(n-1)\text{Var}(Z_{ji}^2)+2]$.

Proof. From Theorem 5.1.1, the adjusted variance of the level-1 population variance is

$$\text{Var}_{\hat{\sigma}_\epsilon^2}(\mathcal{M}(V_\epsilon)) = \frac{V_{M_\epsilon}V_{T_\epsilon}}{V_{M_\epsilon}+V_{T_\epsilon}},$$

where $V_{T_\epsilon} = \frac{1}{gn}V_{R(V_\epsilon)} + \frac{2}{gn(n-1)}(V_{M_\epsilon} + V_{R_\epsilon}^2)$ (see 5.10).

Using (5.14), $V_{R(V_\epsilon)} = (V_{M_\epsilon} + V_{R_\epsilon}^2)\text{Var}(Z_{ji}^2)$, and setting $V_{M_\epsilon} = cV_{R_\epsilon}^2$ as in Goldstein & Woof (2006) we have

$$\begin{aligned} V_{T_\epsilon} &= \frac{1}{gn}[V_{M_\epsilon} + V_{R_\epsilon}^2)\text{Var}(Z_{ji}^2) + \frac{2}{gn(n-1)}(V_{M_\epsilon} + V_{R_\epsilon}^2)] \\ &= \frac{1}{gn}[cV_{R_\epsilon}^2 + V_{R_\epsilon}^2)\text{Var}(Z_{ji}^2) + \frac{2}{gn(n-1)}(cV_{R_\epsilon}^2 + V_{R_\epsilon}^2)] \\ &= \frac{(c+1)V_{R_\epsilon}^2}{gn(n-1)}[(n-1)\text{Var}(Z_{ji}^2) + 2] \\ &= \frac{(c+1)V_{R_\epsilon}^2}{(n-1)}\kappa, \end{aligned}$$

where $\kappa = \frac{1}{gn}[(n-1)\text{Var}(Z_{ji}^2)+2]$ for $c > 0$. Substituting V_{T_ϵ} in the above expression for $\text{Var}_{\hat{\sigma}_\epsilon^2}(\mathcal{M}(V_\epsilon))$, we obtain

$$\begin{aligned} \frac{\text{Var}_{\hat{\sigma}_\epsilon^2}(\mathcal{M}(V_\epsilon))}{V_{M_\epsilon}} &= \frac{\frac{(c+1)V_{R_\epsilon}^2}{(n-1)}\kappa}{V_{M_\epsilon} + \frac{(c+1)V_{R_\epsilon}^2}{(n-1)}\kappa} \\ &= \frac{1}{1 + \frac{n-1}{\kappa}\frac{c}{c+1}}. \end{aligned}$$

The final result is obtained after substituting $V_{R_\epsilon}^2 = \frac{1}{c}V_{M_\epsilon}$ ■

5.4 Some implementation issues

We can now use Theorem 5.3.1 to assess our beliefs about V_{M_ϵ} and $V_{R_\epsilon}^2$ by varying c and the sample size n for a chosen value of $\kappa = \frac{1}{gn}[(n-1)\text{Var}(Z_{ji}^2)+2]$. We note that κ depends on the number of level-2 groups g. For example, setting $\kappa = \frac{2}{g}$ results in $\text{Var}(Z_{ji}^2) = 2$ which is consistent with the assumption of a Gaussian distribution. This is a simplifying assumption just made for the purposes of the present account, but in real applications we would think carefully and consider the effect of varying the kurtosis on our answers. Hence, the proportion of variance resolved relative to the prior in (5.15) simplifies to

$$\frac{\text{Var}_{\hat{\sigma}_\epsilon^2}(\mathcal{M}(V_\epsilon))}{\text{Var}(\mathcal{M}(V_\epsilon))} = \frac{1}{1 + \frac{(n-1)g}{2}\frac{c}{c+1}} \tag{5.16}$$

From (5.16) the proportion of variance resolved decreases monotonically from one ($c = 0$) to $[1 + \frac{(n-1)g}{2}]^{-1}$ (c is large so that $c \approx c + 1$). We may now choose V_{M_ϵ} by exploring our beliefs to the implications of various level-1 sample sizes n given κ. The dependence on the number of level-2 groups g is not a problem as it is fixed in the multilevel design being considered. For instance the STAT1010 example has $g = 7$ classes.

Figure 5.1 shows a nomogram of the relationship between various level-1 sample sizes n, the scaling choice c and the corresponding proportion of variance remaining in $\mathcal{M}(V_\epsilon)$ for the case $g = 7$. Using the graph, we will choose a small c if we feel that the data is sufficiently informative so as to reduce the remaining variance as a proportion of prior rapidly, else we will choose a larger value for c.

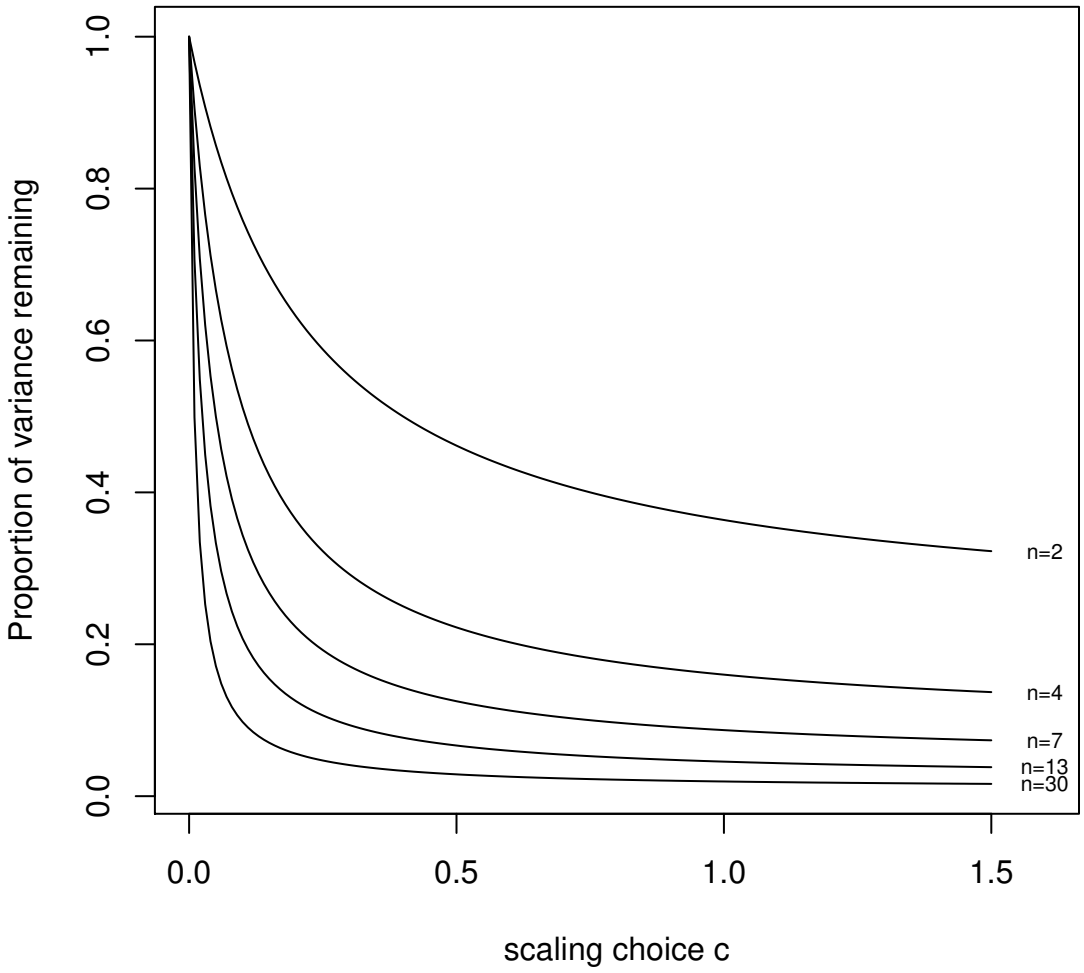


Figure 5.1: *The proportion of prior variance remaining in $\mathcal{M}(V_\epsilon)$ after adjusting $\mathcal{M}(V_\epsilon)$ by the ANOVA estimator $\hat{\sigma}^2_\epsilon$ for $\kappa = 2/g$ for group $g = 7$ and various level-1 sample sizes n, as a function of c. For $\kappa \neq 2$, replace n by $n' = (n-1)g\kappa/2 + 1$. For this balanced case, the total sample sizes gn varies from 14 to 210.*

For $\kappa \neq 2$, we replace n by $n' = (n-1)g\kappa/2 + 1$. From (5.16) we note that if g is small and n is large then $g(n-1) \approx gn$ that is the total sample size. Hence, it is the total sample size that determines the proportion of variance remaining in the level-1 variance $\mathcal{M}(V_\epsilon)$.

Goldstein & Wooff (2006) page 269 uses the notion of equivalent sample size to provide an alternative method for assessing the prior information. For adjusting the

level-1 variance, we write (5.5) as follows.

$$\mathrm{E}_{\hat{\sigma}^2_\epsilon}(\mathcal{M}(V_\epsilon)) = \alpha\hat{\sigma}^2_\epsilon + (1-\alpha)\mathrm{E}(\mathcal{M}(V_\epsilon)) \tag{5.17}$$

where $\alpha = \frac{V_{M_\epsilon}}{V_{M_\epsilon}+V_{T_\epsilon}}$.

If we judge that our prior information is equivalent to a notional sample of size m, then the adjusted expectation in (5.17) can be written in terms of the notional prior and actual sample sizes with

$$\alpha = n/(m+n) \tag{5.18}$$

Goldstein & Wooff (2006) states that the two methods are equivalent and gives the following expression for the relationship between the two methods.

$$m = \frac{\kappa n(c+1)}{(n-1)c} \tag{5.19}$$

We derive (5.19) as follows. Combining (5.6) and (5.16) and substituting α from (5.18) we obtain

$$\begin{aligned}
\frac{\mathrm{Var}_{\hat{\sigma}^2_\epsilon}(\mathcal{M}(V_\epsilon))}{\mathrm{Var}(\mathcal{M}(V_\epsilon))} &= \frac{V_{T_\epsilon}}{V_{M_\epsilon}+V_{T_\epsilon}} \\
&= 1 - \frac{V_{M_\epsilon}}{V_{M_\epsilon}+V_{T_\epsilon}} \\
&= 1-\alpha \\
&= \frac{m}{m+n} = \frac{1}{1+\frac{(n-1)}{\kappa}\frac{c}{c+1}}
\end{aligned}$$

Solving for m in the last line above gives (5.19). Also, for sufficiently large n, (5.19) simplifies to $m \approx \frac{\kappa(c+1)}{c}$ from which $c \approx \frac{\kappa}{m-\kappa}$. The methods we developed above may be viewed as an extension of that of Goldstein & Wooff (2006) for a scalar quantity; substituting $g = 1$ gives similar results.

5.5 Application to STAT1010 data

In section 3.7.2 we carefully assessed the variance of the level-1 residual $\mathcal{R}(y_{ji})$ and obtained

$$Var(\mathcal{R}(y_{ji})) = 237 = V_{R_\epsilon},$$

hence $\mathrm{E}(\mathcal{M}(V_\epsilon))=V_{R_\epsilon} = 237$. In our experience the distribution of examinations marks do not have a thick tail. Hence, we choose $\kappa = 2/g$ (see section 5.4) corresponding to the fourth moment of a Gaussian distribution. Using $V_{R_\epsilon} = 237$ and $V_{M_\epsilon} = cV_{R_\epsilon}^2$ we specify V_{M_ϵ} by choosing the value of c using Figure 5.1 which gives the relationship between c and various sample sizes n for $g = 7$.

Since we are quite satisfied with our judgement of the prior level-1 variance $V_{R_\epsilon} = 237$, we wish to give the prior a reasonable weight in the adjustment. On the other hand, the ANOVA estimator of the level-1 variance is efficient (see section 2.7.2), and therefore we also want to give the sample enough weight. We judge that $c = 0.03$, corresponding to fast variance learning, is appropriate. This choice results in $m \approx 10$, giving a weight of about one third to our prior in the adjusted expectation.

The observed estimate of $\hat{\sigma}_\epsilon^2$, based on $g = 7$ classes and $n = 23$ students (a total of 161 observations), is

$$\hat{\sigma}_\epsilon^2 \;=\; \frac{1}{g(n-1)}\sum_j\sum_i(\epsilon_{ji}-\bar{\epsilon}_{j.})^2 = \frac{1}{g(n-1)}\sum_j\sum_i(y_{ji}-\bar{y}_{j.})^2 = 213.70$$

Before carrying out the adjustment $\mathrm{E}_{\hat{\sigma}_\epsilon^2}(\mathcal{M}(V_\epsilon))$, we examine the 161 squared residuals. Each $(y_{ji}-\bar{y}_{j.})^2/V_{R_\epsilon}$ has a $\chi^2(1)$ (chi-square distribution with one degree of freedom) if y_{ji} are each i.i.d Gaussian. Almost all of the squared residuals are close to their expected value of 237, but 11 are three or more times larger that the expected value $V_{R_\epsilon} = 237$, as shown in Table 5.1 below. Since, the probability $\mathrm{P}(\chi^2_{0.92}(1)) > 3$=0.08, we would expect about 13 such large residuals out of the 161.

Although here we have only 11 large squared residuals, there is a pattern in Table 5.1: the Faculty of Law and Management (C1 to C3) has seven of the large residuals, and these are 4 to 6 times larger than $V_{R_\epsilon} = 237$, while the Faculty of Engineering (C4 to C7) has only four large residuals, these being 3 to 4 times larger than $V_{R_\epsilon} = 237$. We have applied partial diagnostics to explore how data from the two faculties combine to give the final adjusted expectation and found no substantial differences when analysing the adjustments by the two sources separately.

Large squared residuals are mostly due to students scoring very good marks in

Faculty	Law & Management			Engineering			
Class	C1	C2	C3	C4	C5	C6	C7
No. of squared residuals ≥ 3 times $V_{R_\epsilon} = 237$	1	3	3	1	1	1	1
Largest squared residuals (compared to $V_{R_\epsilon} = 237$)	1280.4 (5.4)	1470.6 (6.2)	1018.4 (4.3)	708.0 (3.0)	830.9 (3.5)	838.5 (3.5)	955.6 (4.0)

Table 5.1: *Large observed squared residuals for each of the 7 classes of the STAT1010 data by faculty. The rows show the number of cases exceeding 3 or more times the expected value* $V_{R_\epsilon} = 237$*, the largest squared residuals in each class, and the factor by which they exceed 237 in brackets.*

an otherwise average class. For example, the largest value 1470.6 is due to a student scoring 83% while the class average is 44.65%. We could have anticipated having a few very good students in average classes and specify a tail distribution somewhat thicker than for the normal distribution. In section 5.5.1 we consider the effect of specifying a higher kurtosis on our adjusted expectation for $\mathrm{M}(\mathrm{V}_\epsilon)$.

For the chosen value of $c = 0.03$, we obtain $V_{M_\epsilon} = 1685.07$ and $V_{T_\epsilon} = 751.35$. Using (5.5) and (5.6) from Theorem 5.1.1, the observed adjusted expectation and variance of the level-1 variance $\mathcal{M}(V_\epsilon)$ are as follows.

$$\mathrm{E}_{\hat{\sigma}^2_\epsilon}(\mathcal{M}(V_\epsilon)) = 220.88$$
$$\mathrm{Var}_{\hat{\sigma}^2_\epsilon}(\mathcal{M}(V_\epsilon)) = 519.65,$$

representing a reduction of 1165.42, about 69% of the prior V_{M_ϵ}. The standardized change in expectation is -0.472, indicating that the change in expectation is only marginally smaller than the prior expectation, and hence, unsurprising. The remaining prior variation is about 31%. We conclude that the level-1 variance in examinations marks is as we expected.

Varying c impacts the adjusted expectation for $\mathcal{M}(V_\epsilon)$ marginally but has a greater effect on the adjusted variation. Choosing $c = 0.01$, the adjusted expectation is 226.92 with adjusted variation down to 318.72 and remaining prior variation of 57%. For $c = 0.1$, corresponding to slower variance learning, the adjusted expectation is 216.61, but the adjusted variation increases to 702.11 with 12.5% prior variation remaining.

5.5.1 Effects of a higher kurtosis

Following our discussion above, we now specify a higher kurtosis, consistent with a heavy tail distribution. We intend to compare the results of our adjustments here with those for the above case where $\text{Var}(Z_{ji}^2) = 2$, which for reference, we term the normal case. Given we observed only a few large residuals, we judge a kurtosis leading to about twice the value of $\text{Var}(Z_{ji}^2)$ for the normal case would be suitable. We choose $\nu = 7$ in accordance with the t distribution (see section 5.3), giving $\text{Var}(Z_{ji}^2) = 2(\nu - 1)/(\nu - 4) = 4$. For ease of comparison we keep the same scaling choice $c = 0.03$. Thus $V_{M_\epsilon} = 1685.07$ stays the same, but $V_{T_\epsilon} = 1470.04$, almost double the normal value. This increase results in the observed value $\hat{\sigma}_\epsilon^2$ receiving less weight, 0.53 compared to 0.69 for the normal case, in the adjusted expectation (5.5). The adjusted expectation and variance of the level-1 variance $\mathcal{M}(V_\epsilon)$ is as follows.

$$\begin{aligned} \text{E}_{\hat{\sigma}_\epsilon^2}(\mathcal{M}(V_\epsilon)) &= 224.56 \\ \text{Var}_{\hat{\sigma}_\epsilon^2}(\mathcal{M}(V_\epsilon)) &= 785.11, \end{aligned}$$

a reduction of 899.96 representing 53.41% of the prior, which is less than for the normal case. The standardized change in expectation is -0.415 hence, unsurprising. There is relatively more prior variation remaining, about 46.59%. To conclude, the increased kurtosis, compared to the normal case, results in an increase in the estimated level-1 variance, as well as in its adjusted variation, while the change in expectation is in line with what we expected.

5.6 Adjusting the level-1 variance - unbalanced situation

In section 5.1 we revisited the unbalanced SOEREF model and derived various theorems and results for the adjustment of the population level-1 variance $\mathcal{M}(V_\epsilon)$ for the balanced case. The representation for the squared residuals ϵ_{ji}^2 and the assumptions for the fourth order uncorrelated properties we used in our derivations, also apply to the unbalanced case. However, the proofs of the theorems and some of the results

for unbalanced data are somewhat more complex (due to more complex expressions for T_ϵ and $\text{Var}(T_\epsilon)$) , and may be viewed as extensions of the balanced case as shown below.

5.7 The adjusted expectation and variance of $\mathcal{M}(V_\epsilon)$ for the unbalanced situation

The adjusted mean and variance for $\mathcal{M}(V_\epsilon)$ given $\hat{\sigma}^2_\epsilon$ for the unbalanced case is similar to the balanced situation but with a more complex expression for $\text{Var}(T_\epsilon)$ as the MSE estimator of $\hat{\sigma}^2_\epsilon$ is based on n_j observations in group j for unbalanced data. To adjust the level-1 variance of the SOEREF model for unbalanced data, we have the following theorem.

Theorem 5.7.1. *The Bayes linear adjusted mean and variance of the level-1 population variance $\mathcal{M}(V_\epsilon)$ for unbalanced data based on the mean squared error (MSE) $\hat{\sigma}^2_\epsilon$ is given by*

$$E_{\hat{\sigma}^2_\epsilon}(\mathcal{M}(V_\epsilon)) = \frac{V_{M_\epsilon}\hat{\sigma}^2_\epsilon + V_{T_\epsilon}V_{R_\epsilon}}{V_{M_\epsilon} + V_{T_\epsilon}}, \tag{5.20}$$

with the corresponding adjusted variance

$$Var_{\hat{\sigma}^2_\epsilon}(\mathcal{M}(V_\epsilon)) = \frac{V_{M_\epsilon}V_{T_\epsilon}}{V_{M_\epsilon} + V_{T_\epsilon}}, \tag{5.21}$$

with

$$V_{T_\epsilon} = \frac{1}{(N-g)^2}\left[\sum_j^g \frac{(n_j-1)^2}{n_j}V_{R(V_\epsilon)} + 2\sum_j^g \frac{(n_j-1)}{n_j}(V_{M_\epsilon} + V^2_{R_\epsilon})\right]. \tag{5.22}$$

Proof. The proof proceeds as for Theorem 5.1.1. We prove (5.22) first. For unbalanced data the MSE $\hat{\sigma}^2_\epsilon$ is as follows.

$$\begin{aligned}\hat{\sigma}^2_\epsilon &= \frac{1}{(N-g)}\text{SSE}\\ &= \frac{1}{(N-g)}\sum_j^g\sum_i^{n_j}(y_{ji} - \bar{y}_{j.})^2\\ &= \frac{1}{(N-g)}\sum_j^g\sum_i^{n_j}(\epsilon_{ji} - \bar{\epsilon}_{j.})^2,\end{aligned}$$

where $N = \sum n_j$.

The representation for $\hat{\sigma}^2_\epsilon$ is

$$\begin{aligned}
\hat{\sigma}^2_\epsilon &= \frac{1}{(N-g)} \sum_j^g \left[\sum_i^{n_j} \epsilon^2_{ji} - \frac{1}{n_j} (\sum_i^{n_j} \epsilon_{ji})^2 \right] \\
&= \frac{1}{(N-g)} \sum_j^g \left[\sum_i^{n_j} \frac{(n_j-1)}{n_j} \epsilon^2_{ji} - \frac{2}{n_j} \sum_{i<i'}^{n_j} \epsilon_{ji} \epsilon_{ji'} \right] \\
&= \frac{1}{(N-g)} \sum_j^g \left(\frac{n_j - 1}{n_j} \right) \sum_i^{n_j} [\mathcal{M}(V_\epsilon) + \mathcal{R}_{ji}(V_\epsilon)] \\
&- \frac{2}{(N-g)} \sum_j^g \frac{1}{n_j} \sum_{i<i'}^{n_j} \mathcal{R}_{ji}(Y) \mathcal{R}_{ji'}(Y) \\
&= \mathcal{M}(V_\epsilon) + \frac{1}{(N-g)} \sum_j^g \left(\frac{n_j - 1}{n_j} \right) \sum_i^{n_j} \mathcal{R}_{ji}(V_\epsilon) \\
&- \frac{2}{(N-g)} \sum_j^g \frac{1}{n_j} \sum_{i<i'}^{n_j} \mathcal{R}_{ji}(Y) \mathcal{R}_{ji'}(Y) \\
&= \mathcal{M}(V_\epsilon) + T_\epsilon
\end{aligned} \tag{5.23}$$

where

$$T_\epsilon = \frac{1}{(N-g)} \sum_j^g \left(\frac{n_j - 1}{n_j} \right) \sum_i^{n_j} \mathcal{R}_{ji}(V_\epsilon) - \frac{2}{(N-g)} \sum_j^g \frac{1}{n_j} \sum_{i<i'}^{n_j} \mathcal{R}_{ji}(Y) \mathcal{R}_{ji'}(Y). \tag{5.24}$$

Using the fourth-order uncorrelated properties (5.7) and (5.8):

$$Cov(\mathcal{M}(V_\epsilon), \mathcal{R}_{ji}(y) \mathcal{R}_{ji'}(y)) = Cov(\mathcal{R}_{ji}(V_\epsilon), \mathcal{R}_{ji}(y) \mathcal{R}_{ji'}(y)) = 0,$$

and

$$Cov(\mathcal{R}_{ji}(y) \mathcal{R}_{ji'}(y), \mathcal{R}_{j'i}(y) \mathcal{R}_{j'i'}(y)) = 0, \text{ for } j \neq j',$$

we obtain

$$\begin{aligned}
\text{Var}(T_\epsilon) = V_{T_\epsilon} &= \frac{1}{(N-g)^2} \sum_j^g \left(\frac{n_j - 1}{n_j} \right)^2 n_j V_{R(V_\epsilon)} \\
&+ \frac{4}{(N-g)^2} \sum_j^g \frac{1}{n_j^2} \sum_{i<i'}^{n_j} (V_{M_\epsilon} + V^2_{R_\epsilon}) \\
&= \frac{1}{(N-g)^2} \left[\sum_j^g \frac{(n_j-1)^2}{n_j} V_{R(V_\epsilon)} + 2 \sum_j^g \frac{(n_j-1)}{n_j} (V_{M_\epsilon} + V^2_{R_\epsilon}) \right],
\end{aligned}$$

which proves (5.22). One way to verify the above expressions is to put $n_j = n$ and compare with the balanced case.

Using the properties of T_ϵ, that are similar to the balanced case (see (5.9) to (5.11)) albeit the more complex expression for V_{T_ϵ}, and the decomposition (5.23) for $\hat{\sigma}^2_\epsilon$, we derive the following joint prior assessments

$$\mathrm{E}(\hat{\sigma}^2_\epsilon) = V_{R_\epsilon}, \quad \mathrm{Var}(\hat{\sigma}^2_\epsilon) = V_{M_\epsilon} + V_{T_\epsilon}, \quad \mathrm{Cov}(\hat{\sigma}^2_\epsilon, \mathcal{M}(V_\epsilon)) = V_{M_\epsilon}.$$

The proofs for (5.20) and (5.21) follow from the application of the Bayes linear equations for updating a mean given $\hat{\sigma}^2_\epsilon$. ■

5.8 Choice of priors V_{M_ϵ} and $V_{R(V_\epsilon)}$ for the unbalanced data

In order to choose priors for V_{M_ϵ} and $V_{R(V_\epsilon)}$ we follow the development in Theorem 5.3.1 that gives the proportion of variance resolved in $\mathrm{Var}_{\hat{\sigma}^2_\epsilon}(\mathcal{M}(V_\epsilon))$, relative to the prior V_{M_ϵ}. For the unbalanced case, we have the following theorem.

Theorem 5.8.1. *The proportion of variance resolved in $Var_{\hat{\sigma}^2_\epsilon}(\mathcal{M}(V_\epsilon))$, relative to the prior V_{M_ϵ} for unbalanced data is as follows:*

$$\frac{Var_{\hat{\sigma}^2_\epsilon}(\mathcal{M}(V_\epsilon))}{Var(\mathcal{M}(V_\epsilon))} = \frac{1}{1 + \frac{N-g}{\kappa}\frac{c}{c+1}}. \tag{5.25}$$

Proof. Using (5.14), $V_{R(V_\epsilon)} = (V_{M_\epsilon} + V^2_{R_\epsilon})\mathrm{Var}(Z^2_{ji})$, and setting $V_{M_\epsilon} = cV^2_{R_\epsilon}$ in (5.22) we obtain

$$\begin{aligned} V_{T_\epsilon} &= \frac{1}{(N-g)^2}\left[\sum_j^g \frac{(n_j-1)^2}{n_j} V_{R(V_\epsilon)} + 2\sum_j^g \frac{(n_j-1)}{n_j}(V_{M_\epsilon} + V^2_{R_\epsilon})\right] \\ &= \frac{(c+1)V^2_{R_\epsilon}}{(N-g)^2}\left[\sum_j^g \frac{(n_j-1)^2}{n_j}\mathrm{Var}(Z^2_{ji}) + 2\sum_j^g \frac{(n_j-1)}{n_j}\right] \\ &= \frac{(c+1)V^2_{R_\epsilon}}{(N-g)}\kappa. \end{aligned} \tag{5.26}$$

Expression (5.26) for V_{T_ϵ} is similar to its counterpart in Theorem 5.3.1 but with

$$\kappa = \frac{1}{(N-g)}\left[\sum_j^g \frac{(n_j-1)^2}{n_j}\mathrm{Var}(Z^2_{ji}) + 2\sum_j^g \frac{(n_j-1)}{n_j}\right], \tag{5.27}$$

thence the proof follows. ■

We note that the proportion of variance resolved (5.25) is similar to the balanced case with $N = ng$. For Gaussian kurtosis, substituting $\text{Var}(Z_{ji}^2) = 2$ gives $\kappa = 2$. The nomogram relating the scaling choice c and sample size n for the balanced case in figure 5.1 also applies to the unbalanced case because in both cases the total level-1 sample size N determines the proportion of variance resolved in $\mathcal{M}(V_\epsilon)$. For unbalanced data and $\kappa \neq 2$, we replace N by $N' = (N - g)\kappa/2 + g$.

The alternative method of direct assessment of our prior information described in section (5.4) is also based on the total level-1 sample size. We thus write the equivalent notional prior and actual sample sizes in terms of their respective totals M and N. Therefore in the adjustment $\text{E}_{\hat{\sigma}_\epsilon^2}(\mathcal{M}(V_\epsilon)) = \alpha\hat{\sigma}_\epsilon^2 + (1 - \alpha)\text{E}(\mathcal{M}(V_\epsilon))$ we have

$$\alpha = N/(M + N). \tag{5.28}$$

The equivalence between the two methods is given by the following expression:

$$M = \frac{\kappa N(c + 1)}{(N - g)c}. \tag{5.29}$$

If N is large relative to g, then $M \approx \kappa(c + 1)/c$ from which $c \approx \kappa/(M - \kappa)$.

5.9 Application to the STAT1010 data

We shall use our earlier prior assessment V_{R_ϵ}=237. We wish to compare the balanced and unbalanced analyses so we choose $\kappa = 2$ in line with the fourth moment of a Gaussian distribution as in the balanced case. If we keep our scaling choice at its previous value of $c = 0.03$, then using $M \approx \kappa(c + 1)/c$ our prior is equivalent to a sample size $M = 69$, giving a weight of about one fifth to the prior in the adjustment of $\mathcal{M}(V_\epsilon)$, which is somewhat less compared to the one third in the balanced situation. This seems reasonable, given the relatively large total unbalanced sample size of $N = 269$ compared to 161 in the balanced case.

The observed value of $\hat{\sigma}_\epsilon^2$ based on $g = 7$ classes and a total of $N = 269$ students is 227.29, not far from the 213.70 for the balanced case. We examine the 269 squared residuals comparing each to the $\chi^2(1)$ as in section 5.5. There are 23 squared

residuals that are larger than $3 \times V_{R_\epsilon}$ which is very close to the 22 (out of 269) we would expect from the chi-square distribution.

For the unbalanced data, table 5.2 below does not show any pattern in extreme squared residuals between the two faculties, unlike the balanced case. The apparent differences between faculties for the balanced case may be considered as a statistical artifact resulting from the non-random method we used to select cases from the complete unbalanced data to create the balanced data, namely by removing data from each of six classes so that they all have 23 students as in the smallest class. It is noteworthy that our partial diagnostics in section 5.5 found no substantial differences by faculty. Nevertheless, we still have a small percentage of extreme values and it is important that we check our analysis by specifying a higher kurtosis.

Faculty	Law & Management			Engineering			
Class	C1	C2	C3	C4	C5	C6	C7
No. of squared residuals ≥ 3 times $V_{R_\epsilon} = 237$	3	3	4	2	3	6	2
Largest squared residuals (compared to $V_{R_\epsilon} = 237$)	1243.8 (5.2)	1470.6 (6.2)	1005.8 (4.2)	1502.8 (6.3)	819.6 (3.5)	1083.3 (4.6)	1108.4 (4.7)

Table 5.2: *Large observed squared residuals for each of the 7 classes of the STAT1010 data (unbalanced) by faculty. The rows show the number of cases exceeding 3 or more times the expected value $V_{R_\epsilon} = 237$, the largest squared residuals in each class, and the factor by which they exceed 237 in brackets.*

For the scaling choice $c = 0.03$, $V_{M_\epsilon} = 1685.07$ as for the balanced situation. Using (5.26), $V_{T_\epsilon} = 441.63$ about 40% smaller than the corresponding value of 751.35 for the balanced case. This will result in less weight given to the prior in the adjusted expectation of $\mathcal{M}(V_\epsilon)$. Using (5.20) and (5.21) of Theorem 5.7.1, the observed adjusted expectation and variance of the level-1 variance $\mathcal{M}(V_\epsilon)$ is as follows.

$$\mathrm{E}_{\hat{\sigma}^2_\epsilon}(\mathcal{M}(V_\epsilon)) = 229.31$$
$$\mathrm{Var}_{\hat{\sigma}^2_\epsilon}(\mathcal{M}(V_\epsilon)) = 349.92,$$

The adjusted variation has been reduced by 1335.15, about 79% of the prior V_{M_ϵ}.

The standardized change in expectation of -0.211 is unsurprising and indicates that the adjusted expectation is slightly smaller than the prior expectation. The prior variation remaining is about 21% (10% points less than the balanced case), hence the level-1 variation in examination marks is as we expected. We are less uncertain about the level-1 variance compared to the balanced case due to the larger sample size in the unbalanced data.

As for balanced data, varying the scaling choice c has little effect on the adjusted expectation; the adjusted variation, however, is more sensitive to the variations in c. For $c = 0.01$ the adjusted expectation is 231.52 with 43.53% prior variation remaining while for $c = 0.1$ the adjusted expectation is 228.04 with a mere 7.75% of prior variation remaining.

We consider the effect of a higher kurtosis on our analysis. We choose $\nu = 7$ in accordance with the t distribution as in the balanced case. Hence, $\text{Var}(Z_{ji}^2) = 2(\nu - 1)/(\nu - 4) = 4$. We also keep the same scaling choice $c = 0.03$. Thus $V_{M_\epsilon} = 1685.07$ stays the same, but $V_{T_\epsilon} = 871.80$, almost double the value for Gaussian kurtosis. Despite this increase, the observed value $\hat{\sigma}_\epsilon^2$ receives only marginally less weight, 0.66 compared to 0.69 for the Gaussian case, in the adjusted expectation (5.5). The adjusted expectation and variance of the level-1 variance $\mathcal{M}(V_\epsilon)$ is as follows.

$$\begin{aligned} \text{E}_{\hat{\sigma}_\epsilon^2}(\mathcal{M}(V_\epsilon)) &= 230.60 \\ \text{Var}_{\hat{\sigma}_\epsilon^2}(\mathcal{M}(V_\epsilon)) &= 574.55, \end{aligned}$$

a reduction of 1110.52 representing 65.90% of the prior, which is less than for the Gaussian case. The standardized change in expectation is -0.192 hence, unsurprising. There is relatively more prior variation remaining, about 34.1%. To conclude, increasing the kurtosis to twice that of the Gaussian distribution results in an increase in the estimated level-1 adjusted variation but leaves the adjusted expectation virtually unchanged, while the change in expectation is in line with what we expected.

Chapter 6

The Bayes Linear Minimum Variance Estimator and Two-stage Bayes linear analysis

In Chapter 5 we adjusted the level-1 variance of the SOEREF model. In this chapter, we shall adjust the level-2 variance which is more difficult to estimate and could even be negative. We develop a new methodology which we term Bayes linear Minimum Variance Estimator (BLIMVE) and apply it to adjust beliefs about the population level 2 variance in the SOEREF model. We shall also consider adjustment of variances in the more complex SOEREG model.

Having learned about the variance components in a multilevel model, it seems intuitively sensible to use the adjusted variances to learn about the population means. Such Bayes linear assessment of the means is termed *variance-modified Bayes linear assessments* by Goldstein (1979), while the procedure for assessing the variances first, and then using them to assess means in a second stage, is termed two-stage Bayes linear analysis in Goldstein & Woof (2007, pg. 288). Below we develop and apply two-stage analysis for both the SOEREF and SOEREG models.

6.1 Adjustment of the level-2 variance and the development of a Bayes Linear Minimum Variance Estimator (BLIMVE)

So far we have adjusted the population level-1 variance of the two-level SOEREF model. We shall now adjust our beliefs about the population level-2 variance, or more specifically the variance of the level-2 residual $Var(\mathcal{R}_j(\mathcal{M}(y)))=(\sigma_u^2-\gamma)$, where γ is the prior variance of the population grand mean $\mathcal{M}(y)$. The level-2 variance plays a very important part in the Bayes linear analysis of the SOEREF model: the adjusted expectation and variance, and the canonical analysis of the population grand mean $\mathcal{M}(y)$ and the group j mean $\mathcal{M}(y_j)$, as well as the choice of sample sizes are all largely dependent on $(\sigma_u^2-\gamma)$ (see Chapter 4). Estimation of the level-2 variance though, can be problematic especially if the number of level-2 groups J is small and/or the level-2 variance is close to zero; it could be negative in the classical approach, while in the fully Bayesian approach the choice of a suitable prior for the level-2 variance is not straightforward (see Chapter 2).

In this section we shall develop a Bayes linear estimator of the population level-2 variance of the SOEREF model for unbalanced data which we term Bayes Linear Minimum Variance Estimator (BLIMVE). We begin by considering the simplest case where the population group j mean $\mathcal{M}(y_j)$ is known.

6.2 Assessing the population level-2 variance with known population mean

The Bayes Linear Minimum Variance Estimator involves some complex expressions in the calculations of fourth order quantities. Hence, for ease of exposition, we use the simpler notation of the Second Order Exchangeable Regression (SOEREG) model of which the SOEREF model is a special case (see Chapter 3) as follows.

Definition 6.2.1. *Let y_{ji} represent univariate outcome measurements on each individual i nested in group j. A* **Second-order exchangeable random effects**

(SOEREF) *model is given by either of the following representation form: Hierarchical form:*

$$\textit{Level-1}: y_{ji} = \beta_j + \epsilon_{ji} \tag{6.1}$$

$$\textit{Level-2}: \beta_j = \mathcal{M}(\beta) + \mathcal{R}_j(\beta). \tag{6.2}$$

Single-equation form:

$$y_{ji} = \mathcal{M}(\beta) + \mathcal{R}_j(\beta) + \epsilon_{ji} \tag{6.3}$$
$$i = 1, 2, \ldots, n_j, \quad \textit{and} \quad j = 1, 2, \ldots, J.$$

where $\mathcal{M}(\beta)$ is the population grand mean, β_j is the population group j mean, $\mathcal{R}_j(\beta)$ are the level 2 residuals, and ϵ_{ji} are the level 1 residuals.

The mean and variance of $\mathcal{M}(\beta)$ are

$$E(\mathcal{M}(\beta)) = \mu_\beta, \quad Var(\mathcal{M}(\beta)) = \gamma_\beta, \quad \gamma_\beta \geq 0. \tag{6.4}$$

The mean, variance and covariance of β_j are

$$E(\beta_j) = \mu_\beta, \quad Var(\beta_j) = \sigma^2_\beta, \quad \sigma^2_\beta \geq 0, \quad Cov(\beta_j, \beta_{j'}) = \gamma_\beta \quad \forall j \neq j'. \tag{6.5}$$

The collection of level-2 residuals $\mathcal{R}_1(\beta), \mathcal{R}_2(\beta), \ldots$ are SOE with

$$E(\mathcal{R}_j(\beta)) = 0, \quad Var(\mathcal{R}_j(\beta)) = \sigma^2_\beta - \gamma_\beta = V_{R_\beta}, \quad Cov(\mathcal{R}_j(\beta), \mathcal{R}_{j'}(\beta)) = 0,$$
$$\forall j \neq j', \tag{6.6}$$

and each $\mathcal{R}_j(\beta))$ is uncorrelated with $\mathcal{M}(\beta)$.

The collection of level-1 residuals $\epsilon_{11}, \epsilon_{12} \ldots$ are SOE with

$$E(\epsilon_{ji}) = 0, \quad Var(\epsilon_{ji}) = \sigma^2_\epsilon \quad \forall i, j, \tag{6.7}$$

and $\epsilon_{11}, \epsilon_{12} \ldots$ are mutually uncorrelated and are also uncorrelated with the level-2 residuals $\mathcal{R}_j(\beta))$.

Representation (6.3) corresponds to the form of the SOREF model $y_{ji} = \mathcal{M}(y) + \mathcal{R}_j(\mathcal{M}(y)) + \mathcal{R}_i(y_j)$ we have used so far.

Exchangeability representation for estimating the population level-2 variance

To learn about the population level-2 variance, we could use the level-2 representation (6.2) as follows. The group means β_j are not observable. If, however they were observable we could follow the same procedure as in the adjustment of the population level-1 variance. Let the squared level-2 residuals $[R_j(\beta)]^2 = V_{\beta_j}$, and suppose that we judge that the sequence $V_{\beta_1}, V_{\beta_2}, \ldots$ is second-order exchangeable. Hence, we may write the representation for V_{β_j} as follows

$$[\mathcal{R}_j(\beta)]^2 = V_{\beta_j} = \mathcal{M}(V_\beta) + \mathcal{R}_j(V_\beta) \tag{6.8}$$

where $\mathcal{M}(V_\beta)$ may be regarded as the underlying population level-2 variance with $E(\mathcal{M}(V_\beta)) = V_{R_\beta}$ and constant variance V_{M_β}. The sequence $\mathcal{R}_1(V_\beta), \mathcal{R}_2(V_\beta), \ldots$ is uncorrelated with zero mean and constant variance $V_{R(V_\beta)}$. Also each element of $\mathcal{R}_k(V_\beta)$ is uncorrelated with $\mathcal{M}(V_\beta)$. To learn about the level-2 population variance, we could construct the appropriate squared quantities using (5.31) as follows:

$$(\beta_j - \bar{\beta})^2 = (\mathcal{R}_j(\beta) - \bar{\mathcal{R}})^2, \tag{6.9}$$

where

$$\bar{\beta} = \frac{1}{J}\sum_{j=1}^{J} \beta_j \quad \text{and} \quad \bar{\mathcal{R}} = \frac{1}{J}\sum_{j=1}^{J} \mathcal{R}_j(\beta) \tag{6.10}$$

We could, in principle, decompose the right-hand side of (6.9) and obtain the joint prior assessments necessary to adjust the level 2 variance, but since the β_j's are not observable, this procedure is not feasible in practice. We therefore need to find estimates of the β_j's.

6.3 Construction of (within-group) estimators

We must therefore construct appropriate combinations of our 'observables', namely $\hat{\beta}_j$, which are informative for the population level-2 variance $\mathcal{M}(V_\beta)$.

Suppose we obtain ordinary least squares estimators $\hat{\beta}_j$ of β_j based on the data in group j, that is $y_{j\cdot} = \beta_j + \epsilon_{ji}$ (6.1) of Definition 5.11.1. Hence, for each group j the least squares estimator $\hat{\beta}_j = \bar{y}_{ji} = \sum_i y_{ji}/n_j$. Then we may write

$$\hat{\beta}_j = \beta_j + \delta_j,$$

where the residuals δ_j are uncorrelated with zero mean, and unequal variances as follows.

$$\begin{aligned} E(\delta_j) &= 0 \\ Var(\delta_j) &= \sigma_j^2 \\ Cov(\delta_j, \delta_{j'}) &= 0 \quad \forall j \neq j' \end{aligned} \tag{6.11}$$

We note that $Var(\hat{\beta}_j|\beta_j) = Var(\bar{y}_{ji}|\beta_j) = \sigma_\epsilon^2/n_j = \sigma_j^2$. Hence, we may substitute our Bayes linear estimator of the level-1 variance as obtained in Section 5.7. It is necessary to allow unequal variances (σ_j^2) to deal with unbalanced data, different groups having different sample sizes, n_j. Otherwise, the data may also be inherently heteroscedastic due to some classes having students of about the same ability (less variable performances) and other classes with students of mixed abilities (more variable performances). Whatever the reason for differences of the within-group variances, we wish to account for these differences in our Bayes linear estimator of $\mathcal{M}(V_\beta)$.

We define the within-group j squared quantity as follows

$$z_j = (\hat{\beta}_j - \bar{\hat{\beta}}_a)^2$$

where $\bar{\hat{\beta}}_a = \sum_j a_j \hat{\beta}_j$ and $\sum_j a_j = 1$. The weights a_j will be determined so that more informative groups, with relatively larger sample sizes for example, contribute more weight (larger values of a_j) in estimating $\mathcal{M}(V_\beta)$. The calculations of the weights a_j will be discussed in the sections that follows.

Replacing $\beta_j = \mathcal{M}(\beta) + \mathcal{R}_j(\beta)$ from the level-2 representation in the estimator $\hat{\beta}_j = \beta_j + \delta_j$ yields

$$\hat{\beta}_j = \mathcal{M}(\beta) + \mathcal{R}_j(\beta) + \delta_j \tag{6.12}$$

From (6.12), $\bar{\hat{\beta}}_a = \mathcal{M}(\beta) + \bar{\mathcal{R}}_a + \bar{\delta}_a$, where $\bar{\mathcal{R}}_a = \sum_j a_j \mathcal{R}_j(\beta)$ and $\bar{\delta}_a = \sum_j a_j \delta_j$. Hence, the within group quantities z_j that are informative for estimating $\mathcal{M}(V_\beta)$ can be written as follows

$$\begin{aligned} z_j &= (\hat{\beta}_j - \bar{\hat{\beta}}_a)^2 \\ &= [(\mathcal{R}_j(\beta) - \bar{\mathcal{R}}_a) + (\delta_j - \bar{\delta}_a)]^2 \\ &= (\mathcal{R}_j(\beta) - \bar{\mathcal{R}}_a)^2 + (\delta_j - \bar{\delta}_a)^2 + 2(\mathcal{R}_j(\beta) - \bar{\mathcal{R}}_a)(\delta_j - \bar{\delta}_a). \end{aligned} \tag{6.13}$$

The decomposition (6.13) shows that z_j above comprises the squared level-2 residual $(\mathcal{R}_j(\beta) - \bar{\mathcal{R}}_a)^2$ suitable to learn about $\mathcal{M}(V_\beta)$ via our representation (6.8), the squared level-1 residual $(\delta_j - \bar{\delta}_a)^2$ suitable to learn about σ_ϵ^2 and a cross-product term of level-2 and level-1 residuals. We group the sequence $z_1, z_2, \ldots, z_J$ in the vector $\mathbf{Z}$.

6.4 The Bayes Linear Minimum Variance Estimator

We now have a representation for the population level-2 variance and the corresponding observable quantities $\mathbf{Z}$ to adjust our prior belief about $\mathcal{M}(V_\beta)$. Applying the Bayes linear rule, we obtain the adjusted expectation of the level 2 variance as

$$E_{\mathbf{Z}}(\mathcal{M}(V_\beta)) = E(\mathcal{M}(V_\beta)) + Cov(\mathcal{M}(V_\beta), \mathbf{Z}) Var(\mathbf{Z})^{-1} (\mathbf{Z} - E(\mathbf{Z})) \tag{6.14}$$

with the corresponding adjusted variance

$$Var_{\mathbf{Z}}(\mathcal{M}(V_\beta)) = Var(\mathcal{M}(V_\beta)) - Cov(\mathcal{M}(V_\beta), \mathbf{Z}) Var(\mathbf{Z})^{-1} Cov(\mathbf{Z}, \mathcal{M}(V_\beta)) \tag{6.15}$$

Any choices of weights a_j with $\sum_j a_j = 1$ in $\mathbf{Z}$ would give an estimator of $\mathcal{M}(V_\beta)$ with its accuracy assessed by the adjusted variance but it seems intuitive to choose weights

that minimize this variance. Hence, we obtain the estimated population variance $E_{\mathbf{Z}}(\mathcal{M}(V_\beta))$ by finding the weights a_j that minimises the variance $Var_{\mathbf{Z}}(\mathcal{M}(V_\beta))$ or equivalently, that maximises $Cov(\mathcal{M}(V_\beta)), \mathbf{Z})Var(\mathbf{Z})^{-1}Cov(\mathbf{Z}, \mathcal{M}(V_\beta))$, subject to $\sum_j a_j = 1$.

Our approach belongs to the criteria-based procedures in that we specified a criterion first, namely the minimum adjusted variation, and then we developed the BLIMVE to satisfy this criterion. Criteria-based procedures for estimating variance components are discussed in Searle et al. (1992). One such procedure is the Minimum Variance Quadratic Unbiased Estimator (MINQUE) (Rao, 1971a) which, in principle, is analogous to BLIMVE.

For MINQUE, Rao considers the model $\mathbf{y} = \mathbf{X}\beta + \sum_{\mathbf{i}} \mathbf{Z_i u_i}$ and estimates a linear function of the variance components $\mathbf{p}'\boldsymbol{\sigma}^{\mathbf{2}}$ (e.g. σ^2_ϵ and σ^2_α) using a quadratic function of the data, namely $\mathbf{y}'\mathbf{A}\mathbf{y}$ (BLIMVE uses the collection of quadratic forms z_j). If the random vectors in $\mathbf{u_i}$ were known, then Rao states that a "natural" estimator of σ^2_i would be $\mathbf{u_i'u_i}/q_i$ where q_i is the order of $\mathbf{u_i}$. Thus the estimator of $\mathbf{p}'\boldsymbol{\sigma}^{\mathbf{2}}$ would be $\mathbf{p}'\tilde{\boldsymbol{\sigma}}^{\mathbf{2}} = \sum_i p_i \mathbf{u_i'u_i}/q_i$. However, using the quadratic form, the estimator of $\mathbf{p}'\boldsymbol{\sigma}^{\mathbf{2}}$ is $\mathbf{p}'\hat{\boldsymbol{\sigma}}^{\mathbf{2}} = \mathbf{y}'\mathbf{A}\mathbf{y} = \mathbf{u}'\mathbf{Z}'\mathbf{A}\mathbf{Z}\mathbf{u}$. Rao minimizes a weighted Euclidean norm of the difference $\mathbf{p}'\hat{\boldsymbol{\sigma}}^{\mathbf{2}} - \mathbf{p}'\tilde{\boldsymbol{\sigma}}^{\mathbf{2}}$. So both MINQUE and BLIMVE derive estimators by minimizing a variance. In contrast to BLIMVE, MINQUE does not require inverting a variance-covariance matrix. However, one major problem of MINQUE is that the solutions to the resulting equations are functions of the variance components (i.e. σ^2_ϵ and σ^2_α). Rao uses pre-assigned values of the variance components, say $\sigma^2_{\epsilon 0}$ and $\sigma^2_{\alpha 0}$ to calculate the MINQUE estimators. The latter estimators, however, are only minimum variance if the pre-assigned values are the correct variance components, i.e. $\sigma^2_{\epsilon 0} = \sigma^2_\epsilon$ and $\sigma^2_{\alpha 0} = \sigma^2_\alpha$ (see Swallow and Searle, 1978). In contrast, for Bayes linear estimation of variances there are suitable methods to specify prior variances and fourth-order quantities, as well as to check the sensitivity of our variance estimators to changes in our prior specifications (see Section 5.8 for example).

In the fully Bayesian approach for estimating the level-2 variance, there are problems in specifying suitable priors when the number of level-2 groups J is small (Gelman,2006). We next investigate the BLIMVE estimator for the two groups

situation.

6.5 BLIMVE for the two-group case

In order to investigate the BLIMVE for the two group situation, we need to calculate $E_{\mathbf{Z}}(\mathcal{M}(V_\beta)) = E(\mathcal{M}(V_\beta)) + Cov(\mathcal{M}(V_\beta), \mathbf{Z})Var(\mathbf{Z})^{-1}(\mathbf{Z} - E(\mathbf{Z}))$ for $\mathbf{Z} = (z_1, z_2)$. We start by calculating $Var(\mathbf{Z})^{-1}$, since it is an important part of the maximization criterion. We use the constraint $a_1 + a_2 = 1$ on the two groups to simplify our calculations. Substituting $a_1 + a_2 = 1$ in $z_j = (\bar{y}_j. - \bar{y}_a)^2$, the jth element of $\mathbf{Z}$, gives $z_1 = a_2^2(\bar{y}_1. - \bar{y}_2.)^2$ and $z_2 = a_1^2(\bar{y}_1. - \bar{y}_2.)^2$. Hence, we only need to calculate the variance of z_1 and multiply it by a_1^2 / a_2^2 to obtain the corresponding variance of z_2.

Using representation (6.13):

$$\begin{aligned}
z_1 &= [(\mathcal{R}_1(\beta) - (a_1\mathcal{R}_1(\beta) + a_2\mathcal{R}_2(\beta)) + (\delta_1 - (a_1\delta_1 + a_2\delta_2)]^2 \\
&= a_2^2[(\mathcal{R}_1(\beta) - \mathcal{R}_2(\beta)) + (\delta_1 - \delta_2)]^2 \quad (\text{since} \quad a_2 = 1 - a_1) \\
&= a_2^2[(\mathcal{R}_1(\beta) - \mathcal{R}_2(\beta))^2 + (\delta_1 - \delta_2)^2 + 2(\mathcal{R}_1(\beta) - \mathcal{R}_2(\beta))(\delta_1 - \delta_2)] \\
&= a_2^2[(\mathcal{R}_1(\beta)^2 + \mathcal{R}_2(\beta)^2 - 2\mathcal{R}_1(\beta)\mathcal{R}_2(\beta) + (\delta_1 - \delta_2)^2 \\
&+ 2(\mathcal{R}_1(\beta) - \mathcal{R}_2(\beta))(\delta_1 - \delta_2)] \\
&= 2a_2^2\mathcal{M}(V_\beta) + T_1 + T_1^\delta + T_1^{\delta\beta} \quad (\text{using representation (6.8)})
\end{aligned} \tag{6.16}$$

where,

$$\begin{aligned}
T_1 &= a_2^2[\mathcal{R}_1(V_\beta) + \mathcal{R}_2(V_\beta) - 2\mathcal{R}_1(\beta)\mathcal{R}_2(\beta)] \\
T_1^\delta &= a_2^2[\delta_1^2 + \delta_2^2 - 2\delta_1\delta_2] \\
T_1^{\delta\beta} &= 2a_2^2[(\mathcal{R}_1(\beta) - \mathcal{R}_2(\beta))(\delta_1 - \delta_2)].
\end{aligned} \tag{6.17}$$

We have the following expectations:

$$E(T_1) = 0, \quad E(T_1^\delta) = a_2^2(\sigma_1^2 + \sigma_2^2), \quad E(T_1^{\delta\beta}) = 0. \tag{6.18}$$

In addition to the uncorrelated properties of the residuals $\mathcal{R}_k(\beta), \mathcal{R}_k(V_\beta)$ (see Definition 6.2.1) and δ_k (see 6.11), we assume the following higher order uncorrelated

properties

$$Cov(\mathcal{M}(V_\beta), \mathcal{R}_j(V_\beta) = Cov(\mathcal{M}(V_\beta), \mathcal{R}_j(\beta)) = Cov(\mathcal{M}(V_\beta), \delta_j) = 0$$
$$Cov(\mathcal{R}_j(V_\beta), \mathcal{R}_j(\beta)) = Cov(\mathcal{R}_j(V_\beta), \mathcal{R}_j(\beta)\mathcal{R}_{j'}(\beta)) = Cov(\delta_j, \mathcal{R}_j(\beta)) = 0. \tag{6.19}$$

Using the above properties and the second-order specifications in (6.8) and writing $\mu_4 = E(\delta_j^4)$, the variances are:

$$\begin{aligned} Var(T_1) &= a_2^4[2V_{R(V_\beta)} + 4(V_{M_\beta} + V_{R_\beta}^2)] \\ Var(T_1^\delta) &= a_2^4[(\mu_4 - \sigma_1^4) + (\mu_4 - \sigma_2^4) + 4\sigma_1^2\sigma_2^2] \\ Var(T_1^{\delta\beta}) &= 8a_2^4(\sigma_1^2 + \sigma_2^2)V_{R_\beta} \end{aligned} \tag{6.20}$$

Hence, we obtain the variances of z_1 and z_2 (with a_1 in place of a_2) as follows.

$$\begin{aligned} Var(z_1) &= 4a_2^4 V_{M_\beta} + Var(T_1) + Var(T_1^\delta) + Var(T_1^{\delta\beta}) \\ &= 4a_2^4 V_{M_\beta} + a_2^4[2V_{R(V_\beta)} + 4(V_{M_\beta} + V_{R_\beta}^2)] \\ &\quad + a_2^4[(\mu_4 - \sigma_1^4) + (\mu_4 - \sigma_2^4) + 4\sigma_1^2\sigma_2^2] \\ &\quad + 8a_2^4(\sigma_1^2 + \sigma_2^2)V_{R_\beta} \\ Var(z_2) &= 4a_1^4 V_{M_\beta} + Var(T_2) + Var(T_2^\delta) + Var(T_2^{\delta u}) \\ &= 4a_1^4 V_{M_\beta} + a_1^4[2V_{R(V_\beta)} + 4(V_{M_\beta} + V_{R_\beta}^2)] \\ &\quad + a_1^4[(\mu_4 - \sigma_1^4) + (\mu_4 - \sigma_2^4) + 4\sigma_1^2\sigma_2^2] \\ &\quad + 8a_1^4(\sigma_1^2 + \sigma_2^2)V_{R_\beta} \end{aligned}$$

Replacing a_2^2 by a_1^2 in the representation for z_1 in (6.16) we obtain the representation for z_2.

$$z_2 = 2a_1^2\mathcal{M}(V_\beta) + T_2 + T_2^\delta + T_2^{\delta\beta}, \tag{6.21}$$

where all the terms in T_2 are similar to those of T_1 with a_1 in place of a_2. We calculate the following covariances:

$$\begin{aligned} Cov(T_1, T_2) &= a_1^2 a_2^2[2V_{R(V_\beta)} + 4(V_{M_\beta} + V_{R_\beta}^2)] \\ Cov(T_1^\delta, T_2^\delta) &= a_1^2 a_2^2[(\mu_4 - \sigma_1^4) + (\mu_4 - \sigma_2^4) + 4\sigma_1^2\sigma_2^2)] \\ Cov(T_1^{\delta\beta}, T_2^{\delta\beta}) &= 8a_1^2 a_2^2(\sigma_1^2 + \sigma_2^2)V_{R_\beta}. \end{aligned} \tag{6.22}$$

Hence,

$$\begin{aligned} Cov(z_1, z_2) &= 4a_1^2 a_2^2 V_{M_\beta} + Cov(T_1, T_2) + Cov(T_1^\delta, T_2^\delta) + Cov(T_1^{\delta\beta}, T_2^{\delta\beta}) \\ &= 4a_1^2 a_2^2 V_{M_\beta} + a_1^2 a_2^2 [2V_{R(V_\beta)} + 4(V_{M_\beta} + V_{R_\beta}^2)] \\ &+ a_1^2 a_2^2 [(\mu_4 - \sigma_1^4) + (\mu_4 - \sigma_2^4) + 4\sigma_1^2\sigma_2^2] \\ &+ 8a_1^2 a_2^2 (\sigma_1^2 + \sigma_2^2) V_{R_\beta} \end{aligned}$$

We may write the above expressions more simply as:

$$\begin{aligned} Var(z_1) = a_2^4 4(V_{M_\beta} + V_T), \quad Var(z_2) = a_1^4 4(V_{M_\beta} + V_T), \quad \text{and} \\ Cov(z_1, z_2) = a_1^2 a_2^2 4(V_{M_\beta} + V_T), \end{aligned}$$

giving

$$Var(\mathbf{Z}) = 4(V_{M_\beta} + V_T) \begin{pmatrix} a_2^4 & a_1^2 a_2^2 \\ a_1^2 a_2^2 & a_1^4 \end{pmatrix} \tag{6.23}$$

where

$$\begin{aligned} 4(V_{M_\beta} + V_T) &= 4V_{M_\beta} + [2V_{R(V_\beta)} + 4(V_{M_\beta} + V_{R_\beta}^2)] \\ &+ [(\mu_4 - \sigma_1^4) + (\mu_4 - \sigma_2^4) + 4\sigma_1^2\sigma_2^2] + 8(\sigma_1^2 + \sigma_2^2) V_{R_\beta}, \end{aligned}$$

with

$$V_T = \frac{1}{4}(Var(T_j) + Var(T_j^\delta) + Var(T_j^{\delta\beta})) \quad \forall j = 1, 2.$$

Clearly, $Var(\mathbf{Z})$ is singular. In such a case we use the Moore-Penrose generalized inverse $Var(\mathbf{Z})^\dagger$, namely the generalized inverse constructed from the space of positive eigenvectors as in Goldstein and Wooff (2006).

We write

$$\mathbf{\Sigma} = \begin{pmatrix} a_2^4 & a_1^2 a_2^2 \\ a_1^2 a_2^2 & a_1^4 \end{pmatrix}$$

Since $\mathbf{\Sigma}$ is of rank one it can be shown that the singular value decomposition method for calculating the Moore-Penrose generalized inverse simplifies to

$$\mathbf{\Sigma}^\dagger = trace(\mathbf{\Sigma}^T \mathbf{\Sigma})^{-1} \mathbf{\Sigma}^T.$$

Since $\mathbf{\Sigma}$ is symmetric, we have

$$\mathbf{\Sigma}^{\dagger} = \frac{1}{(a_1^4 + a_2^4)^2} \begin{pmatrix} a_2^4 & a_1^2 a_2^2 \\ a_1^2 a_2^2 & a_1^4 \end{pmatrix}$$

The existence and uniqueness of the Moore-Penrose inverse was established by Penrose (1955) via four conditions as follows.

Corresponding to any $m \times n$ matrix $\mathbf{A}$ there is a unique $n \times m$ matrix $\mathbf{A}^{\dagger}$ such that

1. $\mathbf{A}\mathbf{A}^{\dagger}\mathbf{A} = \mathbf{A}$ (i.e. $\mathbf{A}^{\dagger}$ is a generalized inverse of $\mathbf{A}$);
2. $\mathbf{A}^{\dagger}\mathbf{A}\mathbf{A}^{\dagger} = \mathbf{A}^{\dagger}$ (i.e. $\mathbf{A}$ is a generalized inverse of $\mathbf{A}^{\dagger}$);
3. $(\mathbf{A}\mathbf{A}^{\dagger})^T = \mathbf{A}\mathbf{A}^{\dagger}$ (i.e. $\mathbf{A}\mathbf{A}^{\dagger}$ is symmetric);
4. $(\mathbf{A}^{\dagger}\mathbf{A})^T = \mathbf{A}^{\dagger}\mathbf{A}$ (i.e. $\mathbf{A}^{\dagger}\mathbf{A}$ is symmetric).

It is straightforward to verify that $\mathbf{\Sigma}^{\dagger}$ satisfies the above four conditions. In particular, satisfaction of conditions (1) and (4) above implies that $\mathbf{\Sigma}^{\dagger}$ is a minimum norm generalized inverse of $\mathbf{\Sigma}$ (see Harville, 1997).

We now maximize the BLIMVE condition using the generalized inverse, that is replacing $Var(\mathbf{Z})^{-1}$ in (6.23) by $(4(V_{M_\beta} + V_T))^{-1}\mathbf{\Sigma}^{\dagger}$. From representation (6.16)

$$Cov(\mathcal{M}(V_\beta)), \mathbf{Z}) = 2V_{M_\beta} \begin{pmatrix} a_2^2 & a_1^2 \end{pmatrix}.$$

Maximization of $Cov(\mathcal{M}(V_\beta)), \mathbf{Z})Var(\mathbf{Z})^{-1}Cov(\mathbf{Z}, \mathcal{M}(V_\beta))$ gives

$$\begin{aligned} &\frac{4V_{M_\beta}^2(4(V_{M_\beta} + V_T))^{-1}}{(a_1^4 + a_2^4)^2} \begin{pmatrix} a_2^2 & a_1^2 \end{pmatrix} \begin{pmatrix} a_2^4 & a_1^2 a_2^2 \\ a_1^2 a_2^2 & a_1^4 \end{pmatrix} \begin{pmatrix} a_2^2 \\ a_1^2 \end{pmatrix} \\ &= \frac{V_{M_\beta}^2}{(V_{M_\beta} + V_T)}. \end{aligned} \tag{6.24}$$

Since the above result is free of a_j, there is no minimum adjusted variance. However, we can still make inferences on the level-2 variance as follows.

The adjusted mean is

$$\begin{aligned}
E_{\mathbf{Z}}(\mathcal{M}(V_\beta)) &= E(\mathcal{M}(V_\beta)) + Cov(\mathcal{M}(V_\beta), \mathbf{Z}) Var(\mathbf{Z})^\dagger (\mathbf{Z} - E(\mathbf{Z})) \\
&= V_{R_\beta} + \frac{2V_{M_\beta}(4(V_{M_\beta} + V_T))^{-1}}{(a_1^4 + a_2^4)^2} \begin{pmatrix} a_2^2 & a_1^2 \end{pmatrix} \begin{pmatrix} a_2^4 & a_1^2 a_2^2 \\ a_1^2 a_2^2 & a_1^4 \end{pmatrix} \\
&\quad \begin{pmatrix} a_2^2(\bar{y}_{1.} - \bar{y}_{2.})^2 - a_2^2(2V_{R_\beta} + \sigma_1^2 + \sigma_2^2) \\ a_1^2(\bar{y}_{1.} - \bar{y}_{2.})^2 - a_1^2(2V_{R_\beta} + \sigma_1^2 + \sigma_2^2) \end{pmatrix} \\
&= V_{R_\beta} + 2(4(V_{M_\beta} + V_T))^{-1} V_{M_\beta}((\bar{y}_{1.} - \bar{y}_{2.})^2 - (2V_{R_\beta} + \sigma_1^2 + \sigma_2^2)) \\
&= \frac{\frac{1}{2}V_{M_\beta}[(\bar{y}_{1.} - \bar{y}_{2.})^2 - (\sigma_1^2 + \sigma_2^2)] + V_{R_\beta} V_T}{V_{M_\beta} + V_T}. \qquad (6.25)
\end{aligned}$$

Using (6.24) we calculate the adjusted variance as

$$Var_{\mathbf{Z}}(\mathcal{M}(V_\beta)) = \frac{V_{M_\beta} V_T}{V_{M_\beta} + V_T} \qquad (6.26)$$

Since here the adjusted mean depends on only two groups, we would consider giving more weight to our priors.

The choice of a prior distribution is also an issue in the full Bayes analysis of the random effects model $y_{ji} = \mu + \alpha_j + \epsilon_{ji}$ for a small number of groups J. For example, if the level-2 variance σ_α^2 is close to zero, it is appropriate to use a uniform prior on σ_α. However, for $J = 2$ this leads to an improper posterior density with the undesirable consequence that $\sigma_\alpha = \infty$, resulting in no shrinkage. The level-2 variance primarily controls by how much the adjusted population group j mean is shrunk towards the overall mean (Section 4.4.2). (Gelman, 2006) argues that this lack of shrinkage is to be expected of the Bayes posterior estimator when the number of groups J is small, and is consistent with the conclusion of James and Stein (1960) that unshrunken estimators are admissible if the number of groups $J < 3$.

6.6 BLIMVE for two or more groups

The calculations of variances and covariances of the elements of $Var(\mathbf{Z})$ for the two-group case in section (6.5) are complex. To avoid such complexity in extending the BLIMVE to the general J group case, we shall make use of vectors of weights $\mathbf{a_j}$ in

deriving the elements of $Var(\mathbf{Z})$. We state the following theorem for the calculation of $Var(\mathbf{Z})$ for two or more groups.

Theorem 6.6.1. *The variances and covariances of the elements z_j of $Var(\mathbf{Z})$ for $J \geq 2$ groups are:*

$$\begin{aligned} Cov(z_i, z_j) &= (\mathbf{a_i}^T\mathbf{a_i})(\mathbf{a_j}^T\mathbf{a_j})V_{M_\beta} + Cov(T_i, T_j) + Cov(T_i^\delta, T_j^\delta) \\ &\quad + Cov(T_i^{\delta\beta}, T_j^{\delta\beta}), \quad \forall i, j = 1, 2, \ldots, J, \end{aligned} \tag{6.27}$$

with

$$\begin{aligned} Cov(T_i, T_j) &= (\mathbf{a_i}^2)^T(\mathbf{a_j}^2)(V_{R(V_\beta)} - 2(V_{M_\beta} + V_{R_\beta}^2)) + (\mathbf{a_i}^T\mathbf{a_j})^2 2(V_{M_\beta} + V_{R_\beta}^2), \\ Cov(T_i^\delta, T_j^\delta) &= 2[(\mathbf{a_i}^2)^T G S^2 \mathbf{a_j}^2 + (\mathbf{a_i}^T S \mathbf{a_j})^2], \\ Cov(T_i^{\delta\beta}, T_j^{\delta\beta}) &= 4V_{R_\beta}(\mathbf{a_i}^T\mathbf{a_j})(\mathbf{a_i}^T S \mathbf{a_j}), \end{aligned} \tag{6.28}$$

where $\mathbf{a_i}$ is a vector of weights $(a_1, a_2, \ldots, (a_i - 1), \ldots, a_J)$ with each element $0 < a_i < 1$ and $\sum_{i=1}^{J} a_i = 1$, and $\mathbf{a_i}^2$ are the squared elements of $\mathbf{a_i}$. G and S are diagonal matrices of excess kurtosis $0.5(\gamma_1, \gamma_2, \ldots, \gamma_J)$ and level-1 variances $(\sigma_1^2, \sigma_2^2, \ldots, \sigma_J^2)$ respectively. S^2 contains the squared level-1 variances.

We prove the theorem for $J = 2$ which can then be extended to $J = 3$ or more groups. We first calculate representation (6.16) but without replacing $a_1 + a_2 = 1$.

Proof.

$$\begin{aligned} z_1 &= [(\mathcal{R}_1(\beta) - \sum_{j=1}^{2} a_j\mathcal{R}_j(\beta)) + (\delta_1 - \sum_{j=1}^{2} a_j\delta_j)]^2. \\ &= [(\mathcal{R}_1(\beta) - \sum_{j=1}^{2} a_j\mathcal{R}_j(\beta))]^2 + [(\delta_1 - \sum_{j=1}^{2} a_j\delta_j)]^2 \\ &\quad + 2[(\mathcal{R}_1(\beta) - \sum_{j=1}^{2} \mathcal{R}_j(\beta))][(\delta_1 - \sum_{j=1}^{2} \delta_j)]. \end{aligned}$$

We consider each of the squared and cross-product terms in turn.

$$\begin{aligned} [(\mathcal{R}_1(\beta) - \sum_{j=1}^{2} \mathcal{R}_j(\beta))]^2 &= (a_1 - 1)^2\mathcal{R}_1^2(\beta) + a_2^2\mathcal{R}_2^2(\beta) + 2a_2(a_1 - 1)\mathcal{R}_1(\beta)\mathcal{R}_2(\beta) \\ &= ((a_1 - 1)^2 + a_2^2)\mathcal{M}(V_\beta) + T_1, \end{aligned}$$

using $[\mathcal{R}_j(\beta)]^2 = \mathcal{M}(V_\beta) + \mathcal{R}_j(V_\beta)$, and where

$$T_1 = (a_1 - 1)^2\mathcal{R}_1(V_\beta) + a_2^2\mathcal{R}_2(V_\beta) + 2a_2(a_1 - 1)\mathcal{R}_1(\beta)\mathcal{R}_2(\beta). \tag{6.29}$$

For the squared quantity involving δ, we have

$$T_1^\delta = [(\delta_1 - \sum_{j=1}^{2}\delta_j)]^2 = (a_1 - 1)^2\delta_1^2 + a_2^2\delta_2^2 + 2a_2(a_1 - 1)\delta_1\delta_2,$$

and for the cross-product of δ and $\mathcal{R}_j(\beta))$,

$$\begin{aligned} T_1^{\delta\beta} &= 2[(\mathcal{R}_1(\beta) - \sum_{j=1}^{2}\mathcal{R}_j(\beta))][(\delta_1 - \sum_{j=1}^{3}\delta_j)] \\ &= 2((a_1 - 1)\mathcal{R}_1(\beta) + a_2\mathcal{R}_2(\beta))((a_1 - 1)\delta_1 + a_2\delta_2) \end{aligned}$$

Hence,

$$\begin{aligned} z_1 &= ((a_1 - 1)^2 + a_2^2 + a_3^2)\mathcal{M}(V) + T_1 + T_1^\delta + T_1^{\delta\beta} \\ &= (\mathbf{a_1}^T\mathbf{a_1})\mathcal{M}(V) + T_1 + T_1^\delta + T_1^{\delta\beta}, \end{aligned} \tag{6.30}$$

where the vector $\mathbf{a_1}^T = [(a_1 - 1), a_2]$. In general, for J groups we have $\mathbf{a_i}^T = [a_1, a_2, \ldots, (a_i - 1), a_{i+1}, \ldots, a_J]$.

Similar derivations can easily be obtained for z_2, which is the same as z_1 with $\mathbf{a_1}^T$ replaced by $\mathbf{a_2}^T = [a_1, (a_2 - 1)]$ in all the components as shown below.

$$\begin{aligned} z_2 &= (a_1^2 + (a_2 - 1)^2)\mathcal{M}(V) + T_2 + T_2^\delta + T_2^{\delta\beta} \\ T_2 &= a_1^2\mathcal{R}_1(v) + (a_2 - 1)^2\mathcal{R}_2(v) + 2a_1(a_2 - 1)\mathcal{R}_1(\beta)\mathcal{R}_2(\beta) \\ T_2^\delta &= a_1^2\delta_1^2 + (1 - a_2)^2\delta_2^2 + 2a_1(a_2 - 1)\delta_1\delta_2 \\ T_2^{\delta\beta} &= 2(a_1\mathcal{R}_1(\beta) + (a_2 - 1)\mathcal{R}_2(\beta))(a_1\delta_1 + (a_2 - 1)\delta_2) \end{aligned} \tag{6.31}$$

To calculate the elements of $Var(\mathbf{Z})$, we calculate the following expectations:

$$\begin{aligned} &E(T_1) = 0, \quad E(T_2) = 0, \quad E(T_1^{\delta\beta}) = 0, \quad E(T_2^{\delta\beta}) = 0 \\ &E(T_1^\delta) = (a_1 - 1)^2\sigma_1^2 + a_2^2\sigma_2^2, \quad E(T_2^\delta) = a_1^2\sigma_1^2 + (a_2 - 1)^2\sigma_2^2. \end{aligned} \tag{6.32}$$

Hence, $Cov(T_1, T_2) = E(T_1T_2)$, $Cov(T_1^\delta, T_2^\delta) = E(T_1^\delta T_2^\delta)$ and $Cov(T_1^{\delta\beta}, T_2^{\delta\beta}) = E(T_1^{\delta\beta}T_2^{\delta\beta}) - E(T_1^{\delta\beta})E(T_2^{\delta\beta})$.

We calculate $Cov(T_1, T_2)$ which will also give the variances. Below we write the cross-product in a_1 and a_2 as a squared term.

$$\begin{aligned}
Cov(T_1, T_2) &= (a_1^2(a_1-1)^2 + a_2^2(a_2-1)^2)V_{R(v)} \\
&\quad + 4(a_1a_2(a_1-1)(a_2-1))(V_M + V_R^2) \\
&= (a_1^2(a_1-1)^2 + a_2^2(a_2-1)^2)(V_{R(V_\beta)} - 2(V_{M_\beta} - V_{R_\beta}^2)) \\
&\quad + 2(V_{M_\beta} - V_{R_\beta}^2)\Big((a_1^2(a_1-1)^2 + a_2^2(a_2-1)^2) \\
&\quad + 2(a_1a_2(a_1-1)(a_2-1))\Big) \\
&= (\mathbf{a_1}^2)^T\mathbf{a_2}^2\big(V_{R(V_\beta)} - 2(V_{M_\beta} - V_{R_\beta}^2)\big) + (\mathbf{a_1}^T\mathbf{a_2})^2 2(V_{M_\beta} - V_{R_\beta}^2)) \quad (6.33)
\end{aligned}$$

Below we also write the cross-product in a_1 and a_2 as a squared term.

$$\begin{aligned}
Cov(T_1^\delta, T_2^\delta) &= a_1^2(a_1-1)^2(\mu_4 - \sigma_1^4) + a_2^2(a_2-1)^2(\mu_4 - \sigma_2^4) \\
&\quad + 4(a_1a_2(a_1-1)(a_2-1)\sigma_1^2\sigma_2^2 \\
&= 2\Big(a_1^2(a_1-1)^2(\gamma_1\sigma_1^4)/2 + a_2^2(a_2-1)^2(\gamma_2\sigma_2^4)/2 \\
&\quad + a_1^2(a_1-1)^2\sigma_1^4 + a_2^2(a_2-1)^2\sigma_2^4 + 2(a_1a_2(a_1-1)(a_2-1)\sigma_1^2\sigma_2^2\Big) \\
&= 2((\mathbf{a_1}^2)^T GS^2\mathbf{a_2}^2) + (\mathbf{a_1}^T S\mathbf{a_2})^2, \quad (6.34)
\end{aligned}$$

after substituting $(\mu_4 - \sigma_j^4) = 2(1 + \frac{\gamma_j}{2})\sigma_j^4$, where γ_j is the excess kurtosis in group j.

$$\begin{aligned}
Cov(T_1^{\delta\beta}, T_2^{\delta\beta}) &= 4(a_1(a_1-1) + a_2(a_2-1)) \\
&\quad + (a_1(a_1-1)\sigma_1^2 + a_2(a_2-1)\sigma_2^2)V_{R_\beta} \\
&= 4(\mathbf{a_1}^T\mathbf{a_2})(\mathbf{a_1}^T S\mathbf{a_2}) \quad (6.35)
\end{aligned}$$

Finally, all the variances and covariances of $Var(\mathbf{Z})$ can be obtained from:

$$\begin{aligned}
Cov(z_1, z_2) &= ((a_1-1)^2 + a_2^2)(a_1^2 + (a_2-1)^2)V_{M_\beta} \\
&\quad + Cov(T_1, T_2) + Cov(T_1^\delta, T_2^\delta) + cov(T_1^{\delta\beta}, T_2^{\delta\beta}) \\
&= (\mathbf{a_1}^T\mathbf{a_1})(\mathbf{a_2}^T\mathbf{a_2})V_{M_\beta} + Cov(T_1, T_2) + Cov(T_1^\delta, T_2^\delta) + cov(T_1^{\delta\beta}, T_2^{\delta\beta}),
\end{aligned}$$

where the covariances are obtained from (6.32),(6.33) and (6.34).

Extending the proof for more than two groups follows the same principles as above. For $J = 3$ for example, the representation for z_1 below is the same as for $J = 2$ summing over 3 instead of 2 groups,

$$z_1 = [(\mathcal{R}_1(\beta) - \sum_{j=1}^{3} a_j \mathcal{R}_j(\beta)) + (\delta_1 - \sum_{j=1}^{3} a_j \delta_j)]^2.$$

The variances and covariances are functions of the weight vectors. For example, $\mathbf{a_1}^T = ((a_1 - 1, a_2, a_3)$, of which $J = 2$ is a special case. ■

6.7 Stochastic Optimization of the Bayes Linear Minimum Variance Estimator

BLIMVE requires maximizing the objective function (see Section 5.13)

$$\phi(\mathbf{a}) = Cov(\mathcal{M}(V_\beta)), \mathbf{Z})Var(\mathbf{Z})^{-1}Cov(\mathbf{Z}, \mathcal{M}(V_\beta))$$

over the elements of the vector $\mathbf{a} = (a_1, a_2, \ldots, a_J)$ subject to $\sum_{j=1}^{J} a_j = 1$ where $0 < a_j < 1$. This is a non-linear constrained optimization problem involving the inverse of a matrix function. We solve this problem by designing a stochastic local search algorithm (see for example, Gilli et al., 2019)as follows.

1. generate an initial solution $\mathbf{a}^i$ using a suitable criterion
2. **while** stopping condition not met **do**
3. create new solution $\mathbf{a}^n = N(\mathbf{a}^i)$
4. **if** $\phi(\mathbf{a}^n) \geq \phi(\mathbf{a}^i)$ **then**
5. $\mathbf{a}^i = \mathbf{a}^n$
6. **end while**
7. return $\mathbf{a}^i$

We experiment with the search algorithms to allow us to decide on each of the above steps and also to prevent it getting stuck near a local maximum. For step 1, an initial solution proportional to the level-1 sample sizes performed well. Groups with higher sample sizes, hence lower level-1 variances $\sigma_j^2 = \sigma_\epsilon^2/n_j$, get more weight in the adjustment of the level-2 population variance. For example, for J=3 groups, with level-1 sample sizes $n_j = (10, 20, 30)$ and total sample 60, $\mathbf{a}^i = (\frac{1}{6}, \frac{2}{6}, \frac{3}{6})$ ensuring $\sum_{j=1}^{J} a_j = 1$ and $0 < a_j < 1$.

For step 2, we found that the surface of $\phi(\mathbf{a})$ is rather flat and changes marginally as a function of $\mathbf{a}$, especially near the maximum. Also, searching in the neighbourhood of a potential solution speeds up the search procedure, hence the inclusion of step 3. The stopping condition of 500 iterations is enough to locate the maximum.

In step 3, we modify the initial solution $\phi(\mathbf{a}^i)$ using a neighbourhood function $N(.)$ as follows. We generate a perturbation vector $\mathbf{p}$ of J random uniform variates $U(-\omega, +\omega)$, scale $\mathbf{p}$ to sum to zero and calculate a new solution $\mathbf{a}^n = \mathbf{a}^i + \mathbf{p}$. To continue the above example, $\mathbf{a}^n = (\frac{1}{6}, \frac{2}{6}, \frac{3}{6}) + (0.03, -0.04, 0.01)$; where $\mathbf{p}$ was generated from $U(-0.1, 0.1)$ scaled to leave the condition on $\mathbf{a}^n$ unchanged. Experimentation allows the choice of a suitable value for ω.

Steps 4 and 5 ensure that a maximum is found and stored.

6.8 Prior specification of level 2 quantities

Calculation of $\phi(\mathbf{a})$ also requires that we specify the level-2 quantities V_{R_β}, $V_{R(V_\beta)}$ and V_{M_β}. We consider these in turn.

Specification of the prior level-2 variance V_{R_β} for the STAT1010 data was discussed in detail in Section 3.7.3 leading to a value of $V_{R_\beta} = 59$.

To specify $V_{R(V_\beta)}$, we adopt the same procedure as for the corresponding level-1 quantity $V_{R(V_\epsilon)}$ as in Section 5.3. That is we assume that the level-2 population variance $\mathcal{M}(V_\beta)$ acts like a scale parameter so that $\mathcal{R}_j(\beta) = \sqrt{\mathcal{M}(V_\beta)} Z_j \quad \forall j$. The sequence $Z_1, Z_2, \ldots$ are mutually independent with mean zero and constant variance one and are also independent of $\mathcal{M}(V_\beta)$. Using $[\mathcal{R}_j(\beta)]^2 = \mathcal{M}(V_\beta) + \mathcal{R}_j(V_\beta)$ and following the same calculations as in Section 5.3, we obtain

$$V_{R(V_\beta)} = (V_{M_\beta} + V_{R_\beta}^2)(Kur(Z_j) - 1), \tag{6.36}$$

where we may choose the kurtosis of Z_j in accordance with a Gaussian distribution. For a higher kurtosis, we may use a t distribution with a small ν degrees of freedom, whereas for a smaller kurtosis, a uniform distribution may be suitable as explained in Section 5.3.

To specify V_{M_β}, we use $V_{M_\beta} = cV_{R_\beta}^2$ and follow the same principle as in Section 5.8. We write (5.29) as for a single-level data set as we are focusing on data z_j for each of the J groups as follows.

$$m = \frac{\kappa J(c+1)}{(J-1)c}, \tag{6.37}$$

where m is the notional equivalent level-2 sample size, i.e. we consider our prior about $\mathcal{M}(V_\beta)$ to be worth m observations.

As for the level-1 quantities in $\phi(\mathbf{a})$, they are as specified and estimated in the adjustment of the level-1 population variance $\mathcal{M}(V_\epsilon)$ (see Section 5.9). Hence, we take $\sigma_j^2 = \hat{\sigma}_\epsilon^2/n_j$ and the excess kurtosis $\gamma_j = 0$ for example, as we chose for adjusting $\mathcal{M}(V_\epsilon)$.

6.9 Validation of the algorithm for BLIMVE

Before applying BLIMVE to the STAT1010 data, we use simulation to check whether the method performs as it is intended to.

But first we check for programming error in $\phi(\mathbf{a})$ by comparing its output against the exact calculation for the two-group case in Section 6.5. We select the first two classes of the STAT1010 data and calculate the squared observations $z_j = (\bar{y}_{1.} - \bar{y}_{2.})^2$ and choose the level-1 estimate $\hat{\sigma}_\epsilon^2 = \mathrm{E}_{\hat{\sigma}_\epsilon^2}(\mathcal{M}(V_\epsilon)) = 229.31$. We specify the prior level-2 variance $V_{R_\beta} = 59$ and the fourth-order quantities V_{M_β} and $V_{R(V_\beta)}$ as explained in Section 5.17. Using these values in (6.25) and (6.26) we obtain $E_{\mathbf{Z}}(\mathcal{M}(V_\beta)) = 43.5$ and $Var_{\mathbf{Z}}(\mathcal{M}(V_\beta)) = 5335.8$, which are exactly the same as returned by our R function $\phi(\mathbf{a})$. Incidentally, the ANOVA estimate of the level-2 variance for this same data is negative.

To check whether BLIMVE performs as it is intended to, we use simulated-data experimentation, also referred to as "fake" data simulation (see Chapter 8 of Gelman and Hill, 2007), as follows. We fix "true" values of the parameters in our SOEREF model $y_{ji} = \mathcal{M}(y) + \mathcal{R}_j(\beta) + \epsilon_{ji}$ and use these values to simulate unbalanced data with the same number of classes J and students n_j as in the STAT1010 data. We assume a multivariate Gaussian distribution for y_{ji}. To ensure that the true parameter values are consistent with the STAT1010 data, we fit a multilevel model using lme4 in R to the actual STAT1010 data and use the estimated parameters as our true values. Hence, we fix the overall mean $\mathcal{M}(y) = 54$, the level-1 variance $Var(\epsilon_{ji}) = \sigma_\epsilon^2 = 227$ and the level-2 variance $Var(\mathcal{R}_j(\beta)) = V_{R_\beta} = 80$.

To adjust the level-2 variance using the simulated data, we choose $\hat{\sigma}_\epsilon^2 = 229.31$ and the prior variance $V_{R_\beta} = 59$. With only seven groups (classes), we downgrade the prior by putting $c = 2$ in $V_{M_\beta} = cV_{R_\beta}^2$ and $\kappa = 2$ in (6.37) resulting in the nominal prior level-2 sample size $m = 4$ while the actual sample size is $J = 7$. We specify $V_{R(V_\beta)}$ using (6.36). These specifications will fix the parameters of $\phi(\mathbf{a})$ so that we need to find the optimum $\mathbf{a}$ using the search algorithm of Section 6.7 only once for all our simulated datasets. Here the optimal $\mathbf{a}$ allocates equal weight 1/7 to all of the seven classes.

The results from one thousand simulations are shown in Figure 6.1. The adjusted

expectation has moved from the prior 59 towards the true value 80 with the mean of the simulated $E_{\mathbf{Z}}(\mathcal{M}(V_\beta))$ at 72.8. Hence, BLIMVE performs as intended.

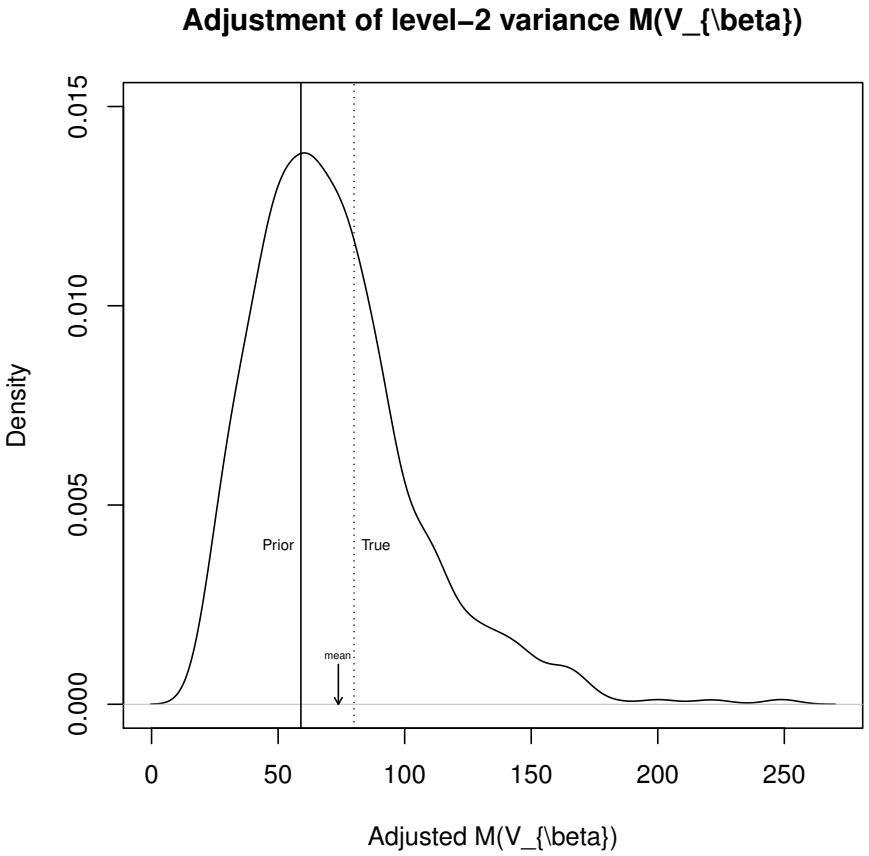


Figure 6.1: *The distribution of adjusted expectations of the level-2 variance $\mathcal{M}(V_\beta)$ using 1000 simulations. Each simulation has 269 students nested in 7 classes as in the STAT1010 data. The full line shows the prior variance $V_{R_\beta} = 59$, the dotted line the true variance 80 and the arrow, the mean of the 1000 adjusted variances $E_{\mathbf{Z}}(\mathcal{M}(V_\beta)) = 72.8$. BLIMVE estimates the population level-2 variance $\mathcal{M}(V_\beta)$ further from the prior and closer to the true variance.*

Figure 6.1 also shows a few large values of $E_{\mathbf{Z}}(\mathcal{M}(V_\beta))$. These result from adjustments based on large squared observations $z_j = (\bar{y}_{j\cdot} - \bar{y}_a)^2$, which themselves result from disparities between one specific class mean $\bar{y}_{j\cdot}$ and the weighted average of all seven class means $\bar{y}_a$. As an example, $E_{\mathbf{Z}}(\mathcal{M}(V_\beta)) = 161.8$ resulted from $z_j = (497.3, 1.4, 88.0, 73.6, 76.3, 74.8, 518.5)$. Of the 1000 $E_{\mathbf{Z}}(\mathcal{M}(V_\beta))$, there are only 21 (2.1%) larger than 160, i.e. twice the true value of $V_{R_\beta} = 80$.

With only $J = 7$ groups and $m = 4$, the prior constrains the distribution of

$E_{\mathbf{Z}}(\mathcal{M}(V_\beta))$ to be positive as in Figure 6.1. We have simulated data with $J = 30$ groups (classes) in similar configuration to the STAT1010 data and chose V_{M_β} with $c = 2$ allowing the data more influence ($m = 3$ v/s $J = 30$) on $E_{\mathbf{Z}}(\mathcal{M}(V_\beta))$. We found that it is possible to obtain a negative BLIMVE estimate of level-2 variance. While this is undesirable, in practice it may be indicative of problems with the multilevel model; a finite population model, for example may be more appropriate (see the discussion in Searle et al. (1992)).

6.10 Application of BLIMVE to the STAT1010 data

We now apply BLIMVE to estimate the population level-2 variance in the STAT1010 data, and in doing so we use the knowledge gained from the simulations in Section 6.9. Our simulations have revealed that the adjusted expectations $E_{\mathbf{Z}}(\mathcal{M}(V_\beta))$ vary a lot and may possibly be larger than our prior $V_{R_\beta} = 59$ despite our careful elicitation in Section 3.7.3. We feel quite uncertain about V_{R_β} and, although we have only seven data points (groups), we wish to let the sample dominate the prior. Thus we judge our prior information is worth $m = 4$ compared to the sample $J = 7$. Our simulations have also shown the possibility of large squared observations and therefore, we are also uncertain about the choice of a suitable kurtosis. To guide our choice, we consider a range of distributions for Z_j in the scaling $\mathcal{R}_j(\beta) = \sqrt{\mathcal{M}(V_\beta)}Z_j$. Thus we calculate the adjusted expectation and variance as shown below.

Table 6.1 shows that there is not much difference between the adjusted expectations since we allowed the sample information to dominate the prior. The adjusted variances, on the other hand, differ substantially; the smallest resulting from the uniform distribution. All three adjusted expectations are quite close to the average of the adjusted expectations of 72.8 from our 1000 simulated STAT1010 datasets (see Figure 6.1). Thus the population level-2 variance may be larger than our elicited prior value of 59 which should be increased by at least 20%, but we are still uncertain about this increase.

As a comparison, the full Bayes analysis of the random effects model for $J = 8$

	Distribution of scaled effects Z_j		
	Uniform	Gaussian	Scaled t_{10}
$Var(Z_j^2)$	0.8	2	3
Scaling factor c in $V_{M_\beta} = cV_{R_\beta}^2$	0.41	1.4	7
$Var(\mathcal{M}(V_\beta))$	1427.21	4873.4	24367
$E_{\mathbf{Z}}(\mathcal{M}(V_\beta))$	71.47	71.99	72.4
$Var_{\mathbf{Z}}(\mathcal{M}(V_\beta))$	606.67	1875.11	8697.54

Table 6.1: *Adjusted expectations and variances of the population level-2 variance $\mathcal{M}(V_\beta)$ of the STAT1010 data for varying kurtosis resulting from the Uniform, Gaussian and scaled t_{10} distributions. The prior variance is $E(\mathcal{M}(V_\beta)) = 59$. The prior information is judged to be worth a notional sample size $m = 4$ against the actual sample $J = 7$ classes.*

schools (groups) described in Section 5.14 also shows some quite large variations in the posterior density estimate of the level-2 variance σ_α^2 using a uniform prior density. Gelman (2006) concludes that for this data with only $J = 8$ groups, it is difficult to rule out the possibility of large values of σ_α^2 and that the uniform prior distribution seems closer to noninformative.

6.11 Two-stage Bayes linear analysis

In Chapter 4 we calculated the adjusted mean and variance of the population grand mean $\mathcal{M}(y)$ of the SOEREF model and showed that they depend on the prior level-1 and 2 variances. We may now perform a **two-stage Bayes linear analysis** by replacing the prior variances with their Bayes linear estimates; the revised estimates are termed **variance-modified Bayes linear assessments** (see Goldstein (1979,1983)). Below we illustrate the application of the two-stage analysis to estimate the population grand mean and the population group j means of the SOEREF model.

6.12 Two-stage Bayes linear analysis of the population grand mean

For our SOEREF model, if our judgement about the population grand mean $\mathcal{M}(y)$ is independent of our beliefs about the population level-1 and 2 variances, then the variance-modified Bayes linear assessments of the mean $E_{\bar{D}_n^*}(\mathcal{M}(y))$ in (6.38) and the variance $Var_{\bar{D}_n^*}(\mathcal{M}(y))$ in (6.39) will lead to improved adjustments over the corresponding original adjustment of the mean (4.10) and the variance (4.15) of $\mathcal{M}(y)$.

$$E_{\bar{D}_n^*}(\mathcal{M}(y)) = \frac{\gamma^{-1}\mu + \sum_{j=1}^{J}(\widehat{V}_{R_\beta} + \frac{\hat{\sigma}_\epsilon^2}{n_j})^{-1}\bar{y}_{j\cdot}}{\gamma^{-1} + \sum_{j=1}^{J}(\widehat{V}_{R_\beta} + \frac{\hat{\sigma}_\epsilon^2}{n_j})^{-1}}, \tag{6.38}$$

and

$$Var_{\bar{D}_n^*}(\mathcal{M}(y)) = \frac{1}{\gamma^{-1} + \sum_{j=1}^{J}(\widehat{V}_{R_\beta} + \frac{\hat{\sigma}_\epsilon^2}{n_j})^{-1}}. \tag{6.39}$$

$\widehat{V}_{R_\beta}$ and $\hat{\sigma}_\epsilon^2$ are our Bayes linear adjusted expectations of the population level-1 and 2 variances. For the STAT1010 data we have

$$E_{\bar{D}_n^*}(\mathcal{M}(y)) = 53.87 \quad Var_{\bar{D}_n^*}(\mathcal{M}(y)) = 9.33 \quad \text{Resolution} = 0.834.$$

compared to the original adjustments in Table 4.1

$$E_{\bar{D}_n}(\mathcal{M}(y)) = 53.86 \quad Var_{\bar{D}_n}(\mathcal{M}(y)) = 8.03 \quad \text{Resolution} = 0.857.$$

There is thus little change in the expectation of $\mathcal{M}(y)$, while the resolution has decreased slightly due to the increase in the estimate $\widehat{V}_{R_\beta}$= 71.99 from the prior $V_{R_\beta} = 59$ (while $\hat{\sigma}_\epsilon^2 = 229.31$ decreased from the prior $\sigma_\epsilon^2 = 237$). Hence, the two-stage analysis better accounts for the higher variability at level-2, that is the variation between classes. This indicates that in our prior judgements we over specified the prior level-1 variance (among students) and under specified the level-2 variance (between classes).

6.13 Two-stage Bayes linear analysis of the population group j means

In learning about the population group j means, and indeed in all subsequent analyses in Chapter 4, the shrinkage factor η played a pivotal role. If we now calculate the variance-modified shrinkage factor $\eta^* = \widehat{V}_{R_\beta}/(\widehat{V}_{R_\beta} + \hat{\sigma}^2_\epsilon/n_j)$, we obtain $\eta^* = 0.2389$ compared to $\eta = 0.1993$, which is based on prior variances. This difference is reflected in the variance-modified Bayes linear assessments for the population group j means as shown in Table 6.2 below.

Class	Management			Engineering			
Population means	$\mathcal{M}(y_1)$	$\mathcal{M}(y_2)$	$\mathcal{M}(y_3)$	$\mathcal{M}(y_4)$	$\mathcal{M}(y_5)$	$\mathcal{M}(y_6)$	$\mathcal{M}(y_7)$
Original	44.64	46.02	47.61	68.51	55.24	57.75	56.08
Variance-modified	44.46	45.77	47.44	68.76	55.27	57.81	56.12

Table 6.2: *Comparisons of the original and the variance-modified adjusted population group j means $\mathcal{M}(y_j)$ for three management and four engineering classes.*

A comparison of the adjustments between the classes reveals that the variance-modified adjusted means for Management classes are less than their corresponding original values, while for Engineering classes they are larger than the original adjustments. This is due to replacing η by the larger estimate of η^* in $E_{\bar{D}_n}(\mathcal{M}(y_j)) = \eta\bar{y}_{j\cdot} + (1-\eta)E_{\bar{D}_n}(\mathcal{M}(y))$ (see 4.27). Hence, relatively more weight is given to the sample class means and less to the population grand mean. Since Engineering students are required to have better A-level mathematics results than Management students, they perform better in STAT1010 examinations and hence, the effect of the variance-modified assessment on the class means seems in the right direction.

The two-stage analysis above is an improvement over our analyses in Chapter 4. We can thus apply the two-stage Bayes linear analysis to obtain improved means and variances, as well as to the canonical resolutions that we applied to design problems for level-1 and level-2 sample sizes. We can also apply the two-stage analysis to the adjustments of our finite SOEREF model in Chapter 4.

6.14 Two-stage Bayes linear analysis of more complex multilevel models

We now consider the more challenging problem of a two-stage analysis of more complex multilevel models, the Second-Order Exchangeable Regression (SOEREG) models defined in Chapter 3. The main problem we are likely to face relates to learning about the group-level variance-covariance matrices in the SOEREG model. In the SOEREF model we allowed only the regression intercept parameter to vary between groups, giving rise to a single variance at the group level. In a SOEREG model however, we may have two or more varying regression parameters, an intercept and a slope for example, hence resulting in a variance-covariance matrix at the group level (see Section 3.4.2).

To simplify our exposition, below we consider a two-stage Bayes linear analysis of only the basic SOEREG model (Section 3.4.2) and make the necessary prior specifications. We discuss some of the difficulties in learning about population variance matrices and motivate the use of a semi-adjusted residual variance matrix before applying it to the two-stage analysis of the STAT1010 data.

6.15 The basic SOEREG model and prior specifications

In Chapter 3 we formulated the basic SOEREG model using notations that can be generalized to more complex models. For ease of reference, we present the basic SOEREG model below and we simplify the notations since we have only one varying-intercept and one varying-slope (α_j, β_j).The regression predictor x_{ji} is the centered A-level score and includes A-Level mathematics. The reasons for centering the A-Level scores around their overall mean to get x_{ji} include ease of interpretation of the model intercept and also it may remove a strong correlation among (α_j, β_j) (see Gelman and Hill (2007)). All Engineering students have A-Level mathematics in

their A-Level scores while many Management students do not. Hence, our model is

$$y_{ji} = \alpha_j + \beta_j x_{ji} + \epsilon_{ji} \qquad \forall j, i, \tag{6.40}$$

with the following exchangeability representations for the intercepts and slopes

$$\alpha_j = \mathcal{M}(\alpha) + \mathcal{R}_j(\alpha) \tag{6.41}$$

$$\beta_j = \mathcal{M}(\beta) + \mathcal{R}_j(\beta). \tag{6.42}$$

The level-1 residuals are uncorrelated with mean zero and variance σ_ϵ^2 and they are also uncorrelated with the level-2 residual terms $\mathcal{R}_j(\alpha)$ and $\mathcal{R}_j(\beta)$. Our specifications for the regression coefficients are as follows:

$$E(\alpha_j) = \mu_\alpha, \quad E(\beta_j) = \mu_\beta \quad Var(\alpha_j) = \sigma_\alpha^2, \quad Var(\beta_j) = \sigma_\beta^2, \tag{6.43}$$

$$Cov(\alpha_j, \alpha_{j'}) = \gamma_\alpha, \quad Cov(\beta_j, \beta_{j'}) = \gamma_\beta, \quad Cov(\alpha_j, \beta_j) = \rho_{\alpha\beta}\sigma_\alpha\sigma_\beta. \tag{6.44}$$

The intercepts α_j and slopes β_j are judged to be correlated with correlation coefficient $\rho_{\alpha\beta}$. These specifications together with (6.41) and (6.42) imply the following:

$$\begin{aligned} E(\mathcal{M}(\alpha)) = \mu_\alpha, \quad E(\mathcal{M}(\beta)) = \mu_\beta, \quad Var(\mathcal{M}(\alpha)) = \gamma_\alpha, \\ Var(\mathcal{M}(\beta)) = \gamma_\beta, \quad Cov(\mathcal{M}(\alpha), \mathcal{M}(\beta)) = 0, \end{aligned} \tag{6.45}$$

with the following residual variance-covariance matrix:

$$\begin{aligned} Var(\mathcal{R}_j(\alpha)) = \sigma_\alpha^2 - \gamma_\alpha = V_{R_\alpha}, \quad Var(\mathcal{R}_j(\beta)) = \sigma_\beta^2 - \gamma_\beta = V_{R_\beta}, \\ Cov(\mathcal{R}_j(\alpha), \mathcal{R}_j(\beta)) = \rho_{\alpha\beta}\sqrt{V_{R_\alpha} V_{R_\beta}} \quad \forall j. \end{aligned} \tag{6.46}$$

6.16 Prior specifications for the SOEREG model

The detailed prior elicitation methods for the SOEREF model discussed in Chapter 3 also apply to the SOEREG model except that there are additional difficulties in specifying a residual variance-covariance matrix such as the level-2 effects (6.46) (see Goldstein and Wooff (2006), page 288). These difficulties include making exchangeability judgements for unobserved quantities as in (6.41) and (6.42), as well

as making unfamiliar uncertainty judgements for the variances and covariance as for $\mathcal{R}_j(\alpha)$ and $\mathcal{R}_j(\beta)$. Finally, variance matrices need to be non-negative definite which constrains the related correlation coefficients in a complicated way (see Gelman and Hill (2007), page 286), besides each being restricted to the interval $(-1, 1)$.

Below we shall elicit priors for the level-1 residual error V_{R_ϵ} and for the means, variances and the correlation of the group-level intercept and slope in (6.43) and (6.44) of the SOEREG model. Though we make use of various external sources and auxiliary data in our elicitation methods, they should be considered as basic - our aim here is to illustrate the two-stage analysis only. In more important applications, we would use more formal methods such as a meta-analysis of the literature or factor analysis of auxiliary data to elicit the priors of interest.

6.16.1 Prior for V_{R_ϵ}

In linear regression, adding a predictor reduces the residual variance. For the SOEREF model we judged $V_{R_\epsilon} = 237$, thus we should reduce our prior expectation of V_{R_ϵ} in the SOEREG model. To guide us, we re-consider the compilation of several studies of multilevel models on mathematics achievements by Hedges and Hedberg (2007) we mentioned in Chapter 3. These multilevel models also report the estimated level-1 residual variances for models without and with a predictor, hence similar to our SOEREF and SOEREG models. Including a demographic predictor such as Socio-Economic Status, reduces V_{R_ϵ} by about 10%, while including a pretest predictor such as prior achievement in mathematics, reduces V_{R_ϵ} by about 60%. For our data, the A-Level score predictor x_{ji} includes A-Level mathematics and two other non-mathematics A-Levels. For management students very few have A-Level mathematics. Also, since the STAT1010 syllabus is only about 50% calculation, we do not expect a strong correlation between A-Level score x_{ji} and STAT1010 examinations y_{ji}. Thus we judge a reduction of about 15% reasonable and specify $V_{R_\epsilon} = 202$.

6.16.2 Priors for the intercept

We assume that adding the predictor x_{ji} to the SOEREF model giving the SOEREG model, does not add any information to cause us to revise the prior for the intercept and its uncertainty, thus we keep $\mu_\alpha = 55$ and $\gamma_\alpha = 56.3$.

To specify a prior for $\text{Var}(\mathcal{R}_j(\alpha))=V_{R_\alpha}$, the residual variance of the intercept, we consider the following argument. Adding a level-1 predictor always reduces the level-1 error variance V_{R_ϵ}but not necessarily a level-2 error variance. For instance, the variance of the group intercept residual error V_{R_α} may actually increase if the predictor, here the A-Level score x_{ji}, is negatively correlated with the response variable y_{ji}, see (Gelman and Hill (2007) page 480) . For example, if Management students have high A-Level scores x_{ji} due to high grades in non-mathematical subjects and their performance in STAT1010 y_{ji} is low, then x_{ji} and y_{ji} will be negatively correlated. Since we do not know whether V_{R_α} will increase or decrease, we also keep $V_{R_\alpha} = 59$ as assessed for the SOEREF model.

6.16.3 Priors for slope

To specify priors for the population slope $\text{E}(\mathcal{M}(\beta))=\mu_\beta$, its uncertainty γ_β and the residual slope variance $\text{Var}(\mathcal{R}_j(\beta))=V_{R_\beta}$, we use $\beta_j = \mathcal{M}(\beta) + \mathcal{R}_j(\beta)$ and proceed as follows. We assess β_j for each of a typical Management class, say β_{mgt}, and a typical Engineering class, β_{eng}, and use these two slopes as a guide to specify the required priors.

Within each class (ignoring subscripts), the OLS estimator of the slope β in a simple linear regression $y_i = \alpha + \beta x_i + e_i$ can be written as $r_{xy}(\sigma_y/\sigma_x)$, where r_{xy} is the correlation coefficient between STAT1010 examinations score y and A-Level score x, σ_y and σ_x are their respective standard deviations.

To assess r_{xy}, we studied a number of multilevel research analyses reporting the strength of correlations between performance in A-level mathematics or its equivalent (x) and performance in mathematics and quantitative subjects at University level (y) and found weak correlations, about $r_{xy} = 0.3$. Because STAT1010 requires only about 50% basic mathematical skills, we judge a moderate correlation $r_{xy} = 0.5$

for Management classes and a stronger correlation $r_{xy} = 0.8$ for Engineering classes more suitable.

To assess σ_y, we take the range of STAT1010 examinations marks (40% to 70%) as a 95% interval and, assuming a Normal distribution and the overall mean score of 55%, we obtain $\sigma_y = 7.5$ for both classes.

To assess σ_x, we require the distribution of A-Level scores for Management and Engineering students. We gather data from the Mauritius Examinations Syndicate Report (2019) on 13,448 students that took A-Levels in subjects typical of those taken by students of Management and Engineering. We obtain the distribution of grades E to A+, corresponding to A-Level scores 2 to 12, by the 13,448 students in A-Level Accounts, Economics and Business for Management students and Mathematics, Chemistry and Physics for Engineering students. From these distributions, we estimate σ_x for each of the two classes. We find little difference in the standard deviations between Management and Engineering classes. We thus take $\sigma_x = 9.09$ for both classes.

Using the above assessments, we calculate $\beta_{mgt} = 0.3 \times 7.5/9.09 = 0.248$ and $\beta_{eng} = 0.8 \times 7.5/9.09 = 0.660$. We take the midpoint (0.248,0.660) as our assessment for $\mu_\beta = 0.454$ and specify $\gamma_\beta = 0.02$ directly. We assess the variance in the slope as $\sigma_\beta^2 = (0.660 - 0.248)^2 = 0.170$ from which we obtain $V_{R_\beta} = (\sigma_\beta^2 - \gamma_\beta) = 0.15$.

6.16.4 Priors for the correlation between intercept and slope

In the multilevel studies on mathematics achievements we have reviewed, the correlation coefficients between α and β were positive and low, in the range of 0.02 to 0.05. Often, a strong $\rho_{\alpha\beta}$ may be due to the center of a predictor x being far from zero; centering the predictor may remove any such high correlation (Gelman and Hill, 2007). Since x_{ji} is centered and we judge $\rho_{\alpha\beta}$ to be somewhat stronger than the higher end of the above range, we assess $\rho_{\alpha\beta} = 0.5$.

We summarize the above prior specifications as follows:

$$\mu_\alpha = 55, \quad \gamma_\alpha = 56.3, \quad V_{R_\alpha} = 59, \quad \mu_\beta = 0.454, \quad \gamma_\beta = 0.02, \quad V_{R_\beta} = 0.15,$$
$$\rho_{\alpha\beta} = 0.5, \quad V_{R_\epsilon} = 202. \tag{6.47}$$

6.17 Bayes linear update of the SOEREG and more complex models

In Section 3.5 we extended the basic SOEREG model to the more general SOEREG model. The latter can be written as the General Bayesian Linear (GBL) model see Smith (1973) and Dempster et al. (1981). De Leeuw and Kreft (1986) discuss ordinary least squares estimators for the GBL; these will provide the sample information required for our two-stage Bayes linear analysis. Furthermore, Bryk and Raudenbush (1992) show how multilevel models with more than two levels can be written in terms of the GBL. Hence, the Bayes linear analysis we shall develop for the basic SOEREG model will also apply to more complex models. Below we define the basic SOEREG model similar to a GBL.

Definition 6.17.1. *Suppose we stack the SOEREG model* $y_{ji} = \alpha_j + \beta_j x_{ji} + \epsilon_{ji}$ *for each group* j *in a vector* $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_J)$*. Then we have the following:*

$$\mathbf{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{6.48}$$

$$\boldsymbol{\beta} = \mathbf{W}\mathcal{M}(\boldsymbol{\beta}) + \mathcal{R}(\boldsymbol{\beta}), \tag{6.49}$$

where $\mathbf{Y}$ *is a vector of level-1 response variables,* $\mathbf{X}$ *and* $\mathbf{W}$ *are predictor matrices at level-1 and 2 respectively, and* $\boldsymbol{\epsilon}$ *and* $\mathcal{R}(\boldsymbol{\beta})$ *are vectors of level-1 and 2 residuals respectively.* $\mathcal{M}(\boldsymbol{\beta})$ *is a vector of population mean intercept and slopes.*

As a simple example, consider data on two classes each with two and three students respectively. Thus, $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2) = ((y_{11}, y_{12}), (y_{21}, y_{22}, y_{23}))$ and using (6.40) we have the level-1 matrix equation .

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{23} \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & 0 & 0 \\ 1 & x_{12} & 0 & 0 \\ 0 & 0 & 1 & x_{21} \\ 0 & 0 & 1 & x_{22} \\ 0 & 0 & 1 & x_{23} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \beta_1 \\ \alpha_2 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \end{bmatrix}, \tag{6.50}$$

corresponding to (6.48) above. The level-2 matrix equation corresponding to (6.49)

is .

$$\begin{bmatrix} \alpha_1 \\ \beta_1 \\ \alpha_2 \\ \beta_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \mathcal{M}(\alpha) \\ \mathcal{M}(\beta) \end{bmatrix} + \begin{bmatrix} \mathcal{R}_1(\alpha) \\ \mathcal{R}_1(\beta) \\ \mathcal{R}_2(\alpha) \\ \mathcal{R}_2(\beta) \end{bmatrix} \tag{6.51}$$

Upon substituting (6.51) in (6.50) we obtain the general SOEREG model of Section 3.5. The above matrix equations correspond to (6.40), (6.41) and (6.42) of the basic SOEREG model. Thus using the specifications (6.45) and (6.46) we have for this example.

$$\begin{aligned} E(\boldsymbol{\epsilon}) &= \mathbf{0} & Var(\boldsymbol{\epsilon}) &= \sigma_\epsilon^2 \mathbf{I}_5 \\ E(\mathcal{M}(\boldsymbol{\beta})) &= [\mu_\alpha, \mu_\beta]^T & Var(\mathcal{M}(\boldsymbol{\beta})) &= \begin{bmatrix} \gamma_\alpha & 0 \\ 0 & \gamma_\beta \end{bmatrix} \\ E(\mathcal{R}(\boldsymbol{\beta})) &= \mathbf{0} & Var(\mathcal{R}(\boldsymbol{\beta})) &= \mathbf{I}_2 \otimes \boldsymbol{\Omega}, \quad \boldsymbol{\Omega} = \begin{bmatrix} V_{R_\alpha} & \sigma_{\alpha\beta} \\ \sigma_{\alpha\beta} & V_{R_\beta} \end{bmatrix}, \end{aligned} \tag{6.52}$$

where $\sigma_{\alpha\beta} = \rho_{\alpha\beta}\sqrt{V_{R_\alpha} V_{R_\beta}}$.

6.17.1 Updating the mean components

We wish to compare a single stage analysis using prior variances only and the two-stage analysis with updated variances. To adjust population means $\mathcal{M}(\boldsymbol{\beta})$ and $\mathcal{R}(\boldsymbol{\beta})$, we make use of Bayes linear sufficiency. Thus we begin by deriving the representation for $\overline{\mathbf{Y}}$ and then we construct the required beliefs.

The representation for $\overline{\mathbf{Y}}$

Consider the basic SOEREG model with J groups and n_j observations in group j. We calculate $Var(\overline{\mathbf{Y}})$ where $\overline{\mathbf{Y}} = \{\bar{y}_{1.}, \bar{y}_{2.}, \ldots, \bar{y}_{J.}\}$. Replacing (6.49) in (6.48) we obtain

$$\mathbf{Y} = \boldsymbol{X}\boldsymbol{W}\mathcal{M}(\boldsymbol{\beta}) + \mathbf{X}\mathcal{R}(\boldsymbol{\beta}) + \boldsymbol{\epsilon}. \tag{6.53}$$

Averaging within each group j gives the representation for $\overline{\mathbf{Y}}$.

$$\overline{\mathbf{Y}} = \overline{\boldsymbol{X}\boldsymbol{W}}\mathcal{M}(\boldsymbol{\beta}) + \overline{\mathbf{X}}\mathcal{R}(\boldsymbol{\beta}) + \overline{\boldsymbol{\epsilon}}, \tag{6.54}$$

with elements as follows. .

$$\begin{bmatrix} \bar{y}_{1.} \\ \bar{y}_{1.} \\ \vdots \\ \bar{y}_{J.} \end{bmatrix} = \begin{bmatrix} 1\ \bar{x}_{1.} \\ 1\ \bar{x}_{2.} \\ \vdots\ \ \vdots \\ 1\ \bar{x}_{J.} \end{bmatrix} \begin{bmatrix} \mathcal{M}(\alpha) \\ \mathcal{M}(\beta) \end{bmatrix} + \begin{bmatrix} 1\ \bar{x}_{1.}\ 0\ \ 0\ \ 0\ \ 0\,0\ \ 0 \\ 0\ \ 0\ \ 1\ \bar{x}_{2.}\ \ 0\ \ 0\,0\ \ 0 \\ \vdots \qquad\qquad \ddots \qquad \ldots \\ 0\ \ 0\ \ 0\ \ 0\ \ \ 0\ \ 0\,1\ \bar{x}_{J.} \end{bmatrix} \begin{bmatrix} \mathcal{R}_1(\alpha) \\ \mathcal{R}_1(\beta) \\ \vdots \\ \mathcal{R}_J(\alpha) \\ \mathcal{R}_J(\beta \end{bmatrix} + \begin{bmatrix} \bar{\epsilon}_{1.} \\ \bar{\epsilon}_{2.} \\ \vdots \\ \bar{\epsilon}_{J.2} \end{bmatrix}, \quad (6.55)$$

The two regressor matrices in (6.55) are of simple forms; $\overline{\boldsymbol{XW}}$ is similar to the regressor matrix in simple linear regression with one intercept but with group means of the explanatory variable i.e. $(1\ ,\ \bar{x}_{j.})$ in column j and $\overline{\boldsymbol{X}}$ is the direct sum of $(1\ ,\ \bar{x}_{j.}\ )$ for each j.

Beliefs for the data quantities

Using (6.54),

$E(\overline{\mathbf{Y}}) = \overline{\boldsymbol{XW}}\boldsymbol{\mu_\beta}$, where $\boldsymbol{\mu_\beta} = E(\mathcal{M}(\boldsymbol{\beta})) = [\mu_\alpha, \mu_\beta]^T$,

$Var(\overline{\mathbf{Y}}) = \boldsymbol{\Sigma}_{\overline{\mathbf{Y}}}$, where

$$\boldsymbol{\Sigma}_{\overline{\mathbf{Y}}} = (\overline{\boldsymbol{XW}})\boldsymbol{\Gamma}(\overline{\boldsymbol{XW}})^T + \overline{\mathbf{X}}(\mathbf{I}_J \otimes \boldsymbol{\Omega})\overline{\mathbf{X}}^T + \boldsymbol{\Psi}. \quad (6.56)$$

$\boldsymbol{\Gamma} = Var(\mathcal{M}(\boldsymbol{\beta}))$ as in (6.52), $\boldsymbol{\Psi}$ is a diagonal matrix with jth element $Var(\epsilon_j) = \sigma_\epsilon^2/n_j$ and $Var(\mathcal{R}(\boldsymbol{\beta})) = \mathbf{I}_J \otimes \boldsymbol{\Omega}$, with

$$\boldsymbol{\Omega} = \begin{bmatrix} V_{R_\alpha} & \sigma_{\alpha\beta} \\ \sigma_{\alpha\beta} & V_{R_\beta} \end{bmatrix} \quad (6.57)$$

Beliefs between regression coefficients, level-2 residuals and the data quantities

Using (6.54) the beliefs between the population mean intercept and slope and the data is

$$Cov(\mathcal{M}(\boldsymbol{\beta}), \overline{\mathbf{Y}}) = \boldsymbol{\Gamma}(\overline{\boldsymbol{XW}})^T, \quad (6.58)$$

while the corresponding beliefs for the level-2 residuals are

$$Cov(\mathcal{R}(\boldsymbol{\beta}), \overline{\mathbf{Y}}) = (\mathbf{I}_J \otimes \boldsymbol{\Omega})\boldsymbol{X}^T. \tag{6.59}$$

Applying the Bayes linear rule, we obtain the adjusted expectation of mean intercept and slope with the corresponding adjusted variance as follows.

$$E_{\overline{\mathbf{Y}}}(\mathcal{M}(\boldsymbol{\beta})) = \boldsymbol{\mu}_{\boldsymbol{\beta}} + \boldsymbol{\Gamma}(\overline{\boldsymbol{XW}})^T \boldsymbol{\Sigma}_{\overline{\mathbf{Y}}}^{-1}[\overline{\mathbf{Y}} - \overline{\boldsymbol{XW}}\boldsymbol{\mu}_{\boldsymbol{\beta}}] \tag{6.60}$$

$$Var_{\overline{\mathbf{Y}}}(\mathcal{M}(\boldsymbol{\beta})) = \boldsymbol{\Gamma} - \boldsymbol{\Gamma}(\overline{\boldsymbol{XW}})^T \boldsymbol{\Sigma}_{\overline{\mathbf{Y}}}^{-1} \overline{\boldsymbol{XW}}\boldsymbol{\Gamma}. \tag{6.61}$$

Since, $E(\mathcal{R}(\boldsymbol{\beta})) = 0$, the adjusted mean level-2 residuals and the corresponding variance are

$$E_{\overline{\mathbf{Y}}}(\mathcal{R}(\boldsymbol{\beta})) = (\mathbf{I}_J \otimes \boldsymbol{\Omega})\boldsymbol{X}^T \boldsymbol{\Sigma}_{\overline{\mathbf{Y}}}^{-1}[\overline{\mathbf{Y}} - \overline{\boldsymbol{XW}}\boldsymbol{\mu}_{\boldsymbol{\beta}}] \tag{6.62}$$

$$Var_{\overline{\mathbf{Y}}}(\mathcal{R}(\boldsymbol{\beta})) = (\mathbf{I}_J \otimes \boldsymbol{\Omega}) - (\mathbf{I}_J \otimes \boldsymbol{\Omega})\boldsymbol{X}^T \boldsymbol{\Sigma}_{\overline{\mathbf{Y}}}^{-1} \boldsymbol{X}(\mathbf{I}_J \otimes \boldsymbol{\Omega}). \tag{6.63}$$

6.18 Application to the STAT1010 data

Using our prior specifications in (6.47), we adjust the collection of population mean and residual intercepts and slopes using (6.60) to (6.63). Our main goal is to compare the results below with the two-stage analysis. We thus only consider some basic interpretation rather than the full interpretative methods for individual and collection of adjusted expectations described in Chapter 3 and applied in Chapter 4.

6.18.1 Adjustment of $\mathcal{M}(\boldsymbol{\beta})$

Firstly, we adjust the overall population mean intercept and slopes $\mathcal{M}(\boldsymbol{\beta})$ for the STAT1010 data using the group means examinations scores $\bar{y}_{j\cdot}$ in the seven classes. The prior and adjusted expectation are

$$E(\mathcal{M}(\boldsymbol{\beta})) = [55, 0.454]^T$$

$$E_{\overline{\mathbf{Y}}}(\mathcal{M}(\boldsymbol{\beta})) = [53.91, 0.465]^T,$$

with prior and adjusted variation

$$Var(\mathcal{M}(\boldsymbol{\beta})) = \begin{bmatrix} 56.3 & 0 \\ 0 & 0.02 \end{bmatrix} \quad Var_{\overline{\mathbf{Y}}}(\mathcal{M}(\boldsymbol{\beta})) = \begin{bmatrix} 7.98 & 0.00 \\ 0.00 & 0.019 \end{bmatrix}$$

The prior expectation of the intercept has decreased from 55 to the adjusted expectation of 53.91. This is unsurprising as confirmed by the standardized adjustment of -0.16 . The increase in expectation of the slope from the prior of 0.454 to adjusted value of 0.465 is also unsurprising as the standardized adjustment is 0.35.

The adjustment of the uncertainty shows that the prior variance for the intercept has been reduced from 56.3 to the adjusted variance of 7.98, representing a variance resolution of 85.83%. However, for the slope, the variance resolution is significantly less, 5% for a reduction from prior variance of 0.02 to an adjusted variance of 0.019. So, the data is much more informative in learning about the intercept and rather uninformative in updating the slope. We may have been overconfident about the priors for the slope and we will consider this issue when we update the individual class intercepts and slopes.

6.18.2 Adjustment of $\boldsymbol{\beta}$ and $\mathcal{R}(\boldsymbol{\beta})$

Using the representation $\boldsymbol{\beta} = \mathbf{W}\mathcal{M}(\boldsymbol{\beta}) + \mathcal{R}(\boldsymbol{\beta})$, we calculate adjusted population intercepts and slopes $E_{\overline{\mathbf{Y}}}(\boldsymbol{\beta}) = \mathbf{W}E_{\overline{\mathbf{Y}}}(\mathcal{M}(\boldsymbol{\beta})) + E_{\overline{\mathbf{Y}}}(\mathcal{R}(\boldsymbol{\beta}))$ and compare these to ordinary least squares estimates (OLS) of intercepts and slopes in each class.

The results in Figure 6.2. reveal that there is little shrinkage in adjusted intercepts as they are quite close to their corresponding OLS estimates. The adjusted slopes however, are all shrunk towards the prior slope of 0.454. Unlike the OLS slope estimates, we do not have a negative adjusted slope which is more sensible for the STAT1010 data.

Comparing the variance resolutions in Section 6.18.1 above, we learn more from the data about the intercepts than the slope. We may have been overconfident in specifying the prior for the slope and we are somewhat unsatisfied with the over shrinkage of the slope. We consider the effect of revising our prior for the slope next.

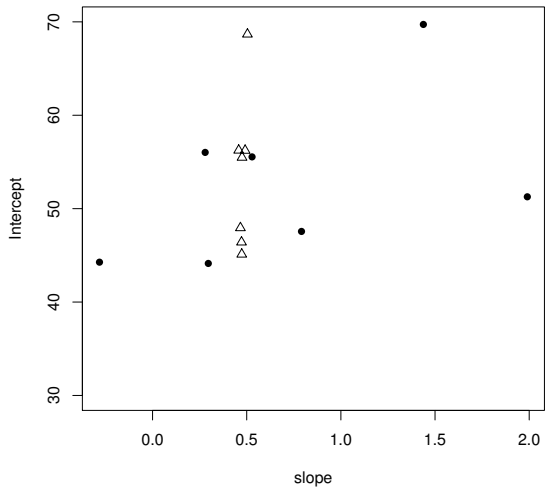


Figure 6.2: *Scatterplot of group-level intercepts and slopes shown in Table 6.3. The dots indicate OLS estimates and the triangles show the adjusted quantities. There is little shrinkage in adjusted intercepts but considerably more shrinkage in slopes towards the prior of 0.454.*

6.18.3 The effect of revising the prior for the slope

We consider the over shrinkage of the slopes to be a result of our tight variance specifications. We thus increase our priors for the slope as follows.

$$\mu_\beta = 1, \quad \gamma_\beta = 0.9, \quad V_{R_\beta} = 1. \tag{6.64}$$

Using the above, the prior variance of the slope $\sigma^2_\beta = V_{R_\beta} + \gamma_\beta = 1.9$, implying that we admit slopes in the interval $(1 \pm 2\sqrt{1.9}) = (-1.76, 3.76)$. We are admitting the possibility of a negative slope between examination score and A-Level score; our priors for the slope are more skeptical now. We keep the remaining priors in (6.47) unchanged. The effect of our increased uncertainty is shown in Figure 6.3 below.

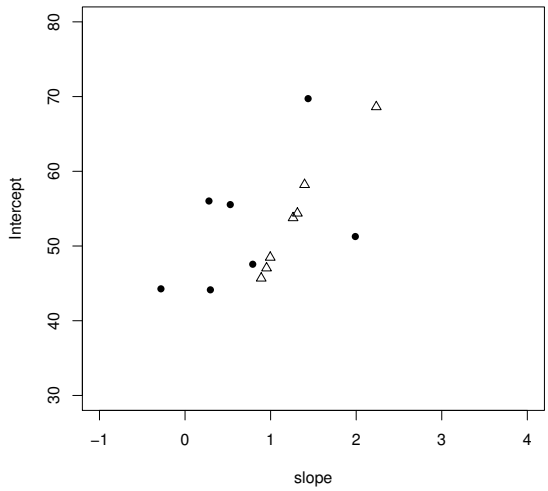


Figure 6.3: *Scatterplot of group-level intercepts and slopes with priors for the slope revised. The dots indicate OLS estimates and the triangles show the adjusted quantities. There is little shrinkage in adjusted intercepts and also less shrinkage in slopes compared to Figure 6.2.*

The small shrinkage in the intercepts is as in Figure 6.2 but now there is equally less shrinkage in the slopes. The first three triangles with intercepts between (40,50) and slopes (0,1) are Management classes, the rest are for Engineering classes. We deem the pattern in adjusted quantities in Figure 6.2 more plausible for the STAT1010 data. The increase in our uncertainty specifications for the slope has increased the variance resolution from 5% to 22.22%.

Table 6.3 below gives the data used to plot Figure 6.3 and also the adjusted residuals. The latter shows that the adjusted intercepts $E_{\overline{\mathbf{Y}}}(\mathcal{R}(\alpha))$ and slopes effects $E_{\overline{\mathbf{Y}}}(\mathcal{R}(\beta))$ are all negative for Management classes (C1 to C3), implying the intercepts and slopes are revised downward from the respective priors, while for Engineering classes they are revised upward.

6.19 Variance update of the SOEREG model

We now consider updating the level-1 variance and the level-2 variance-covariance matrix. We shall use ordinary least squares estimates (OLS) to obtain unbiased sample information for our updates. Rao (1965a) and Swamy (1970) develop Gauss-

	Intercept			Slope		
Class	$\hat{\alpha}_{OLS}$	$E_{\overline{\mathbf{Y}}}(\mathcal{M}(\alpha))$	$\mathrm{E}_{\overline{\mathbf{Y}}}(\mathcal{R}(\alpha))$	$\hat{\beta}_{OLS}$	$E_{\overline{\mathbf{Y}}}(\mathcal{M}(\beta))$	$\mathrm{E}_{\overline{\mathbf{Y}}}(\mathcal{R}(\beta))$
C1	44.13	44.75	-8.77	0.30	0.10	-1.14
C2	44.27	46.36	-7.16	-0.28	0.30	-0.93
C3	47.56	48.05	-5.47	0.79	0.53	-0.71
C4	69.72	68.58	15.05	1.44	3.20	1.96
C5	56.02	59.35	5.82	0.28	2.00	0.76
C6	51.27	53.74	0.21	1.99	1.27	0.03
C7	55.54	54.33	0.80	0.53	1.34	0.10

Table 6.3: *Comparisons between ordinary least squares (OLS) estimates of group-level intercepts and slopes in each of the seven classes of the STAT1010 data with the corresponding adjusted intercepts and slopes.*

Markov theory for multilevel (random coefficient) models and prove that OLS methods yield minimum variance unbiased linear estimates of regression coefficients and also of the level-1 and level-2 variances. They fit linear regressions within each group j to obtain OLS estimates of unknown parameters and residuals and, in turn, use these to derive estimates of the variances. Their methods apply to general SOEREG models. Below we shall adapt these methods, simplifying them for the basic SOEREG model.

6.19.1 Adjustment of the level-1 variance V_{R_ϵ}

To adjust V_{R_ϵ}, we first derive an unbiased estimator for σ_ϵ^2 using the level-1 regression of the basic SOEREG model (see Definition 6.17.1) for each group j as follows.

$$y_{ji} = \alpha_j + \beta_j x_{ji} + \epsilon_{ji}. \tag{6.65}$$

We fit (6.65) and obtain the residual vector for each group j

$$\hat{r}_j = \hat{y}_j - \hat{\alpha}_j - \hat{\beta}_j x_{ji}, \tag{6.66}$$

from which we obtain the unbiased estimator of $\hat{\sigma}_j^2$ as

$$\hat{\sigma}_j^2 = \frac{1}{n_j - p} \hat{r}_j^T \hat{r}_j \qquad \forall j \tag{6.67}$$

(6.67) is suitable in case of variance heterogeneity at level-1. For instance, we may use (6.67) to adjust V_{R_ϵ} partially for each group based on $\hat{\sigma}_j^2$ and assess the effect on the final adjustment. If the partial analysis reveals group differences in variance, we may consider changing our prior specifications by modeling the heterogeneity as a function of a level-1 predictor (see Section 3.7.2) or we may set $Var(\boldsymbol{\epsilon}) = diag(\sigma_j^2)$ instead of $\sigma_\epsilon^2 \boldsymbol{I}_J$.

An estimate of the homogeneous level-1 variance is given by

$$\hat{\sigma}_\epsilon^2 = \frac{1}{N-p} \sum_j \hat{r}_j^T \hat{r}_j, \quad N = \sum_j n_j, \tag{6.68}$$

which is equivalent to the residual mean square for a single level regression ignoring all groups. This is applicable to our SOEREG model where the level-1 residuals ϵ_{ji} are second-order exchangeable over all students i and groups j, and has mean zero and constant variance σ_ϵ^2. To adjust V_{R_ϵ} based on $\hat{\sigma}_\epsilon^2$, we use the following results from Goldstein and Wooff (2006), page 275.

A representation for $\hat{\sigma}_\epsilon^2$ is

$$\begin{aligned} \hat{\sigma}_\epsilon^2 &= \frac{1}{N-p} \sum_j \hat{r}_j^T \hat{r}_j \\ &= \frac{1}{N-p} \boldsymbol{\epsilon}^T (\boldsymbol{I} - \boldsymbol{H}) \boldsymbol{\epsilon} \\ &= \mathcal{M}(V_\epsilon) + T_\epsilon, \end{aligned} \tag{6.69}$$

where $\boldsymbol{\epsilon}$ is a vector of level-1 residuals and $\boldsymbol{H}$ is the hat matrix. Representation (6.69) is similar to Theorem 5.5.1 in our adjustment of $\mathcal{M}(V_\epsilon)$ for the SOEREF model from which, after making the necessary specifications and fourth order uncorrelated assumptions about $\mathcal{M}(V_\epsilon)$ and T_ϵ, we obtain the adjusted mean and variance of $\mathcal{M}(V_\epsilon)$ given σ_ϵ^2

$$\mathrm{E}_{\hat{\sigma}_\epsilon^2}(\mathcal{M}(V_\epsilon)) = \frac{V_{M_\epsilon} \hat{\sigma}_\epsilon^2 + V_{T_\epsilon} V_{R_\epsilon}}{V_{M_\epsilon} + V_{T_\epsilon}}, \tag{6.70}$$

$$\mathrm{Var}_{\hat{\sigma}_\epsilon^2}(\mathcal{M}(V_\epsilon)) = \frac{V_{M_\epsilon} V_{T_\epsilon}}{V_{M_\epsilon} + V_{T_\epsilon}}. \tag{6.71}$$

6.19.2 Choice of prior values and application to STAT1010

We make the simplifying assumptions as in Chapter 5 and use the representation $\epsilon_i = \sqrt{\mathcal{M}(V_\epsilon)}Z_i$ and choose the kurtosis $Var(Z_i^2) = 2$ leading to

$$V_{R(V_\epsilon)} = 2(V_{M_\epsilon} + V_{R_\epsilon}^2),$$

and

$$V_{T_\epsilon} = \frac{1}{N-p}[2(V_{M_\epsilon} + V_{R_\epsilon}^2)] = \frac{1}{N-p}V_{R(V_\epsilon)}.$$

We note that for the STAT1010 data $N = 269$ and $p = 2$, so we use $N - p + 1 \approx 269$ to determine the equivalent sample size m.

Given $N = 269$, we expect the sample to resolve a large percentage of the prior uncertainty in $\mathcal{M}(V_\epsilon)$. We shall consider our prior to be worth about $m = 6$ observations which leads to $c \approx \kappa/m - \kappa = 0.5$ for a choice of κ according to a normal distribution. We specified $V_{R_\epsilon} = 202$ in (6.47) and, together with $m = 6$ and $c = 0.5$, we have the specifications

$$V_{M_\epsilon} = cV_{R_\epsilon}^2 = 20402.$$

$$V_{T_\epsilon} = \frac{1}{N-p}[2(V_{M_\epsilon} + V_{R_\epsilon}^2)] = 458.47.$$

OLS gives $\hat{\sigma}_\epsilon^2 = 281.23$. Our adjusted mean and variance are thus

$$\mathrm{E}_{\hat{\sigma}_\epsilon^2}(\mathcal{M}(V_\epsilon)) = \frac{V_{M_\epsilon}\hat{\sigma}_\epsilon^2 + V_{T_\epsilon}V_{R_\epsilon}}{V_{M_\epsilon} + V_{T_\epsilon}} = \frac{20402 \times 281.23 + 458.47 \times 202}{20402 + 458.47} = 279.49.$$

$$\mathrm{Var}_{\hat{\sigma}_\epsilon^2}(\mathcal{M}(V_\epsilon)) = \frac{V_{M_\epsilon}V_{T_\epsilon}}{V_{M_\epsilon} + V_{T_\epsilon}} = 448.39.$$

The updated residual level-1 variance is 279.49, larger than our prior specification $V_{R_\epsilon} = 202$. The high value of the adjusted variance shows that we remain quite uncertain about the population level-1 variance.

6.19.3 Adjustment of the level-2 residual variance matrix $\boldsymbol{\Omega}$

To adjust $\boldsymbol{\Omega}$, the population residual variance matrix of the level-2 intercept $\mathcal{R}_j(\alpha)$ and slope $\mathcal{R}_j(\beta)$, we plan to use the direct method as explained in Goldstein and Wooff (2006), page 283. Their main result is as follows. The adjustment of a residual variance matrix $\mathcal{M}(V)$ by the space spanned by the sample variance matrix S_n^2 and the constant matrices is given by

$$\mathrm{E}_{S_n^2}(\mathcal{M}(V)) = (1-\alpha)\mathrm{E}(\mathcal{M}(V)) + \alpha S_n^2, \tag{6.72}$$

where, using the equivalent sample size heuristic as in our previous variance adjustments (see Section 6.8), we may choose $\alpha = J/(m+J)$ with m and J being the notional and actual level-2 sample sizes respectively.

To calculate the unbiased OLS estimate S_n^2, we follow Rao (1965a) and Swamy (1970) who substitute the OLS estimate $\hat{\boldsymbol{\beta}}$ obtained from the level-1 regression $\mathbf{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ in the level-2 model giving $\hat{\boldsymbol{\beta}} = \mathbf{W}\mathcal{M}(\boldsymbol{\beta}) + \mathcal{R}(\boldsymbol{\beta})$. A second regression of $\hat{\boldsymbol{\beta}}$ on $\mathbf{W}$ yield the OLS estimate $\widehat{\mathcal{M}(\boldsymbol{\beta})}$ and the residual $\widehat{\mathcal{R}(\boldsymbol{\beta})}$. For the basic SOEREG model, we do not have any level-2 predictors so the level-2 regressions in each group j are:

$$\alpha_j = \mathcal{M}(\alpha) + \mathcal{R}_j(\alpha) \tag{6.73}$$

$$\beta_j = \mathcal{M}(\beta) + \mathcal{R}_j(\beta). \tag{6.74}$$

Substituting the OLS estimates $\hat{\alpha}_j$ and $\hat{\beta}_j$ from the level-1 regression (6.65) in (6.73) and (6.74), we derive the OLS estimates of $\widehat{\mathcal{M}(\alpha)} = \bar{\hat{\alpha}}$ and $\widehat{\mathcal{M}(\beta)} = \bar{\hat{\beta}}$. Hence, our estimate of the level-2 residual variance matrix is

$$S_n^2 = \frac{1}{J-p}\begin{bmatrix} \sum_j(\hat{\alpha}_j - \bar{\hat{\alpha}})^2 & \sum_j(\hat{\alpha}_j - \bar{\hat{\alpha}})(\hat{\beta}_j - \bar{\hat{\beta}}) \\ \sum_j(\hat{\alpha}_j - \bar{\hat{\alpha}})(\hat{\beta}_j - \bar{\hat{\beta}}) & \sum_j(\hat{\beta}_j - \bar{\hat{\beta}})^2 \end{bmatrix}, \tag{6.75}$$

with observed value

$$\begin{bmatrix} 96.36 & 4.02 \\ 4.02 & 0.71 \end{bmatrix} \tag{6.76}$$

For our STAT1010 data we have sufficient level-1 observations n_j, the range is $(23, 47)$, to reliably estimate $\hat{\alpha}_j$ and $\hat{\beta}_j$. However, S_n^2 is estimated based on only

$(J-p)=(7-2)=5$ degrees of freedom. Thus we do not expect to resolve much of the variation. We choose our prior information as worth $m=4$ observations while the actual sample size is 5 giving $\alpha=5/9$. Hence, our updated variance matrix is

$$\frac{4}{9}\begin{bmatrix} 59 & 3.84 \\ 3.84 & 1 \end{bmatrix}+\frac{5}{9}\begin{bmatrix} 96.36 & 4.02 \\ 4.02 & 0.71 \end{bmatrix}=\begin{bmatrix} 79.76 & 3.94 \\ 3.94 & 0.84 \end{bmatrix}$$

There does not appear to be any contradiction between the updated and prior residual variance matrices. We learn most about the residual intercept which has been reduced to 79.76, which is consistent with the value 71.99 that we obtained when updating the intercept only in the SOEREF model using BLIMVE (see Table 6.1). There is little change in the adjusted covariance. The adjusted residual slope has increased significantly compared to its prior value.

6.20 Two-stage analysis of the SOEREG model

For the two-stage analysis, the prior and adjusted expectation are

$$E(\mathcal{M}(\boldsymbol{\beta}))=[55,1]^T$$
$$E_{\overline{\mathbf{Y}}}(\mathcal{M}(\boldsymbol{\beta}))=[53.86,1.21]^T,$$

with prior and adjusted variation

$$Var(\mathcal{M}(\boldsymbol{\beta}))=\begin{bmatrix} 56.3 & 0 \\ 0 & 0.9 \end{bmatrix}\quad Var_{\overline{\mathbf{Y}}}(\mathcal{M}(\boldsymbol{\beta}))=\begin{bmatrix} 10.25 & 0.28 \\ 0.28 & 0.74 \end{bmatrix}$$

The prior expectation of the intercept has decreased from 55 to the adjusted expectation of 53.86. This is unsurprising as confirmed by the standardized adjustment of -0.17, which is almost the same as in the single stage analysis. The increase in expectation of the slope from the prior of 1 to the adjusted value of 1.21 is also unsurprising as the standardized adjustment is 0.525.

One major difference is the higher adjusted variances of the two-stage analysis compared to the single-stage analysis, especially for the slope. This results in somewhat more shrinkage in the slopes, while the shrinkage in the intercepts remains more or less the same as in the single-stage analysis (compare Figure 6.3 and Figure 6.4 below).

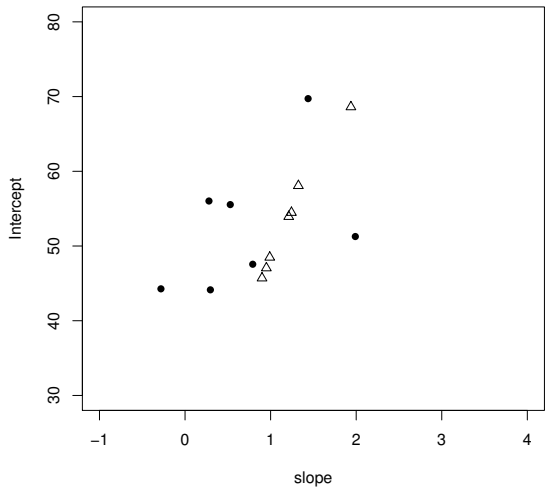


Figure 6.4: *Scatterplot of group-level intercepts and slopes for the two-stage analysis. The dots indicate OLS estimates and the triangles show the adjusted quantities. There is little shrinkage in adjusted intercepts and also less shrinkage in slopes compared to Figure 6.2.*

Table 6.4 below also shows that the increased uncertainty has led to an increase in shrinkage, particularly in the slope, compared to Table 6.3. The Engineering class C4 appears to be an exceptional class; it has a higher intercept and slope compared to all the other classes.

	Intercept		Slope	
Class	$E_{\overline{\mathbf{Y}}}(\mathcal{M}(\alpha))$	$\mathrm{E}_{\overline{\mathbf{Y}}}(\mathcal{R}(\alpha))$	$E_{\overline{\mathbf{Y}}}(\mathcal{M}(\beta))$	$\mathrm{E}_{\overline{\mathbf{Y}}}(\mathcal{R}(\beta))$
C1	45.70	-8.16	0.90	-0.31
C2	47.06	-6.81	0.95	-0.26
C3	48.48	-5.38	0.99	-0.22
C4	68.61	14.75	1.94	0.73
C5	58.07	4.20	1.32	0.12
C6	53.93	0.07	1.21	0.01
C7	54.48	0.62	1.24	0.04

Table 6.4: *Two-stage update of group-level intercepts and slopes in each of the seven classes of the STAT1010 data. The analysis separates Management (C1 to C3) and Engineering (C4 to C7) classes in two homogeneous groups.*

Chapter 7

Discussions and further study

In this thesis we have applied Bayes linear methods to analyse multilevel models. In Chapter 2 we reviewed the concepts underlying multilevel data structures and the need for multilevel modeling in the context of some important applications. We discussed a number of classical estimation methods and also the fully Bayesian hierarchical modeling approach. The difficulties in making full prior specifications, as well as in computing posterior densities were presented and used to motivate a Bayes linear approach that requires limited beliefs specifications only.

In Chapter 3 we used second-order exchangeability (SOE) judgements to formulate our versions of multilevel models. We defined the SOE random effects (SOEREF) model and extended it to a more general SOE regression (SOEREG) model which was shown to encompass models with more levels and complex error structures. In the context of the STAT1010 data, we discussed and illustrated some methods to specify priors for mean and variance components in our models.

Bayes linear estimation of population overall and population group means in the SOEREF model were developed and discussed in Chapter 4. The closed form expressions we derived for the adjusted mean and the resolution transform proved useful to understand the relationships between adjusted quantities and also to address sample design issues and sample size determination with cost constraints for both level-1 and level-2 units in the SOEREF model. Bayes linear diagnostics were used to assess our estimates and applied to the STAT1010 data using specially written codes in the R statistical programming language. A finitely exchangeable version

of our model was formulated and analysed comparatively to the infinite version of our model.

In Chapter 5, we discussed the difficulties in learning about population variances and variance matrices in particular, and developed Bayes linear methods to estimate the level-1 variance in both the balanced and unbalanced cases. We applied these methods to the STAT1010 data and illustrated the choice of priors for fourth order quantities. The sensitivity of our adjusted variances to a higher kurtosis was also investigated.

We developed a new method, the Bayes Linear Minimum Variance Estimator (BLIMVE) to estimate the level-2 variance of the SOEREF model in Chapter 6. The method is applicable to two or more groups and we validated it using simulation. We also developed methods, based on OLS estimates, to estimate variances in more complex multilevel models and applied these in a two-stage analysis to update intercepts and slopes in our SOEREG model.

In Chapter 2 we explained the importance of multilevel models and presented examples of some important real world applications. Currently, researchers face many challenges in applying multilevel modeling in these important areas. Some of these complications result from too small or too large data sets, leading to problems in convergence of estimation algorithms. The Bayes linear methods we have developed would be promising in analyzing more complex multilevel models both in situations where data is limited or there is too much data. The limited beliefs specifications requirements also give the Bayes linear approach an added advantage, especially in multilevel models where it is difficult to specify probabilistic priors and hyperpriors, not to mention that they may have hidden consequences.

The objectives of the research underlying this thesis have been achieved. The development of a Bayes linear simulation approach could further be turned into a practical methodology, as well as functioning as a method for giving insights into the comparative strengths of different methods and widen its range of application as we explain below.

7.1 Bayes linear simulation

Simulation is widely used in multilevel modeling with the twin purposes of parameter estimation and studying properties of (complex) estimators. In the fully Bayesian approach for example, MCMC simulation is routinely applied to obtain estimates of mean and variance parameters. In the frequentist approach, parametric and non-parametric bootstrap simulation techniques are often used to obtain robust estimates of parameters and their uncertainties.

A key advantage of simulation is that it provides a unifying framework for the estimation of a wide range of ever more complex models, including "*models for multivariate mixtures of Gaussian, ordered or unordered categorical responses and continuous distributions that are not Gaussian, each of which can be defined at any level of a multilevel data hierarchy.*", see, for example, Goldstein et al. (2009). In both the Bayesian and frequentist approaches, simulation experiments have been designed to study, evaluate and compare the properties of alternative multilevel estimation techniques. Simulation experiments are also widely used to study design and sample size determination issues in multilevel modeling.

Simulation may also be adopted in the Bayes linear approach to derive estimates in complex models and also to study the properties of more complex estimation methods such as for the two-stage Bayes linear estimates that are particularly complex, having no analytical solution. Using simulation for Bayes linear estimation has the added advantage that only first and second-order moments need to be simulated. If variances are to be estimated as well, then we also require fourth-order moments. Simulating moments rather than full distributions as in MCMC represent a considerable gain in computing time and power.

The development of Bayes linear simulation methods has thus the potential to broaden its application to a wider range of statistical methodologies and application areas.

Bibliography

[1] M. Aitkin, D. Anderson and J. Hind (1981), *Statistical modelling of data on teaching styles.*, Journal of the Royal Statistical Society. Series A (General), 144(4):419-461, 1981.

[2] Barnett, V. and T. Lewis (1994). *Outliers in Statistical Data, 3rd Edition.* Wiley.

[3] Bayes, T. (1763). *An essay towards solving a problem in the doctrine of chances*, Philos. Trans. R. Soc. London, 53, 370418. Reprinted in Biometrika, 45, 1958, 293-315.

[4] Bennett, N. (1976), *Teaching styles and pupil progress.*, London, Open Books, 1976.

[5] Bernardo, J.M. and Smith, A.F.M. (1994). *Bayesian Theory.* John Wiley and Sons Inc., New York.

[6] Bryk A. S. and S. W. Raudenbush. (1992) *Hierarchical Linear Models.* Sage Publications.

[7] Chung Y., S. Rabe-Hesketh, A. Gelman, J. Liu, and V. Dorie (2011), *Avoiding Boundary Estimates in Linear Mixed Models Through Weakly Informative Priors*, U.C. Berkeley Division of Biostatistics Working Paper Series. Working Paper 284.

[8] Cochran, W. G. (1999). *Sampling Techniques.* Wiley.

[9] Congdon, P. (2003) *Applied Bayesian Modelling.* Wiley series in probability and statistics.

[10] Cronbach, L. J., Furby, L. (1970). *How we should measure "change": Or should we?* Psychological Bulletin, 74(1), 6880.

[11] Davis, J. A. (1966). *The Campus as a Frog Pond: An Application of the Theory of Relative Deprivation to Career Decisions of College Men* . American Journal of Sociology, Vol. 72, No. 1 (Jul., 1966), pp. 17-31

[12] de Finetti (1937) *Foresight: Its Logical Laws, Its Subjective Sources*, Annales de l′Institut Henri Poincar, 7, 168. (English translation by H.E. Kyburg, Jr. and H.E. Smokier, eds. in Studies in Subjective Probability. (1964, 2nd ed. 1980), Robert E. Krieger, Huntington, New York.

[13] de Finetti, B. (1974) , *Theory of Probability*, (translation by A Machi and AFM Smith of 1970 book) 2 volumes, New York: Wiley, 1974-5.

[14] Diaconis, P. and Freedman, D. (1984). *Partial exchangeability and sufficiency.*, Proceedings of the Indian Statistical Institute Golden Jubilee International Conference on Statistics: Applications and New Directions. J.K. Ghosh and J. Roy (eds.), Indian Statistical Institute, Calcutta, pp. 205236.

[15] Gelfand A. E., Adrian F. M. Smith (1990). *Sampling-Based Approaches to Calculating Marginal Densities.*, Journal of the American Statistical Association, Volume 85, Issue 410 (Jun., 1990),. 398-409

[16] Gelman A. (2005), *Analysis of variance: why it is more important than ever.*, Ann. Statist. 33(1): 1-53

[17] Gelman A. (2006), *Prior distributions for variance parameters in hierarchical models*, Bayesian Analysis (2006): 1(3),pp 515-533

[18] Gelman A. (2006), *Multilevel (Hierarchical) modeling: What it Can and Cannot Do*, Technometrics, August(2006): 48(3),pp 432-435

[19] Gelman, A. and J Hill. *Data Analysis Using Multilevel/Hierarchical Models.* Cambridge University Press, 2007.

[20] Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari and D. B. Rubin (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC.

[21] Goldstein, H. and M.J.R Healy (1995) *The Graphical Presentation of a Collection of Means*. Journal of the Royal Statistical Society. Series A, Vol 158, No.1, 175 - 177

[22] Goldstein, H. (1986) *Multilevel Mixed Linear Model Analysis using Iterative Generalized Least Squares*. Biometrika, Vol. 73, No.1, 43 - 56.

[23] Goldstein, H. (1991) *Multilevel Modelling of Survey Data*. The Statistician, Vol 4, No.2, 235 - 244.

[24] Goldstein, H. *Multilevel Statistical Models*. John Wiley and Sons Ltd, 2010.

[25] Goldstein, H., J. Carpenter, M.G. Kenward and K. A. Levin (2009) *Multilevel models with multivariate mixed response types*. Statistical Modelling 9(3): 173197

[26] Goldstein, M.(2006) *"Subjective Bayesian Analysis: Principles and Practice."*, Bayesian Anal. 1 (3) 403 - 420

[27] Goldstein, M. (1981) *Revising Previsions: a Geometric Interpretation (with Discussion).*, Journal of the Royal Statistical Society, Series B, 43(2), 105-130

[28] Goldstein, M. David Wooff (2007), *Bayes Linear Statistics, Theory Methods*, Wiley.

[29] Goldstein, M. and Wooff, D. A. (1995), *Bayes linear computation: concepts, implementation and programming environment*. Statistics and Computing. 5, 327-341.

[30] Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953) *Sample Survey Methods and Theory*. Vol. 1, Wiley, New York.

[31] Henderson, C.R., Kempthorne, O., Searle, S.R. and von Krosigk, C.M. (1959). *The Estimation of Environmental and Genetic Trends from Records Subject to Culling*. Biometrics, 15, 192-218.

[32] Hill, B. M. (1965), *Inference about Variance Components in the One-Way Model*. Journal of the American Statistical Association, Vol. 60, No. 311. (Sep., 1965), pp. 806-825.

[33] Hodges, J. S (1998). *Some algebra and geometry for hierarchical models, applied to diagnostics*. Journal of the Royal Statistical Society. Series B, Vol 60, No.3, 497 - 536

[34] Hox, J (2002) *Multilevel Analysis: Techniques and Applications*. Lawrence Erlbaum Associates, Inc., Publishers, (2002).

[35] Kreft, I. G. G. and J. de Leeuw (1998). *Introducing multilevel modeling*, Thousand Oaks: Sage.

[36] Lad, F. (1996). *Operational Subjective Statistical Methods: A Mathematical, Philosophical, and Historical Introduction*. John Wiley, New York.

[37] Lauritzen, Steffen, (2003), *Rasch models with exchangeable rows and columns*. Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting

[38] Lindley, D. V. (2000). *The Philosophy of Statistics*. Journal of the Royal Statistical Society. Series D. Vol. 49, No.3, 293 - 337.

[39] Lindley, D.V. and Novick, M.R. (1981). *The Role of Exchangeability in Inference*. Ann. of Statistics, 9, 4558.

[40] Patterson, H. D. and R. Thompson *Recovery of Inter-Block Information when Block Sizes are Unequal*. Biometrika, Vol. 58, No. 3 (Dec., 1971), pp. 545-554

[41] Perry, C. and Greig, I. D. *Estimating the Mean and Variance of Subjective Distributions in PERT and Decision Analysis*. Management Science, Vol. 21 (1975), pp. 14771480.

[42] Pukelsheim, F. (1994) *The three sigma rule*. The American Statistician 48 (1994) 8891.

[43] Ramsey, F. P. (1926) *Truth and Probability*. The Foundations of Mathematics and other Logical Essays, Ch. VII, p.156-198, edited by R.B. Braithwaite, London: Kegan, Paul, Trench, Trubner Co., New York: Harcourt, Brace and Company.

[44] Rao, J. N. K. (2003). *Small Area Estimation*. John Wiley and Sons Ltd, 2003.

[45] Scheff, H. (1959). *The analysis of variance*. John Wiley and Sons Ltd, 1959.

[46] Searle, S. R., C. E. McCulloch and G. Casella. (1992) *Variance Components*. John Wiley and Sons Ltd, 1992.

[47] Shouls S, P. Congdon and S. Curtis (1996) *Modelling inequality in reported long term illness in the UK: combining individual and area characteristics*. Journal of Epidemiology and Community Health, 1996; 50(3), 336 - 376

[48] Singer J. D. and J. B Willett (1993) *It's about time: Using discrete-time survival analysis to study duration and the timing of events* . Journal of Educational Statistics, 18(2), 155195

[49] Robert, C. P. (2001), *The Bayesian Choice*. Second Edition, New York: Springer-Verlag.

[50] Rodriguez A, Dunson DB, Gelfand AE. (2008) *The nested Dirichlet process, with Discussion.*, Journal of the American Statistical Association. 2008;103:11311144. 279.

[51] Savage, L.J. (1954). *The Foundations of Statistics*. John Wiley and Sons, New York

[52] Snijners, T. A. B. and R.J Bosker. *Multilevel analysis: an introduction to basic and advanced multilevel modeling*. Sage Publications 1999.

Appendix A

Table of Notations

The table below lists some of the more important notations used in this thesis. The notation, a brief description, and the section in which it was first defined is provided.

Notation	Description	Section
y_i	response variable for unit i in a single level regression	2.5.4
β_0	population intercept in a single level regression	2.5.4
β_1	population slope in a single level regression	2.5.4
x_i	predictor variable for unit i	2.5.4
ϵ_i	residual error term for unit i in a single level regression	2.5.4
y_{ji}	response variable for unit i nested in group j a two-level regression	2.5.5
β_{0j}	random intercept in a two-level regression	2.5.5
ϵ_{ji}	level 1 residual error term for unit i nested in group j	2.5.5
n_j	No. of level 1 units in group j	2.5.5
J	No. of level 2 groups	2.5.5
C	a collection of random quantities	2.13.1
B	a collection of beliefs	2.13.1
D	a collection of data quantities	2.13.1
Σ	variance of vector Z_i in second order exchangeability	2.13.2
Γ	Covariance of vector Z_i and Z_j in second order exchangeability	2.13.2

Continued ...

Notation	Description	Section
$\mathcal{M}(Z)$	the population mean of vector Z in the representation theorem	2.13.2
$\mathcal{R}_i(Z)$	the population residual of vector Z in the representation theorem	2.13.2
$\bar{Z}_n$	sample mean vector in Bayes linear sufficiency	2.14.1
V_R	Variance of $\mathcal{R}_k(Z)$ in adjustment of variance	2.14.3
V_k	sequence of squared residuals $V_k = [\mathcal{R}_k(Z)]^2$ in adjustment of variance	2.14.3
$\mathcal{M}(V)$	underlying population mean in the representation theorem for variance adjustment	2.14.3
V_M	variance of $\mathcal{M}(V)$	2.14.3
$\mathcal{R}_k(V)$	population residual of V in the representation theorem for variance adjustment	2.14.3
s_n^2	sample variance	2.14.3
T	population (squared) residual terms in the representation for s_n^2	2.14.3
V_T	variance of T	2.14.3
$S(z)$	a standardized observation	2.15.2
$Dis(z)$	discrepancy for an observation	2.15.2
$Dr(Y)$	discrepancy ratio for a multivariate (or multilevel) quantity Y	2.15.3
$Dis_Y(B)$	discrepancy of the adjustment vector $Dis(E_Y(B))$ for the collection of parameters in vector $B = (B_1, B_2, \ldots, B_r)$	2.15.4
$Dr_Y(B)$	adjustment discrepancy ratio of $B = (B_1, B_2, \ldots, B_r)$	2.15.4
$E_{[F/D]}(B)$	partial adjustment of B by F given D	2.15.5
$Size_d(B)$	size of an adjustment of the collection B when the observed value of $D = d$	2.15.5
$Size_{[f/d]}(B)$	size of the partial adjustment, or the partial size	2.15.5

Continued ...

Notation	Description	Section
$\mathbb{Z}_d(B)$	bearing of the adjustment of B when we observe $D = d$	2.15.5
$\mathbb{Z}_{[f/d]}(B)$	partial bearing	2.15.5
$Sr_{[f/d]}(B)$	partial size ratio	2.15.5
$PC(d, [f/d])$	path correlation	2.15.5
y_{ji}	outcome or response variable of unit (student) i in group (class) j	3.1
n_j	level 1 sample size (students) in group (class) j; $n_j = n$ in balanced designs	3.1
J	level 2 sample size (number of classes)	3.1
μ	prior expectation for y_{ji}	3.1
σ_y^2	variance of y_{ji}	3.1
$\mathcal{M}(y)$	population grand mean	3.1
$\mathcal{M}(y_j)$	population group j mean	3.1
$\mathcal{R}_i(y_j)$	level 1 residual	3.1
$\mathcal{R}_j(\mathcal{M}(y))$	level 2 residual	3.1
σ_ϵ^2	level 1 variance, $Var(\mathcal{R}_i(y_j)) = \sigma_\epsilon^2$	3.1
$\sigma_u^2 - \gamma$	level 2 variance, $Var(\mathcal{R}_j(\mathcal{M}(y))) = \sigma_u^2 - \gamma$	3.1
γ	variance of $\mathcal{M}(y)$	3.1
ρ	intra-class or intra-cluster correlation	3.3
z_{ji}	level 1 predictor of a SOEREG model	3.4.2
$\beta_{[0]j}$	intercept of regression in group j	3.4.2
$\beta_{[1]j}$	slope of regression in group j	3.4.2
ϵ_{ji}	level 1 residual error term	3.4.2
$\mathcal{M}(\beta_0)$	underlying population mean intercept	3.4.2
$\mathcal{M}(\beta_1)$	underlying population mean slope	3.4.2
$\mathcal{R}_j(\beta_0)$	residual for regression intercept in group j	3.4.2
$\mathcal{R}_j(\beta_1)$	residual for regression slope in group j	3.4.2
μ_0	prior mean for intercept $\beta_{[0]j}$	3.4.3
μ_1	prior mean for slope $\beta_{[1]j}$	3.4.3

Continued ...

Notation	Description	Section
σ_0^2	variance of intercept $\beta_{[0]j}$	3.4.3
σ_1^2	variance of slope $\beta_{[1]j}$	3.4.3
γ_0	covariance between intercepts, i.e $Cov(\beta_{[0]j}, \beta_{[0]j'})$	3.4.3
γ_1	covariance between slopes, i.e $Cov(\beta_{[1]j}, \beta_{[1]j'})$	3.4.3
ρ_{01}	correlation between intercepts and slopes	3.4.3
$\bar{y}_{j.}$	group j sample mean of y_{ji} based on a sample size n_j	4.2
$\bar{D}_n$	collection of group means $\{\bar{y}_{1.}, \bar{y}_{2.}, \ldots, \bar{y}_{J.}\}$	4.2
$E_{\bar{D}_n}(\mathcal{M}(y))$	adjusted expectation of $\mathcal{M}(y)$ by $\bar{D}_n$	4.3
$Var_{\bar{D}_n}(\mathcal{M}(y))$	adjusted variance of $\mathcal{M}(y)$ by $\bar{D}_n$	4.3.1
$R_{\bar{D}_n}(\mathcal{M}(y))$	resolution for the adjustment of $\mathcal{M}(y)$	4.3.2
$\mathcal{C}_{R_j}$	vector of level 2 residuals	4.4.1
$\mathbb{T}_{B:D}$	resolution transform matrix for the adjustment of B by D	4.5
λ_r	rth canonical resolution, the rth (ordered) eigenvalue of $\mathbb{T}_{B:D}$	4.5
W_i	ith canonical quantity	4.5
n_{opt}, J_{opt}	optimal level 1 and level 2 sample sizes respectively	4.10.3
G	size of finite population of level 2 units (groups)	4.13
N	size of finite population of level 1 units (individuals within groups for a balanced design)	4.13
J	sample size of level 2 units, $J \leq G$	4.13
n	sample size of level 1 units from each group (balanced data), $n \leq N$	4.13
$\mathcal{M}^{[N]}(y_j)$	finite population group j mean	4.13
$\mathcal{M}^{[G]}(y)$	finite population grand mean	4.13
$\mathcal{R}_i^{[N]}(y_j)$	finite level 1 residual	4.13
$\mathcal{R}_j^{[G]}(\mathcal{M}^{[N]}(y))$	finite level 2 residual	4.13
$\mathcal{M}(V_\epsilon)$	population level-1 residual variance	5.1

Continued ...

Notation	Description	Section
$\mathcal{R}_{ji}(V_\epsilon)$	uncorrelated sequence with variance equal to shape of population distribution	5.1
$V_{R(V_\epsilon)}$	prior variance of $\mathcal{R}_{ji}(V_\epsilon)$	5.1
V_{R_ϵ}	prior expectation E($\mathcal{M}(V_\epsilon)$)	5.1
V_{M_ϵ}	prior variance Var($\mathcal{M}(V_\epsilon)$)	5.1
T_ϵ	fourth order quantity for updating level-1 variance	5.1
$BLIMVE$	Bayes Linear Minimum Variance Estimator	6.1
β_j	population intercept in group j	6.2
$\mathcal{M}(\beta)$	population mean intercept in a SOEREF model	6.2
$\mathcal{R}_j(\beta)$	residual intercept in a SOEREF model	6.2
$\mathcal{M}(V_\beta)$	population intercept residual variance	6.2
V_{R_β}	prior expectation for population intercept residual variance	6.2
δ_j	residual from OLS fit	6.3
α_j	group j intercept in a basic SOEREG model	6.15
β_j	group j slope in a basic SOEREG model	6.15
$\mathcal{M}(\alpha)$	population mean intercept	6.15
$\mathcal{R}_j(\alpha)$	population residual intercept	6.15
$\mathcal{M}(\beta)$	population mean slope	6.15
$\mathcal{R}_j(\beta)$	population residual slope	6.15
V_{R_α}	prior expectation for population intercept residual variance	6.15
V_{R_β}	prior expectation for population slope residual variance	6.15
$\rho_{\alpha\beta}$	population correlation between residual intercept and slope	6.15
$\mathcal{M}(\boldsymbol{\beta})$	a vector of population mean intercept and slopes	6.17
$\boldsymbol{\epsilon}$	a vector of level-1 residuals	6.17
$\mathcal{R}(\boldsymbol{\beta})$	a vector of level-2 residual intercepts and slopes	6.17

End of table

Appendix B

R and [B/D] codes for Chapter 4

Here we present the R codes used to calculate and plot the discrepancy measures in Section 4.7 , as well as an R function, BALM (BAyes Linear Modeling), used to check our [B/D] outputs for the Bayes linear analysis of the SOEREF model. We also present our [B/D] codes for adjustment of the SOEREF model using the (unbalanced) STAT1010 data.

```
#Reads the examinations marks of the STAT1010 data in R
exams ←read.table("exams.txt",header=T)
#Calculates the group j means that are Bayes linear sufficient for the
    adjustments.
ybar_j ←sapply(exams,mean,na.rm=T)
#A function to calculate length of columns in a dataframe by ignoring NA's
lenna ←function(x){sum(!is.na(x))}
n_j ←apply(exams,2,lenna)#Group j sample sizes
J ←ncol(exams)
#Prior specifications
mu ←55
gamma ←56.3
sigma_sq_e ←237
V2 ←59 #The variance of the level_2 residual=(sigma^2_u - gamma).
#Variance of data means ybar_j only, ignoring M(y)
V3 ←V2+sigma_sq_e/n_j
```

```
#Calculate standardized observations (equation 4.44) & box-plot (Figure
  4.1(a))
sdfun ←function(x){(x-mu)/sqrt(gamma+sigma_sq_e+V2)}
S_y ←apply(exams,2,sdfun)
boxplot(S_y,ylab="Standardized observation")
mtext("Management class  Engineering class",1,line=2,adj=0)
#Calculate discrepancy (equation 4.45) & box-plot (Figure 4.1(b))
disfun ←function(x){(x-mu)^2/(gamma+sigma_sq_e+V2)}
Dis_y ←apply(exams,2,disfun)
boxplot(Dis_y,ylab="Discrepancy")
mtext("Management class  Engineering class",1,line=2,adj=0)
text(locator(1),"Best")
#Calculates the discrepancy ratio (equation 4.46). Calculating V(D)^{-1},
    denoted VDinv.
m ←matrix(1,nrow=J,ncol=J)
VD ←diag(V3)+m*gamma
VDinv ←solve(VD)
Dr_ybar_j ←(ybar_j-mu)\%*\%VDinv\%*\%(ybar_j-mu)/J
```

The [B/D] codes for adjusting mean components in the SOEREF model. It is easier to initially formulate a balanced design as this simplifies both the setting of indices j and i and construction of the variance-covariance structures and then we delete the redundant elements to create an unbalanced design.

```
@stat1010
@Input prior mean and variances
element: m=55
var: v(1,m)=56.3
fvar: v(1,e.r.s,e.t.k)= 237*(.r.s=.t.k)
fvar: v(1,u.r,u.k)=59*(.r=.k)
@Formulate the level 1 and level 2 SOE models as in Definition 3.1.1 of
    the SOEREF model
assign: y.r.s=m.r+e.r.s
assign: m.r=m + u.r
```

```
  @Specify indices for a balanced design temporarily
42 index:j=1,1,7
  index:i=1,1,47
44 @Construct variance-covariance structures simultaneously for y.j.i and m.j
  cobuild: y.j.i, m.j
46 @Group quantities in bases for adjustments
  base y=y.1.$,y.2.$,y.3.$,y.4.$,y.5.$,y.6.$,y.7.$
48 base m0=m$
  data: <ydat.1,ydat.2,ydat.3,ydat.4,ydat.5,ydat.6,ydat.7> 47 @data
50 @Find the number of data in each group j and input the data
  for: j=1,1,7
52 c: %m=maxcase(ydat.[j])
  for: i=1,1,%m
54 data: y.[j].[i](1)=ydat.[j]([i])
  end:
56 @Delete the redundant elements created in lines 42 and 43 to obtain an
     unbalanced design.
  @This makes [B/D] modify the variance-covariance structures accordingly
     for the resulting unbalanced design
58 for: i=%m+1,1,47
  xelement: y.[j].[i]
60 end:
  end:
62 return:
  @data
64 28 55 31 97 77 33 64
  35 36 55 73 61 52 57
66 59 51 51 69 33 44 65
  49 32 43 59 52 64 52
68 47 36 63 80 69 43 89
  65 71 34 81 61 80 40
70 79 35 54 93 49 65 40
  35 47 52 71 59 71 73
```

```
72 23 45 33 67 64 67 72
   31 44 63 64 31 65 59
74 23 17 32 64 73 59 71
   64 30 15 87 65 75 32
76 36 39 36 67 49 75 65
   31 55 78 61 52 35 56
78 48 47 42 48 71 88 63
   33 59 61 61 63 47 73
80 19 36 33 67 28 67 48
   61 16 47 80 52 72 55
82 57 44 78 75 55 64 67
   40 44 41 73 55 48 55
84 56 36 49 68 44 59 40
   40 69 31 59 77 40 39
86 35 83 38 55 67 45 61
   48 -999 34 55 56 43 75
88 41 -999 49 56 33 49 53
   44 -999 35 79 36 65 44
90 36 -999 77 71 41 40 76
   63 -999 53 65 47 91 33
92 40 -999 -999 51 83 71 40
   68 -999 -999 81 52 53 47
94 36 -999 -999 67 47 40 64
   56 -999 -999 83 69 63 64
96 32 -999 -999 31 40 55 63
   24 -999 -999 81 45 67 53
98 71 -999 -999 87 43 52 52
   17 -999 -999 88 67 51 73
100 37 -999 -999 83 72 57 75
   29 -999 -999 65 47 47 23
102 45 -999 -999 87 57 29 33
   55 -999 -999 81 43 52 47
104 57 -999 -999 53 59 40 57
```

```
-999 -999 -999 55 84 67 -999
-999 -999 -999 83 53 79 -999
-999 -999 -999 45 -999 85 -999
-999 -999 -999 87 -999 27 -999
-999 -999 -999 59 -999 91 -999
-999 -999 -999 67 -999 -999 -999
@Read the file in [B/D] and perform the adjustments
BD>m:@stat1010
BD>adjust:[m0/y]
BD>show:a+
BD>show:v+
```

The R codes and outputs of BALM, a function for the adjustment of the unbalanced SOEREF model based on our closed-form calculations in Chapter 4.

```
BALM ←function(data,mu,gamma,sigma_sq_e,V2,FUN=lenna){
ybar_j ←sapply(data,mean,na.rm=T)
n_j ←apply(data,2,lenna)
V3 ←V2+sigma_sq_e/n_j
adjvar ←1/(1/gamma+sum(1/V3))
adjmean ←adjvar*(mu/gamma+sum(ybar_j/V3))
resolution←(1-1/(1+V0*sum(1/V3)))*100
adjres ←V2/V3*( (ybar_j-mu)-adjvar*sum((ybar_j-mu)/V3))
adjgpmean ←adjmean+adjres
nu ←V2/V3# nu  is the shrinkage factor
adjvar2 ←nu*sigma_sq_e/n_j + nu^2*adjvar
cv ←V2/(1/gamma+sum(1/V3))/V3
#Adjusted variance of M(Y_j)
adjvargpmean ←adjvar+adjvar2-2*cv
#Resolution for group means
resolgpmean←(V0+V2-adjvargpmean)/(V0+V2)*100
output ←list(Adjusted_Mean=adjmean,Adjusted_Variance=adjvar,
Resolution=resolution,Adjusted_Group_Mean=adjgpmean,
Adjusted_Variance_Group_Mean=adjvargpmean,Resolution_Group_Mean=
```

```
   resolgpmean)
return(output)
}
#Output from application of BALM to STAT1010 data
> BALM(exams,55,56.3,237,59)
$Adjusted_Mean
[1] 53.86296
$Adjusted_Variance
[1] 8.026269
$Resolution
[1] 85.74375
$Adjusted_Group_Mean
      C1       C2       C3       C4       C5       C6       C7
44.63574 46.02166 47.61118 68.51379 55.24316 57.74772 56.07587
$Adjusted_Variance_Group_Mean
      C1       C2       C3       C4       C5       C6       C7
5.328592 8.949704 7.528671 4.695275 5.099322 4.790163 5.328592
$Resolution_Group_Mean
      C1       C2       C3       C4       C5       C6       C7
95.37850 92.23790 93.47036 95.92778 95.57734 95.84548 95.37850
```

R codes and [B/D] programme for the partial sequential adjustments of Section 4.10.

```
#Partial Bayes linear analysis:Sequential adjustment
av←1/(1/gamma+cumsum(1/V3))#R's cumsum() function simplifies the
   calculation of the adjusted variances
#av: Sequentially adjusted variances. We check that the last value is
   equal to the full adjusted variance.
#       C1        C2        C3        C4        C5        C6        C7
#30.121628 20.996121 16.012681 12.809818 10.687620  9.161357  8.026269
 adjmean←adjvar*(mu/gamma+sum(ybar_j/V3))
 amean←av*(mu/gamma+cumsum(ybar_j/V3))
 # We check that the last value is equal to the full adjusted mean.
#      C1       C2       C3       C4       C5       C6       C7
#49.76047 48.21288 47.85719 52.23939 52.75839 53.51934 53.86296
```

```
plot(amean,type="l",ylim=c(40,70),ylab="Sequentially adjusted overall mean
    ",xlab="Class")
points(ybar_j)
abline(55,0,lty=2)
text(2,55.7,labels="Prior",cex=0.7)
text(2,49.5,labels="Adjusted",cex=0.7)
text(2,46,labels="Data",cex=0.7)
#Partial Resolution
resolution←1-1/(1+gamma*cumsum(1/V3))
>resolution
       C1        C2        C3        C4        C5        C6        C7
0.4649800 0.6270671 0.7155829 0.7724721 0.8101666 0.8372761 0.8574375
#Check: this final value of 0.8574375 = resolution whole data
> diff(resolution)# Incremental changes in resolution
        C2         C3         C4         C5         C6         C7
0.16208717 0.08851580 0.05688922 0.03769446 0.02710947 0.02016141
#Checked in [B/D] same results as above
channel:i3=unbal74.txt
m:@stat1010
adjust:[m/y1]
show:a+,v+
adjust:[+/y2]
show:a+,v+
....until adjust:[+/y7]
```