This is to certify that the

dissertation entitled

THE APPLICATION OF GIBBS SAMPLING TO NESTED VARIANCE COMPONENTS MODELS WITH HETEROGENOUS WITHIN-GROUP VARIANCE

presented by

RAFA M. KASIM

has been accepted towards fulfillment of the requirements for

Ph.D. degree in Measurement & Qunt. Mthd.

Dr. Stephen W. Raudenbush

Major professor

Stephen W Raudenbush

Date March 15, 1994

THE APPLICATION OF GIBBS SAMPLING TO NESTED VARIANCE COMPONENTS MODELS WITH HETEROGENOUS WITHIN-GROUP VARIANCE

By

Rafa M. Kasim

A DISSERTATION

Submitted to
Michigan State University
College of Education
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Counseling, Educational
Psychology and Special Education

1994

ABSTRACT

THE APPLICATION OF GIBBS SAMPLING TO NESTED VARIANCE COMPONENTS MODELS WITH HETEROGENOUS WITHIN-GROUP VARIANCE

By

Rafa M. Kasim

Research in social sciences such as education, psychology, and sociology often involves analyzing hierarchically structured data. The use of Bayesian procedures with hierarchical linear regression models to analyze such data helped researchers obtain answers to questions where standard approaches seem to fail. It also improved the estimates of the regression coefficients by obtaining their empirical Bayes estimates.

Empirical Bayes estimates of the regression coefficients at two levels of hierarchy are usually conditioned on the maximum likelihood estimates of the variance components, and often assuming homogeneity of within-group variance. There are times however, where research interest is focused on studying the variance components, especially when there exists a clear evidence of heterogeneity of variance, and there is a great concern about the effect of the uncertainty in estimating these variances on the empirical Bayes estimates of the regression coefficients. A mixed model with random intercept, which permits heterogenous within-group variances, $\{\sigma_j^2\}$ for $j=1,\cdots,k$, across the k units is introduced within the Bayesian approach. Information about the within-group

variance and the intercept from these k units represent their prior distributions in a second level of the analysis. The marginal posterior distributions of the variance components and regression coefficients parameters are obtained via Gibbs sampling. The goal is to obtain Bayes estimates of the parameters of interest, especially those of the within-group variances, and study the effect of adjusting for the uncertainty in estimating the variance components on the estimation of the regression coefficients. Bayes estimates of the parameters for the specified model, with heterogenous within-group variance obtained via Gibbs sampling, are compared to their empirical Bayes estimates obtained via HLM analysis assuming homogeneity of variance.

The process of Gibbs sampling applied on several artificial data sets, and on one real data set to obtain the marginal posterior distributions of the parameters for the specified model. The real data set represents a nationwide random sample from the high schools in the U.S. Eighteen artificial data sets were generated to represent three models. The variation between the 18 data sets covers the complexity of the model used (number of predictors at the two levels), the degree of heterogeneity of within-group variance, and the number of groups in level-2. These data sets were generated to reflect the different models used in analyzing the real data set.

There appears to be no substantial difference between Bayes estimates (posterior means) and empirical Bayes estimates of the regression coefficients. Note that empirical Bayes estimates are based on the assumption of homogeneity of within-group variance, $\sigma_j^2 = \sigma_\circ^2$. When it comes to the estimation of the variance components, HLM estimates of the within-group variance component, $\sigma_\circ^2$ are found to be positively biased, especially when there exists clear evidence of heterogeneity of variance.

ACKNOWLEDGEMENT

I would like to thank my dissertation director Dr. Stephen Raudenbush for his constant help, support, and advice through my study and the writing of this dissertation. I am very fortunate to have him as a teacher, as an advisor, and best of all as friend. I would also like to express my appreciation to Dr. Dennis Gilliland for his valuable suggestion and comments on my work. Having him as a member in my dissertation committee helped me learn a lot from him. I would also like to thank the rest of the dissertation committee members Dr. Betsy Becker and Dr. Richard Houang for their support and encouragement. I would also like to express my appreciation to Dr. Michael Seltzer for his insightful comments on my work. My wife Hripsime and my kids Nader, Neda, and Nabeel deserve special thanks for their understanding and patience so I can finish my dissertation. Finally, but most importantly, I would like to express my deepest gratitude to my parents and my brothers and sister for their endless love and support.

TABLE OF CONTENTS

CHAPTER 1

Introduction

Social and educational research is usually conducted in natural settings such as schools or communities rather than in a controlled experimental setting. Natural settings often have a hierarchical structure. In education, for example, students are nested within classrooms, and classrooms are nested within schools. This hierarchical structure is reflected in the data collected from these nested levels. Observations at one level share common characteristics. In teacher-effectiveness studies, for example, students within a classroom share common characteristics of the teacher and his/her teaching method.

Student learning may be viewed as a result of social interaction within the classroom and the school system (Bandura, 1977). Students are assigned to classrooms, and they are taught in a planned and structured manner. Even though individual students respond differently to the same teaching process, their responses will have commonality. Therefore, observations on students learning cannot often be assumed independent. However, independence is one of the assumptions of many statistical procedures that researchers often use in their analyses.

Ignoring the hierarchical structure of the data leads to other statistical problems (Burstein, 1980; Cooley, Bond and Mao, 1981; Cronbach and Webb, 1975; Knapp, 1977; and Raudenbush and Bryk, 1988), such as aggregation bias. Aggregation bias occurs when the relationship between two variables varies (in magnitude and sometimes even in direction) for different levels of data analysis (Cooley, Bond and Mao, 1981; Cronbach and Webb, 1975; Robinson, 1950). For example, the relationship between socioeconomic status (SES) and academic achievement is more likely to be higher in school level data than it is in student level data. This difference in the relationship is attributed to the effect of aggregation of the data in each school. In general, school level membership is related to SES; schools with higher levels of SES are likely to have higher quality educational programs which lead to higher achievement in those schools.

Information loss is another substantive problem that results from ignoring the hierarchical structure in data analysis. After the research or the evaluation design is determined, a type of data analysis is usually chosen to answer questions related to one level of data. If the research questions were asked about students, it is likely that they will be answered with a student-level analysis. If questions were asked about classrooms, then it is likely that they will be answered with a classroom-level analysis. When the analysis is done at the classroom-level (i.e., using the classroom as unit of analysis), it is very possible that some important relationships between variables that were measured on the student level will be ignored.

However, there has been recent development of statistical methods which are appropriate for analyzing a hierarchically structured data based on multi-stage (often two-stage) linear models (Aitkin and Longford, 1986; Burstein, 1980; Goldstein, 1986; Lindley and Smith, 1972; Smith, 1973; Rao, 1972; Rosenberg, 1973; Raudenbush and Bryk, 1986). General linear mixed models or hierarchical linear models (HLM) are some of the names used to describe the statistical models developed for analyzing multilevel data.

The multi-stage methods share a common characteristic of treating the within-unit regression parameters as outcomes for the between-unit models. They differ, however, in the procedure for estimating variance components. Some of these procedures use the iterative generalized least squares method suggested by Goldstein (1986), some use the EM algorithm method as applied by Raudenbush and Bryk (1986), and other use the Fisher scoring approach (Longford, 1987). The estimates produced by all these numerical procedures are maximum likelihood point estimates. They are consistent, efficient, asymptotically unbiased, and normally distributed (Harville, 1977).

A two stage HLM can be presented in the following general form. Within unit j, for $j = 1, \cdots, k$ units, we have

$$Y_j = X_j\beta_j + \epsilon_j \ , \tag{1.1}$$

where Y_j is an $n_j \times 1$ vector of observations, X_j is an $n_j \times p$ matrix of within-unit predictor variables, β_j is a $p \times 1$ parameter vector which captures the relationships between Y_j and X_j within each unit, and ϵ_j is an $n_j \times 1$ vector of random errors which is assumed to have a multivariate normal distribution with mean vector 0 and variance covariance matrix $I\sigma_j^2$; that is $\epsilon_j \sim N(0, I\sigma_j^2)$. For X_j with rank $(X_j) = p$ the Ordinary Least Square (OLS) estimate for β_j is given by $\hat{\beta}_j = (X_j'X_j)^{-1}X_j'Y_j$ with a sampling variance of $Var(\hat{\beta}_j | \beta_j) = \sigma_j^2(X_j'X_j)^{-1}$.

By allowing the within-unit regression model coefficient to vary as a function of between-unit variables, we have a multi-level model that describes hierarchically structured data. Therefore, in the second (between-unit) stage, we have

$$\beta_j = W_j\gamma + U_j \ , \tag{1.2}$$

where W_j is a $p \times q$ matrix of known between-unit predictors and γ is a $q \times 1$ vector of parameters that capture the effect of the between-unit predictors W_j on the within-units parameters. U_j is assumed to be $N(0, \mathrm{T})$, where T is the residual variance-covariance matrix for β_j after accounting for the effects of W_j. Thus, we can have a more general

representation of the multilevel model by substituting 1.2 into 1.1 to obtain

$$Y_j = X_j W_j \gamma + \xi_j \text{ ,} \tag{1.3}$$

where $\xi_j = X_j U_j + \epsilon_j$. One might estimate the effects in 1.3 by using an ordinary least squares (OLS) multiple regression approach to the estimation of multiple simultaneous effects. However, one of the basic statistical assumptions for OLS is the homogeneity of variance for ξ_j. Because ξ_j is equal to $X_j U_j + \epsilon_j$, this assumption can only be valid when either U_j is set to equal zero or if $X_j = X$ for all j. Setting either U_j equal to zero or $X_j = X$, however, implies holding all the within-unit regression model parameters as a constant across all units and ignoring the hierarchical structure of the data. This leads to the violation of the assumption of independence.

The total variance $Var(\hat{\beta}_j | W_j)$ of the within-unit regression estimates is made of their sampling variance $Var(\hat{\beta}_j | \beta_j)$ plus their residual variance $Var(\beta_j | W_j)$. That is

$$Var(\hat{\beta}_j | W_j) = \sigma_j^2 (X_j' X_j)^{-1} + \mathbf{T} \text{ .} \tag{1.4}$$

Many applications have been found which can utilize the statistical model in 1.1 and 1.2. In longitudinal studies (Bryk and Raudenbush, 1987; Laird and Ware, 1982; Strenio, Weisberg and Bryk, 1983) the growth model parameters for each subject become the multivariate outcomes for the between-unit

models. Subject characteristics (either observed or controlled) can serve as predictors for these outcomes.

In research synthesis (Raudenbush and Bryk, 1985), a simple measurement model for the observed effect size is set equal to the true effect size plus some error in each study. This measurement model is a special form of the within-unit model in 1.1. The parameter β_j becomes the true effect size in study j and X_j becomes a scaler with a value of one. The true effect sizes vary across studies as a function of known study characteristics plus error in the form of the between-unit model in 1.2.

A major benefit of fitting models 1.1 and 1.2 to hierarchical data is the possibility of getting empirical Bayes estimates for the within-unit regression model parameters. Empirical Bayes estimates are more stable and outperform the classical estimates with regard to expected mean squared error. Information from other groups can be used to get improved estimates for the within-unit regression model parameters. In providing a prediction equation, Braun, Jones, Rubin, and Thayer (1983) presented a situation where there are several predictors and few cases in subgroups of the population which makes it hard to obtain least squares prediction equations. The objective of their study was to provide separate predictive equations for the Graduate Management Admission Test (GMAT) for the white and minority students in each business school. Only 4% of 8500 students

sampled from 59 business schools were minority students. The number of minority students within each school ranged from 2 to 10; this made it hard to obtain a prediction equation for minority students for a particular school. With the HLM analysis, information from other schools was borrowed to obtain the predictive equation for the minority students within each school.

Assuming that σ_j^2 and $\mathbf{T}$ are known, empirical Bayes estimates of β_j and γ are given by

$$\beta_j^* = \lambda_j \hat{\beta}_j + (1-\lambda_j) W_j \gamma^* \quad , \tag{1.5}$$

where

$$\lambda_j = \mathbf{T}(V_j+\mathbf{T})^{-1} \quad , \tag{1.6}$$

$$V_j = \sigma_j^2 (X_j' X_j)^{-1} \tag{1.7}$$

and

$$\gamma^* = \left\{ \sum_j W_j' [Var(\hat{\beta}_j | W_j)]^{-1} W_j \right\}^{-1} \left\{ \sum_j W_j' [Var(\hat{\beta}_j | W_j)]^{-1} \hat{\beta}_j \right\} . \tag{1.8}$$

The first component of β_j^* in 1.5 is the OLS estimate of β_j from the data within each unit weighted by its reliability λ_j. The second component, γ^* is an empirical Bayes estimate of the between-unit parameters, γ from 1.2. More comprehensive presentations of the HLM and different methods of estimation of the parameters of the model are given by

Lindley and Smith (1972), Smith (1973), and Raudenbush (1984, 1988).

Although empirical Bayes methods have been around for a long time, they became increasingly popular and were applied to many types of problems in the early 1970s. Efron and Morris (1972, 1973, 1975, 1977) discussed many applications of empirical Bayes methods. Empirical Bayes estimates of the parameters in hierarchical linear models are obtained when the true values for σ_j^2 and $\mathbf{T}$ are known. In practice, however, these values are rarely known to practitioners and often are estimated from the data. Common estimates often used for these two parameters are their maximum likelihood (ML) estimates. Since estimates of the regression coefficients of the HLM are functions of ML estimates of σ_j^2 and $\mathbf{T}$, they are also considered to be ML estimates (Raudenbush, 1988).

In classical ANOVA procedures σ_j^2 and $\mathbf{T}$ represent the within- and between-group variance components, respectively. In balanced designs these variance components can be estimated by solving a set of simultaneous linear equations obtained from equating the observed mean squares (which are quadratic forms of the observations) to their expected values (Searle, 1971). For unbalanced designs, there is more than one set of quadratic forms to use for estimating these variance components. Depending on the particular set used, inconsistent estimates of the variance components may be obtained (Searle, 1971).

Introducing HLM to the analysis of educational data can improve the quality of the data analysis in education. It, also helps applied statisticians obtain sensible answers to questions where standard approaches seem to fail. However, there continue to be obstacles related to HLM applications and theoretical derivations.

In some applications where the assumption of homogeneity of variance cannot be verified, drawing inferences on the variance components becomes part of the research interest, especially the variances of the within-unit residuals (Leonard, 1975; Rao, 1970; Raudenbush and Bryk, 1987). The motivation for making such inferences on the variances can be attributed to the intrinsic interest in the variances themselves and to the fact that they are part of the standard errors of estimates for the regression coefficients. In other applications, where we have small samples within each group and a small number of groups, maximum likelihood estimates for the variance components σ_j^2 and $\mathbf{T}$ might become unstable. Substituting these unstable estimates as true values for the variance components will distort the regression effect estimates.

Purpose of the Study

The purpose of this study is to obtain marginal posterior distributions for all effect parameters and variance

components in a linear mixed model with random intercepts without relying on the assumption of homogeneity of the residual variance σ_j^2 across groups. Non-informative priors will be posed on the hyper-parameters that describe the distributions for random parameters. The marginal posterior distribution for each parameter will then be approximated using the Gibbs sampling method (Tanner, 1993).

When research interest is focused on making inferences on variances, finding the entire posterior distribution for each of the variance components is more informative than having only a point estimate of the variance. Having the entire posterior distribution at her/his disposal will allow the researcher to have the flexibility in choosing an appropriate estimate for the variance. Further, s/he can establish probability intervals around the estimate.

Objectives of the Study

The idea of approximating the marginal posterior distributions for all parameters in the model leads to the question of how feasibly to do so. Therefore, the questions of this study are:

1- *How do the posterior mean (Gibbs) estimates of the parameters of a mixed model with heterogenous within-group variance differ from their empirical Bayes estimates which are based on the homogeneity of within-*

group variance assumption? (Empirical Bayes estimates for the within- and between-unit regression coefficients were conditioned on ML estimates of the variance components with homogeniety of within-group variance).

2- How do inferences about regression coefficients change when taking into account the uncertainty in estimating the variance components?

3- Can a single estimate be used as typical value for all the different residual variances $\{\sigma_j^2\}$?

4- How precise are the posterior mean (Gibbs) estimates of the residual variances $\{\sigma_j^2\}$ in estimating their single typical value?

The last two questions focus on the parameters of the prior distribution of the residual variances $\{\sigma_j^2\}$. Inferences can be made on these two parameters from their approximated marginal posterior distributions.

Bayes Solution

If reasonable priors are posed for all the unknown parameters in the model, theoretically, the marginal posterior distributions of the parameters of interest can be obtained by integrating out the other nuisance parameters in the model. Bayes estimates for these parameters can then be derived from their respective marginal posterior distributions based on a certain loss function.

From a Bayesian viewpoint, estimating a parameter θ, given its sufficient estimate $\hat{\theta}$ from the sample implies selecting a decision function, say D, so that $D(\hat{\theta})$ is a predicted value of θ when both the computed value $\hat{\theta}$ and the posterior density function $f(\theta|\hat{\theta})$ are known. Often, the researcher predicts an experimental value of any random variable to be its expected value; however, the median or the mode can also be used as predicted values. For many cases, it is desirable that the choice of the decision function D should depend upon the loss function say $L[\theta, D(\hat{\theta})]$, that we try to minimize (Hogg and Tanis, 1977). For example, when the loss function is given by:

$$L[\theta, D(\hat{\theta})] = [\theta - D(\hat{\theta})]^2 \tag{1.9}$$

then Bayes' solution that minimizes this loss function is given by $D(\hat{\theta}) = E(\theta|\hat{\theta})$. It is also possible to obtain Bayes' probability confidence intervals for the estimates directly from their respective marginal posterior distributions.

A problem associated with the application of Bayesian methods to hierarchical models is the difficulty of the mathematics involved in the derivation of the posterior distributions (Lindley and Smith, 1972). The multiple dimensions of the parameter space in the model make it difficult to express some of the required equations in closed form. Therefore, certain marginal posterior distributions

cannot be obtained analytically through direct evaluation of an integral, and in many applications, integrating out the nuisance parameters is a difficult job. Consequantlly, empirical Bayes estimates are used to approximate truly Bayesian estimates.

In some situations we might find that the focus of the research is centered around making inferences about σ_j^2 and **T** rather than the regression coefficients in the model. Therefore, obtaining the marginal posterior distributions for these parameters is more informative than obtaining their single point estimates. Even when concern focuses on the regression coefficients it will be useful to insure that posterior uncertainty regarding these coefficients fully reflects posterior uncertainty about the variance-covariance components.

Importance of the Study

In educational settings, where we have several classrooms or several schools, interest is mainly focused on studying the factors that affect students' academic achievement. Traditionally, student achievement is evaluated by comparing schools or classrooms on their means on some measure of academic achievement, usually students' test scores. Often, important decisions such as school funding or teacher promotion are based on this evaluation. The fairness

of these decisions depends on whether the test score means in different schools or classrooms truly represent students' academic achievement. Many factors that could affect these means can be controlled statistically through model specification. However, it is still possible that for a slightly higher test score mean, one school will get more funds than others or a specific teacher may get promoted. It is also possible that the school that got extra funds or the teacher who got promoted produced higher variability in their students' academic achievement.

A successful educational program or an effective teacher should arguably strive not only to increase and facilitate students' academic achievement, but also to reduce the gap between students' learning and produce equality in their achievement (Bloom, 1984). Therefore, the evaluation process should be based not only on average achievement, but also on the equity of the achievement. Consequently, any improvement on the methods of estimating these two criteria certainly helps the decision maker in his/her evaluation of the educational system.

Another example is drawn from research syntheses (Raudenbush and Bryk, 1985; Rubin, 1981) where a simple model for the observed effect size is set equal to unknown true effect size plus some error for each study. The objective is to obtain an efficient estimate of each study effect size as well as its precision. It is also common in some cases to

have a number of groups with a small number of observations for each group which were obtained under the same conditions. For example, in longitudinal studies, the growth of subjects from the same age can be observed over several points of time. One might become interested in studying the growth rate of a particular subject in comparison to other subjects. The replicate observations on the subject can then be used for variance estimation.

In all of above examples, there are two sets of parameters (means and variances) with several parameters of the same type in each set. Within each set, these parameters are related by common circumstances; schools or classrooms in the first example, studies in the second example, and the individual subject in the last example. Within the Bayesian framework, it is logical to assume that these many parameters that share common circumstances would have a common parametric prior distribution. This distribution summarizes the information about these parameters prior to data collection. This allows the researcher to use such prior knowledge about these parameters to get their improved estimates.

With the help of modern computers and developments in simulation theory, research interest is being directed towards approximating the posterior distributions of the parameters in the model being investigated and obtaining their Bayesian estimates (Tanner and Wong, 1987). In his work on hierarchical linear models, Seltzer (1988) adopted the data

augmentation method to obtain Bayesian estimates for variance components. He also investigated Bayes' estimates for the within-school regression coefficients when they are assumed to have a t-distribution (see also Seltzer, 1993). Fotiu (1989) also applied the method of data augmentation in estimating the joint posterior distributions for the effect parameters (within- and between-units) when analyzing many groups. Both Seltzer and Fotiu approximated the posterior distribution of σ_j^2 under the assumption that it is equal to σ^2 across all groups. This assumption is usually made for convenience, rather than because it necessarily holds true. By making this assumption, the researcher can pool all the observations from all the groups to get a large-sample point estimate of σ^2. However, when the marginal posterior distributions of all the parameters in the model can be approximated, this assumption becomes no longer essential to the analysis. In fact, sometimes the main objectives of the statistical analysis is to estimate the posterior distribution for each σ_j^2.

In conclusion, the objective of this study is to investigate the use of a mixed linear model with heterogenous variances $\{\sigma_j^2\}$ across groups. A linear regression model with a random intercept will be used to represent the relationship between the criterion variable and the predictors in each group. By obtaining the marginal posterior distribution for all the parameters in the model, the practitioner can obtain Bayes estimates for the effect parameters as well as variance-

covariance components. Based on the theoretical formulation of the model a number of simulated data sets will be generated to verify the accuracy of the estimation under different parameter specifications. These specifications reflect the degree of heterogeneity of variance, the different numbers of group sizes, and the complexity of the regression model for each group. The estimation process will be applied to a national sample of US high schools where math achievement has been studied as a function of school and student characteristics. One important feature of this data set is the heterogeneity of the residual variance across schools. Radenbush and Bryk (1987) investigated the heterogeneity of variance in this data set through the application of hierarchical linear models.

Organization of the Study

Chapter 2 will present a statement of the problem and review of literature conducted on the topic. That chapter will also highlight some of the problems in variance estimation, and its relation to the HLM. In chapter 3, a general linear mixed model with a random intercept will be presented to represent the several applications presented in this study. In addition, a comprehensive description of the assumptions associated with the model will be provided within the Bayesian framework. Chapter 4 will explain the Gibbs

sampler and its application in achieving the objectives of the study. An example will be provided to illustrate the iteration process of Gibbs sampling. Chapter 5 will present the derivation of the conditional distributions for the parameters needed in the iteration process. It will also provide a description to the computational steps for the iteration process presented in chapter 4. Chapter 6 will presents the results of the small simulation study as well as the results of an analysis of mathematic achievement data from US high schools. Discussion of the results for the generated data and the US high school data, as well as conclusions, will be presented in chapter 7.

CHAPTER 2

Statement of the Problem

Variance estimation plays an important role in almost every kind of quantitative research. A basic requirement of nearly all forms of analysis is that a measure of precision be provided for each estimate derived from the data. The most commonly used measure of precision of an estimator is the reciprocal of its variance. The sample variance is used to estimate the precision of the mean or other location estimates. It is also used to check for gross errors affecting a single observation. Further, it is used to provide an estimate of a variance component such as a pooled within sample variance in several kinds of statistical procedures.

In HLM analysis, the two variance-covariance components σ_j^2 and $\mathbf{T}$ are used in obtaining empirical Bayes estimates for the effect parameters in equations 1.1 and 1.2. Compared to the Bayesian approach, the empirical Bayes approach does not require the specification of prior distributions for σ_j^2 and $\mathbf{T}$. The two-level hierarchical linear model and its assumptions for the empirical Bayes approach can be restated as

$$Y_j = X_j\beta_j + \epsilon_j\ , \qquad \text{where} \qquad \epsilon_j \sim N(0, I\sigma_j^2)\ , \tag{2.1}$$

for the first level, and

$$\beta_j = W_j\gamma + U_j\ , \ \text{where}\ U_j \sim N(0,\mathrm{T}),\ \text{and}\ \gamma \sim N(\Theta,\Gamma), \tag{2.2}$$

for the second level. The joint posterior distribution for the effect parameters β_j and γ is conditioned on the values of σ_j^2 and T (Lindley and Smith, 1972). The joint density function $f_1(\beta_j, \gamma \mid Y_j, \sigma_j^2, \mathrm{T})$ for this posterior distribution can be expressed as

$$f_1(\beta_j,\gamma \mid Y_j,\sigma_j^2,\mathrm{T}) \propto f_2(Y_j \mid \beta_j,\sigma_j^2)\, f_3(\beta_j \mid \sigma_j^2,\ \gamma,\mathrm{T})\, f_4(\gamma)\ , \tag{2.3}$$

where $f_2(Y_j \mid \beta_j, \sigma_j^2)$ is the likelihood function of the data, $f_3(\beta_j \mid \sigma_j^2,\ \gamma, \mathrm{T})$ represents the conditional prior density function of β_j, and $f_4(\gamma)$ is the density function for a noninformative prior for γ. The values for σ_j^2 and T are often estimated by one of several methods of numerical estimation. One common method is the EM algorithm (Dempster, Laird and Rubin, 1977) which produces their maximum likelihood estimates. These estimates are asymptotically normally distributed, unbiased and efficient estimates.

Problems Associated with the Estimation of σ^2 and T

In some applications, the sample size within each group as well as the number of the groups might not be large enough to allow the specific estimates of σ_j^2 and $\mathbf{T}$ to obtain the asymptotic properties. Nevertheless, practitioners often use maximum likelihood estimates of σ_j^2 and $\mathbf{T}$ to represent their true values to obtain the conditional posterior distributions for the effect parameters.

Problems Associated with Estimating T

In 1.2 of the HLM, $\mathbf{T}$ represents the residual variance-covariance matrix for the within-group parameters $\boldsymbol{\beta}_j$ across all the groups. The shape of the sampling distribution of this variance-covariance matrix depends on the number of groups in the study. Consider the model investigated in this study, for example, where only the intercept is considered random. The matrix $\mathbf{T}$, then, becomes scalar τ^2 , which is the variance of the intercept across all the groups. When the number of groups is relatively small, say $k \leq 10$, it is possible that the distribution of an estimate of τ^2 becomes highly skewed. Using the mode of that distribution (maximum likelihood estimate) as an estimate for τ^2 might not be a good representation of its true value.

In a validation study conducted on eight law schools, Rubin (1983) pointed out the problem of using maximum

likelihood estimates for τ^2 when there are small number of schools in the study. He stated that

> "The specific problem here is that the likelihood of τ^2 is not at all symmetric about the maximum likelihood estimate, and thus this estimate is not representative of reasonable values for τ^2. Integrating over τ^2 in such cases is a much more reasonable way to summarize evidence than to fix τ^2 at some value" (Rubin, 1983, p. 15).

Notice that the estimated value of τ^2 is being substituted in 1.4 to 1.8 to obtain the precision weight for the empirical Bayes estimates for the regression coefficient of the HLM model. Consequently, invalid estimation of this parameter causes the estimates of the regression coefficients of the HLM model in 1.5 and 1.8 and their reliability estimates in 1.6 to be distorted.

Problems Associated with an Invalid Assumption of Homogeneity of Variance

When maximum likelihood estimates are obtained for variance components, $\{\sigma_j^2\}$ for $j = 1, \cdots, k$, are often assumed homogeneous. Researchers often assume homogeneity of variances across all groups out of convenience rather than conviction. This assumption allows the pooling of the observations from all the groups to get one large sample maximum likelihood estimate of σ^2 for all the groups. Berlin (1984) stated that

> "large data sets are rarely homogeneous in their precision and that usual statistical analysis fails if it does not take such differences into consideration." (p. 209).

When the number of observations in each group is large enough to allow for consistent maximum likelihood estimation, Mason, Wong and Entwisle (1984) advised that a separate variance estimate should be obtained for each group.

However, when the number of observations in each group is not large enough to allow for consistent maximum likelihood estimation and the parameter variances are not equal, pooling the observations from all the groups to obtain a single ML estimate for σ^2 might give an inaccurate estimate of the variance. Assuming homogeneity of within-group variance Bassiri (1988) showed that the within-group variance estimate becomes unstable when the groups' sample sizes are relatively small. This could lead to invalid estimates of the regression coefficients β_j and their precision estimates (see equations 1.4 and 1.5). Consequently, inferences about or confidence intervals for the regression effects would be invalid.

Equation 1.4 shows when the true values of σ_j^2 and $\mathbf{T}$ were being used, part of the variability in the estimated variance of the regression effect β_j stems from the variability in X across groups. Another part stems from the variability in σ_j^2's when no homogeneity of variance assumption is being made. By assuming homogeneity of variance, $\sigma_j^2 = \sigma^2$, we are ignoring part of the variability in the variances of the regression effects. The part being ignored is attributable to the variability in the residual variances.

For illustrative purposes, consider the case where X_j is assumed equal to X across the k groups for $j=1, \ldots, k$. Thus, the within-unit regression coefficient estimates become $\hat{\beta}_j = (X'X)^{-1}X'Y_j$. When homogeneity of variance holds, the variance-covariance matrix for these estimates can be derived from 1.4 as

$$Var(\hat{\beta}_j \mid W_j) = \sigma^2 (X'X)^{-1} + \mathbf{T} \ . \tag{2.4}$$

Equation 2.4 implies that the within-unit effects are estimated with the same precision across all the k groups. However, if there is a considerable variability in the residual variances across groups, the estimation of this precision becomes invalid and those estimates of the regression effects no longer have the same precision.

In a classical analysis, when researchers study any treatment effect, they compare the means of the criterion variable from different treatments. One of the assumptions they often make is homogeneity of variance. The justification of this assumption is to get a single pooled variance estimate for the within treatment variance to represent the error term in the analysis so that treatment effect can be tested. Bryk and Raudenbush (1988) showed that heterogeneity of variance in the traditional analysis of treatment effects can be seen as evidence of an interaction between treatment and subject-specific characteristics. They warned researchers against the problem of obtaining biased estimates for treatment effect

when ignoring heterogenous variances in the classical analysis.

Consider the classical case, often referred to as the Behrens-Fisher case, of testing the hypothesis $\mu_a - \mu_b = 0$ with unequal population variances, $\sigma_a^2 \neq \sigma_b^2$. Using $\overline{X}_a - \overline{X}_b$ as an estimate for $\mu_a - \mu_b$, the sampling distribution of the statistic

$$Z = \frac{(\overline{X}_a - \overline{X}_b) - (\mu_a - \mu_b)}{\sqrt{(\frac{\sigma_a^2}{n_a}) + (\frac{\sigma_b^2}{n_b})}} \tag{2.5}$$

is $N(0,1)$. The given hypothesis cannot be tested without knowing σ_a^2 and σ_b^2. When the sample sizes n_a and n_b are both large enough, we can substitute the unbiased estimates of the variances, S_a^2 and S_b^2, for the corresponding parameters, and the resulting statistic has a sampling distribution approximated by $N(0,1)$. In practice, however, when n_a and n_b are both relatively small and noticeably unequal, the resulting statistic

$$Z^* = \frac{(\overline{X}_a - \overline{X}_b) - (\mu_a - \mu_b)}{\sqrt{(\frac{S_a^2}{n_a}) + (\frac{S_b^2}{n_b})}} \tag{2.6}$$

would have what is known by Behrens-Fisher distribution (Winer, 1971). Therefore, when there are both heterogeneity of variance and a noticeable difference in the sample sizes between treatments, the homogeneity of variances assumption becomes questionable.

The Analysis of Many Group Variances σ_j^2

A common situation where many variances need to be estimated arises when there is a heterogeneity of variance in linear models. Rao (1970) introduced a method called Minimum Norm Quadratic Unbiased Estimation (MINQUE) for estimating the residual variances in linear models when these variances are found to be heterogeneous. MINQUE estimates are a linear function of the distinct variances under certain conditions. These conditions are related to the choice of the matrix that creates a quadratic form in the outcome variable. If some of the variances are the same, their corresponding coefficients in the linear function can be chosen to be the same. The variances estimated using the MINQUE method can be used to obtain improved estimates of the coefficients of the linear model. Further, they can be used in obtaining the estimate of the precision of the simple least squares estimator of the regression coefficients or any linear function of these coefficients.

Cook and Weisberg (1983) modeled the variability in variances as a function of some explanatory variables. The n residuals from a linear regression model are assumed to have a multivariate normal distribution, with mean zero and variance-covariance matrix $\sigma^2 W$, where W is a diagonal matrix with all diagonal elements $w_i > 0$ for $i=1,\ldots,n$. Estimating the residual variances implies estimating the elements of W.

The variability in w_i, which represents the variability in the variances, is expressed as a function of a $1 \times q$ row vector of known explanatory variables $z_i = (z_{ij})$ for $j = 1, \ldots, q$ and $q \times 1$ vector of unknown parameters λ as in the following two models

$$w_i = \exp\left(\sum_{j=1}^{q} \lambda_j z_{ij}\right), \tag{2.7}$$

or

$$w_i = \exp\left(\sum_{j=1}^{q} \lambda_j \log(z_{ij})\right). \tag{2.8}$$

Note that the variables z_{ij} can have negative values in the first model but not in the second one. The predicted values of w_i, from the above models, can then be used to estimate the residual variances from their variance-covariance matrix $\sigma^2 W$.

Cox and Solomon (1986) have also dealt with the issue of estimation of many variances. They provided examples illustrating the problem of estimating many variances from small samples. One of the examples covers the situation where there is a systematic difference in variance between samples. Observations within a sample are assumed to come from a $N(\mu_i, \sigma_i^2)$ population where $i = 1, \ldots, k$. The ith variance σ_i^2 is a function either of an explanatory variable z_i, characterizing the ith population, or of μ_i. They adopted similar versions of the models in 2.7 and 2.8 to represent the systematic change in the variances.

They also presented another example for the case where

the changes in the variances are considered random. They consider it a complement to the above case (where changes in the variances are systematic) and called it the "overdispersion model", (Cox and Solomon, 1986, p. 544). In this case, different population variances are considered to be independent unobserved values of a random variable σ_i^2. This variable has an inverse gamma distribution with density

$$\left(\tfrac{1}{2}\nu_\circ\sigma_\circ^2\right)^{\frac{1}{2}} \left(\sigma_i^2\right)^{\frac{-1}{2\nu_\circ}} \exp\left(\frac{-\nu_\circ\sigma_\circ^2}{\sigma_i^2}\right) \left\{\Gamma\left(\frac{1}{2\nu_\circ}\right)\right\}^{-1}, \qquad \textbf{(2.9)}$$

where $\nu_\circ$ represents the degrees of freedom, and $\sigma_\circ^2$ some constant.

Given the inverse Gamma density in 2.9 and considering the *i*th population mean μ_i as a nuisance parameter, the estimation of the *i*th population variance σ_i^2 is based on the marginal likelihood of the *i*th sample. The underlying idea of the above model suggested by Cox and Solomon (1986) coincides very closely with Bayesian thinking.

Bayesian Approach

Lindley (1971) used the Bayesian approach in the estimation of many different means and variances. In the case where both means and variances were unknown but have exchangeable distributions, he used the conjugate priors for these distributions to derive the joint posterior density for the means and variances. The joint density of all the parameters was based on a three-stage model of hierarchy. The

first stage describes the data, and its likelihood function given the means and the variances. That is, given $\{\mu_j\}$ and $\{\sigma_j^2\}$ the data Y_{ij} are independent and normally distributed with $E(Y_{ij}) = \mu_j$ and $Var(Y_{ij}) = \sigma_j^2$ for $i = 1,\cdots,n_j$; $j = 1,\cdots,k$. In the second stage, the means $\{\mu_j\}$ and the variances $\{\sigma_j^2\}$ are assumed to be independent with two different exchangeable distributions. The exchangeable distribution for the means $\{\mu_j\}$ is assumed to be normal with mean θ and variance τ. The exchangeable distribution for the variances $\{\sigma_j^2\}$ is an inverse chi-square, where the variable $\nu_\circ\sigma_\circ^2/\sigma_j^2$ is distributed as a χ^2 with $\nu_\circ$ degrees of freedom and $\sigma_\circ^2$ as a typical value for σ_j^2. The third stage describes the prior knowledge about the parameters in the second stage. For the parameters in the normal distribution a conjugate prior distribution for the mean θ with vague prior knowledge produces locally uniform prior. Similar to σ_j^2, the variance τ is assumed to have an inverse chi-square where $\nu_\tau\sigma_\tau^2/\tau$ distributed as χ^2 with ν_τ degrees of freedom and σ_τ^2 as a typical value for τ. Since σ_j^2 is distributed as an inverse chi-square, a conjugate prior for $\sigma_\circ^2$ given $\nu_\circ$ is a function of χ^2 with r and λ as two constants describing the distribution. The parameter $\nu_\circ$ is assumed to have a uniform prior on $(0,\infty)$.

The joint posterior density for the means $\{\mu_j\}$ and the variances $\{\sigma_j^2\}$ was derived by integrating out all other parameters (which are θ, τ, $\nu_\circ$ and $\sigma_\circ^2$) from the overall joint posterior density function. The derived estimates were based

on a simple model where there were several groups with different means and variances. If a more complex linear model with several predictors were to be used the estimation process then becomes more complex. The added coefficients for the predictors in the linear model make integrating the joint density function of the parameters analytically, extremely difficult.

Leonard (1975) provided a procedure for modeling the variability in the means and the variances from several normal populations using the Bayesian approach. His model for the means, when the variances are known, represents a general case of the one presented by Lindley (1971). When populations means are known, Leonard expressed the log-transformed variances as a function of some explanatory variables using a linear model similar to the one for the means. The log-transformed variances were assumed to have an asymptotic normal exchangeable distribution similar to the one for the means. Expressing the variances or any transformation of them as function of other variables is similar to what Cook and Weisberg (1983) did as shown in equation 2.7. The major difference here is that Cook and Weisberg consider λ_j in 2.7 to be fixed unknown parameters that need to be estimated, while Leonard proposed the use of prior knowledge about a similar parameter for the linear model of the log-transformed variances.

The resulting estimates of the means and variances are

shrinkage estimates. They represent a weighted average of the standard estimate from the sample plus a weighted average of all the samples estimates. The weights are related to the reliability of the standard estimate from the sample. The shrinkage estimates of the variances are found by taking the exponential of the shrinkage estimates of the log-transformed variances.

When both means and variances are unknown, Leonard proposed an iterative method for estimating the joint posterior mode vectors of the means and the variances. He substitutes the shrinkage variance estimates in the expression for the means to obtain their new estimates. The new estimates of the means are then used to find new shrinkage estimates for the variances.

Raudenbush and Bryk (1987) tackled the issue of estimating variances from many groups as a special case of a more general problem of estimating the parameters of a two-stage HLM. They provided two different methods for obtaining empirical Bayes estimates for the variances. The conceptualization of their models is similar to those presented by Lindley (1971) and Leonard (1975). In the first method, estimates were derived using the exact distribution (chi-square) of the variance estimate within each group. That is,

$$s_i^2 \sim \sigma_i^2 \chi^2_{n_i-1} / (n_i - 1) \ . \qquad \textbf{(2.10)}$$

Assuming that the variances have an exchangeable prior distribution, the variable $\nu_\circ\sigma_\circ^2/\sigma_i^2$ has a chi-square distribution with $\nu_\circ$ degrees of freedom and $\sigma_\circ^2$ represents the typical value for the variances prior to observing the data. This distribution is the natural conjugate prior for the one in 2.10. Thus, the variance σ_i^2 is distributed as $\nu_\circ\sigma_\circ^2\chi^{-2}_{(\nu_\circ)}$, where $\chi^{-2}_{(\nu_\circ)}$ denotes a variable with an inverse chi-square distribution.

In the second method, however, estimates were derived using the asymptotic normal approximation to the sampling distribution of the logarithmic transformed variance estimate within each group. For the first stage of the HLM

$$d_i=\delta_i+e_i\ ,\quad \text{where } e_i \sim N(0,(n_i-1)^{-1}/2)\quad \text{for } i=1,\ldots,k, \qquad \textbf{(2.11)}$$

where $d_i=\frac{1}{2}[\log(s_i^2)-c_i]$, c_i is a bias correction, and $\delta_i=\frac{1}{2}\log(\sigma_i^2)$. For the second stage of the HLM

$$\delta_i = W_i'\gamma + U_i\ ,\quad \text{where } U_i \sim N(0,\tfrac{v^{-1}}{2})\quad \text{for } i=1,\ldots,k\ , \qquad \textbf{(2.12)}$$

where W_i is a $M\times1$ vector of known predictors, γ is a $M\times1$ vector of effect parameters, and U_i is a random error normally distributed with mean equals zero and variance equal $\frac{v^{-1}}{2}$. The use of the normal approximation to the sampling distributions of the variances permits the use of the above formulation and allows for the hypothesis testing associated with it.

Estimates from both methods (exact distribution and normal approximation) are shrinkage estimates. They are

weighted averages of two components. For the exact distribution, the conditional posterior mean estimate for the variance is given by

$$\tilde{\sigma}_i^2 = \frac{(n_i - 1)\, s_i^2 + \nu_\circ \sigma_\circ^2}{n_i + \nu_\circ - 3} \quad . \tag{2.13}$$

The first component is the usual estimate of the variance based on the observations within each group weighted proportionally by its degrees of freedom. The second is the typical value $\sigma_\circ^2$ of the variances in their prior distribution weighted by the concentration of these variances $\nu_\circ$ around that typical value.

For the normal theory, however, the empirical Bayes estimate for the logarithmic transformed variance is given by

$$\tilde{\delta}_i = \hat{\lambda}_i d_i + (1 - \hat{\lambda}_i)\, W_i'\hat{\gamma} \ , \tag{2.14}$$

where $\hat{\lambda}_i = (n_i - 1)/(n_i + \hat{v} - 1)$, and $\hat{v}$ is estimated numerically.

The first component is the ordinary estimate of the logarithmic transformation of the variance estimate within a group, weighted proportionally by its degrees of freedom. The second component is a predicted value based on information from all groups, weighted proportionally by the concentration of the parameters estimated by the first component around that predicted value.

The resulting estimates from both methods are conditional estimates. In the method where exact sampling distributions of the variance estimates are being used, the empirical Bayes

estimates for the variances are conditioned on knowing the true values for the parameters $\sigma_\circ^2$ and $\nu_\circ$. In practice, however, the true values of these two parameters are usually unknown. To obtain the shrinkage estimates of the variances in 2.13, $\sigma_\circ^2$ and $\nu_\circ$ have to be estimated. Based on the EM algorithm approach by (Dempster, Laird and Rubin, 1977) for numerical estimation, Kasim (1986) developed a procedure for estimating $\sigma_\circ^2$ and $\nu_\circ$. The complete data for estimating $\sigma_\circ^2$ and $\nu_\circ$ are $\{s_j^2\}, \{\sigma_j^2\}$ for $j = 1, \cdots, k$, where only $\{s_j^2\}$ have been observed. The parameters $\sigma_\circ^2$ and $\nu_\circ$ are estimated by maximizing their likelihood function $l(\{\sigma_j^2\} | \nu_\circ, \sigma_\circ^2)$ (M-step), assuming that $\{\sigma_j^2\}$ have been observed. The sufficient statistics for estimating $\sigma_\circ^2$ and $\nu_\circ$ are functions of $\{\sigma_j^2\}$. Since $\{\sigma_j^2\}$ cannot be observed, the posterior expectations (E-step) of these sufficient statistics are used to estimate $\sigma_\circ^2$ and $\nu_\circ$. Finding the posterior expectations of the sufficient statistics, however, depends on knowing the values of $\sigma_\circ^2$ and $\nu_\circ$ as well as the data $\{s_j^2\}$. This dependency between the posterior expectations and $\sigma_\circ^2$ and $\nu_\circ$ is used in an iterative process to estimate $\sigma_\circ^2$ and $\nu_\circ$. Therefore, given the observed data $\{s_j^2\}$ and initial estimates of $\sigma_\circ^2$ and $\nu_\circ$, the values of sufficient statistics are estimated by their posterior expectations. The estimated sufficient statistics are then used in maximizing $l(\{\sigma_j^2\} | \nu_\circ, \sigma_\circ^2)$ to obtain new estimates of $\sigma_\circ^2$ and $\nu_\circ$. Going back and forth between the expectation step and the maximization step, we get reasonable estimates of $\sigma_\circ^2$ and

$v_\circ$. In fact, each step of this iteration process increases $l(\{\sigma_j^2\}|v_\circ, \sigma_\circ^2)$, which provides us with better estimates of $\sigma_\circ^2$ and $v_\circ$ than those from the step before. The iteration process is terminated when the absolute change in the values of $\sigma_\circ^2$ and $v_\circ$, between any two steps, is sufficiently small.

In the method where the normal approximation to the sampling distribution of the variance estimates is being used, the empirical Bayes estimates of the variances are conditioned on the residual variance v in the second stage model (see 2.15). The true value of this parameter is unknown, and it must be estimated. Similar to the way that $\sigma_\circ^2$ and $v_\circ$ are estimated, the maximum likelihood estimate v is obtained numerically via the EM algorithm.

The work of Lindley (1971), Leonard (1975) and Raudenbush and Bryk (1987) on estimating variances of many groups share the characteristic of borrowing information from other groups to get an improved estimate of the variance for a particular group. As an application of this idea Singh and Sedransk (1988) provided an example for obtaining improved estimates of strata variances when strata sample sizes are small. The improved variance estimator uses data borrowed from other strata, initially thought to be similar. The resulting estimate is a shrinkage variance estimate for the stratum, which is similar to the one in 2.14.

Comments on the Analysis of Many Variances

There are still several theoretical and practical weakness in estimating and investigating the variability in many variances. Empirical Bayes estimates of the variances, which were suggested by Raudenbush and Bryk (1987), for example, are not adjusted for the uncertainty in estimating the conditioning parameters $\sigma_{\circ}^2$ and $\nu_{\circ}$ in the exact method, and v in the normal approximation method. This is similar to the problem (Tanner and Wong, 1987) of obtaining empirical Bayes estimate for the effect parameters in HLM without adjusting these estimates for the uncertainty in estimating the variance components.

The focus on obtaining conditional point estimates for the variances in the model rather than deriving the entire posterior distribution for each variance can cause some problems. In some cases, the misrepresentation of the true values of the conditioning (given) parameters will lead to obtaining inaccurate conditional estimates of the variances. For example, in the case of the exact theory of Raudenbush and Bryk (1987), the variability of the group variances is inversely proportional to the conditioning parameter $\nu_{\circ}$ plus some constant. Therefore, if there is a clear evidence of heterogeneity of variance then the likelihood of $\nu_{\circ}$ is not at all symmetric around its maximum likelihood estimate. Consequently, this estimate is an inaccurate representation

for the true value of ν_{o}. Conditioning the variance estimates for the different groups on the estimated value of this parameters will distort these variance estimates. This problem is similar to the one presented earlier by Rubin (1983) for estimating τ^2 when the number of groups is small.

Normal approximation for the distribution of the estimated variances is another crucial point in the analysis of many variances. Bartlett and Kendall (1946) reported that the normal approximation is sufficiently accurate when the sample size in each group is at least equal to 10. For smaller samples, the normal approximation to the distribution of the estimated variances might be less accurate. The inaccuracy in the normal approximation can lead to inappropriate representation of the model. A severe loss of efficiency in estimation can occur when sample sizes are less than 5 (Cox and Solomon, 1986).

Adjusting for The Uncertainty in Estimating σ_j^2 and T

Standard methods do not exist for adjusting the conditional posterior distribution for the effect parameters to reflect increased variability due to the uncertainty of estimating σ_j^2 and T (Tanner and Wong, 1987). The estimates of the regression coefficients given in (1.5) to (1.8) are conditioned on knowing the true values of σ_j^2 and T . In practice, however, these parameters are usually unknown, and

their ML estimates are often used to allow for empirical Bayes estimation for the regression coefficients. Therefore, part of the variability in the empirical Bayes estimates for the regression coefficients should be attributed to the uncertainty in estimating σ_j^2 and $\mathbf{T}$. Having the marginal posterior distributions for the regression coefficients will facilitate obtaining their estimates without conditioning on σ_j^2 and $\mathbf{T}$. Consequently, the estimates will be correctly adjusted for the uncertainty of estimating the variance components σ_j^2 and $\mathbf{T}$.

In summary, the problem in getting empirical Bayes estimates for the effect parameters in multilevel models is centered on getting appropriate estimates for σ_j^2 and $\mathbf{T}$. When the sample sizes and the number of groups are large enough, maximum likelihood estimates of σ_j^2 and $\mathbf{T}$ are a good solution to the problem of getting those empirical Bayes estimates. However, when there is heterogeneity of variances and the number of observations within each group as well as the number of groups are not large enough to allow asymptotic estimation, maximum likelihood estimates of σ_j^2 and $\mathbf{T}$ may not be appropriate; it is essential that one should move toward being fully Bayesian in this analysis.

CHAPTER 3

Model Specification

This chapter provides a description of the statistical model used in this study and its underlying theory. The basic assumptions for this model are provided within the Bayesian framework. The joint posterior distribution for the parameters in the model is also presented.

A common characteristic often shared by evaluation studies that involve social interventions, like a new educational program, drug rehabilitation, or health care promotion, is that the data collection process is based on more than one level of hierarchy. The most common case is where there are measures for program or group characteristics and other measures for characteristics of subjects who are nested within the groups or programs. Any model set up for investigating relationships between measures from this kind of data needs to acknowledge the hierarchical structure of the data.

Statistical Model

This section presents a general linear model that captures the hierarchical structure of nested data. The model and its assumptions can be described in three stages. The first stage describes data obtained from k groups, where $j=1,\cdots,k$. Given the set of parameters λ, U_j, and σ_j^2, let Y_j be a vector of n_j independent observations from a normal distribution with a mean $Z_j\lambda + 1_jU_j$ and a variance $I_j\sigma_j^2$. A linear model with random intercept and fixed effect parameters for both group and subject level variables can be specified as

$$Y_j = Z_j\lambda + 1_jU_j + \epsilon_j \quad , \tag{3.1}$$

where

$\epsilon_j \sim N(0,\ I_j\sigma_j^2)$,

U_j is a random error for the intercept,

$$Z_j = \begin{bmatrix} 1 & W_{j1} & \cdots & W_{jq} & X_{1j1} & \cdots & X_{1jp} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ 1 & W_{j1} & \cdots & W_{jq} & X_{n_jj1} & \cdots & X_{n_jjp} \end{bmatrix}, \quad \text{and} \quad \lambda = \begin{bmatrix} \gamma_\circ \\ \gamma_1 \\ \vdots \\ \gamma_q \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}. \tag{3.2}$$

The part $[W_{j1} \cdots W_{jq}]$ of the matrix Z_j represents q known group level variables, the part $[X_{ij1} \cdots X_{ijp}]$ represents p known fixed effect subject level variables, and λ is $p+q+1$ vector of parameters that capture the effect of group and subject level variables on the outcome variable Y_j.

In the second stage both the intercept $\gamma_\circ + U_j$ and the residual variance σ_j^2 are assumed to be independent and vary randomly across the k groups. Given λ and τ^2, the random part of the intercept U_j is distributed as normal with mean zero and variance τ^2. Given $\sigma_\circ^2$ as a typical value of σ_j^2, the variable $\frac{\sigma_\circ^2}{\theta\sigma_j^2}$ is distributed as $\chi^2_{(\frac{1}{\theta})}$ with $\frac{1}{\theta}$ degrees of freedom.

The last stage represents vague prior knowledge about the hyper-parameters λ, τ^2, θ, and $\sigma_\circ^2$ for the two distributions presented in the second stage. The prior for each one of these four parameters is assumed proportional to some constant.

The model in 3.1 and its associated assumptions can be summarized as follows:

1- The observed data

Given λ, U_j and σ_j^2 the data Y_j is distributed as

$$Y_j \mid \lambda, U_j, \sigma_j^2 \quad \sim \quad N(Z_j'\lambda + 1_j U_j \ , \ I_j \sigma_j^2) \ . \tag{3.3}$$

2- The parameters of exchangeable distributions

Here σ_j^2 and U_j are assumed independent,

$$U_j \mid \lambda, \tau^2 \quad \sim \quad N(0, \tau^2) \ , \tag{3.4}$$

and

$$\frac{\sigma_\circ^2}{\theta\sigma_j^2} \quad \sim \quad \chi^2_{(\frac{1}{\theta})} \ . \tag{3.5}$$

3- The prior knowledge about the hyper-parameters.

The joint prior distribution $p(\lambda, \tau^2, \theta, \sigma_\circ^2)$ can be written as

$$
\begin{aligned}
p(\lambda, \tau^2, \theta, \sigma_\circ^2) &\propto p(\lambda)\, p(\tau^2)\, p(\theta, \sigma_\circ^2)\ , \quad \text{where} \\
p(\lambda) &\propto C_\lambda \\
p(\tau^2) &\propto C_{\tau^2} \\
p(\theta, \sigma_\circ^2) &\propto p(\sigma_\circ^2 | \theta)\, p(\theta)\ , \quad \text{where} \\
p(\sigma_\circ^2 | \theta) &\propto C_{\sigma_\circ^2}\ , \quad \text{and} \\
p(\theta) &\propto C_\theta\ .
\end{aligned}
\qquad \textbf{(3.6)}
$$

The hierarchical structure of the models in (3.1) and (3.2) fits nicely within the Bayesian approach. The next section provides more discussion about the specification of the parametric forms for the distributions of the exchangeable parameters U_j and σ_j^2 and the hyper-parameters λ, τ^2, θ, and $\sigma_\circ^2$. These specifications are required to obtain Bayes estimates for the parameters of interest.

Assumptions of the Model

There are three sets of assumptions associated with the model in 3.1. The first assumption is about the data. Given λ, U_j and σ_j^2, we assume that the data Y_j are normal with mean equal to $Z_j\lambda + 1_j U_j$ and variance equal to $I_j\sigma_j^2$.

In the second stage of the model, the intercept $\gamma_\circ + U_j$ and the residual variance σ_j^2 are allowed to vary randomly across the groups. The parametric forms of the distributions for U_j and σ_j^2 are based on the hierarchical structure of the model. More explicitly, there are two sets of parameters with several parameters of the same type in each set (i.e., several

intercepts and several variances). The parameters within each set usually describe a common variable that is shared by all the groups. In the example of evaluating educational program in different schools this common variable could be the effectiveness of an educational program adopted by each school within a district. For the meta-analysis example, the common variable is the effect size that varies in value across all studies used in meta-analysis. Within each set of parameters (intercepts or residual variances), one can think of the value of each parameter in the set as a single realization of a random variable that has a common distribution with the rest of the parameters in the set. When this conceptualization is translated to a probability structure, each parameter can be viewed as a random variable having an exchangeable distribution with the rest of the parameters in the set. Another way of saying this is that these parameters are independently identically distributed random variables with a prior distribution yet to be specified.

Under the Bayesian approach, the prior knowledge about the parameters can be used to specify the parametric form of their prior distributions. Bretthorst (1988) provided general guidelines for choosing prior distributions. He stated that

> "There are two questions one may consider to help in this. First, one should ask 'Are the parameters logically connected?' That is, if we gain additional information about one of the parameters does it change the estimates we would make about the others? If the answer is yes, then the parameters are not logically independent. It will be useful to find a representation where the

parameters are independent.
Another useful question is 'What are the invariances that the prior probability must obey?' That is, what transformation would convert the present problem into one where we have the state of prior knowledge." (Bretthorst, 1988, p.183).

The first question deals with. issue of independence between the parameters for which we seek prior distributions. For the model in 3.1 we assumed U_j and σ_j^2 independent. It implies that knowing the residual variance in each group does not depend on knowing the random error associated with intercept. The second question deals with the choice of conjugate prior distribution for the parameter of interest. Therefore, given λ and τ^2, we assume that the conjugate prior for the random part of the intercept U_j is normal with mean equal to zero and variance equal to τ^2. Further, given θ and $\sigma_\circ^2$, we assume that $\sigma_\circ^2/(\theta\sigma_j^2)$ has a chi-square distribution with $\frac{1}{\theta}$ degrees of freedom. The hyper-parameter $\sigma_\circ^2$ describes the value that one expects the residual variance σ_j^2 to have, prior to observing the data. Therefore, σ_j^2 is distributed as $\frac{\sigma_\circ^2}{\theta}\chi^{-2}_{\left(\frac{1}{\theta}\right)}$, where $\chi^{-2}_{\left(\frac{1}{\theta}\right)}$ is an inverse chi-square with $\frac{1}{\theta}$ degrees of freedom. Whileθ is a function of the degrees of freedom, it can also be thought of as a measure of variability of the variances $\{\sigma_j^2\}$. The probability density function $p(\sigma_j^2|\theta,\sigma_\circ^2)$ is found to be

$$p(\sigma_j^2|\theta,\sigma_\circ^2) = \frac{\left(\frac{\sigma_\circ^2}{2\theta}\right)^{\frac{1}{2\theta}}}{\Gamma\left(\frac{1}{2\theta}\right)} (\sigma_j^2)^{-\left(\frac{1}{2\theta}+1\right)} \exp\left(\frac{-\sigma_\circ^2}{2\theta\sigma_j^2}\right). \qquad (3.7)$$

The above density function is similar to the one given by Lindley (1965), with a minor difference of $\theta=1/\nu_\circ$. As Lindley described it,

> "For all $\nu_\circ > 0$ the above density function increases from zero at $\sigma_j^2 = 0$ to a maximum at $\sigma_j^2 = \sigma_\circ^2 \nu_\circ / (\nu_\circ + 2)$ and then tends to zero as $\sigma_j^2 \to \infty$. The density has therefore the same general shape for all degrees of freedom as the chi-square distribution itself has for degrees of freedom in excess of two." (Lindley, 1965, p. 28).

In his interpretation of the above density as a prior density for σ_j^2, Lindley said:

> "To take it as a prior density for σ_j^2 is equivalent to saying that values that are very large or very small are improbable, the most probable value is $\sigma_j^2 = \sigma_\circ^2 \nu_\circ / (\nu_\circ + 2)$ and, because the decrease from the maximum as σ_j^2 increases is less than the corresponding decrease as σ_j^2 diminishes, the values above the maximum are more probable than those a similar distance below the maximum." (Lindley, 1965 p. 28).

The above interpretation to the prior distribution of σ_j^2 makes choosing such a prior more realistic than assuming equal likelihood for all values of σ_j^2 (i.e., a locally uniform prior).

The mean and the variance of σ_j^2 can be easily found by realizing that the density in 3.7 is an inverse Gamma function with $\alpha = 1/2\theta$ and $\beta = 2\theta/\sigma_\circ^2$. Therefore,

$$E(\sigma_j^2|\theta, \sigma_\circ^2) = \frac{1}{\beta(\alpha - 1)} = \frac{\sigma_\circ^2}{1 - 2\theta}, \quad \text{and}$$
$$Var(\sigma_j^2|\theta, \sigma_\circ^2) = \frac{1}{\beta^2(\alpha - 1)^2(\alpha - 2)} = \frac{2\theta\sigma_\circ^4}{(1 - 2\theta)^2(1 - 4\theta)} . \qquad (3.8)$$

The results in 3.8 reveal that θ has to be less than .25 for the variance of σ_j^2 to be defined. As θ gets smaller, the mean and the variance are approximated to $\sigma_\circ^2$ and $2\theta\sigma_\circ^4$.

In many analyses σ_j^{-2} is used as a measure of precision for each estimate derived from the data. From 3.7, the prior distribution of σ_j^{-2} is Gamma with $\alpha = 1/2\theta$ and $\beta = 2\theta/\sigma_\circ^2$. Its mean and variance are found to be

$$E(\sigma_j^{-2}|\theta, \sigma_\circ^2) = \alpha\beta = \sigma_\circ^{-2}, \quad \text{and}$$
$$Var(\sigma_j^{-2}|\theta, \sigma_\circ^2) = \alpha\beta^2 = 2\theta\sigma_\circ^{-4} . \qquad (3.9)$$

Another measure that describes σ_j^2 and has the characteristics of being invariant to the change in the mean of its distribution, and depending only on θ, is the coefficient of variation (Linhart, 1965). From 3.8 and 3.9 this coefficient is found to be $\sqrt{\frac{2\theta}{1-4\theta}}$ for σ_j^2 and $\sqrt{2\theta}$ for σ_j^{-2}. Small values of θ indicate precise knowledge about σ_j^2 prior to observing the data.

The last set of assumptions concerns the prior knowledge about the parameters λ, τ^2, θ, and $\sigma_\circ^2$. They are the hyper-parameters for the exchangeable distributions for the parameters $\{U_j\}$ and $\{\sigma_j^2\}$.

The two questions presented earlier by Bretthorst (1988) might serve as general guidelines for choosing prior distributions. However, when the model in 3.1 is being

applied, prior knowledge about the state of the hyper-parameters λ, τ^2, θ, and $\sigma_\circ^2$ is usually limited or non-existent. That is, the higher we move in the hierarchical structure of the parameters in the model, the less information we have about those parameters prior to observing the data. In other words, one is usually more informed about the distribution of the $\{U_j\}$ and the $\{\sigma_j^2\}$, prior to observing the data, than about the distribution of λ, τ^2, θ, and $\sigma_\circ^2$. Therefore, the prior distributions for λ, τ^2, θ, and $\sigma_\circ^2$ should have the characteristic of being non-informative priors. This implies that the posterior distributions for these parameters should rely heavily on the information obtained from the data, with little or no weight given to the prior knowledge about the parameters.

For the parameter λ, we take its conjugate prior distribution, (which is normal) with vague prior knowledge about the parameter itself. Therefore, λ is distributed uniformly over its parameter space R_{q+p+1}; that is $p(\lambda) \propto C_\lambda$, where C_λ is constant. The subscript λ is being used with the constant to identify it with the parameter.

Similar to σ_j^2, a natural conjugate prior for τ^2 would be an inverse chi-square with ν_{τ^2} degrees of freedom. A vague prior for τ^2 implies $\nu_{\tau^2} = 0$, which produces what is often called Jeffery's prior (i.e., $p(\tau^2) \propto 1/\tau^2$). This prior distribution is found to cause some difficulties when the number of groups is relatively small (Seltzer, 1988). The

effect of this prior with τ^2 near zero has a great influence on the density function of the data such that the marginal posterior distribution of τ^2 becomes heavily concentrated around $\tau^2 = 0$ (Morris, 1983 and Lindley, 1983). As an alternative, a uniform prior distribution is chosen for τ^2 (Seltzer 1989). That is $p(\tau^2) \propto C_{\tau^2}$ for $(\tau^2 > 0)$, where C_{τ^2} is a constant. The prior distributions for both λ and τ^2 will then become insignificant in their contribution to their posterior distributions.

For θ and $\sigma_\circ^2$, we write their joint prior distribution proportional to the product of the conditional prior distribution of $\sigma_\circ^2$ given θ and the prior distribution of θ. That is

$$p(\sigma_\circ^2, \theta) \propto p(\sigma_\circ^2 | \theta)\, p(\theta) . \qquad \textbf{(3.10)}$$

We specify $p(\sigma_\circ^2|\theta)$ to have conjugate prior to the density in 3.7, which is of a chi-square form as

$$p(\sigma_\circ^2 | \theta, r, \omega) \propto (\sigma_\circ^2)^{r-1} \exp\left(\frac{-\sigma_\circ^2 \omega}{2\theta}\right) , \qquad \textbf{(3.11)}$$

where r and ω are two parameters describing the prior for $\sigma_\circ^2$ (Lindley, 1971). A vague prior for $\sigma_\circ^2$ implies setting $r=0$ and $\omega=0$ which will produce what is known as Jeffrey's prior. However, the fact that the marginal posterior distribution of $\sigma_\circ^2$ is heavily dominated by the data undermines the effect of the prior of $\sigma_\circ^2$ on its marginal posterior distribution. Thus, for mathematical convenience, a constant prior $C_{\sigma_\circ^2}$ is specified

for $\sigma_\circ^2$.

In the second set of the assumptions, the hyper-parameter θ is presented as a function of the degrees of freedom for the exchangeable distributions for $\{\sigma_j^2\}$. It also, represents the variability of $\{\sigma_j^2\}$, and mathematically there is no reason why it should not be any positive number (Lindley, 1971). Therefore, we assume that $p(\theta)$ has a uniform distribution on the real line. That is $p(\theta) \propto C_\theta$, where C_θ is a constant.

The Joint Density Function of the Parameters in the Model

Applying the Bayesian approach to the model in 3.1 allowed the parameters in the last two stages of the model to be treated as unknown parameters. In the first stage $\{U_j\}$ and $\{\sigma_j^2\}$ for $j = 1,\cdots,k$ are the parameters with exchangeable priors. The second stage covers the hyper-parameters λ, τ^2, θ, and $\sigma_\circ^2$, where λ is defined in 3.2. They are the parameters of the prior distributions for $\{U_j\}$ and $\{\sigma_j^2\}$. Vague prior distributions for λ, τ^2, θ, and $\sigma_\circ^2$ are taken from their usual conjugate families.

Treating all the parameters in the model as unknown parameters, allows their joint distribution along with the data Y_j to be distributed as a product of conditional distributions as follows:

$$\prod_{j=1}^{k} p(Y_j\,|\,\lambda, U_j, \sigma_j^2)\, p(U_j\,|\,\lambda, \tau^2)\, p(\sigma_j^2\,|\,\theta, \sigma_\circ^2)\, p(\lambda, \tau^2, \theta, \sigma_\circ^2)\,. \qquad \textbf{(3.12)}$$

Substituting the density functions for the assumed

distributions from the previous section for each of the above conditional distributions produces a joint density function proportional to

$$\left[\prod_{j=1}^{k}(\sigma_j^2)^{\frac{-n_j}{2}}\right]\exp\left[-\frac{1}{2}\sum_{j=1}^{k}(Y_j - Z_j\lambda - 1_jU_j)'\sigma_j^{-2}(Y_j - Z_j\lambda - 1_jU_j)\right]$$

$$\times \quad (\tau^2)^{-\frac{k}{2}}\exp\left[-\frac{1}{2\tau^2}\sum_{j=1}^{k}U_j^2\right]$$

$$\times \quad \frac{\left(\frac{\sigma_\circ^2}{2\theta}\right)^{\frac{k}{2\theta}}}{\left(\Gamma\left(\frac{1}{2\theta}\right)\right)^k}\left[\prod_{j=1}^{k}(\sigma_j^2)^{-(\frac{1}{2\theta}+1)}\right]\exp\left[-\frac{\sigma_\circ^2}{2\theta}\sum_{j=1}^{k}\frac{1}{\sigma_j^2}\right]. \tag{3.13}$$

Theoretically, the marginal posterior distribution of any parameter in the model can be obtained by integrating the joint density function in 3.13 over the space of the other parameters.

Being able to obtain the marginal posterior distribution of any parameter in the model implies being able to obtain interval and point posterior estimates of that parameter. For example, an estimate of λ based on its marginal posterior distribution will not be affected by the uncertainty in estimating the variance components, avoiding an undesirable characteristic of empirical Bayes estimates of λ. Further, individual group estimates of $\{\sigma_j^2\}$ for $j = 1,\cdots,k$ become useful when we are investigating groups' heterogeneity of variance. Also, individual estimates of $\{U_j\}$ become useful for checking the normality assumption for the error term associated with each group intercept. Moreover, inferences on the hyper-

parameters τ^2, σ_o^2, and θ can be made using their respective marginal posterior distributions.

With the help of modern computers and developments in simulation theories, we are now able to approximate these marginal posterior distribution numerically, to the desired degree of accuracy. The methods of Data Augmentation (Tanner and Wong, 1987) and Gibbs sampling (Tanner, 1993) are being used for numerical integration to obtain marginal distributions. Morris (1987) shows how one can use these methods for hierarchical Bayes models.

CHAPTER 4

Obtaining Marginal Posterior Distributions Via Gibbs Sampling

This chapter presents the method of Gibbs sampling as an iterative procedure for calculating marginal posterior distributions. The process of Gibbs sampling is best understood within the context of data augmentation (Tanner and Wong, 1987). Therefore, the basic idea of the data augmentation procedure will be explained. An example, using the normal distribution, will illustrate its application. A simple modification to the idea of data augmentation will facilitate the understanding of Gibbs sampling procedure. This procedure will then be used in chapter 5 to approximate the marginal posterior distributions of the parameters in the model in 3.1.

Data Augmentation

The argument that Box and Tiao (1964) made in their investigation of the importance of the assumptions applied to the comparison of variances, using the Bayesian approach, can facilitate our understanding of data augmentation. Very often the distribution of observations Y depends on more than one parameter, say two parameters $\mathbf{Z}_1$ and $\mathbf{Z}_2$, one of which, say Z_1, is of immediate interest to us. For non-Bayesian methods,

this could cause an extremely difficult problem in dealing with the other parameter $\mathbf{Z}_2$. However, under the Bayesian approach making inferences about $\mathbf{Z}_1$ is simplified by finding the marginal posterior distribution $p(\mathbf{Z}_1|Y)$. This marginal posterior distribution can be obtained by integrating out the other set of parameters $\mathbf{Z}_2$ from the joint posterior distribution $p(\mathbf{Z}_1, \mathbf{Z}_2|Y)$. In the data augmentation process, $\mathbf{Z}_2$ can be thought of as latent variable or missing data augmenting the observed data Y (Tanner and Wong, 1987).

In the following discussion all distributions are posterior distributions, therefore, the term "posterior" is deleted from the names of the distributions for simplicity. If we write the joint distribution $p(\mathbf{Z}_1, \mathbf{Z}_2|Y)$ as a product of the conditional distribution of $\mathbf{Z}_1$ and the marginal distribution of $\mathbf{Z}_2$,

$$p(\mathbf{Z}_1, \mathbf{Z}_2|Y) = p(\mathbf{Z}_1|\mathbf{Z}_2, Y)\ p(\mathbf{Z}_2|Y)\ , \tag{4.1}$$

the marginal distribution of $\mathbf{Z}_1$ can then be written as

$$p(\mathbf{Z}_1|Y) = \int_T p(\mathbf{Z}_1|\mathbf{Z}_2, Y)\ p(\mathbf{Z}_2|Y)\ \partial\mathbf{Z}_2\ , \tag{4.2}$$

where T is the parameter space for $\mathbf{Z}_2$. As Box and Tiao (1964) pointed out:

> "... the marginal posterior distribution of the parameter $p(\mathbf{Z}_2|Y)$ acts as a weight function multiplying the conditional distribution $p(\mathbf{Z}_1|\mathbf{Z}_2, Y)$ of the parameter of interest. It is frequently helpful in understanding the problem and the nature of the conclusions which can safely be drawn to consider not only $p(\mathbf{Z}_1|Y)$ but also the components of the integral on the right-hand side of equation [4.2]. One is thus led to consider the conditional distribution of $\mathbf{Z}_1$ for particular

> values of the nuisance parameter $\mathbf{Z}_2$ in relation to the probability of occurrence of the postulated values of the nuisance parameter." (p. 153).

These two basic ideas, integrating out the nuisance parameter $\mathbf{Z}_2$ from the joint distribution $p(\mathbf{Z}_1, \mathbf{Z}_2 | Y)$, and using the marginal distribution of the nuisance parameter $p(\mathbf{Z}_2 | Y)$ as a weight function in that integral, can be generalized to the process of data augmentation when a direct solution to the integral in 4.2 can not be found. Tanner (1993) defines $p(\mathbf{Z}_2 | Y)$ in the integral in 4.2 as the predictive distribution which can be used with the observed data Y to obtain the posterior distribution $p(\mathbf{Z}_1 | Y)$. In many cases however, $p(\mathbf{Z}_2 | Y)$ is not known, which makes it impossible to obtain $p(\mathbf{Z}_1 | Y)$. The joint distribution of $\mathbf{Z}_1$ and $\mathbf{Z}_2$ as in 4.1 and 4.2 can then be used to obtain the marginal distribution $p(\mathbf{Z}_2 | Y)$. In other words, the joint distribution of $\mathbf{Z}_1$ and $\mathbf{Z}_2$ in 4.1 can be written as:

$$p(\mathbf{Z}_1, \mathbf{Z}_2 | Y) = p(\mathbf{Z}_2 | \mathbf{Z}_1, Y)\ p(\mathbf{Z}_1 | Y)\ . \tag{4.3}$$

The marginal distribution for the second parameter $p(\mathbf{Z}_2 | Y)$ can be obtained by integrating the joint distribution in 4.3 over the parameter space R of the first parameter $\mathbf{Z}_1$:

$$p(\mathbf{Z}_2 | Y) = \int_R p(\mathbf{Z}_2 | \mathbf{Z}_1, Y)\ p(\mathbf{Z}_1 | Y)\, \partial \mathbf{Z}_1\ . \tag{4.4}$$

Just as in 4.2, the marginal distribution of the parameter $p(\mathbf{Z}_1 | Y)$ in 4.4 can be thought of as a weight function multiplying the conditional distribution of the parameter $p(\mathbf{Z}_2 | \mathbf{Z}_1, Y)$. Integrating this weighted conditional distribution over all admissible values of $\mathbf{Z}_1$ will give us the marginal distribution of $\mathbf{Z}_2$.

Carrying out the integration in equation 4.4 analytically requires knowing the parametric form of the marginal distribution of the first parameter $p(\mathbf{Z}_1|Y)$. Likewise, carrying out the integration in equation 4.2 analytically requires knowing the parametric form of the marginal distribution of the second parameter $p(\mathbf{Z}_2|Y)$. But these two marginal distributions are unknown to us and we are interested in finding them. Therefore, this dependency between the two marginal distributions for the two parameter becomes the key point of the iteration process of data augmentation.

In 4.1 and 4.3, the joint distribution of $\mathbf{Z}_1$ and $\mathbf{Z}_2$ is being expressed as a function of four other distributions, two of which have unknown parametric forms. Therefore, it is hard to sample from them. They are the marginal distributions $p(\mathbf{Z}_1|Y)$ and $p(\mathbf{Z}_2|Y)$. The other two have known parametric forms, and one can easily sample from them. They are the conditional distributions $p(\mathbf{Z}_1|\mathbf{Z}_2, Y)$ and $p(\mathbf{Z}_2|\mathbf{Z}_1, Y)$. The iteration process, therefore, involves repeated sampling of $\mathbf{Z}_1$ and $\mathbf{Z}_2$ from their conditional distributions $p(\mathbf{Z}_1|\mathbf{Z}_2, Y)$ and $p(\mathbf{Z}_2|\mathbf{Z}_1, Y)$, which accomplishes the numerical integration of the joint distributions in 4.2 and 4.4, and produces numerical approximations to the marginal distributions for $\mathbf{Z}_1$ and $\mathbf{Z}_2$.

To exploit the dependency between the two distributions in 4.2 and 4.4, we start with an initial estimate of $\mathbf{Z}_1$. One

can think of this initial estimate as being sampled from a poor approximation of its marginal posterior distribution, $p(\mathbf{Z}_1|Y)$. A sample of M values of $\mathbf{Z}_2$ can then be drawn from its conditional distribution $p(\mathbf{Z}_2|\mathbf{Z}_1, Y)$, where the initial estimate of $\mathbf{Z}_1$ is used along with data Y. For each of the$M$ values of $\mathbf{Z}_2$, we form a conditional distribution of $\mathbf{Z}_1$ given $\mathbf{Z}_2=\mathbf{Z}_{2m}$ defined as $p(\mathbf{Z}_1|\mathbf{Z}_{2m}, Y)$, for $m=1,\cdots,M$. The weighing of each of these conditional distributions is done empirically by the sampling process from the marginal posterior distribution $p(\mathbf{Z}_2|Y)$ in 4.2. That is, a large proportion of M of these conditional distributions are conditioned on values of $\mathbf{Z}_2$ which are near the mode of $p(\mathbf{Z}_2|Y)$. Rubin (1987) refers to the process of sampling $\mathbf{Z}_2$ from $p(\mathbf{Z}_2|Y)$, in order to form the conditional distribution of the first set of parameters $p(\mathbf{Z}_1|\mathbf{Z}_2, Y)$ as a multiple imputation process. The mixture of the M conditional distributions of $\mathbf{Z}_1$ given $\mathbf{Z}_{2m}$, $m=1,\cdots,M$ represented by their weighted average, where the weighing process is done empirically (as explained previously), produces the marginal posterior distribution $p(\mathbf{Z}_1|Y)$ at the t<u>th</u> iteration

$$P^t(\mathbf{Z}_1|Y) \approx M^{-1}\sum_{m=1}^{M} p(\mathbf{Z}_1|\mathbf{Z}_{2m}, Y) \quad . \tag{4.5}$$

To simulate the marginal distribution of $\mathbf{Z}_1$ given in 4.5, we simply sample M values of $\mathbf{Z}_1$; one from each of the M conditional distributions given in the mixture. These M values of $\mathbf{Z}_1$ represent an approximation to the marginal

distribution of the first parameter $p^1(\mathbf{Z}_1|Y)$ at time t=1 of the iteration process. However, this approximation is a poor one since the sampled values of $\mathbf{Z}_2$ came from a rough guess of its marginal distribution $p(\mathbf{Z}_2|Y)$. Representing the M weighted conditional distributions by their average captures the idea of integrating the joint distribution of $\mathbf{Z}_1$ and $\mathbf{Z}_2$ in 4.2 over the parameter space of $\mathbf{Z}_2$ to produce the marginal distribution of $\mathbf{Z}_1$, when M is relatively large (say M = 4000). An estimate of the density function for the marginal posterior distribution of $\mathbf{Z}_1$ can be obtained by averaging the conditional densities in 4.5 (Gelfand and Smith, 1990).

For each of the M values of $\mathbf{Z}_1$, which are sampled from 4.5, a conditional distribution of $\mathbf{Z}_{2m}$ given $\mathbf{Z}_1$=$\mathbf{Z}_{1m}$, can be formed as $p(\mathbf{Z}_2|\mathbf{Z}_{1m}, Y)$, for $m=1,\cdots,M$. By adopting the same logic for approximating $p(\mathbf{Z}_1|Y)$, the marginal distribution for the second parameter can be approximated by

$$p^{t}(\mathbf{Z}_2|Y) \approx M^{-1}\sum_{m=1}^{M} p(\mathbf{Z}_2|\mathbf{Z}_{1m}, Y) \ , \qquad \textbf{(4.6)}$$

where $p^1(\mathbf{Z}_2|Y)$ is an approximation of the marginal distribution of $\mathbf{Z}_2$ at iteration t=1. We then sample M new values of $\mathbf{Z}_2$ for iteration t=2 from the resulting distribution in 4.6 to get a new approximation to the marginal distribution of $\mathbf{Z}_1$ in 4.5.

By iterating between 4.5 and 4.6, the two marginal distributions $p(\mathbf{Z}_1|Y)$ and $p(\mathbf{Z}_2|Y)$ stabilize (Tanner and Wong, 1987). Sampling the values of $\mathbf{Z}_1$ and $\mathbf{Z}_2$ from their respective

stabilized distributions is virtually identical to sampling from the true marginal distributions of $\mathbf{Z}_1$ and $\mathbf{Z}_2$.

Tanner and Wong (1987) pointed out that one can select any value for M (the sample size for each $\mathbf{Z}_1$ and $\mathbf{Z}_2$ in each iteration) to carry out the iteration process. They stated:

> "Even when M is as small as 1, the iteration is still 'in the right direction' in the sense that the average of $p(\mathbf{Z}_1|\mathbf{Z}_{2m}, Y)$ over the augmented data patterns generated across iterations will converge to $p(\mathbf{Z}_1|Y)$." (Tanner and Wong, 1987, p. 530).

Moreover, they argued that the value of M can be changed between iterations. In fact they recommended that one should start with a small value for M in the first few iterations and increase this value as the number of iterations increases. They stated:

> "In practice, however, it is inefficient to take M large during the first few iterations when the estimated posterior distribution is far from the true distribution. Rather, it is suggested that M initially be small and then increased with successive iterations." (Tanner and Wong, 1987, p. 539).

Example

Let Y be a random sample of n independent observations drawn from a normal distribution with mean μ and variance σ^2, where both parameters are unknown. The sample mean $\bar{y}=n^{-1}\sum_1^n y_i$ and the sample variance $s^2 = (n-1)^{-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$ are jointly sufficient for (μ, σ^2). The likelihood for (μ, σ^2), therefore, is given by (Box and Tiao, 1973)

$$\ell(\mu, \sigma^2 | \bar{y}, s^2) \propto (\sigma^2)^{-\frac{n}{2}} \exp\left\{\frac{-1}{2\sigma^2}[(n-1)s^2 + n(\mu-\bar{y})^2]\right\} . \qquad \textbf{(4.7)}$$

The likelihood in 4.7 reveals that the distribution of Y depends on the two parameters μ and σ^2. From the Bayesian viewpoint, making inferences on any one of these two parameters requires finding its marginal posterior distribution by integrating out the other parameter from their joint posterior distribution.

The joint posterior distribution of μ and σ^2 can be written proportional to the product of their likelihood function $\ell(\mu, \sigma^2 | \bar{y}, s^2)$ and the prior knowledge about μ and σ^2. Mathematically stated

$$p(\mu, \sigma^2 | Y) \propto \ell(\mu, \sigma^2 | \bar{y}, s^2) p(\mu, \sigma^2) . \qquad \textbf{(4.8)}$$

The second part of the right-hand side of 4.8 represents the prior knowledge about the two parameters μ and σ^2. It can be assumed priori that μ and σ^2 are independent, and the form of this prior knowledge can be specified by the usual conjugate priors of these two parameters. A vague prior knowledge about μ produces a prior distribution which is proportional to a constant C. Similarly, a vague prior knowledge about σ^2 produces a prior distribution proportional to σ^{-2} (Box and Tiao, 1973).

Substituting the assumed forms for the likelihood function $\ell(\mu, \sigma^2 | \bar{y}, s^2)$ and the prior distributions for μ and σ^2 in 4.8 produces the joint posterior density for μ and σ^2.

$$p(\mu, \sigma^2 | Y) \propto (\sigma^2)^{-(\frac{n}{2}+1)} \exp\left\{\frac{-1}{2\sigma^2}[(n-1)s^2 + n(\mu-\bar{y})^2]\right\}. \qquad \textbf{(4.9)}$$

Based on 4.9, and where σ^2 is known the conditional distribution $p(\mu|\sigma^2, Y)$ is normal with mean $\bar{y}$ and variance σ^2/n. When σ^2 is unknown, however, $p(\mu|y)$ represents a t-distribution with mean $\bar{y}$ and variance S^2/n and $(n-1)$ degrees of freedom, resulting from the integration of 4.9 with respect to σ^2 (Box and Tiao, 1973).

When μ is known, the sample variance is defined to be $S^2 = n^{-1}\sum_{i=1}^{n}(y_i - \mu)^2$. The variable nS^2/σ^2 is therefore distributed as chi-square with n degrees of freedom. Consequently, $\sigma^2|\mu, Y$ is distributed as $nS^2\chi^{-2}_{(n)}$ where $\chi^{-2}_{(n)}$ is an inverted chi-square variable with n degrees of freedom. When μ is unknown however, $p(\sigma^2|Y)$ represents the probability density of $(n-1)S^2\chi^{-2}_{(n-1)}$, where $\chi^{-2}_{(n-1)}$ is an inverted chi-square variable with $(n-1)$ degrees of freedom, resulting from the integration of 4.9 with respect to μ.

For illustrative purpose, let us assume that the forms for the distributions of $p(\mu|y)$ and $p(\sigma^2|Y)$ are unknown, and need to be approximated. Our objective in this example is to show how these two distributions can be numerically approximated, using the process of data augmentation presented in the previous section.

The joint posterior distribution of μ and σ^2 is written as a product of the conditional distribution of μ and the marginal posterior distribution of σ^2

$$p(\mu, \sigma^2 | Y) = p(\mu | \sigma^2, Y)\ p(\sigma^2 | y) \ . \qquad \textbf{(4.10)}$$

Then the marginal posterior distribution of μ can be defined by

$$p(\mu | Y) = \int_{\sigma^2} p(\mu | \sigma^2, Y)\ p(\sigma^2 | Y)\, d\sigma^2 \ . \qquad \textbf{(4.11)}$$

Similarly, the joint posterior distribution in 4.10 is written as a product of two new distributions

$$p(\mu, \sigma^2 | Y) = p(\sigma^2 | \mu, Y)\ P(\mu | Y) \ . \qquad \textbf{(4.12)}$$

Then the marginal posterior distribution of σ^2 can be defined by

$$p(\sigma^2 | Y) = \int_{\mu} p(\sigma^2 | \mu, Y)\, p(\mu | Y)\, d\mu \ . \qquad \textbf{(4.13)}$$

Examining 4.11 and 4.13 reveals that obtaining the marginal posterior distribution of one parameter depends on obtaining the marginal posterior distribution of the other parameter. Therefore, this dependency between the two equations can be used to show how the method of data augmentation can be used to approximate the two marginal posterior distributions in the following steps.

1. As a starting point for the iteration process, μ can be estimated from the data Y by $\hat{\mu} = \sum_{i=1}^{n} y_i/n$. This estimate can be thought of as μ, which has been sampled from the current approximation of its marginal posterior distribution $p(\mu | Y)$.

2. Using the current estimate $\hat{\mu}$ to represent the true value of μ, find the sample variance as $S^2 = n^{-1}\sum_{i=1}^{n}(y_i - \mu)^2$. This variance estimate will be used, in conjunction with the sampled values of a chi-square random variable, in the next step to sample σ^2 from its conditional distribution $p(\sigma^2|\mu, Y)$.
3. Sample M values of $\chi^2_{(n)m}$ for $m=1,\cdots,M$ from a chi-square distribution with n degrees of freedom. For each sampled value of $\chi^2_{(n)m}$ find $\hat{\sigma}^2_m$ as $\hat{\sigma}^2_m = nS^2/\chi^2_{(n)m}$. Applying steps 2 and 3 resembles the process of sampling M values of σ^2 from its conditional posterior distribution $p(\sigma^2|\mu, y)$. The mixture of these M conditioned values of σ^2 forms an initial approximation to the marginal posterior distribution $p(\sigma^2|Y)$. This approximation, however, is an inaccurate one since the sampled values of σ^2 are conditioned on a poor estimate of μ.
4. Given the data Y and the M values of σ^2_m which are sampled in step 3, we sample M new values of μ_m from $p(\mu|\sigma^2_m, Y)$, which is $N(\bar{y}, \sigma^2_m/n)$, where $\bar{y}$ is the sample mean from the data. The mixture of the M sampled values of the mean μ_m represents an approximation to its marginal posterior distribution $p(\mu|Y)$.
5. These new sampled values of μ_m from step 4 can then be used to get M new estimates of S^2_m as in step 2 for a new cycle of approximation. As we continue iterating between steps 2 to 5, the mixture of the M values of σ^2_m which

are sampled in step 3, and the mixture of the M values of μ_m which are sampled in step 4 become increasingly accurate in representing the marginal posterior distributions of μ and σ^2.

Gibbs Sampling

The process of data augmentation was presented for the case of approximating the marginal posterior distributions for only two parameters, Z_1 and Z_2 with $M>1$. When $M=1$, Tanner (1993) defines the iteration process as "chained data augmentation". In the normal data example given above, the two parameters were represented by μ and σ^2 with $M>1$. There are cases however, where there are more than two parameters for which we require numerical approximations to their marginal posterior distributions. To obtain them, a simple modification to the logic of data augmentation process can be used.

When the iteration process of data augmentation is generalized to more than two parameters with M set equal to one, the process is called "multivariate chained data augmentation" (Tanner, 1993) or Gibbs sampling (Gelfand, Hills, Racine-Poon and Smith, 1990).

Consider the previous case, where the distribution of the observations Y depends not only on Z_1 and Z_2 but also on a third parameter say Z_3. Under the Bayesian approach, making inferences about any of the three parameters is simplified by finding its marginal posterior distribution, by integrating out the other sets of parameters from the joint distribution of $\mathbf{Z}_1$, $\mathbf{Z}_2$ and $\mathbf{Z}_3$.

The joint distribution of $\mathbf{Z}_1$, $\mathbf{Z}_2$ and $\mathbf{Z}_3$ can be written as a product of the conditional distribution of $\mathbf{Z}_1$ and the joint distribution of $\mathbf{Z}_2$ and $\mathbf{Z}_3$,

$$p(\mathbf{Z}_1, \mathbf{Z}_2, \mathbf{Z}_3 | Y) = p(\mathbf{Z}_1 | \mathbf{Z}_2, \mathbf{Z}_3, Y)\ p(\mathbf{Z}_2, \mathbf{Z}_3 | Y)\ . \tag{4.14}$$

Then, the marginal posterior distribution of $\mathbf{Z}_1$ can be written as

$$p(\mathbf{Z}_1 | Y) = \int_T \int_W p(\mathbf{Z}_1 | \mathbf{Z}_2, \mathbf{Z}_3, Y)\ p(\mathbf{Z}_2, \mathbf{Z}_3 | Y)\ \partial\mathbf{Z}_3\ \partial\mathbf{Z}_2\ , \tag{4.15}$$

where T and W are the parameter spaces for $\mathbf{Z}_2$ and $\mathbf{Z}_3$, respectively.

Similar to 4.2, the joint distribution $p(\mathbf{Z}_2, \mathbf{Z}_3 | Y)$ can be thought of as a weight function multiplying the conditional distribution of the parameter of interest $p(\mathbf{Z}_1 | \mathbf{Z}_2, \mathbf{Z}_3, Y)$. In other words, we consider the conditional distribution of $\mathbf{Z}_1$ for particular values of parameters $\mathbf{Z}_2$ and $\mathbf{Z}_3$ in relation to the probability of getting those values of $\mathbf{Z}_2$ and $\mathbf{Z}_3$.

Similarly, the marginal posterior distribution for each of $\mathbf{Z}_2$ and $\mathbf{Z}_3$ can be obtained by integrating out the other parameters from the joint posterior distribution of $\mathbf{Z}_1$, $\mathbf{Z}_2$ and $\mathbf{Z}_3$ as follows :

$$p(\mathbf{Z}_2 | Y) = \int_R \int_W p(\mathbf{Z}_2 | \mathbf{Z}_1, \mathbf{Z}_3, Y)\ p(\mathbf{Z}_1, \mathbf{Z}_3 | Y)\ \partial\mathbf{Z}_3\ \partial\mathbf{Z}_1\ , \tag{4.16}$$

and

$$p(\mathbf{Z}_3 | Y) = \int_R \int_T p(\mathbf{Z}_3 | \mathbf{Z}_1, \mathbf{Z}_2, Y) \; p(\mathbf{Z}_1, \mathbf{Z}_2 | Y) \; \partial \mathbf{Z}_2 \; \partial \mathbf{Z}_1 \; . \qquad \textbf{(4.17)}$$

As in data augmentation, the three equations 4.15, 4.16 and 4.17 define the iteration process of Gibbs sampling. Further, it is assumed that there are only three distributions that are easy to sample from. They are the three conditional distributions of $\mathbf{Z}_1$, $\mathbf{Z}_2$ and $\mathbf{Z}_3$, represented by the first part of the right-hand side in each of 4.15, 4.16, and 4.17. The iteration process therefore involves repeated sampling of $\mathbf{Z}_1$, $\mathbf{Z}_2$, and $\mathbf{Z}_3$ from these conditional distributions to accomplish the numerical approximations of their marginal posterior distributions.

Starting with initial values of $\mathbf{Z}_2^{(0)}$, $\mathbf{Z}_3^{(0)}$, and the data Y, sample $\mathbf{Z}_1^{(1)}$ from its conditional distribution $p(\mathbf{Z}_1 | \mathbf{Z}_2^{(0)}, \mathbf{Z}_3^{(0)}, Y)$. Given the values of $\mathbf{Z}_1^{(1)}$, $\mathbf{Z}_3^{(0)}$, and the data Y, sample $\mathbf{Z}_2^{(1)}$ from its conditional distribution $p(\mathbf{Z}_2 | \mathbf{Z}_1^{(1)}, \mathbf{Z}_3^{(0)}, Y)$. Finally, given the values of $\mathbf{Z}_1^{(1)}$, $\mathbf{Z}_2^{(1)}$, and the data Y, sample $\mathbf{Z}_3^{(1)}$ from its conditional distribution $p(\mathbf{Z}_3 | \mathbf{Z}_1^{(1)}, \mathbf{Z}_2^{(1)}, Y)$. The three sampling processes complete one iteration of Gibbs sampling. After a large number of iterations say X, sampling any of the three parameters resemble sampling that parameter from its marginal posterior distribution. To simulate a marginal distribution for $\mathbf{Z}_3$, for example, after an X iterations, sample m values of $\mathbf{Z}_3$ from $(\mathbf{Z}_3 | \mathbf{Z}_1^{(X)}, \mathbf{Z}_2^{(X)}, Y)$. The mixture of these m sampled values of $\mathbf{Z}_3$ represents a numerical approximation to its marginal distribution. That is, the sampled values $\mathbf{Z}_{3,1}^{(X)}, \ldots\ldots, \mathbf{Z}_{3,m}^{(X)}$ can be viewed as a numerical approximation to the marginal posterior distribution $p(\mathbf{Z}_3 | Y)$. Similarly, the marginal posterior distributions for $\mathbf{Z}_1$ and $\mathbf{Z}_2$ can be approximated.

Gibbs sampling can be generalized to as many parameters as the investigated model has. The basic idea is to write the joint distribution of all parameters in forms similar to 4.15 to 4.17. Repeated sampling of the parameters from their corresponding conditional distributions results in approximation of their marginal distribution.

A special case of Gibbs sampling is found for certain applications, where some of the parameters depend only on a selected number of other parameters. For instance, in the case of approximating the marginal posterior distributions for the three sets of parameters $\mathbf{Z}_1$, $\mathbf{Z}_2$ and $\mathbf{Z}_3$, it is possible that the conditional distribution of say, $\mathbf{Z}_3$ depends only on $\mathbf{Z}_2$ and the data Y. While equations 4.15 and 4.16 for the marginal distributions for $\mathbf{Z}_1$ and $\mathbf{Z}_2$ do not change, equation 4.17, however, becomes

$$p(\mathbf{Z}_3 | Y) = \int_T p(\mathbf{Z}_3 | \mathbf{Z}_2, Y)\ p(\mathbf{Z}_2 | Y)\ \partial \mathbf{Z}_2 \quad . \qquad \textbf{(4.18)}$$

A repeated sampling of $\mathbf{Z}_1$, $\mathbf{Z}_2$ and $\mathbf{Z}_3$ from their conditional distributions in 4.15, 4.16 and 4.18 will approximate their marginal posterior distributions. Chapter 5 will show how the special case of Gibbs sampling is being used for approximating the marginal posterior distributions of the parameters in the model presented in 3.1.

CHAPTER 5

The Application of Gibbs Sampling to The Model of The Study

Two sections make up this chapter. The first section presents the application of Gibbs sampling for obtaining the marginal posterior distributions of the parameters of the model in 3.1. The conditional distributions of these parameters are also derived so they can be used in the application of Gibbs sampling. The second section presents the steps taken to empirically test the application of Gibbs sampling. These steps include setting the model, specifying and generating artificial data sets, assessing the heterogeneity of variance, obtaining initial estimates for the parameters to start the iteration process, assessing convergence of the iterative program, and describing the real data used in this study.

Obtaining Marginal Posterior Distributions of Parameters of the Model in 3.1

The proposed model in 3.1 involves two sets of parameters: those which have exchangeable prior distributions $\{U_j\}$ and $\{\sigma_j^2\}$ for $j=1,\cdots,k$ (i.e., $\{U_j\} = U_1, U_2, \cdots, U_k$ and $\{\sigma_j^2\} = \sigma_1^2, \sigma_2^2, \cdots, \sigma_k^2$) and the hyper-parameters τ^2, λ, θ, and $\sigma_\circ^2$ where $\lambda' = \{\gamma_\circ, \gamma_1, \cdots, \gamma_q, \beta_1, \cdots, \beta_p\}$. Our objective is to approximate,

numerically, the marginal posterior distribution for each one of these parameters using Gibbs sampling.

In general, the marginal posterior distribution for each one of these parameters can be found by integrating its joint posterior distribution with the other parameters over the parameter spaces of all other parameters. Mathematically stated, given the joint distribution in 3.12, the marginal distribution for each of the parameters of the model in 3.1 can be written as follows:

$$p(\{U_j\} \mid Y) = \int_{\tau^2}\int_{\lambda}\int_{\{\sigma_j^2\}} p(\{U_j\} \mid \tau^2, \lambda, \{\sigma_j^2\}, Y)\, p(\tau^2, \lambda, \{\sigma_j^2\} \mid Y) \times \partial\{\sigma_j^2\}\partial(\lambda)\,\partial(\tau^2) \,, \qquad \textbf{(5.1)}$$

$$p(\{\sigma_j^2\} \mid Y) = \int_{\sigma_\circ^2}\int_{\theta}\int_{\lambda}\int_{\{U_j\}} p(\{\sigma_j^2\} \mid \sigma_\circ^2, \theta, \lambda, \{U_j\}, Y)\, p(\sigma_\circ^2, \theta, \lambda, \{U_j\} \mid Y) \times \partial\{U_j\}\,\partial(\lambda)\,\partial(\theta)\,\partial(\sigma_\circ^2) \,, \qquad \textbf{(5.2)}$$

$$p(\lambda \mid Y) = \int_{\{\sigma_j^2\}}\int_{\{U_j\}} p(\lambda \mid \{\sigma_j^2\}, \{U_j\}, Y)\, p(\{\sigma_j^2\}, \{U_j\} \mid Y)\, \partial\{U_j\}\, \partial\{\sigma_j^2\} \,, \qquad \textbf{(5.3)}$$

$$p(\tau^2 \mid Y) = \int_{\{U_j\}} p(\tau^2 \mid \{U_j\}, Y)\, p(\{U_j\} \mid Y)\, \partial\{U_j\} \,, \qquad \textbf{(5.4)}$$

$$p(\theta \mid Y) = \int_{\sigma_\circ^2}\int_{\{\sigma_j^2\}} p(\theta \mid \sigma_\circ^2, \{\sigma_j^2\}, Y)\, p(\sigma_\circ^2, \{\sigma_j^2\} \mid Y)\, \partial\{\sigma_j^2\}\, \partial(\sigma_\circ^2) \,, \qquad \textbf{(5.5)}$$

$$p(\sigma_\circ^2 \mid Y) = \int_\theta \int_{\{\sigma_j^2\}} p(\sigma_\circ^2 \mid \theta, \{\sigma_j^2\}, Y)\, p(\theta, \{\sigma_j^2\} \mid Y)\ \partial\{\sigma_j^2\}\, \partial(\theta)\ . \tag{5.6}$$

Clearly, any one of the above marginal posterior distributions depends only on some of the parameters in the joint posterior distribution in 3.12. More specifically, the marginal posterior distribution for $\{U_j\}$ in 5.1 for example, does not depend on $\sigma_\circ^2$ and θ. Similarly, the marginal posterior distribution for τ^2 in 5.4 depends only on $\{U_j\}$. This is true for the rest of the parameters.

The first term of each integral in 5.1 to 5.6 represents the conditional distribution for one of the parameter(s) of interest, which usually has a known parametric form and from which we can sample. The second term of the integral can be thought of as a weight function multiplying the conditional distribution of the parameter of interest. Integrating the weighted conditional distribution over all the admissible values of the parameters in the second term will produce the marginal posterior distribution of the parameter of interest.

To exploit the dependency between the marginal posterior distributions in 5.1 to 5.6, two things must be known: the parametric forms of the conditional distributions in the integral in 5.1 to 5.5 and initial approximations to the values of the parameters $\lambda^{(0)}$, $\sigma_\circ^{2\,(0)}$, $\{U_j^{(0)}\}$, and $\{\sigma_j^{2\,(0)}\}$. The superscript "(0)" used with each of those parameters represents the initial value of the parameters. More discussion will be given later on the derivation of the

parametric forms of the conditional distributions of the parameters and the method of obtaining initial values for the specified parameters. Let us now assume that we know the parametric forms for the conditional distribution and we can sample from them. Given current estimates (start with initial values) of the parameters $\lambda^{(0)}$, $\sigma_\circ^{2\,(0)}$, $\{U_j^{(0)}\}$, and $\{\sigma_j^{2\,(0)}\}$, the iteration process starts as follows:

1- Given $\{U_j^{(0)}\}$ and the data Y, sample one value of $\tau^{2\,(1)}$ from $p(\tau^2 | \{U_j^{(0)}\}, Y)$ in 5.4. Notice that the superscript "(1)" used with τ^2 refers to the first cycle of the iteration process.

2- Given $\{\sigma_j^{2\,(0)}\}$, $\sigma_\circ^{2\,(0)}$ and the data Y, sample one value of $\theta^{(1)}$ from its conditional distribution $p(\theta | \sigma_\circ^{2\,(0)}, \{\sigma_j^{2\,(0)}\}, Y)$ in 5.5.

3- Given the sampled value of $\theta^{(1)}$ from step 2, $\{\sigma_j^{2\,(0)}\}$, and the data Y, sample one value of $\sigma_\circ^{2\,(1)}$ from its conditional distribution $p(\sigma_\circ^2 | \theta^{(1)}, \{\sigma_j^{2\,(0)}\}, Y)$ in 5.6.

4- Given the sampled value of $\tau^{2\,(1)}$ from step 1, $\{\sigma_j^{2\,(0)}\}$, $\lambda^{(0)}$, and the data Y, sample one set of values of $\{U_j^{(1)}\} = \{U_1^{(1)}, U_2^{(1)}, \cdots, U_k^{(1)}\}$ for $j=1,\cdots,k$ from its conditional distribution $p(\{U_j\} | \lambda^{(0)}, \tau^{2\,(1)}, \{\sigma_j^{2\,(0)}\}, Y)$ in 5.1.

5- Given the sampled values of $\{U_j^{(1)}\}$, the sampled value of $\theta^{(1)}$, the sampled value of $\sigma_\circ^{2\,(1)}$ from steps 4, 2 and 3 respectively, the initial value of $\lambda^{(0)}$, and the data Y, sample one set of values of $\{\sigma_j^{2\,(1)}\} = \{\sigma_1^{2\,(1)}, \sigma_2^{2\,(1)}, \cdots, \sigma_k^{2\,(1)}\}$ for

$j=1,\cdots,k$ from its conditional distribution $p(\{\sigma_j^2\}|\sigma_\circ^{2\,(1)},\theta^{(1)},\lambda^{(0)},\{U_j^{(1)}\},Y)$ in 5.2.

6- Given the set of $\{U_j^{(1)}\}$, the set $\{\sigma_j^{2\,(1)}\}$, and the data Y, sample one value of $\lambda^{(1)}$ from its conditional distribution $p(\lambda|\{U_j^{(1)}\},\{\sigma_j^{2\,(1)}\},Y)$ in 5.3.

Going through these six steps of sampling from the given conditional distributions finishes one cycle of the iteration process in Gibbs sampling. This cycle produces the first sampled value of each of the parameters, which can then be used to produce new sampled values of the same parameters in the second cycle. In the second cycle, starting with step 1 again and given $\{U_j^{(1)}\}$, generate new value of $\tau^{2\,(2)}$ where the superscript "(2)" used with τ^2 refers to the second cycle of the iteration process. As we continue iterating between the above six steps, the mixture of the values of any one of these six sets of parameters becomes increasingly accurate in representing its marginal posterior distribution.

Finding the Conditional Distributions for the Parameters of the Model in 3.1

One of the requisites for obtaining the marginal posterior distributions in 5.1 to 5.6 through the iteration process of Gibbs sampling is to be able to sample from the conditional distributions for the respective parameters. This implies the need for identifying the parametric forms for

these conditional distributions. Within the Bayesian approach the elements of the set $\{U_j\} = U_1, U_2, \cdots, U_k$ are assumed independent and identically distributed with respect to their conditional distribution as defined in (5.1). Identifying the conditional distribution of any element of this set therefore, can be generalized to the other elements of the set. The same logic applies in determining the conditional distribution for $\{\sigma_j^2\} = \sigma_1^2, \sigma_2^2, \cdots, \sigma_k^2$.

The conditional distribution of U_j given λ, τ^2, $\{\sigma_j^2\}$, and Y

From lines 1 and 2 of the joint density function in 3.13, the conditional distribution of $U_j | \lambda, \tau^2, \{\sigma_j^2\}, Y$ is found to have a density function proportional to

$$\prod_{j=1}^{k}\left[(\sigma_j^2)^{-\frac{n_j}{2}} \exp\left(-\frac{1}{2}(Y_j - Z_j\lambda - 1_j U_j)'\sigma_j^{-2}(Y_j - Z_j\lambda - 1_j U_j)\right)\right] \times (\tau^2)^{-\frac{k}{2}} \exp\left(\frac{-\sum_{j=1}^{k} U_j^2}{2\tau^2}\right). \tag{5.7}$$

Since U_j appears only in the exponential parts of the expression in 5.7, the density function in 5.7, can be written proportional to

$$\exp\left\{-\frac{1}{2}\sum_{j=1}^{k}\left((d_j - 1_j U_j)'\sigma_j^{-2}(d_j - 1_j U_j) + \frac{U_j^2}{\tau^2}\right)\right\}, \tag{5.8}$$

where $Y_j - Z_j\lambda = d_j$. Ignoring the terms that do not depend on U_j, the above expression can be simplified to

$$\exp\left\{-\frac{1}{2}\sum_{1}^{k}\left(-2\,(U_j' 1_j' d_j)\,\sigma_j^{-2} + U_j^2\,(n_j\sigma_j^{-2} + \tau^{-2})\right)\right\} . \tag{5.9}$$

Completing the square in 5.9 results in

$$\exp\left(-\frac{1}{2}\sum_{1}^{k}\left(U_j - (n_j\sigma_j^{-2} + \tau^{-2})^{-1} 1_j' d_j \sigma_j^{-2}\right)^2 (n_j\sigma_j^{-2} + \tau^{-2})\right) . \tag{5.10}$$

Substituting for $1_j' d_j = n_j(\overline{Y}_j - \overline{Z}_j'\lambda)$ in 5.10 results in

$$\exp\left(-\frac{1}{2}\sum_{1}^{k}\left(U_j - n_j\sigma_j^{-2}(n_j\sigma_j^{-2}+\tau^{-2})^{-1}(\overline{Y}_j - \overline{Z}_j'\lambda)\right)^2 (n_j\sigma_j^{-2}+\tau^{-2})\right). \tag{5.11}$$

The expression in 5.11 represents a kernel of a normal density function. Therefore, $U_j|\lambda, \tau^2, \{\sigma_j^2\}, Y$ is distributed as normal with mean U_j^* and variance V_{u_j}, where

$$U_j^* = n_j\sigma_j^{-2}(n_j\sigma_j^{-2}+\tau^{-2})^{-1}(\overline{Y}_j - \overline{Z}_j'\lambda) \ , \quad V_{U_j} = (n_j\sigma_j^{-2}+\tau^{-2})^{-1} . \tag{5.12}$$

The conditional distribution of σ_j^2 given λ, θ, $\sigma_\circ^2$, $\{U_j\}$, and Y

From lines 1 and 3 of the joint density in 3.13, the conditional distribution of $\sigma_j^2|\lambda, \nu_\circ, \sigma_\circ^2, \{U_j\}, Y$ is found to have a density function proportional to

$$(\sigma_j^2)^{-\frac{n_j}{2}} \exp\left(-\frac{1}{2}\,(Y_j - Z_j\boldsymbol{\lambda} - 1_j U_j)'\sigma_j^{-2}\,(Y_j - Z_j\boldsymbol{\lambda} - 1_j U_j)\right)$$
$$\times \quad \frac{1}{\Gamma\left(\frac{1}{2\theta}\right)} \left(\frac{\sigma_\circ^2}{2\theta}\right)^{\frac{1}{2\theta}} (\sigma_j^2)^{-(\frac{1}{2\theta}+1)} \exp\left(\frac{-\sigma_\circ^2}{2\theta\sigma_j^2}\right) . \tag{5.13}$$

Let $(Y_j - Z_j\boldsymbol{\lambda} - 1_j U_j)'(Y_j - Z_j\boldsymbol{\lambda} - 1_j U_j) = n_j S_j^2$. Thus, 5.13 can be rewritten as

$$\frac{\left(\frac{\sigma_\circ^2}{2\theta}\right)^{\frac{1}{2\theta}}}{\Gamma\left(\frac{1}{2\theta}\right)} (\sigma_j^2)^{-\left(\frac{n_j}{2}+\frac{1}{2\theta}+1\right)} \exp\left\{-\left(\frac{n_j S_j^2}{2\,\sigma_j^2} + \frac{\sigma_\circ^2}{2\theta\sigma_j^2}\right)\right\} . \tag{5.14}$$

Ignoring the terms that do not involve σ_j^2, results in writing the expression in 5.14 proportional to

$$(\sigma_j^2)^{-(\frac{n_j}{2}+\frac{1}{2\theta}+1)} \exp\left\{-\frac{1}{\sigma_j^2}\left(\frac{n_j S_j^2\theta + \sigma_\circ^2}{2\theta}\right)\right\} . \tag{5.15}$$

In general, a continuous random variable X with a density function f is said to have an Inverse Gamma distribution with parameters α and β if

$$f(x) = \frac{1}{\Gamma(\alpha)\,\beta^\alpha}\, x^{-(\alpha+1)} \exp\left(-\frac{1}{x\beta}\right), \qquad 0 \le x \le \infty . \tag{5.16}$$

The expression in 5.15 therefore, represents a kernel of a Inverse Gamma density with the two parameters

$$\alpha_{\sigma_j^2} = \frac{n_j}{2} + \frac{1}{2\theta}, \qquad \beta_{\sigma_j^2} = \frac{2\theta}{n_j S_j^2\theta + \sigma_\circ^2} . \tag{5.17}$$

To sample σ_j^2 from its conditional distribution, the following strategy can be adopted: define $\sigma_j^2 = h(\sigma_j^{-2})$, where h

is a one to one function defined as $h(\sigma_j^{-2}) = (\sigma_j^{-2})^{-1}$. The conditional density function of σ_j^{-2} is found to be

$$g(\sigma_j^{-2} \mid \lambda, \theta, \sigma_\circ^2, \{U_j\}, Y) \propto f(\sigma_j^2 \mid \lambda, \theta, \sigma_\circ^2, \{U_j\}, Y) \left| \frac{\partial h(\sigma_j^{-2})}{\partial(\sigma_j^{-2})} \right| , \qquad (5.18)$$

where $\left| \dfrac{\partial h(\sigma_j^{-2})}{\partial(\sigma_j^{-2})} \right| = (\sigma_j^{-2})^{-2}$. Therefore,

$$g(\sigma_j^{-2} \mid \lambda, \theta, \sigma_\circ^2, \{U_j\}, Y) \propto (\sigma_j^{-2})^{\frac{n_j}{2} + \frac{1}{2\theta} - 1} \exp\left\{ -\sigma_j^{-2} \left(\frac{n_j S_j^2 \theta + \sigma_\circ^2}{2\theta} \right) \right\} . \qquad (5.19)$$

In general, a continuous random variable x with a density function f is said to have a Gamma distribution with parameters α and β if

$$f(x) = \frac{1}{\Gamma(\alpha)\,\beta^\alpha} x^{\alpha-1} \exp\left(\frac{-x}{\beta} \right) , \qquad 0 \le x \le \infty . \qquad (5.20)$$

The expected value and the variance of X are given by

$$E(X) = \alpha\beta , \qquad Var(X) = \alpha\beta^2 . \qquad (5.21)$$

The expression in 5.21, therefore has the kernel of Gamma density function. Thus, given $\lambda, \theta, \sigma_\circ^2, \{U_j\}$, and Y, the parameter σ_j^{-2} is distributed as Gamma with

$$\alpha_{\sigma_j^{-2}} = \frac{n_j}{2} + \frac{1}{2\theta} , \qquad \beta_{\sigma_j^{-2}} = \frac{2\theta}{n_j S_j^2 \theta + \sigma_\circ^2} \qquad (5.22)$$

To sample σ_j^2, simply sample σ_j^{-2} from 5.21 then use the relation $h(\sigma_j^{-2}) = (\sigma_j^{-2})^{-1}$ to get σ_j^2.

The conditional expectation and the variance of σ_j^{-2} are easily found by applying 5.21,

$$E(\sigma_j^{-2} \mid \lambda, \theta, \sigma_\circ^2, \{U_j\}, Y) = \left(\omega_j S_j^2 + (1-\omega_j)\,\sigma_\circ^2\right)^{-1}, \quad \text{and}$$

$$Var(\sigma_j^{-2} \mid \lambda, \theta, \sigma_\circ^2, \{U_j\}, Y) = \frac{2\,\omega_j}{n_j}\left[\omega_j S_j^2 + (1-\omega_j)\,\sigma_\circ^2\right]^{-2}, \tag{5.23}$$

where $\omega_j = \dfrac{n_j}{n_j + \theta^{-1}}$.

The conditional expectation in 5.23 has a form similar to Stein's (1956) shrinkage estimate. It is a form of a weighted estimate of the variance S_j^2, and the asymptotic overall mean of the variances $\sigma_\circ^2$. The weights depend, in a natural way, on the degrees of freedom for obtaining S_j^2 and $\sigma_\circ^2$. As θ approaches zero, indicating homogenous variances, so does ω_j leaving the expected value of σ_j^{-2} equal to $\sigma_\circ^{-2}$. When θ gets very large however, the value of ω_j approaches one, leaving the expected value of σ_j^{-2} equal to the inverse of the sample variance S_j^{-2}.

The operational definition of reliability is the ratio of the "true" variance to the observed variance. Therefore, one can think of $1 - \omega_j = \dfrac{\theta^{-1}}{n_j + \theta^{-1}}$ as a reliability index for σ_j^{-2}, expressed in terms of its precision. It represents the proportion of the precision in the OLS estimates of σ_j^{-2} that is parameter precision.

The conditional distribution of λ given $\{U_j\}$, $\{\sigma_j^2\}$, and Y

From line 1 of the joint density in 3.13, the conditional distribution of $\lambda \mid \{U_j\}, \{\sigma_j^2\}, Y$ is found to have a density function proportional to

$$\exp\left(\sum_1^k -\frac{1}{2}\,(Y_j - Z_j\lambda - 1_jU_j)'\sigma_j^{-2}\,(Y_j - Z_j\lambda - 1_jU_j)\right). \tag{5.24}$$

Let $Y_j - 1_jU_j = D_j$. The expression in 5.24 can now be written as

$$\exp\left\{-\frac{1}{2}\sum_1^k \left[(D_j - Z_j\lambda)'\sigma_j^{-2}\,(D_j - Z_j\lambda)\right]\right\}. \tag{5.25}$$

Multiplying the terms inside the exponential, and ignoring the terms that do not depend on λ, the conditional distribution of λ becomes proportional to

$$\exp\left(-\frac{1}{2}\sum_1^k \sigma_j^{-2}\,(\lambda'Z_j'Z_j\lambda \;-\; 2\lambda'Z_j'D_j)\right). \tag{5.26}$$

By completing the square, 5.26 is rewritten as

$$\begin{aligned}\exp\,[-\frac{1}{2}&\left(\lambda - \left(\sum_1^k (Z_j'Z_j)\,\sigma_j^{-2}\right)^{-1}\sum_1^k (Z_j'D_j)\,\sigma_j^{-2}\right)'\sum_1^k (Z_j'Z_j)\,\sigma_j^{-2} \\ &\times\left(\lambda - \left(\sum_1^k (Z_j'Z_j)\,\sigma_j^{-2}\right)^{-1}\sum_1^k (Z_j'D_j)\,\sigma_j^{-2}\right)]\end{aligned} \tag{5.27}$$

When $Y_j - 1_jU_j$ is substituted back for D_j in 5.27, the resulting expression represents a kernel of a normal density. Thus $\lambda\,|\,\{U_j\}, \{\sigma_j^2\}, Y$ is distributed as normal with mean λ^* and variance V_λ, where

$$\begin{aligned}\lambda^* &= \left(\sum_1^k (Z_j'Z_j)\,\sigma_j^{-2}\right)^{-1}\left(\sum_1^k (z_j'Y_j\sigma_j^{-2}) \;-\; \sum_1^k Z_j'1_jU_j\sigma_j^{-2}\right), \\ V_\lambda &= \left(\sum_1^k (Z_j'Z_j)\,\sigma_j^{-2}\right)^{-1}.\end{aligned} \tag{5.28}$$

The mean in 5.28 is a weighted least square estimate with $Y_j - 1_j U_j$ as an outcome, Z_j as predictors, and σ_j^{-2} as the weighing factor.

The conditional distribution of τ^2 given $\{U_j\}$ and Y

From line 2 of the joint density in 3.13, the conditional distribution of $\tau^2 \mid \{U_j\}, Y$ is found to have a density function proportional to

$$(\tau^2)^{-\frac{k}{2}} \exp\left(\frac{-\sum_1^k U_j^2}{2\tau^2} \right) , \tag{5.29}$$

Compared to the general form of an Inverse Gamma function in 5.16, the expression in 5.29 has a similar density function. Therefore, $\tau^2 \mid \{U_j\}, Y$ is distributed as an Inverse Gamma variable with

$$\alpha_{\tau^2} = \frac{k}{2} - 1 \quad , \qquad \beta_{\tau^2} = \frac{2}{\sum_1^k U_j^2} \quad . \tag{5.30}$$

The conditional expectation and the variance of τ^2 are easily found to be

$$E(\tau^2 \mid \{U_j\}, Y) = \frac{\sum_1^k U_j^2}{k-4}, \quad \text{and}$$
$$Var(\tau^2 \mid \{U_j\}, Y) = \frac{2\left(\sum_1^k U_j^2\right)^2}{(k-4)^2 (k-6)} . \tag{5.31}$$

To sample τ^2 from its conditional distribution, adopt the same strategy used for sampling σ_j^2; that is, let

$$\tau^2 = h(\tau^{-2}) = (\tau^{-2})^{-1} . \tag{5.32}$$

The density function for τ^{-2} can then be found as

$$g(\tau^{-2} \mid \{U_j\}, Y) = f(\tau^2 \mid \{U_j\}, Y) \left| \frac{\partial h(\tau^{-2})}{\partial(\tau^{-2})} \right| , \tag{5.33}$$

where $\left| \frac{\partial h(\tau^{-2})}{\partial(\tau^{-2})} \right| = (\tau^{-2})^{-2}$. Therefore,

$$h(\tau^{-2} \mid \{U_j\}, Y) \propto (\tau^{-2})^{\frac{k}{2}-2} \exp\left(\frac{-\tau^{-2} \sum_1^k U_j^2}{2} \right) . \tag{5.34}$$

Compared to the general form of a Gamma function in 5.20, the expression in 5.34 is similar. Therefore, $\tau^{-2} \mid \{U_j\}, Y$ is distributed as Gamma variable with

$$\alpha_{\tau^{-2}} = \frac{k}{2} - 1 , \qquad \beta_{\tau^{-2}} = \frac{2}{\sum_1^k U_j^2} . \tag{5.35}$$

The conditional expectation and the variance of τ^{-2} are

$$E(\tau^{-2}\,|\,\{U_j\},\,Y) = \frac{k-2}{\sum_1^k U_j^2}\,, \quad \text{and}$$

$$Var(\tau^{-2}\,|\,\{U_j\},\,Y) = \frac{2(k-2)}{\left(\sum_1^k U_j^2\right)^2}\,. \qquad (5.36)$$

To sample τ^2, simply sample τ^{-2} from 5.34, then use the relation in 5.32 to get τ^2.

The conditional distribution of $\sigma_\circ^2$ given θ and $\{\sigma_j^2\}$

From line 3 of the joint density in 3.13, the conditional distribution of $\sigma_\circ^2\,|\,\theta,\{\sigma_j^2\}$ is found to have a density function proportional to

$$\frac{\left(\frac{\sigma_\circ^2}{2\theta}\right)^{\frac{k}{2\theta}}}{\left(\Gamma\left(\frac{1}{2\theta}\right)\right)^k}\exp\left(\frac{-\sigma_\circ^2}{2\theta}\sum_1^k\frac{1}{\sigma_j^2}\right)\prod_1^k(\sigma_j^2)^{-\left(\frac{1}{2\theta}+1\right)}\,. \qquad (5.37)$$

Let G and H represent the geometric and the harmonic means of σ_j^2's as follows:

$$G = \prod_1^k\left(\sigma_j^2\right)^{\frac{1}{k}}\,, \quad \text{Thus} \quad \prod_1^k \sigma_j^2 = G^k\,.$$

$$H = \frac{k}{\sum_1^k \frac{1}{\sigma_j^2}}\,, \quad \text{Thus} \quad \sum_1^k\frac{1}{\sigma_j^2} = \frac{k}{H}\,. \qquad (5.38)$$

Substituting the equalities of 5.38 into 5.37 produces

$$\frac{\left(\frac{1}{2\theta}\right)^{\frac{k}{2\theta}}}{\left(\Gamma\left(\frac{1}{2\theta}\right)\right)^{k} G^{k\left(\frac{1}{2\theta}+1\right)}} (\sigma_\circ^2)^{\frac{k}{2\theta}} \exp\left(\frac{-k\sigma_\circ^2}{2\theta H}\right) . \tag{5.39}$$

The first term of the above expression does not involve $\sigma_\circ^2$, therefore, the above expression can be rewritten proportional to

$$(\sigma_\circ^2)^{\frac{k}{2\theta}} \exp\left(\frac{-k\sigma_\circ^2}{2\theta H}\right) . \tag{5.40}$$

The above expression represents the kernel of a Gamma density with

$$\alpha_{\sigma_\circ^2} = \frac{k}{2\theta} + 1 \quad , \qquad \beta_{\sigma_\circ^2} = \frac{2\theta H}{k} . \tag{5.41}$$

The conditional expectation and variance for $\sigma_\circ^2$ therefore, are

$$\begin{aligned} E(\sigma_\circ^2 \,|\, \theta, \{\sigma_j^2\}) &= H + \frac{2\theta}{k} \quad , \qquad \text{and} \\ Var(\sigma_\circ^2 \,|\, \theta, \{\sigma_j^2\}) &= \left(H + \frac{2\theta}{k}\right)\frac{2\theta H}{k} . \end{aligned} \tag{5.42}$$

Notice that the first part of the above expectation represents information depicted by the harmonic mean of the groups variances $\{\sigma_j^2\}$. The second part involves θ, which reflects the variability of these variances. As θ approaches the zero, implying homogeneity of variance situation, the second part in that expectation diminishes, and the conditional expectation

of $\sigma_\circ^2$ becomes equal to the harmonic mean of the group variances.

Similar to σ_j^2 and τ^2, we can find the conditional distribution of $\sigma_\circ^{-2}$ by letting

$$\sigma_\circ^2 = g(\sigma_\circ^{-2}) = \left(\sigma_\circ^{-2}\right)^{-1} . \tag{5.43}$$

The density function for $\sigma_\circ^{-2}$ can then be found as

$$h(\sigma_\circ^{-2} \mid \{\sigma_j^2\}, \theta) = f(\sigma_\circ^2 \mid \{\sigma_j^2\}, \theta) \left| \frac{\partial g(\sigma_\circ^{-2})}{\partial(\sigma_\circ^{-2})} \right| . \tag{5.44}$$

where $\left| \frac{\partial g(\sigma_\circ^{-2})}{\partial(\sigma_\circ^{-2})} \right| = \left(\sigma_\circ^{-2}\right)^{-2}$. Therefore,

$$h(\sigma_\circ^{-2} \mid \{\sigma_j^2\}, \theta) \propto \left(\sigma_\circ^{-2}\right)^{\frac{k}{2\theta}-2} \exp\left(\frac{-k}{2\theta H \sigma_\circ^{-2}}\right) . \tag{5.45}$$

The expression above has the same general form as the inverse Gamma function in 5.16. Therefore, $\sigma_\circ^{-2} \mid \{\sigma_j^2\}, \theta$ is distributed as an Inverse Gamma variable with $\alpha_{\sigma_\circ^{-2}}$ and $\beta_{\sigma_\circ^{-2}}$ given in 5.41. The mean and variance of $\sigma_\circ^{-2}$ are

$$E(\sigma_\circ^{-2} \mid \{\sigma_j^2\}, \theta) = \frac{1}{H} , \quad \text{and}$$
$$Var(\sigma_\circ^{-2} \mid \{\sigma_j^2\}, \theta) = \frac{1}{H^2\left(\frac{k}{2\theta} - 1\right)} . \tag{5.46}$$

The Conditional Distribution of θ given $\sigma_\circ^2$ and $\{\sigma_j^2\}$

From line 3 of the joint density in 3.13, the conditional distribution of $\theta \mid \sigma_\circ^2, \{\sigma_j^2\}$ is found to have a density function proportional to

$$\frac{(\sigma_\circ^2)^{\frac{k}{2\theta}}}{\left(\Gamma\left(\frac{1}{2\theta}\right)\right)^k}\left(\frac{1}{2\theta}\right)^{\frac{k}{2\theta}}\exp\left(\frac{-\sigma_\circ^2}{2\theta}\sum_1^k\frac{1}{\sigma_j^2}\right)\prod_1^k(\sigma_j^2)^{-(\frac{1}{2\theta}+1)}\ . \tag{5.47}$$

Let G and H be defined as in 5.38. Further, to simplify the term $\Gamma(\frac{1}{2\theta})$ in 5.47 we use the first term of Stirling's approximation. The general expression for Stirling's approximation for a variable P is given by

$$\Gamma(p)\approx(2\pi)^{\frac{1}{2}}p^{(p-\frac{1}{2})}\exp(-p)\left[1+\frac{1}{12p}+\frac{1}{288p^2}-\frac{139}{51840p^3}\cdots\right]. \tag{5.48}$$

Substituting $1/2\theta$ for p in the first term of 5.48 will produces the following approximation.

$$\left(\Gamma\left(\frac{1}{2\theta}\right)\right)^k\approx\left((2\pi)^{\frac{1}{2}}\left(\frac{1}{2\theta}\right)^{\frac{1}{2\theta}-\frac{1}{2}}\exp\left(\frac{-1}{2\theta}\right)\right)^k\ . \tag{5.49}$$

Substituting G, H, and 5.49 in 5.47 results in

$$\frac{\left(\frac{1}{2\theta}\right)^{\frac{k}{2\theta}}(\sigma_\circ^2)^{\frac{k}{2\theta}}}{(2\pi)^{\frac{k}{2}}\left(\frac{1}{2\theta}\right)^{\frac{k}{2\theta}-\frac{k}{2}}\exp\left(-\frac{k}{2\theta}\right)}\,G^{-k(\frac{1}{2\theta}+1)}\exp\left(\frac{-k\sigma_\circ^2}{2\theta H}\right)\ . \tag{5.50}$$

The expression in 5.50 reduces to

$$\frac{1}{(2G)^k\pi^{\frac{k}{2}}}\left(\frac{\sigma_\circ^2}{G}\right)^{\frac{k}{2\theta}}\left(\frac{1}{\theta}\right)^{\frac{k}{2}}\exp\left(\frac{-k\sigma_\circ^2}{2\theta H}+\frac{k}{2\theta}\right) \tag{5.51}$$

The first term of 5.51 does not involve θ, therefore, the above expression is rewritten proportional to

$$\left(\frac{1}{\theta}\right)^{\frac{k}{2}}\exp\left[\frac{-k}{2\theta}\left(\frac{\sigma_\circ^2}{H}-1-\log(\sigma_\circ^2)+\log(G)\right)\right] \tag{5.52}$$

The above expression has kernel of an Inverse Gamma density. Therefore, $\theta \mid \sigma_\circ^2, \{\sigma_j^2\}$ has an inverse Gamma distribution with

$$\alpha_\theta = \frac{k}{2} - 1 \, , \qquad \beta_\theta = \frac{2}{k\left(\frac{\sigma_\circ^2}{H} - \log(\sigma_\circ^2) + \log(G) - 1\right)} \, . \tag{5.53}$$

The conditional expectation and variance of θ are found to be

$$E(\theta \mid \sigma_\circ^2, \{\sigma_j^2\}) = \frac{k}{k-4}\left(\frac{\sigma_\circ^2}{H} - \log(\sigma_\circ^2) + \log(G) - 1\right) , \quad \text{and}$$
$$Var(\theta \mid \sigma_\circ^2, \{\sigma_j^2\}) = \frac{2k^2\left(\frac{\sigma_\circ^2}{H} - \log(\sigma_\circ^2) + \log(G) - 1\right)^2}{(k-6)(k-4)^2} \, . \tag{5.54}$$

Notice that as $k \to \infty$, the above conditional expectation and variance will be

$$E(\theta \mid \sigma_\circ^2, \{\sigma_j^2\}) = \mu^* \, , \quad \text{and}$$
$$Var(\theta \mid \sigma_\circ^2, \{\sigma_j^2\}) = \frac{2(\mu^*)^2}{k} \, , \tag{5.55}$$

where $\mu^* = \log(G) - \log(H)$.

Sampling θ from its conditional distribution is done using the same strategy used in sampling the groups' variances from their conditional distributions. That is, define $\theta = h(\theta^{-1})$, where h is a one to one function defined as $h(\theta^{-1}) = (\theta^{-1})^{-1}$. The conditional density function $g(\theta^{-1} \mid \sigma_\circ^2, \{\sigma_j^2\})$ is found as follows:

$$g(\theta^{-1} \mid \sigma_{\circ}^{2}, \{\sigma_{j}^{2}\}) \propto f(\theta \mid \sigma_{\circ}^{2}, \{\sigma_{j}^{2}\}) \left|\frac{\partial h(\theta^{-1})}{\partial(\theta^{-1})}\right| , \tag{5.56}$$

where $\left|\frac{\partial h(\theta^{-1})}{\partial(\theta^{-1})}\right| = (\theta^{-1})^{-2}$. Therefore,

$$f(\theta^{-1} \mid \sigma_{\circ}^{2}, \{\sigma_{j}^{2}\}) \propto (\theta^{-1})^{\frac{k}{2}-2} \exp\left[\frac{-\theta^{-1} k}{2}\left(\frac{\sigma_{\circ}^{2}}{H} - 1 - \log(\sigma_{\circ}^{2}) + \log(G)\right)\right] \tag{5.57}$$

The above expression is the kernel of Gamma density. Thus, $\theta^{-1} \mid \sigma_{\circ}^{2}, \{\sigma_{j}^{2}\}$ has a Gamma distribution with

$$\alpha_{\theta^{-1}} = \frac{k}{2} - 1 , \qquad \beta_{\theta^{-1}} = \frac{2}{k\left(\frac{\sigma_{\circ}^{2}}{H} - \log(\sigma_{\circ}^{2}) + \log(G) - 1\right)} \tag{5.58}$$

To sample θ, simply sample θ^{-1} from 5.51, then use the relation $h(\theta^{-1}) = (\theta^{-1})^{-1}$ to get θ.

Empirical Applications of Gibbs Sampling

The procedures presented in the previous section for obtaining the marginal posterior distributions for the parameters of the model in 3.1 were next tested empirically. This section presents the processes for this test in two phases. The first phase involves the use of artificial data sets with pre-specified values of parameters in a FORTRAN computer program that implements the controlled statistical sampling technique. One of the objectives at this phase was to validate the theoretical aspects of the statistical techniques used. The resulting marginal posterior

distributions of the parameters in the model can be examined against their true values, which are used in generating the artificial data set and their parameter spaces. Another objective is to check the integrity of the computation of the computer program so that it can be used with confidence on real data in the second phase.

A real data set used by Raudenbush and Bryk (1987) is analyzed in the second phase of the empirical testing of the procedure. The data set comes from a sample of 160 U.S. high schools. It includes 83 catholic schools and a random sample of 77 public schools drawn from the High School and Beyond data base. Sample sizes range from 14 to 66 student per school with an average of 44.5 student per school.

Artificial Data Analysis

The general model presented in 3.1 allows for q between-unit variables W and p within-unit variables X. Theoretically, the statistical procedure presented in the previous chapter works with any number of between and within-unit variables as long as $k>q$ and $n_j>p$. However, for practicality, three versions of that model will be presented. In the first version, no between and within-unit variables were specified. This is the simplest case, where the model in 3.1 reduces to a simple ANOVA model. One between and one within-unit variable will be included in the second version of

the model with a random intercept and fixed effects predictors, i.e., $q = p = 1$. The last version covers the case of two between and three within-unit variables with random intercept and fixed effects predictors , i.e., $q = 2$ and $p = 3$ (see Table 5.1).

The above three models range in their complexity to represent conditions often found in educational research projects. The first is a simple version of the model, used to facilitate the understanding of the statistical procedure and its application. The third version of the model is similar to the example of the real data used in the study.

Data Specifications

Two primary but not completely mutually exclusive criteria were used to create the data. The first deals with the degree of heterogeneity of variance. The second criterion deals with the number of groups k to be generated. For each combination of these two criteria along with each of the three versions of the model a data set was generated.

Based on the first criterion (degree of heterogeneity of variance) and given that the coefficient of variation (C.V.) for $\{\sigma_j^2\}$ is equal to $\sqrt{\frac{2\theta}{1-4\theta}}$, where $0<\theta<0.25$, two alternatives were chosen. The first reflects the case of heterogenous variances based on a large value of $\theta=0.2$ with C.V. = 1.41. The other alternative reflects the case of not so heteroge-

nous variances (i.e., small θ=0.02) with C.V. = 0.21. This data set is used to test the applicability of holding θ to a very small value (close to zero) to reflect the homogeneity of variance.

For the second criteria, number of groups (k) generated three values of k, 100, 40 and 15, were chosen. These values reflect large, moderate, and small numbers of groups.

The combination of the three versions of the model with the two choices of the first criterion and the three group sample sizes in the second criterion creates 18 data sets. Variations of the values of the other parameters (like λ, τ^2, and $\sigma_\circ^2$) in the model were arbitrarily controlled based on how many between- and within-unit variables exist in the model (i.e., different versions of the model). Table 5.1 presents a summary of the three models used with eighteen data sets. Except for the intercept, regression coefficients in these three models are assumed to be fixed effects.

Table 5.1

Models used in generating the artificial data

Model	Case	k
Model (1) $Y_j = Z_j\boldsymbol{\lambda} + 1_j U_j + \epsilon_j$, for $j = 1, \cdots, k$; $Z_j = \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$, and $\boldsymbol{\lambda} = \gamma_\circ$.	Heterogeneity case (θ=0.2)	k=100
		k=40
		k=15
	Homogeneity case (θ=0.02)	k=100
		k=40
		k=15
Model (2) $Y_j = Z_j\boldsymbol{\lambda} + 1_j U_j + \epsilon_j$, for ; $Z_j = \begin{bmatrix} 1 & W_j & X_{1j} \\ \vdots & \vdots & \vdots \\ 1 & W_j & X_{n_j j} \end{bmatrix}$, and $\boldsymbol{\lambda} = \begin{bmatrix} \gamma_\circ \\ \gamma_1 \\ \beta \end{bmatrix}$.	Heterogeneity case (θ=0.2)	k=100
		k=40
		k=15
	Homogeneity case (θ=0.02)	k=100
		k=40
		k=15
Model (3) $Y_j = Z_j\boldsymbol{\lambda} + 1_j U_j + \epsilon_j$, for ; $Z_j = \begin{bmatrix} 1 & W_{j1} & W_{j2} & X_{1j1} & X_{1j2} & X_{1j3} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & W_{j1} & W_{j2} & X_{n_j j1} & X_{n_j j2} & X_{n_j j3} \end{bmatrix}$ and $\boldsymbol{\lambda} = \begin{bmatrix} \gamma_\circ \\ \gamma_1 \\ \gamma_2 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$.	Heterogeneity case (θ=0.2)	k=100
		k=40
		k=15
	Homogeneity case (θ=0.02)	k=100
		k=40
		k=15

Data Creation

Given the model in 3.1, all eighteen data sets were generated in two general steps. The first step covered the generation of data for the group level variables. The second step covered the generation of data within each group based on the data generated for the group level variables in the first step. These two steps reflect the inherent nesting of the two levels of the model.

Step one - Between-group data

The following random vectors with k elements (k = number of groups) in each vector were generated:

1. Within-group sample sizes n_j vector, for $j = 1, \ldots, k$. First, k random numbers were generated from a UNIFORM distribution between 5 and 60. Each number was rounded to the nearest integer to represent a sample size n_j for each of the k groups. Variation in sample sizes are considered to include a wide range from small (n_j=5) to large (n_j=60). The following is an example of sample size vector:

Group	**Sample size** n_j
1	$n_1 = 40$
2	$n_2 = 15$
⋮	⋮
k	$n_k = 60$

2. Within-group variances $\{\sigma_j^2\}$ vector. For pre-specified values of θ and $\sigma_\circ^2$ the variances $\{\sigma_j^2\}$ were generated from their respective distribution.

$$f(\sigma_j^2 | \theta, \sigma_\circ^2) \sim \Gamma^{-1}(\alpha, \beta), \quad \text{where}$$
$$\alpha = \frac{1}{2\theta}, \quad \beta = \frac{2\theta}{\sigma_\circ^2} \tag{5.59}$$

3. Based on 3.9, the k elements of the $\{U_j\}$ vector were generated from a normal distribution with mean equal zero and variance equal τ^2.

4. Now, q vectors of $W = [W_1, W_2, \cdots, W_q]$ and p vectors of $\overline{X} = [\overline{X}_1, \overline{X}_2, \cdots, \overline{X}_p]$ each has k elements were generated from a multivariate normal distribution with mean vector of zero and pre-specified variance-covariance matrix Σ. The values of the p vectors of $\overline{X} = [\overline{X}_1, \overline{X}_2, \cdots, \overline{X}_p]$, were then used, in step two, as means of the p vectors for the within-group predictors $X_j = [X_{j1}, X_{j2}, \cdots, X_{jp}]$ in the design matrix Z in 3.2.

Step two - Within-group data

The generated data in the second step is for the within-group variables. It depends on the values of the between-group variables generated in the first step. For each j, where $j = 1, \cdots, k$ groups, an p variables of $X_j = [X_{j1}, X_{j2}, \cdots, X_{jp}]$ within-group predictors and an error term e_j were generated. The number of observations for these variables within each group is equal to the sample size n_j generated in the first data set.

1. Based on the model in 3.1, $e_{ij} \sim N(0, I\sigma_j^2)$. Therefore, given $\{\sigma_j^2\}$ for $j = 1, 2, \cdots, k$, which are generated in the first step, n_j values of e_{ij}, for $i = 1, 2, \cdots, n_j$, are generated for every one of the k groups. The e_{ij}'s are normal with mean zero and variance σ_j^2.

2. For each of the k elements of the p vectors of $\overline{X} = [\overline{X}_1, \overline{X}_2, \cdots, \overline{X}_p]$ generated in the first step, n_j vectors of observations were generated for each of $X_{ij} = [X_{ij1}, X_{ij2}, \cdots, X_{ijp}]$, where $i = 1, 2, \cdots, n_j$ and $j = 1, 2, \cdots, k$, from normal distribution with mean $\overline{X}_{jp}$ and a pre-specified value for the variance $\sigma^2_{X_{ijp}}$.

3. For pre-specified values of the parameter vector $\lambda' = [\gamma_\circ, \gamma_1, \cdots, \gamma_q, \beta_1, \beta_2, \cdots, \beta_p]$ the values of the outcome variable Y_{ij} in a particular data set were then generated by substituting the generated values of all the between- and within-group variables in the right hand side of equation 3.1.

In summary, one artificial data set for a model with one between and one within-group variables, for example, can have the following form:

1- $k = 40$ number of groups , $n_j \sim$ Uniform(5,60)

2- $\theta = 0.2$, $\sigma_\circ^2 = 50$.

3- $\sigma_j^2 \sim \Gamma^{-1}(\alpha = 2.5 \ , \ \beta = .008)$, $\alpha = \dfrac{1}{2\theta}$, $\beta = \dfrac{2\theta}{\sigma_\circ^2}$

4- $\tau^2 = Var(U_j) = 9.0$, Thus $U_j \sim N(0, 9.0)$.

5- $\Sigma = Cov(W,\overline{X}) = \begin{pmatrix} 4 & 1 \\ 1 & 8 \end{pmatrix}$, $\sigma^2_{X_{ij}} = 81$, thus $W \sim N(0,4)$,

$\overline{X} \sim N(0,8)$, $X_{ij} \sim N(\overline{X}_j, 81)$, and $e_{ij} \sim N(0, \sigma^2_j)$.

6- $\gamma_\circ = 5.0$, $\gamma_1 = 3.0$, $\beta = 0.5$, where $p = q = 1$

7- $Y_{ij} \sim N(\ 5+3W_j+.5X_{ij}+U_j\ ,\ \sigma^2_j\ )$

Tables 5.2, 5.3 and 5.4 presents the true values of the parameters used to generate the 18 artificial data sets used in this study. Note that the within-group sample sizes n_j is a random sample from uniform (5, 60) for all the artificial data sets.

Table 5.2

The true values of the parameters used to generate the artificial data sets with k = 100

<table>
<tr><th colspan="2">k = 100</th></tr>
<tr><th>θ = 0.2</th><th>θ = 0.02</th></tr>
<tr><td>Model (1)
$\sigma_\circ^2=30 \quad \tau^2=6.25 \quad \lambda=\gamma_\circ=6$</td><td>Model (1)
$\sigma_\circ^2=15 \quad \tau^2=9 \quad \lambda=\gamma_\circ=5$</td></tr>
<tr><td>Model (2)
$\sigma_\circ^2=30 \quad \tau^2=4.00$
$\lambda=\begin{bmatrix}\gamma_\circ=3.00\\ \gamma_1=1.50\\ \beta_1=3.50\end{bmatrix}$
$\Sigma=\begin{bmatrix}4.00 & \\ 0.21 & 3.00\end{bmatrix}$
$\sigma_{x_1}^2 = 400$</td><td>Model (2)
$\sigma_\circ^2=35 \quad \tau^2=6.25$
$\lambda=\begin{bmatrix}\gamma_\circ=12.5\\ \gamma_1=6.00\\ \beta_1=2.50\end{bmatrix}$
$\Sigma=\begin{bmatrix}8.00 & \\ 0.70 & 4.00\end{bmatrix}$
$\sigma_{x_1}^2 = 400$</td></tr>
<tr><td>Model (3)
$\sigma_\circ^2=30 \quad \tau^2=6.25$
$\lambda=\begin{bmatrix}\gamma_\circ=8.00\\ \gamma_1=3.50\\ \gamma_2=2.75\\ \beta_1=1.00\\ \beta_2=3.50\\ \beta_3=0.75\end{bmatrix}$
$\Sigma=\begin{bmatrix}4.00 & & & & \\ 0.21 & 3.00 & & & \\ 0.72 & 0.33 & 7.00 & & \\ 0.62 & 0.21 & 0.20 & 4.00 & \\ 0.30 & 0.20 & 0.21 & 0.11 & 5.00\end{bmatrix}$
$\sigma_{x_1}^2=676 \quad \sigma_{x_2}^2=400 \quad \sigma_{x_3}^2=484$</td><td>Model (3)
$\sigma_\circ^2=30 \quad \tau^2=3.0625$
$\lambda=\begin{bmatrix}\gamma_\circ=8.00\\ \gamma_1=3.50\\ \gamma_2=2.75\\ \beta_1=1.00\\ \beta_2=3.50\\ \beta_3=0.75\end{bmatrix}$
$\Sigma=\begin{bmatrix}3.00 & & & & \\ 0.11 & 1.50 & & & \\ 0.32 & 0.35 & 2.50 & & \\ 0.09 & 0.22 & 0.20 & 3.00 & \\ 0.42 & 0.20 & 0.14 & 0.11 & 1.00\end{bmatrix}$
$\sigma_{x_1}^2=225 \quad \sigma_{x_2}^2=289 \quad \sigma_{x_3}^2=100$</td></tr>
</table>

Σ is the variance-covariance matrix between the W's and $\bar{X}$'s

Table 5.3

The true values of the parameters used to generate the artificial data sets with k = 40

k = 40	
$\theta = 0.2$	$\theta = 0.02$
Model (1) $\sigma_\circ^2=25$ $\tau^2=2.25$ $\lambda=\gamma_\circ=8$	Model (1) $\sigma_\circ^2=15$ $\tau^2=9$ $\lambda=\gamma_\circ=5$
Model (2) $\sigma_\circ^2=40$ $\tau^2=2.25$ $\lambda=\begin{bmatrix}\gamma_\circ=3.00\\ \gamma_1=1.50\\ \beta_1=3.50\end{bmatrix}$ $\Sigma=\begin{bmatrix}4.00 & \\ 0.21 & 3.00\end{bmatrix}$ $\sigma_{x_1}^2 = 169$	Model (2) $\sigma_\circ^2=30$ $\tau^2=9$ $\lambda=\begin{bmatrix}\gamma_\circ=12\\ \gamma_1=4\\ \beta_1=2\end{bmatrix}$ $\Sigma=\begin{bmatrix}6.00 & \\ 0.30 & 4.00\end{bmatrix}$ $\sigma_{x_1}^2 = 400$
Model (3) $\sigma_\circ^2=25$ $\tau^2=2.25$ $\lambda=\begin{bmatrix}\gamma_\circ=8.00\\ \gamma_1=3.50\\ \gamma_2=2.75\\ \beta_1=1.00\\ \beta_2=3.50\\ \beta_3=0.75\end{bmatrix}$ $\Sigma=\begin{bmatrix}6.00 & & & & \\ 0.11 & 3.00 & & & \\ 0.32 & 0.35 & 3.50 & & \\ 0.12 & 0.22 & 0.20 & 4.00 & \\ 0.30 & 0.20 & 0.21 & 0.11 & 5.00\end{bmatrix}$ $\sigma_{x_1}^2=121$ $\sigma_{x_2}^2=169$ $\sigma_{x_3}^2=196$	Model (3) $\sigma_\circ^2=15$ $\tau^2=6.25$ $\lambda=\begin{bmatrix}\gamma_\circ=8.00\\ \gamma_1=3.50\\ \gamma_2=2.75\\ \beta_1=1.00\\ \beta_2=3.50\\ \beta_3=0.75\end{bmatrix}$ $\Sigma=\begin{bmatrix}2.00 & & & & \\ 0.13 & 2.50 & & & \\ 0.12 & 0.15 & 1.50 & & \\ 0.09 & 0.05 & 0.10 & 3.00 & \\ 0.02 & 0.20 & 0.14 & 0.11 & 1.00\end{bmatrix}$ $\sigma_{x_1}^2=64$ $\sigma_{x_2}^2=121$ $\sigma_{x_3}^2=100$

Σ is the variance-covariance matrix between the W's and $\bar{X}$'s

Table 5.4

The true values of the parameters used to generate the artificial data sets with k = 15

k = 15	
θ = 0.2	θ = 0.02
Model (1) $\sigma_\circ^2=10 \quad \tau^2=1 \quad \lambda=\gamma_\circ=5$	Model (1) $\sigma_\circ^2=10 \quad \tau^2=3.0625 \quad \lambda=\gamma_\circ=5$
Model (2) $\sigma_\circ^2=20 \quad \tau^2=16$ $\lambda=\begin{bmatrix}\gamma_\circ=8.00\\ \gamma_1=3.50\\ \beta_1=1.50\end{bmatrix}$ $\Sigma=\begin{bmatrix}3.00 & \\ 0.11 & 2.00\end{bmatrix}$ $\sigma_{x_1}^2 = 49$	Model (2) $\sigma_\circ^2=10 \quad \tau^2=36$ $\lambda=\begin{bmatrix}\gamma_\circ=8.0\\ \gamma_1=3.5\\ \beta_1=1.5\end{bmatrix}$ $\Sigma=\begin{bmatrix}3.00 & \\ 0.11 & 2.00\end{bmatrix}$ $\sigma_{x_1}^2 = 49$
Model (3) $\sigma_\circ^2=10 \quad \tau^2=2.25$ $\lambda=\begin{bmatrix}\gamma_\circ=8.00\\ \gamma_1=3.50\\ \gamma_2=2.75\\ \beta_1=1.00\\ \beta_2=3.50\\ \beta_3=0.75\end{bmatrix}$ $\Sigma=\begin{bmatrix}6.00 & & & & \\ 0.11 & 3.00 & & & \\ 0.32 & 0.35 & 3.50 & & \\ 0.12 & 0.22 & 0.20 & 4.00 & \\ 0.30 & 0.20 & 0.21 & 0.11 & 5.00\end{bmatrix}$ $\sigma_{x_1}^2=121 \quad \sigma_{x_2}^2=169 \quad \sigma_{x_3}^2=196$	Model (3) $\sigma_\circ^2=30 \quad \tau^2=4$ $\lambda=\begin{bmatrix}\gamma_\circ=6.00\\ \gamma_1=4.50\\ \gamma_2=2.50\\ \beta_1=2.00\\ \beta_2=3.50\\ \beta_3=1.75\end{bmatrix}$ $\Sigma=\begin{bmatrix}2.00 & & & & \\ 0.13 & 2.50 & & & \\ 0.12 & 0.15 & 1.50 & & \\ 0.09 & 0.05 & 0.10 & 3.00 & \\ 0.02 & 0.20 & 0.14 & 0.11 & 1.00\end{bmatrix}$ $\sigma_{x_1}^2=25 \quad \sigma_{x_2}^2=49 \quad \sigma_{x_3}^2=16$

Σ is the variance-covariance matrix between the W's and $\bar{X}$'s

Initial Estimates for Gibbs Sampling

We recall from the section titled "Empirical Application of Gibbs Sampling" on page 87, that initial estimates for λ, $\sigma_\circ^2$, $\{U_j\}$, and $\{\sigma_j^2\}$ are required to start the iteration process of Gibbs sampling. Reasonable initial estimates for these parameters could be their empirical Bayes estimates. However, to avoid the dependency in comparing empirical Bays estimates of these parameters with those produced by Gibbs sampling, the formal estimates will not be used as initial estimates for Gibbs sampling. Alternatively, least square estimates are used as initial estimates for these parameters as follows:

1- Using the data Y and the design matrix Z in 3.2, The parameter λ is estimated by $\lambda^{(0)} = \left(\sum_{j=1}^{k} Z_j' Z_j \right)^{-1} \sum_{j=1}^{k} Z_j' Y_j$. The superscript (0) is used with the parameter λ to represents the initial estimate of the parameter.

2- Given the data Y and the computed value of $\lambda^{(0)}$ from the previous step, each element of the set $\{U_j\}$ is estimated by $U_j^{(0)} = \frac{1}{n_j}\left(\sum_{i=1}^{n_j} Y_j - \sum_{i=1}^{n_j} Z_j \lambda^{(0)} \right)$.

3- Similarly, each element of the set $\{\sigma_j^2\}$ is estimated by $\sigma_j^{2\,(0)} = n_j^{-1}(Y_j - Z_j\lambda^{(0)} - U_j^{(0)})'(Y_j - Z_j\lambda^{(0)} - U_j^{(0)})$. The values of $\lambda^{(0)}$ and $U_j^{(0)}$ were computed from the previous two steps.

4- Finally, given the set of estimated values $\{\sigma_j^{2\,(0)}\}$ from step 3, $\sigma_\circ^2$ is estimated by the harmonic mean H of $\{\sigma_j^{2\,(0)}\}$ as follow $\sigma_\circ^{2\,(0)} = k\left(\sum_{j=1}^{k} \sigma_j^{-2\,(0)} \right)^{-1}$.

The above estimates can be thought of as sampled values

of the parameters from the current approximation to their marginal posterior distributions. These sampled values are then used to simulate the distributions for the rest of the parameters. The simulated distributions however, are poor ones because they are based on poor estimates of λ, $\sigma_\circ^2$, $\{U_j\}$, and $\{\sigma_j^2\}$. Going through several iterations should improve the approximations of all the marginal posterior distributions.

Another parameter estimated from the data is θ. While the estimate of this parameter was not needed to start the iteration process of Gibbs sampling, its estimated value, however, is used for empirical confirmation of its Gibbs (posterior mean) estimate. Except for θ, empirical Bayes estimates for all other parameters of the model in 3.1 can be easily found and compared to their counter part estimates from Gibbs sampling. Therefore, an estimate of θ from Gibbs sampling can be compared to a one that is based on the log transformation of the estimated group variances (Raudenbush and Bryk, 1987). That is, when S_j^2 for $j=1,\cdots,k$ is used to estimate the group residual variance with v_j degrees of freedom, the transformation d_j, where

$$d_j = \frac{1}{2}\left[\log(S_j^2) + v_j^{-1}\right] , \tag{5.60}$$

is used to estimate $\delta_j = \frac{1}{2}\log(\sigma_j^2)$. Each d_j is approximately distributed as $N\left[\delta_j, 1/(2v_j^{-1})\right]$, for $j=1,\cdots,k$. Similarly, using the assumption in 3.5, where $\sigma_\circ^2/(\theta\sigma_j^2)$ is distributed as a chi-square variable with θ^{-1} degrees of freedom, δ_j is distributed as $N(\Delta, \frac{\theta}{2})$, where Δ is the average of all δ_j.

A total variance estimate of d_j approximately consists of a sampling variance and a residual parameter variance $Var(d_j|\Delta) = Var(d_j|\delta_j) + Var(\delta_j|\Delta)$, and

$$E\left(\frac{\sum_1^k (d_j - \Delta)^2}{k-1}\right) = \frac{1}{2v_j} + \frac{\theta}{2}. \tag{5.61}$$

When Δ and $1/(2v_j^{-1})$ in 5.67 are approximated by $\hat{\Delta} = k^{-1}\sum_1^k d_j$ and $V = k^{-1}\sum \frac{1}{2v_j}$ respectively, θ can be estimated by

$$\theta^{(0)} = 2\left(\frac{\sum_1^k (d_j - \hat{\Delta})^2}{k-1} - V\right). \tag{5.62}$$

Assessing the Heterogeneity of Variance

The primary interest of this research is to study the analysis of hierarchically structured data when there is evidence of heterogeneity of variance. Therefore, a measure for assessing homogeneity of variance was needed. Based on the results from the previous section, $\delta_j = \frac{1}{2}\log(\sigma_j^2)$ is approximately distributed as $N(\Delta, \frac{\theta}{2})$, where Δ is an average of δ_j for $j=1,\cdots,k$. Therefore, a natural way of assessing the homogeneity of group variances is to test the hypothesis $H_o: \delta_j = \Delta$, $j=1,\cdots,k$ by the statistic

$$2\sum_{j=1}^{k} v_j (d_j - \hat{\Delta})^2 \tag{5.63}$$

which has a large sample chi-square distribution with $k-1$ degrees of freedom (Raudenbush and Bryk, 1987).

Random number generation

Three computer programs, using the FORTRAN language, were written to implement the generation of the data and the iteration process of Gibbs sampling. The first two programs are for creating the artificial data sets in two steps. The third program applies the iterative steps of Gibbs sampling. In this program sufficient statistics are calculated from an input data file (either the artificial data created by the first two programs or the real data) and used to calculate the initial estimates of the parameters λ, $\sigma_\circ^2$, $\{U_j\}$, and $\{\sigma_j^2\}$, the chi-square statistic for testing the hypothesis of homogeneity of variance as well as the moment estimate of θ. The initial estimates of the hyper-parameters are then treated as being sampled from the current approximation of their marginal posterior distributions and they are used to start the iteration process between six subroutines. Each subroutine is written to sample one value of each of the parameters θ, $\sigma_\circ^2$, τ^2, λ and one set of $\{U_j\}$ and $\{\sigma_j^2\}$ for $j=1,\cdots,k$ from its corresponding conditional distribution. The iteration process is then terminated when the number of iterations required for convergence is reached.

The conditional distributions of the parameters involved in the iteration process of Gibbs sampling have three general parametric forms: normal, gamma and inverse gamma. The International Mathematical and Statistical Library (IMSL),

Version 10 is used with the programs to generate random numbers from these distributions. Sampling a random variable from an inverse gamma distribution with certain parameters was established by inverting a random variable that was sampled from gamma distribution with the same parameters.

Only 600 observations are printed out for each of the parameters as a final representation of its marginal posterior distribution. To avoid the dependency between two successive iterations, the 600 observations are drawn from the last 3600 iteration of the program in a systematic order with a 6 iterations jump between selected observations.

Specifying the Criteria for Convergence

Gelfand, Hills, Racine-Poon and Smith (1990) have listed several procedures for checking convergence in Gibbs sampling. One of these procedures, they recommend, is the overlay plotting of the estimated density of the simulated distribution at several points of the iteration process. One can assume convergence when these plots become equivalent. Another procedure is the use of a Quantile-Quantile plot. Two equal size samples, each drawn several iterations away from the other, are ordered and plotted. As the number of iterations before drawing the two samples increases, the plot of the ordered samples moves toward a 45° line as an indication of convergence. The above two procedures were used

in this study for checking the convergence in approximating the marginal posterior distributions. The number of iterations required for convergence for model 3 with $\theta=0.2$ and $k=15$ was used for the subsequent runs.

Real Data Analysis

Description of the variables available in the High School and Beyond data set is given in table 5.5.

Table 5.5

Description of Variables in High School and Beyond data

Variable	Description
	Student Characteristics
MATHACH	Mathematic achievement (outcome).
SES	A measure of socioeconomic status.
MINORITY	An indicator of minority status.
GENDER	An indicator for females.
	School Characteristics
SECTOR	School affiliation, Public VS. Catholic.
MEANSES	Mean of SES.

Previous research (Raudenbush and Bryk, 1987) has demonstrated the use of hierarchical linear model in studying the effect of the organizational characteristics of schools on dispersion in mathematics achievement (MATHACH). They have used the model for estimating the residual dispersion in mathematics achievement for each school using information from the whole sample. The resulting estimates of the residual dispersions are empirical Bayes estimates, which are conditioned on ML estimates of the hyper-parameters of their assumed distribution.

This study goes beyond getting a single estimate of the residual dispersion for each school by approximating not just the marginal distribution for the dispersions but also the marginal distribution for every parameter in the 3.1 model, including the hyper-parameters. Clearly, this will give the researcher more flexibility in doing her/his inferences about any parameter of the model. Table 5.6 presents the different models used with the High School and Beyond data set.

Table 5.6

Models used with the High School and Beyond data set

No	Model
1	MATHACH = BASE + SECTOR + MEANSES + MINORITY + GENDER + SES
2	MATHACH = BASE + SECTOR + GENDER + SES
3	MATHACH = BASE + MEANSES + GENDER + SES
4	MATHACH = BASE + SECTOR + SES
5	MATHACH = BASE + SECTOR + GENDER
6	MATHACH = BASE + MEANSES + SES
7	MATHACH = BASE + SECTOR
8	MATHACH = BASE + SES
9	MATHACH = BASE

Similar to the models in table 5.1, all school and student level variables in any of the models in table 5.6 are assumed to be fixed effects except the intercept, it is assumed random.

CHAPTER 6

Results

This chapter provides a discussion of the results of the empirical application of Gibbs sampling. The discussion covers the results from the artificial data and the High School and Beyond data set. Answers to the questions concerning the objectives of this study which were presented in chapter 1 are provided through this discussion.

Required Number of Iterations for Convergence

Several procedures have been suggested for checking the convergence of the iteration process of Gibbs sampling (Casella & George, 1992; Gelfand, Hills, Racine-Poon and Smith, 1990 and Lewis and Orav, 1989). In this study two graphical criteria were used in deciding about the convergence of the iteration process of the program. The first criterion is the overlay plotting of the density function for each parameter approximated by different samples. These samples were generated from several runs of the program on the same data set with different number of iterations in each run. The second criterion is based on "Quantile-Quantile" (Q-Q) plots. Observations from two different samples based on the same model, but different number of iterations are sorted and plotted against each other in a scatter plot. Using the first criterion, one can assume convergence when the graphs of the density, from different samples, for the marginal posterior distribution of a particular parameter become very close or identical to each other. Using the second criterion, convergence can be detected when the scatter plot of the sorted observations from two samples forms a 45 degree line.

Model 3 from table 5.1 with $\theta=0.2$ and $k=15$ was used as a reference to decide about the appropriate number of iterations required for convergence for all data sets. It is assumed that this model represents the scenario of the largest number of iterations required for convergence because it has the largest number of parameters and the smallest number of groups, $k=15$, with heterogeneous variances.

Using the specified model the program was run 4 times. The total number of iterations in each run were 4600, 8600, 10600 and 15600 iterations respectively. Four samples, one from each run of the program, with 600 observations in each sample were generated. To avoid dependency between consecutive observations within each sample (Casella & George, 1992) the 600 observations were systematically drawn from the last 3600 iterations in each run. Thus, the first sample is made of 600 observations drawn systematically from the last 3600 iterations after 1000 iterations in the first run. The second sample is made of another 600 observations drawn systematically from the last 3600 iterations after 5000 iterations in the second run. The third sample is also made of 600 observations drawn systematically from the last 3600 iterations but after 7000 iterations in the third run. And finally the fourth sample is made of another 600 observations drawn systematically form the last 3600 iterations after 12000 iterations in the last run.

Figures 1 and 2 show the overlay graphs of the densities for the marginal posterior distributions of the hyper-parameters in model 3 of table 5.1. Four densities, each based on one sample, were drawn for each parameter. These densities are smoothed by the "kernel" method (Silverman, 1986), which superimposes a univariate nonparametric kernel density estimator. The estimator shows areas where the observations are most concentrated in the sample.
Figures 3 and 4 present two Q-Q plots for each hyper-parameter. One plot is based on the samples generated from the runs of 8600 and 10600 iterations and the other plot is based on the samples generated from the runs of 8600 and 15600 iterations.

Density graphs for the exchangeable parameters $\{U_j\}$ and $\{\sigma_j^2\}$ for $j=1,\cdots,15$ of all the 15 groups in the chosen data set are considered in deciding about the convergence of the iteration process. However, only 4 groups, selected at random, were presented here for illustrative purpose. Figures 5 and 6 provide the overlay graphs of the densities of the marginal posterior distributions and the Q-Q plots for σ_j^2 for

the 4 selected groups. The overlay graphs of the densities and the Q-Q plots for U_j are provided in figures 7 and 8.

In general the overlay graphs show that densities from the four samples are almost identical for most of the parameters. Densities for θ, $\sigma_\circ^2$, τ^2 and β_3 based on the 4600 iterations sample show little divergence from the rest of the samples. Considering the overlay graphs and the Q-Q plots of all parameters, it is decided that 8600 iterations is sufficient to achieve convergence. The Q-Q plots show that observations from the sample based on 8600 iterations form a 45 degree line when they are plotted against the observations of the samples based on 10600 and 15600 iterations. Divergence from the 45 degree line occurs only for small number of observations that fall at the tail of the distribution for every parameter. These observations can be thought of as outliers. They have much lower probabilities of being in the distribution than the majority of the observations.

Figure 6.1

Overlay graphs of the estimated densities of the marginal posterior distributions for the hyper-parameters θ, σ_o^2, τ^2, and γ_o

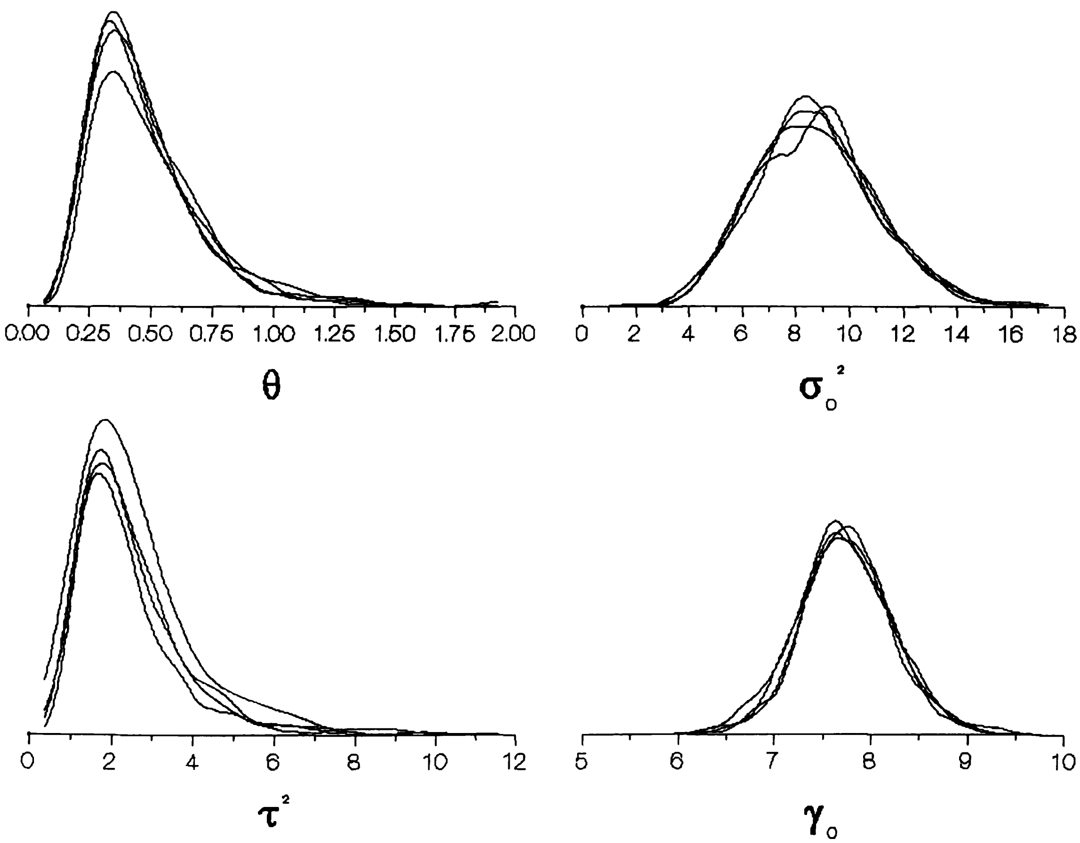


Figure 6.2

Overlay graphs of the estimated densities of the marginal posterior distributions for the hyper-parameters γ_1, β_1, β_2, and β_3

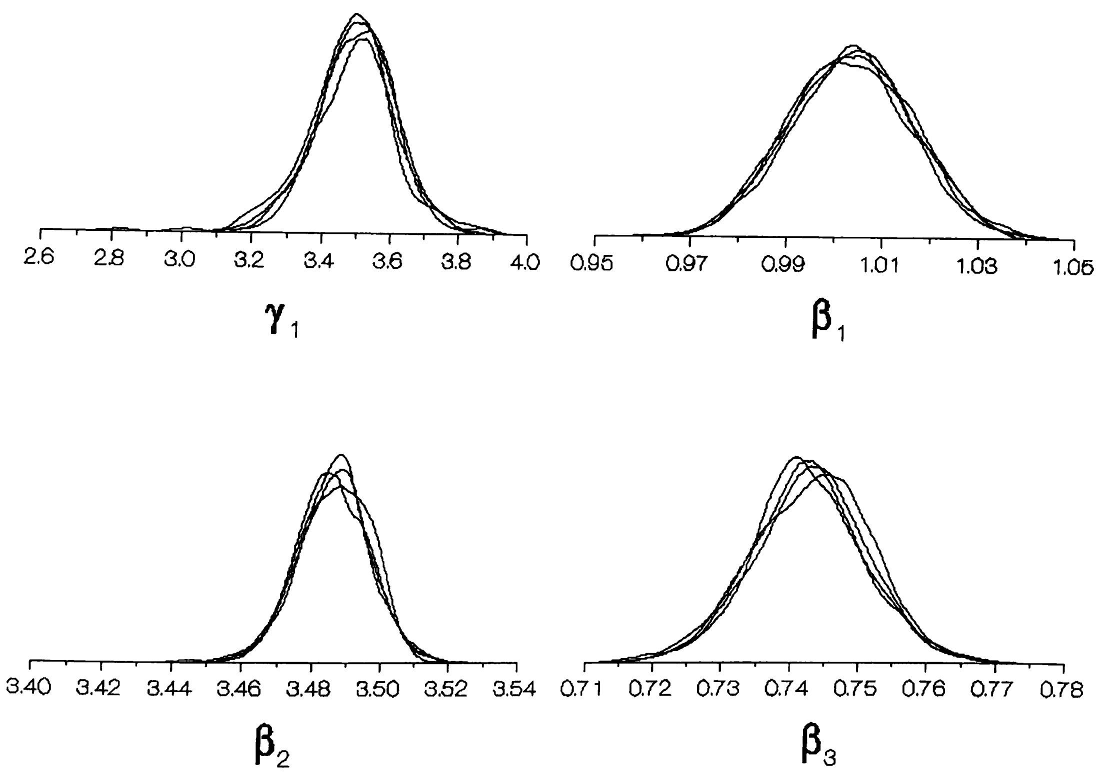


Figure 6.3

Q-Q plots of the observation of the marginal posterior distributions for the hyper-parameters

θ, $\sigma_{\circ}^{2}$, τ^{2}, and $\gamma_{\circ}$

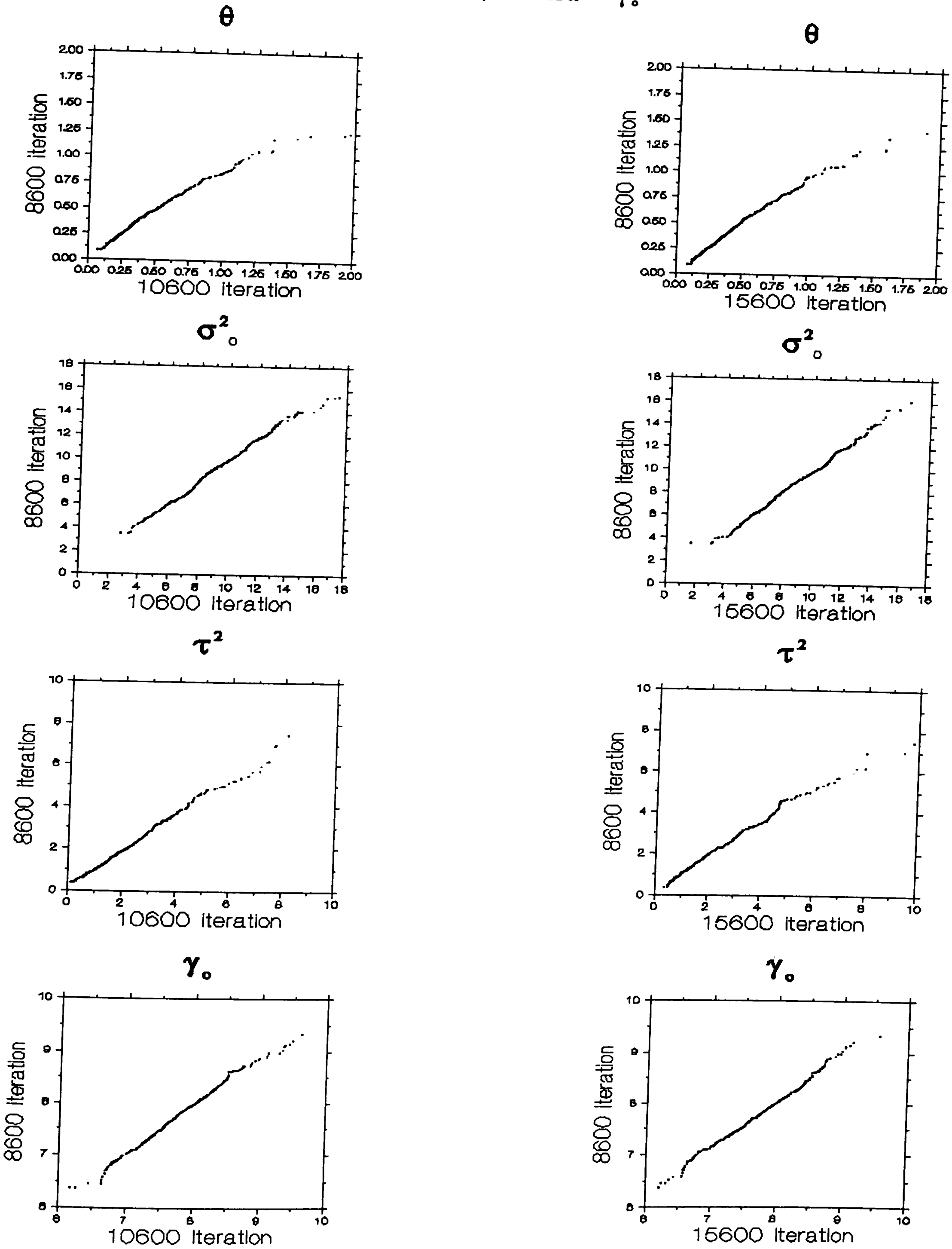


Figure 6.4

Q-Q plots of the observation of the marginal posterior distributions for the hyper-parameters

γ_1, β_1, β_2, and β_3

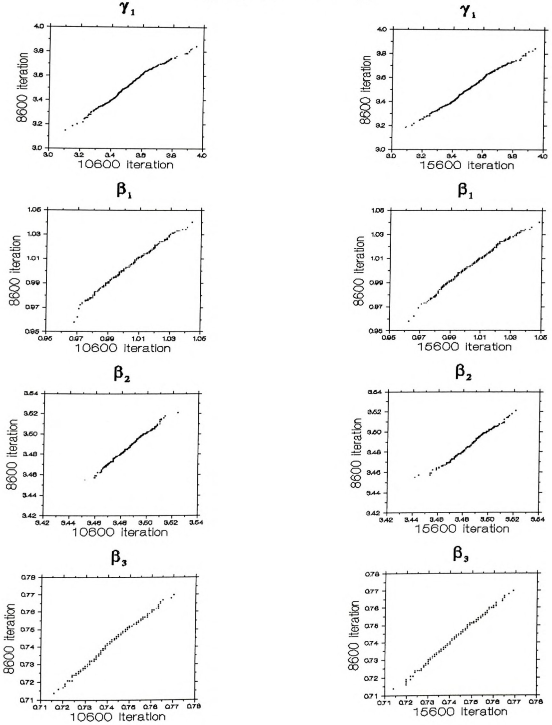


Figure 6.5

Overlay graphs of the estimated densities of the marginal posterior distributions of selected groups' variances $\{\sigma_j^2\}$

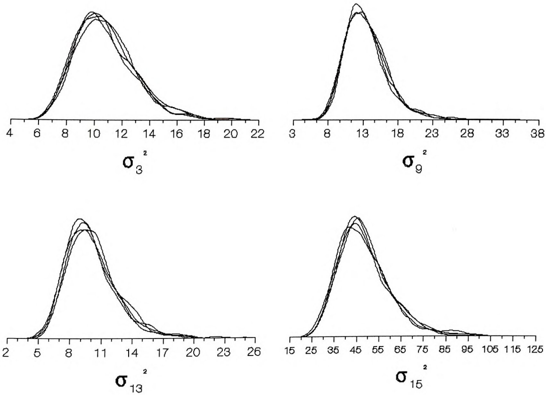


Figure 6.6

Q-Q plots of the points of the marginal posterior distributions of selected groups' variances $\{\sigma_j^2\}$

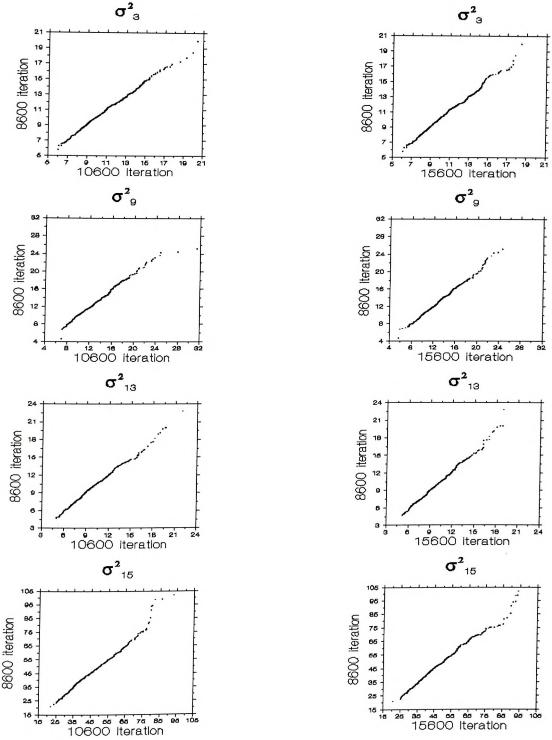


Figure 6.7

Overlay graphs of the estimated densities of the marginal posterior distributions of selected groups' intercept errors $\{U_j\}$

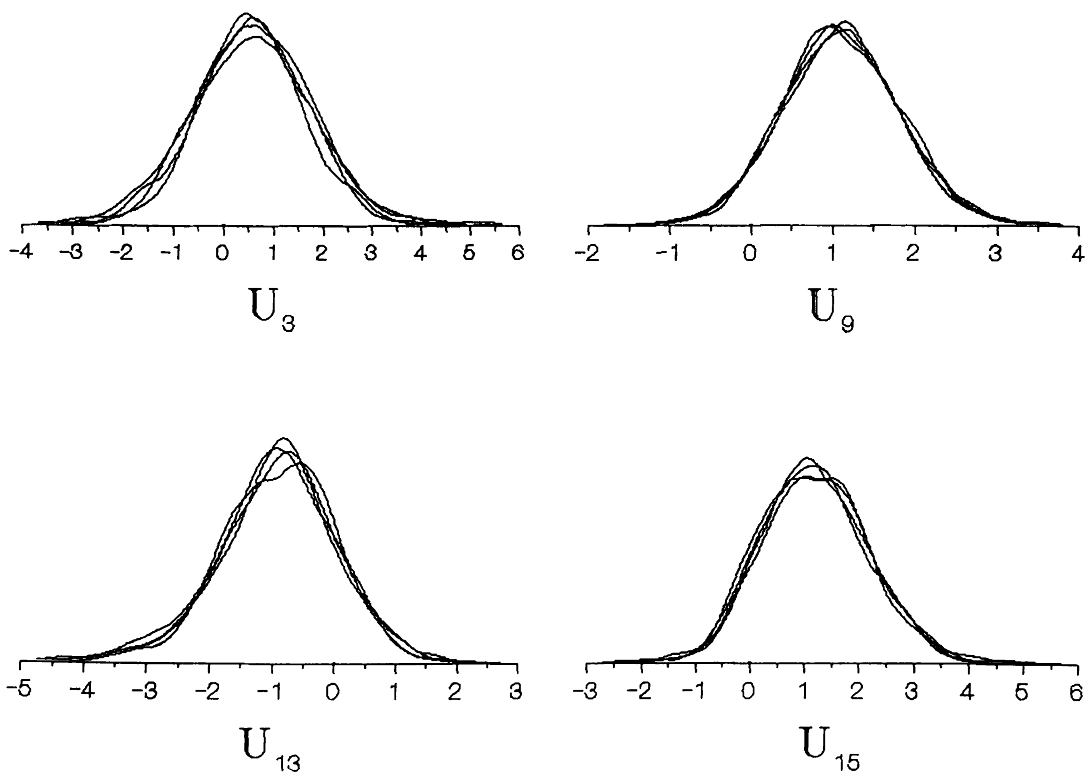
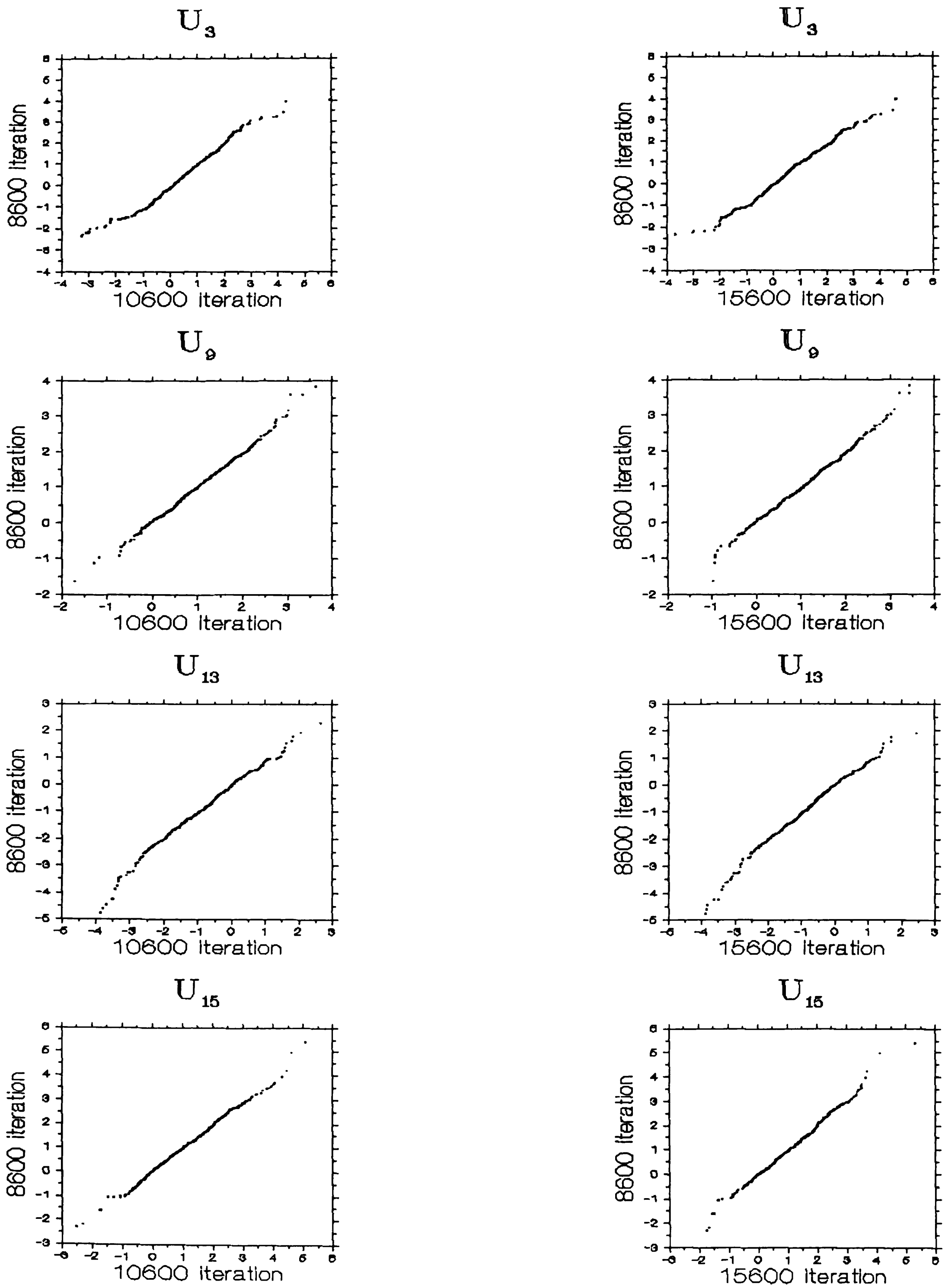


Figure 6.8

Q-Q plots of the points of the marginal posterior distributions of selected groups' intercept errors $\{U_j\}$



Artificial Data

There are 18 data sets (see table 5.1) created by the combination of the three models, the two values chosen for θ, and the three choices of k. The marginal posterior distributions for θ, $\sigma_\circ^2$, τ^2, λ, $\{\sigma_j^2\}$, and $\{U_j\}$ where $j=1,\cdots,k$ in each model are calculated for each data set. Posterior means and standard deviations are calculated from the approximated marginal posterior distributions. Information about the exchangeable parameters $\{\sigma_j^2\}$ and $\{U_j\}$ are presented for only 10 groups selected in systematic sampling for data sets with k=100 or 40. Information about these parameters in data sets with k=15 groups are presented for only 8 groups . That is, for k=100, group1, group11, group21, ..., group91 were presented, for k=40, group1, group5, group9, ..., group37 were presented, and for k=15, group1, group3, group5, ..., group15 were presented.

Empirical Bayes estimates for the regression coefficients (i.e., λ) with their standard error of estimates were obtained for each of the 18 data sets using the HLM program. Maximum likelihood (ML) estimates for the variance components, τ^2 and $\sigma_\circ^2$ were also obtained from the HLM analysis. (Note that the HLM estimate of $\sigma_\circ^2$ represents a pooled within-group variance ML estimate.)

Because of the large cost of analysis, the limited time and resources available, each of the 18 data sets represents

just one sample generated from a population that is different from the other data sets' populations. Therefore, there is a chance that any of the data sets might not be a good representative to its true population and the associated parameters. This has the implication that estimates from a given data set are based on one sample, which limit our assessment of error. A more complete study would involve generating multiple data sets from the same population. Estimates of a given parameter can then be obtained from each generated sample via Gibbs sampling. Given these different estimates, mean squared errors for example, could be calculated for the given parameter.

High School and Beyond data

As pointed earlier, this data is from a probability sample of 160 U.S. high schools. Table 5.5 describes the variables in this data set. The outcome variable used in all of the models is a standardized mathematics achievement score. Nine models (see Table 5.6) were applied to HSB data. Variations between these models are based on the number and type (school vs. student) of variables. Variations in school-level variables are expected to influence mainly the estimated value of τ^2, while variations in student-level variables affect all estimates of variance components. Gibbs sampling was used with 8600 iterations for all of the models. Similar

to the artificial data, information about the exchangeable parameters $\{\sigma_j^2\}$ and $\{U_j\}$ are presented for only 10 schools selected in systematic sampling. Empirical Bayes estimates for the regression coefficients in all models were also obtained via the HLM program.

Comparing Estimates of Variance Components

Part A of Table 6.1 to Table 6.18 provides estimates for the variance components $\sigma_\circ^2$, θ, τ^2 and $\{\sigma_j^2\}$ for each of the 18 artificial data sets. Variance component estimates for the High School and Beyond data set are provided in part A of table 6.19 to table 6.27.

Estimates of $\sigma_\circ^2$

The hyper-parameter $\sigma_\circ^2$ represents the typical value that any of the parameters $\{\sigma_j^2\}$ for $j=1,\cdots,k$ can take in their prior distribution. One objective of this study is to find a way of estimating that typical value without assuming homogeneity of variance. Taking the fully Baysian approach and letting $\sigma_\circ^2$ be random facilitated the derivation of its marginal posterior distribution as well as the marginal posterior distributions for each of the groups' residual variances.

The fact that the marginal posterior distribution of $\sigma_\circ^2$ is not far from being symmetric even when k is small (see

figure 6.1), makes the difference between the posterior mean and the posterior mode estimates insignificant. If the marginal posterior distribution of $\sigma_\circ^2$ is far from being symmetric, it has to be positively skewed, and the posterior mean is larger than the posterior mode. With this in mind, we find that HLM estimates of $\sigma_\circ^2$ are larger than Gibbs posterior mean estimates of $\sigma_\circ^2$ in all data sets including high school and beyond data set. When they are compared to the actual values of $\sigma_\circ^2$ (part A table 1 to table 18) HLM estimates of $\sigma_\circ^2$ are still found to be larger than the actual value of the parameter in all data sets. Furthermore, the over estimation of $\sigma_\circ^2$ in HLM analysis is more pronounced for data sets with extreme heterogeneity of variance (see tables with θ=0.20). This indicates that the HLM estimate of the within-group residual variance $\sigma_\circ^2$ is over estimated when there is clear evidence of heterogeneity of variances.

One obvious implication of over estimating the within-group residual variance is the effect on empirical Bayes estimates of the regression coefficients. As we see in equations 1.5 to 1.7 and under the homogeneity of variance assumption (i.e. $\sigma_j^2=\sigma_\circ^2$, $j=1,\cdots,k$) that larger value of $\sigma_\circ^2$ causes λ_j in equation 1.5 to be smaller therefore giving less weight to $\hat{\beta}_j$ in obtaining β_j^*. Another implication stems from the homogeneity of variance assumption itself, where some of the groups with smaller actual values of their residual variance σ_j^2 are assumed to have an inflated residual variance

which is equal to $\sigma_\circ^2$ because of the homogeneity of variance assumption.

Estimates of θ

The hyper-parameter θ represents a scale parameter of the prior distribution for the exchangeable parameters $\{\sigma_j^2\}$ for $j=1,\cdots,k$. It is inversely proportional to the concentration of $\{\sigma_j^2\}$ around their typical value $\sigma_\circ^2$. Finding the posterior distribution of θ and its estimate provides an answer to the question: "How precise are the posterior mean estimates of the exchangeable parameters $\{\sigma_j^2\}$ in estimating the typical value $\sigma_\circ^2$?" The marginal posterior distribution of θ is found to cover its actual value within 95 percent confidence interval in all of the artificial data sets. In most of the data sets, and especially in the High School and Beyond data set, Gibbs estimates of θ are found to be of the same magnitude of its moment estimates, which are based on the log transformation of the residual variances (see chapter 5 for derivation). Although they are close in their magnitude, Gibbs estimates of θ derived from data sets with k=15 are found to be larger than their moment estimates (see tables with k=15). This finding is expected because when the number of groups is relatively small the marginal posterior distribution of θ becomes more positively skewed which causes the posterior mean (Gibbs estimate) to get larger.

Estimates of τ^2

The hyper-parameter τ^2 represents the variance of the prior distribution of the exchangeable parameters $\{U_j\}$. In general, Gibbs estimates of τ^2 are found to be of the same magnitude of the HLM estimates for both the artificial data sets and the High School and Beyond data set. However, in most of the data sets, Gibbs estimates of τ^2 are found to be slightly higher than HLM estimates. In fact, Gibbs estimates of τ^2 for artificial data sets with small number of groups (k=15) are found to be noticeably higher than HLM estimates. That is because as the number of groups get smaller the marginal posterior distribution τ^2 becomes more positively skewed causing the posterior mean estimate of τ^2 (Gibbs estimate) to be larger than the posterior mode of HLM.

Estimates of $\{\sigma_j^2\}$

One advantage of this study is the ability to obtain the marginal posterior distribution for each of the within-group residual variance $\{\sigma_j^2\}$ for $j=1,\cdots,k$. In all of the artificial data sets the actual values of $\{\sigma_j^2\}$ were found to fall within the 95 percent confidence intervals which were derived from their marginal posterior distributions. The posterior means (Gibbs estimates of $\{\sigma_j^2\}$), the actual values of $\{\sigma_j^2\}$, the standard deviations, the coefficient of variations (c.v.), and

the sample sizes, n_j, for the selected groups in each of the artificial data sets were presented. When compared to their actual values, Gibbs' estimates of $\{\sigma_j^2\}$ were found to have the same magnitude in most of the data sets. We also found that Gibbs estimates of $\{\sigma_j^2\}$ in groups with small sample size have higher c.v. than estimates with larger sample sizes. This indicates that estimates of $\{\sigma_j^2\}$ which are based on small samples are less stable than those based on a large sample sizes.

Comparing Estimates of the Regression Coefficients

Part B of table 6.1 to table 6.18 provide estimates for the regression coefficients in $\lambda' = [\gamma_\circ, \gamma_1, \cdots, \gamma_q, \beta_1, \cdots, \beta_p]$ as well as estimates for $\{U_j\}$ for $j=1, \cdots, k$ in the selected groups for the 18 artificial data sets. Regression coefficients and $\{U_j\}$ estimates for the high school and beyond data set are provided in part B of table 6.19 to table 6.27.

As mentioned earlier, unlike Bayes estimates, which are derived from their marginal posterior distributions via Gibbs sampling, empirical Bayes estimates λ are conditioned on knowing the true values of the variance components $\sigma_\circ^2$ and τ^2. In practice, these components are unknown and need to be estimated. One obvious problem with empirical Bayes estimates is that they do not account for the uncertainty in estimating the variance components. Consequently one can ask: "How do

Gibbs estimates of the regression coefficient differ from their empirical Bayes estimates?" And "How do inferences about regression coefficients change when taking into account the uncertainty about the estimation of variance components, especially when there is a heterogeneity of variance?"

Estimates of λ

The vector λ has two types of regression coefficients. The first includes $\{\gamma_0, \cdots, \gamma_q\}$, for the intercept and between-group variable effects. Variables like SECTOR and MEANSES in the High School and Beyond data are examples of this type. The second includes $\{\beta_1, \cdots, \beta_p\}$, for the within-group variable effects. Variables like MINORITY, GENDER and SES in the High School and Beyond data set are examples of this type. Except for the intercept, between- and within-group regression coefficients are assumed fixed for both the artificial data sets and the High School and Beyond data set.

In general, Gibbs estimates (posterior means) and HLM estimates (posterior Modes) of the regression coefficients are very close to the actual values of the coefficients for the artificial data sets. There are few cases where the two estimates are noticeably different from the actual value of the coefficients, but when compared to each other, Gibbs and HLM estimates are found to be very close in their values. In part B of table 6.8, for example, the actual value for γ_0 is

set to be 3.00, while Gibbs and HLM estimates of that coefficient are 2.598 and 2.609 respectively. Similarly, part B of table 6.17 shows that the actual value for γ_1 is equal to 3.50, while Gibbs and HLM estimates of that coefficient are 4.603 and 4.579 respectively. Large deviations in the estimates of the regression coefficients from their actual values are very few and seem to have a random pattern that could be attributed to the selected random samples. Similar findings are obtained when using high school and beyond data set. Gibbs and HLM estimates for the regression coefficients are found to be of the same magnitude in all the models for High Schools and Beyond data set.

A standard deviation computed from the marginal posterior distribution produced by Gibbs sampling for each regression coefficient is compared to the standard error of the estimate from HLM. In general we find that both the posterior standard deviations of Gibbs sampling (standard error of Gibbs estimates) and the standard error of the estimates from HLM are very close in their values and in some cases they are equal. Since Gibbs estimates of the regression coefficients are close in their values and their standard error of estimates to their counterpart estimates from HLM, inferences (t-test statistics) about these coefficients using Gibbs estimates are not drastically different from the HLM inferences. This is found to be interesting especially for the cases were there is a noticeable heterogeneity of variance

and $\sigma_{\circ}^2$ is being over estimated in HLM analysis.

Estimates of $\{U_j\}$

Based on their marginal posterior distributions, Gibbs estimates (posterior means) of $\{U_j\}$, where $j=1,\cdots,k$ and the standard deviations for the selected groups in each of the artificial data sets are presented with their actual values. Because Gibbs estimates of $\{U_j\}$ are derived from one sample in each data set, qualities of the estimates are difficult to evaluate. However, in most cases we find that these estimates are close in their magnitude and their sign to the actual values of $\{U_j\}$.

Table 6.1-A

Generated data: Model 1, θ = 0.20 and k = 100

Variance components estimates

Hyper-parameters

	θ		σ_o^2	τ^2
GIBBS****	0.242	GIBBS****	29.700	7.079
MOMENT EST.	0.317	HLM	48.512**	7.550
ACTUAL	0.200	ACTUAL	30.000	6.250
GIBBS S.D.*	0.042	GIBBS S.D.*	2.313	1.251
CHI-SQUARE	809.349	D.F. = 99	P-value = 0.000	

Within-groups residual variances

	σ_1^2	σ_{11}^2	σ_{21}^2	σ_{31}^2	σ_{41}^2
GIBBS****	12.144	136.379	29.342	16.480	35.496
ACTUAL	7.018	147.361	23.720	19.950	73.466
GIBBS S.D.*	5.239	26.185	5.600	3.714	20.162
C.V.***	0.431	0.192	0.191	0.225	0.568
n_j	12	55	59	43	9

	σ_{51}^2	σ_{61}^2	σ_{71}^2	σ_{81}^2	σ_{91}^2
GIBBS****	30.851	35.306	30.168	16.077	534.123
ACTUAL	26.693	27.382	30.924	11.768	508.802
GIBBS S.D.*	6.804	16.158	6.200	3.518	202.966
C.V.***	0.221	0.458	0.206	0.219	0.380
n_j	36	9	49	42	14

* Standard deviation of the marginal posterior distribution
** Pooled within-group variance estimate
*** Coefficient of variation
**** Posterior Mean

Table 6.1-B

Generated data: Model 1, θ = 0.20 and k = 100

Regression coefficient and random error U_j estimates

Regression coefficient

	γ_0
GIBBS****	5.978
HLM	5.905
ACTUAL	6.000
GIBBS S.D.*	0.293
HLM S.E.**	0.309

Random error U_j

	U_1	U_{11}	U_{21}	U_{31}	U_{41}
GIBBS****	-0.750	1.399	0.627	-3.044	-3.218
ACTUAL	-0.531	2.102	0.289	-2.077	-3.807
GIBBS S.D.*	0.944	1.298	0.718	0.673	1.637

	U_{51}	U_{61}	U_{71}	U_{81}	U_{91}
GIBBS****	-1.550	0.685	0.405	1.188	-0.793
ACTUAL	-2.935	2.678	-0.795	1.525	-2.614
GIBBS S.D.*	0.917	1.616	0.779	0.647	2.459

* Standard deviation of the marginal posterior distribution
** Standard error of the HLM estimate of the parameter
**** Posterior Mean

Table 6.2-A

Generated data: Model 1, θ = 0.20 and k = 40

Variance components estimates

Hyper-parameters

	θ		$\sigma_{\circ}^2$	τ^2
GIBBS****	0.178	GIBBS****	24.553	3.248
MOMENT EST	0.181	HLM	34.201**	3.524
ACTUAL	0.200	ACTUAL	25.000	2.250
GIBBS S.D.*	0.052	GIBBS S.D.*	2.657	1.124
CHI-SQUARE	229.232	D.F. = 39		P-value = 0.000

Within-groups residual variances

	σ_1^2	σ_5^2	σ_9^2	σ_{13}^2	σ_{17}^2
GIBBS****	23.474	111.119	36.078	32.839	31.858
ACTUAL	37.935	172.278	39.534	29.618	24.233
GIBBS S.D.*	8.440	37.033	8.530	8.794	11.685
C.V.***	0.360	0.333	0.236	0.268	0.367
n_j	12	17	34	29	16

	σ_{21}^2	σ_{25}^2	σ_{29}^2	σ_{33}^2	σ_{37}^2
GIBBS****	19.849	66.245	14.515	16.994	18.957
ACTUAL	15.690	53.687	20.660	12.568	16.566
GIBBS S.D.*	3.647	17.411	3.296	7.929	3.351
C.V.***	0.184	0.263	0.227	0.467	0.177
n_j	59	27	42	10	58

* Standard deviation of the marginal posterior distribution
** Pooled within-group variance estimate
*** Coefficient of variation
**** Posterior Mean

Table 6.2-B

Generated data: Model 1, θ = 0.20 and k = 40

Regression coefficient and random error U_j estimates

Regression coefficient

	γ_0
GIBBS****	7.780
HLM	7.800
ACTUAL	8.000
GIBBS S.D.*	0.349
HLM S.E.**	0.347

Random error U_j

	U_1	U_5	U_9	U_{13}	U_{17}
GIBBS****	-1.804	-1.785	-2.445	3.204	2.490
ACTUAL	-2.259	-2.884	-2.816	3.721	1.602
GIBBS S.D.*	1.143	1.620	0.897	0.948	1.168

	U_{21}	U_{25}	U_{29}	U_{33}	U_{37}
GIBBS****	-1.490	-0.469	1.913	-0.903	-1.351
ACTUAL	-2.116	-0.850	1.459	-0.961	-1.970
GIBBS S.D.*	0.660	1.182	0.646	1.037	0.680

* Standard deviation of the marginal posterior distribution
** Standard error of the HLM estimate of the parameter
**** Posterior Mean

Table 6.3-A

Generated data: Model 1, θ = 0.20 and k = 15

Variance components estimates

Hyper-parameters

	θ		$\sigma_\circ^2$	τ^2
GIBBS****	0.309	GIBBS****	12.755	1.439
MOMENT EST	0.229	HLM	20.544**	1.036
ACTUAL	0.200	ACTUAL	10.000	1.000
GIBBS S.D.*	0.181	GIBBS S.D.*	2.997	0.882
CHI-SQUARE	90.567	D.F. = 14		P-value = 0.000

Within-groups residual variances

	σ_1^2	σ_3^2	σ_5^2	σ_7^2	σ_9^2
GIBBS****	8.226	16.472	6.023	26.187	31.382
ACTUAL	10.489	14.213	4.842	17.763	35.425
GIBBS S.D.*	3.744	3.477	2.494	6.481	7.892
C.V.***	0.455	0.211	0.414	0.247	0.251
n_j	12	42	17	37	34

	σ_{11}^2	σ_{13}^2	σ_{15}^2
GIBBS****	36.008	36.058	17.321
ACTUAL	37.924	35.642	18.583
GIBBS S.D.*	6.959	9.575	4.712
C.V.***	0.193	0.266	0.272
n_j	55	29	29

* Standard deviation of the marginal posterior distribution
** Pooled within-group variance estimate
*** Coefficient of variation
**** Posterior Mean

Table 6.3-B

Generated data: Model 1, θ = 0.20 and k = 15

Regression coefficient and random error U_j estimates

Regression coefficient

	γ_0
GIBBS****	5.129
HLM	5.051
ACTUAL	5.000
GIBBS S.D.*	0.362
HLM S.E.**	0.346

Random error U_j

	U_1	U_3	U_5	U_7	U_9
GIBBS****	1.068	-0.310	-0.854	-0.178	0.125
ACTUAL	2.184	-0.629	-0.526	-0.661	1.001
GIBBS S.D.*	0.722	0.583	0.615	0.699	0.737

	U_{11}	U_{13}	U_{15}
GIBBS****	0.801	-0.281	0.436
ACTUAL	1.491	-1.015	0.205
GIBBS S.D.*	0.713	0.799	0.689

* Standard deviation of the marginal posterior distribution
** Standard error of the HLM estimate of the parameter
**** Posterior Mean

Table 6.4-A

Generated data: Model 1, θ = 0.02 and k = 100

Variance components estimates

Hyper-parameters

	θ		$\sigma_\circ^2$	τ^2
GIBBS****	0.037	GIBBS****	15.147	9.917
MOMENT EST	0.040	HLM	16.162**	9.837
ACTUAL	0.020	ACTUAL	15.000	9.000
GIBBS S.D.*	0.010	GIBBS S.D.*	0.599	1.531
CHI-SQUARE	210.674	D.F. = 99	P-value = 0.000	

Within-groups residual variances

	σ_1^2	σ_{11}^2	σ_{21}^2	σ_{31}^2	σ_{41}^2
GIBBS****	13.155	14.107	17.298	12.285	14.344
ACTUAL	11.426	13.504	14.683	13.404	18.347
GIBBS S.D.*	3.241	2.280	2.721	2.270	3.791
C.V.***	0.246	0.162	0.157	0.185	0.264
n_j	12	55	59	43	9

	σ_{51}^2	σ_{61}^2	σ_{71}^2	σ_{81}^2	σ_{91}^2
GIBBS****	17.973	16.501	18.226	19.889	17.447
ACTUAL	17.665	14.297	20.474	18.957	15.502
GIBBS S.D.*	3.528	4.244	3.030	3.598	4.237
C.V.***	0.196	0.257	0.166	0.181	0.243
n_j	36	9	49	42	14

* Standard deviation of the marginal posterior distribution
** Pooled within-group variance estimate
*** Coefficient of variation
**** Posterior Mean

Table 6.4-B

Generated data: Model 1, θ = 0.02 and k = 100

Regression coefficient and random error U_j estimates

Regression coefficient

	γ_0
GIBBS****	4.986
HLM	4.977
ACTUAL	5.000
GIBBS S.D.*	0.281
HLM S.E.**	0.325

Random error U_j

	U_1	U_{11}	U_{21}	U_{31}	U_{41}
GIBBS****	-0.852	2.395	0.589	-3.340	-4.346
ACTUAL	-0.637	2.522	0.347	-2.493	-4.568
GIBBS S.D.*	1.053	0.584	0.604	0.582	1.140

	U_{51}	U_{61}	U_{71}	U_{81}	U_{91}
GIBBS****	-2.380	1.828	0.001	1.417	-3.302
ACTUAL	-3.522	3.213	-0.954	1.830	-3.137
GIBBS S.D.*	0.745	1.240	0.658	0.705	1.013

* Standard deviation of the marginal posterior distribution
** Standard error of the HLM estimate of the parameter
**** Posterior Mean

Table 6.5-A

Generated data: Model 1, θ = 0.02 and k = 40

Variance components estimates

Hyper-parameters

	θ		$\sigma_\circ^2$	τ^2
GIBBS****	0.026	GIBBS****	15.961	13.485
MOMENT EST	0.020	HLM	16.777**	12.800
ACTUAL	0.020	ACTUAL	15.000	9.000
GIBBS S.D.*	0.014	GIBBS S.D.*	0.934	3.455
CHI-SQUARE	62.049	D.F. = 39	P-value = 0.009	

Within-groups residual variances

	σ_1^2	σ_5^2	σ_9^2	σ_{13}^2	σ_{17}^2
GIBBS****	14.304	14.116	16.287	15.650	19.158
ACTUAL	13.883	11.068	16.891	12.937	20.906
GIBBS S.D.*	3.056	2.850	2.730	2.931	3.976
C.V.***	0.214	0.202	0.168	0.187	0.208
n_j	12	17	34	29	16

	σ_{21}^2	σ_{25}^2	σ_{29}^2	σ_{33}^2	σ_{37}^2
GIBBS****	18.359	21.648	13.340	15.458	17.387
ACTUAL	16.387	22.196	16.551	11.647	16.846
GIBBS S.D.*	2.676	4.415	2.269	3.496	2.558
C.V.***	0.146	0.204	0.170	0.226	0.147
n_j	59	27	42	10	58

* Standard deviation of the marginal posterior distribution
** Pooled within-group variance estimate.
*** Coefficient of variation
**** Posterior Mean

Table 6.5-B

Generated data: Model 1, θ = 0.02 and k = 40

Regression coefficient and random error U_j estimates

Regression coefficient

	γ_0
GIBBS****	4.890
HLM	4.808
ACTUAL	5.000
GIBBS S.D.*	0.589
HLM S.E.**	0.581

Random error U_j

	U_1	U_5	U_9	U_{13}	U_{17}
GIBBS****	-4.450	-5.957	-5.805	7.551	4.774
ACTUAL	-4.518	-5.768	-5.631	7.441	3.205
GIBBS S.D.*	1.218	1.045	0.876	0.873	1.178

	U_{21}	U_{25}	U_{29}	U_{33}	U_{37}
GIBBS****	-3.773	-1.536	3.329	-2.143	-3.455
ACTUAL	-4.232	-1.699	2.919	-1.923	-3.939
GIBBS S.D.*	0.788	1.009	0.801	1.355	0.752

* Standard deviation of the marginal posterior distribution
** Standard error of the HLM estimate of the parameter
**** Posterior Mean

Table 6.6-A

Generated data: Model 1, θ = 0.02 and k = 15

Variance components estimates

Hyper-parameters

	θ		$\sigma_{\circ}^2$	τ^2
GIBBS****	0.050	GIBBS****	11.288	5.771
MOMENT EST	0.039	HLM	12.243**	4.747
ACTUAL	0.020	ACTUAL	10.000	3.063
GIBBS S.D.*	0.051	GIBBS S.D.*	1.253	2.778

CHI-SQUARE 24.469 D.F. = 14 P-value = 0.032

Within-groups residual variances

	σ_1^2	σ_3^2	σ_5^2	σ_7^2	σ_9^2
GIBBS****	9.864	12.575	9.934	14.300	9.744
ACTUAL	9.080	11.087	8.585	10.694	8.563
GIBBS S.D.*	2.724	2.228	2.588	2.800	2.007
C.V.***	0.276	0.177	0.261	0.196	0.206
n_j	12	42	17	37	34

	σ_{11}^2	σ_{13}^2	σ_{15}^2
GIBBS****	14.705	9.318	10.911
ACTUAL	17.018	6.224	10.676
GIBBS S.D.*	2.530	2.120	2.215
C.V.***	0.172	0.228	0.203
n_j	55	29	29

* Standard deviation of the marginal posterior distribution
** Pooled within-group variance estimate
*** Coefficient of variation
**** Posterior Mean

Table 6.6-B

Generated data: Model 1, $\theta = 0.02$ and $k = 15$

Regression coefficient and random error U_j estimates

Regression coefficient

	γ_0
GIBBS****	5.098
HLM	5.092
ACTUAL	5.000
GIBBS S.D.*	0.653
HLM S.E.**	0.590

Random error U_j

	U_1	U_3	U_5	U_7	U_9
GIBBS****	2.915	-0.840	-1.489	-0.815	1.228
ACTUAL	3.822	-1.101	-0.920	-1.157	1.751
GIBBS S.D.*	1.013	0.813	0.981	0.878	0.815

	U_{11}	U_{13}	U_{15}
GIBBS****	2.304	-1.564	0.660
ACTUAL	2.611	-1.776	0.359
GIBBS S.D.*	0.825	0.830	0.878

* Standard deviation of the marginal posterior distribution
** Standard error of the HLM estimate of the parameter
**** Posterior Mean

Table 6.7-A

Generated data: Model 2, θ = 0.20 and k = 100

Variance components estimates

Hyper-parameters

	θ		$\sigma_\circ^2$	τ^2
GIBBS****	0.216	GIBBS****	33.081	4.715
MOMENT EST	0.247	HLM	56.951**	4.116
ACTUAL	0.200	ACTUAL	30.000	4.000
GIBBS S.D.*	0.037	GIBBS S.D.*	2.603	0.922

CHI-SQUARE 863.006 D.F. = 99 P-value = 0.000

Within-groups residual variances

	σ_1^2	σ_{11}^2	σ_{21}^2	σ_{31}^2	σ_{41}^2
GIBBS****	16.329	41.471	24.863	25.736	21.571
ACTUAL	12.933	32.057	24.293	32.472	17.323
GIBBS S.D.*	7.083	8.120	4.858	5.365	11.159
C.V.***	0.434	0.196	0.195	0.208	0.517
n_j	12	55	59	43	9

	σ_{51}^2	σ_{61}^2	σ_{71}^2	σ_{81}^2	σ_{91}^2
GIBBS****	70.313	23.057	414.811	43.554	25.085
ACTUAL	59.509	18.808	341.339	37.248	16.019
GIBBS S.D.*	16.253	12.659	82.279	9.882	10.227
C.V.***	0.231	0.549	0.198	0.227	0.408
n_j	36	9	49	42	14

* Standard deviation of the marginal posterior distribution
** Pooled within-group variance estimate
*** Coefficient of variation
**** Posterior Mean

Table 6.7-B

Generated data: Model 2, θ = 0.20 and k = 100

Regression coefficients and random error U_j estimates

Regression coefficients

	γ_0	γ_1	β_1
GIBBS****	2.880	1.623	3.493
HLM	2.927	1.629	3.493
ACTUAL	3.000	1.500	3.500
GIBBS S.D.*	0.254	0.112	0.004
HLM S.E.**	0.254	0.115	0.005

Random error U_j

	U_1	U_{11}	U_{21}	U_{31}	U_{41}
GIBBS****	-2.281	0.692	3.165	-0.925	3.678
ACTUAL	-2.371	0.245	3.314	-0.418	3.625
GIBBS S.D.*	1.118	0.885	0.667	0.782	1.295

	U_{51}	U_{61}	U_{71}	U_{81}	U_{91}
GIBBS****	1.074	2.349	-0.475	-1.842	0.607
ACTUAL	-0.691	2.290	-1.741	-2.601	1.560
GIBBS S.D.*	1.179	1.277	1.808	0.984	1.109

* Standard deviation of the marginal posterior distribution
** Standard error of the HLM estimate of the parameter
**** Posterior Mean

Table 6.8-A

Generated data: Model 2, θ = 0.20 and k = 40

Variance components estimates

Hyper-parameters

	θ		σ_o^2	τ^2
GIBBS****	0.254	GIBBS****	39.420	1.336
MOMENT EST	0.269	HLM	74.133**	1.041
ACTUAL	0.200	ACTUAL	40.000	2.250
GIBBS S.D.*	0.066	GIBBS S.D.*	4.913	0.636

CHI-SQUARE 319.225 D.F. = 39 P-value = 0.000

Within-groups residual variances

	σ_1^2	σ_5^2	σ_9^2	σ_{13}^2	σ_{17}^2
GIBBS****	27.456	47.465	21.786	38.281	28.705
ACTUAL	35.656	34.510	14.777	24.560	27.952
GIBBS S.D.*	11.696	16.689	5.734	10.126	10.606
C.V.***	0.426	0.352	0.263	0.265	0.369
n_j	12	17	34	29	16

	σ_{21}^2	σ_{25}^2	σ_{29}^2	σ_{33}^2	σ_{37}^2
GIBBS****	75.117	31.557	31.286	51.772	80.495
ACTUAL	76.792	52.517	39.124	85.585	90.053
GIBBS S.D.*	13.517	8.694	7.126	25.840	14.678
C.V.***	0.180	0.275	0.228	0.499	0.182
n_j	59	27	42	10	58

* Standard deviation of the marginal posterior distribution
** Pooled within-group variance estimate
*** Coefficient of variation
**** Posterior Mean

Table 6.8-B

Generated data: Model 2, $\theta = 0.20$ and $k = 40$

Regression coefficients and random error U_j estimates

Regression coefficients

	γ_0	γ_1	β_1
GIBBS****	2.598	1.412	3.489
HLM	2.609	1.575	3.478
ACTUAL	3.000	1.500	3.500
GIBBS S.D.*	0.260	0.154	0.009
HLM S.E.**	0.306	0.167	0.010

Random error U_j

	U_1	U_5	U_9	U_{13}	U_{17}
GIBBS****	-0.792	-0.311	1.713	0.274	0.077
ACTUAL	-1.729	-0.517	2.001	0.608	0.416
GIBBS S.D.*	0.946	0.933	0.778	0.834	0.849

	U_{21}	U_{25}	U_{29}	U_{33}	U_{37}
GIBBS****	1.292	0.502	-0.286	-0.808	0.101
ACTUAL	2.023	1.551	-1.673	-1.383	-1.162
GIBBS S.D.*	0.818	0.827	0.695	1.084	0.841

* Standard deviation of the marginal posterior distribution
** Standard error of the HLM estimate of the parameter
**** Posterior Mean

Table 6.9-A

Generated data: Model 2, θ = 0.20 and k = 15

Variance components estimates

Hyper-parameters

	θ		$\sigma_{\circ}^2$	τ^2
GIBBS****	0.206	GIBBS****	20.978	26.121
MOMENT EST	0.136	HLM	32.276**	21.553
ACTUAL	0.200	ACTUAL	20.000	16.000
GIBBS S.D.*	0.112	GIBBS S.D.*	3.699	12.460

CHI-SQUARE 94.728 D.F. = 14 P-value = 0.000

Within-groups residual variances

	σ_1^2	σ_3^2	σ_5^2	σ_7^2	σ_9^2
GIBBS****	19.828	14.153	32.483	18.957	18.681
ACTUAL	34.028	9.497	25.433	14.237	14.756
GIBBS S.D.*	8.519	3.090	11.107	4.483	4.261
C.V.***	0.430	0.218	0.342	0.237	0.228
n_j	12	42	17	37	34

	σ_{11}^2	σ_{13}^2	σ_{15}^2
GIBBS****	90.083	38.498	15.047
ACTUAL	75.116	27.346	13.960
GIBBS S.D.*	17.875	9.963	4.210
C.V.***	0.198	0.259	0.280
n_j	55	29	29

* Standard deviation of the marginal posterior distribution
** Pooled within-group variance estimate
*** Coefficient of variation
**** Posterior Mean

Table 6.9-B

Generated data: Model 2, θ = 0.20 and k = 15

Regression coefficients and random error U_j estimates

Regression coefficients

	γ_0	γ_1	β_1
GIBBS****	8.360	3.471	1.489
HLM	8.469	3.456	1.496
ACTUAL	8.000	3.500	1.500
GIBBS S.D.*	1.353	0.565	0.032
HLM S.E.**	1.314	0.498	0.037

Random error U_j

	U_1	U_3	U_5	U_7	U_9
GIBBS****	1.199	-2.392	4.201	-4.396	3.303
ACTUAL	2.455	-2.504	5.519	-4.095	3.339
GIBBS S.D.*	2.744	3.532	2.052	1.553	1.535

	U_{11}	U_{13}	U_{15}
GIBBS****	-0.561	2.127	-0.412
ACTUAL	-0.809	2.790	0.433
GIBBS S.D.*	1.952	1.866	1.787

* Standard deviation of the marginal posterior distribution
** Standard error of the HLM estimate of the parameter
**** Posterior Mean

Table 6.10-A

Generated data: Model 2, θ = 0.02 and *k* = 100

Variance components estimates

Hyper-parameters

	θ		σ_o^2	τ^2
GIBBS****	0.017	GIBBS****	35.539	7.145
MOMENT EST	0.011	HLM	36.606**	7.018
ACTUAL	0.020	ACTUAL	35.000	6.250
GIBBS S.D.*	0.007	GIBBS S.D.*	1.238	1.265

CHI-SQUARE 147.231 D.F. = 99 P-value = 0.001

Within-groups residual variances

	σ_1^2	σ_{11}^2	σ_{21}^2	σ_{31}^2	σ_{41}^2
GIBBS****	35.437	43.223	35.010	32.267	35.255
ACTUAL	64.585	39.928	34.088	34.930	29.343
GIBBS S.D.*	5.708	6.309	4.788	4.625	6.840
C.V.***	0.161	0.146	0.137	0.143	0.194
n_j	12	55	59	43	9

	σ_{51}^2	σ_{61}^2	σ_{71}^2	σ_{81}^2	σ_{91}^2
GIBBS****	38.987	34.889	36.209	35.478	40.695
ACTUAL	36.968	43.337	26.444	30.017	49.005
GIBBS S.D.*	5.997	6.333	5.013	4.751	7.505
C.V.***	0.154	0.182	0.138	0.134	0.184
n_j	36	9	49	42	14

* Standard deviation of the marginal posterior distribution
** Pooled within-group variance estimate
*** Coefficient of variation
**** Posterior Mean

Table 6.10-B

Generated data: Model 2, θ = 0.02 and k = 100

Regression coefficients and random error U_j estimates

Regression coefficients

	γ_0	γ_1	β_1
GIBBS****	12.402	6.114	2.494
HLM	12.423	6.103	2.492
ACTUAL	12.500	6.000	2.500
GIBBS S.D.*	0.303	0.092	0.005
HLM S.E.**	0.296	0.094	0.005

Random error U_j

	U_1	U_{11}	U_{21}	U_{31}	U_{41}
GIBBS****	-3.010	0.762	3.940	-1.147	4.319
ACTUAL	-2.963	0.306	4.143	-0.523	4.532
GIBBS S.D.*	1.538	0.893	0.824	0.871	1.665

	U_{51}	U_{61}	U_{71}	U_{81}	U_{91}
GIBBS****	0.891	2.876	-1.608	-2.707	0.419
ACTUAL	-0.863	2.863	-2.176	-3.251	1.949
GIBBS S.D.*	1.088	1.577	0.826	0.938	1.484

* Standard deviation of the marginal posterior distribution
** Standard error of the HLM estimate of the parameter
**** Posterior Mean

Table 6.11-A

Generated data: Model 2, θ = 0.02 and k = 40

Variance components estimates

Hyper-parameters

	θ		$\sigma_\circ^2$	τ^2
GIBBS****	0.035	GIBBS****	29.917	7.263
MOMENT EST	0.032	HLM	32.254**	6.721
ACTUAL	0.020	ACTUAL	30.000	9.000
GIBBS S.D.*	0.018	GIBBS S.D.*	1.948	2.043

CHI-SQUARE 77.033 D.F. = 39 P-value = 0.000

Within-groups residual variances

	σ_1^2	σ_5^2	σ_9^2	σ_{13}^2	σ_{17}^2
GIBBS****	26.545	33.325	27.540	41.700	29.654
ACTUAL	25.764	27.759	20.545	35.204	32.972
GIBBS S.D.*	6.314	7.933	5.057	8.774	6.750
C.V.***	0.238	0.238	0.184	0.210	0.228
n_j	12	17	34	29	16

	σ_{21}^2	σ_{25}^2	σ_{29}^2	σ_{33}^2	σ_{37}^2
GIBBS****	26.641	23.280	26.810	27.529	32.862
ACTUAL	24.737	23.238	31.347	28.753	36.224
GIBBS S.D.*	4.310	5.255	4.580	7.098	5.119
C.V.***	0.162	0.226	0.171	0.258	0.156
n_j	59	27	42	10	58

* Standard deviation of the marginal posterior distribution
** Pooled within-group variance estimate
*** Coefficient of variation
**** Posterior Mean

Table 6.11-B

Generated data: Model 2, θ = 0.02 and k = 40

Regression coefficients and random error U_j estimates

Regression coefficients

	γ_0	γ_1	β_1
GIBBS****	11.437	3.925	1.993
HLM	11.437	3.945	1.992
ACTUAL	12.000	4.000	2.000
GIBBS S.D.*	0.446	0.205	0.006
HLM S.E.**	0.453	0.190	0.006

Random error U_j

	U_1	U_5	U_9	U_{13}	U_{17}
GIBBS****	-2.712	-0.781	4.310	1.104	0.671
ACTUAL	-3.458	-1.034	4.003	1.215	0.832
GIBBS S.D.*	1.580	1.268	1.149	1.192	1.311

	U_{21}	U_{25}	U_{29}	U_{33}	U_{37}
GIBBS****	4.498	2.814	-1.766	-2.587	-1.270
ACTUAL	4.046	3.102	-3.346	-2.766	-2.323
GIBBS S.D.*	0.884	1.012	0.952	1.756	0.996

* Standard deviation of the marginal posterior distribution
** Standard error of the HLM estimate of the parameter
**** Posterior Mean

Table 6.12-A

Generated data: Model 2, $\theta = 0.02$ and $k = 15$

Variance components estimates

Hyper-parameters

	θ		$\sigma_\circ^2$	τ^2
GIBBS****	0.022	GIBBS****	12.254	57.949
MOMENT EST	0.007	HLM	12.661**	50.305
ACTUAL	0.020	ACTUAL	10.000	36.000
GIBBS S.D.*	0.023	GIBSS S.D.*	1.190	26.081
CHI-SQUARE	13.851	D.F. = 14	P-value = 0.266	

Within-groups residual variances

	σ_1^2	σ_3^2	σ_5^2	σ_7^2	σ_9^2
GIBBS****	11.264	14.066	12.354	12.040	13.844
ACTUAL	10.470	12.036	8.624	8.755	13.937
GIBBS S.D.*	2.423	2.263	2.194	1.866	2.199
C.V.***	0.215	0.161	0.178	0.155	0.159
n_j	12	42	17	37	34

	σ_{11}^2	σ_{13}^2	σ_{15}^2
GIBBS****	13.817	13.743	11.593
ACTUAL	12.099	10.924	10.002
GIBBS S.D.*	1.990	2.324	1.927
C.V.***	0.144	0.169	0.166
n_j	55	29	29

* Standard deviation of the marginal posterior distribution
** Pooled within-group variance estimate
*** Coefficient of variation
**** Posterior Mean

Table 6.12-B

Generated data: Model 2, θ = 0.02 and k = 15

Regression coefficients and random error U_j estimates

Regression coefficients

	γ_0	γ_1	β_1
GIBBS****	8.485	3.661	1.493
HLM	8.792	3.471	1.493
ACTUAL	8.000	3.500	1.500
GIBBS S.D.*	2.162	0.710	0.025
HLM S.E.**	1.961	0.740	0.024

Random error U_j

	U_1	U_3	U_5	U_7	U_9
GIBBS****	1.899	-5.080	7.603	-6.932	4.577
ACTUAL	3.683	-3.755	8.279	-6.142	5.008
GIBBS S.D.*	3.494	4.444	2.636	2.207	2.178

	U_{11}	U_{13}	U_{15}
GIBBS****	-1.349	3.609	-0.033
ACTUAL	-1.213	4.185	0.649
GIBBS S.D.*	2.482	2.314	2.528

* Standard deviation of the marginal posterior distribution
** Standard error of the HLM estimate of the parameter
**** Posterior Mean

Table 6.13-A

Generated data: Model 3, θ = 0.2 and k = 100

Variance components estimates

Hyper-parameters

	θ		$\sigma_\circ^2$	τ^2
GIBBS****	0.294	GIBBS****	28.457	6.244
MOMENT EST	0.364	HLM	81.360**	5.433
ACTUAL	0.200	ACTUAL	30.000	6.250
GIBBS S.D.*	0.050	GIBBS S.D.*	2.542	1.097

CHI-SQUARE 1368.057 D.F. = 99 P-value = 0.000

Within-groups residual variances

	σ_1^2	σ_{11}^2	σ_{21}^2	σ_{31}^2	σ_{41}^2
GIBBS****	16.973	23.737	11.577	26.024	20.127
ACTUAL	22.335	22.998	16.267	32.926	12.840
GIBBS S.D.*	7.233	4.809	2.224	5.754	9.068
C.V.***	0.426	0.203	0.192	0.221	0.451
n_j	12	55	59	43	9

	σ_{51}^2	σ_{61}^2	σ_{71}^2	σ_{81}^2	σ_{91}^2
GIBBS****	16.712	23.608	35.107	46.319	30.096
ACTUAL	18.524	32.182	26.896	30.733	23.624
GIBBS S.D.*	4.018	11.181	7.283	10.596	11.934
C.V.***	0.240	0.474	0.207	0.229	0.397
n_j	36	9	49	42	14

* Standard deviation of the marginal posterior distribution
** Pooled within-group variance estimate.
*** Coefficient of variation
**** Posterior Mean

Table 6.13-B

Generated data: Model 3, θ = 0.20 and k = 100

Regression coefficients and random error U_j estimates

Regression coefficients

	γ_0	γ_1	γ_2	β_1	β_2	β_3
GIBBS****	7.694	3.769	2.856	1.001	3.491	0.748
HLM	7.896	3.743	2.782	1.003	3.494	0.742
ACTUAL	8.000	3.500	2.750	1.000	3.500	0.750
GIBBS S.D.*	0.307	0.131	0.180	0.003	0.004	0.004
HLM S.E.**	0.299	0.135	0.186	0.003	0.006	0.005

Random error U_j

	U_1	U_{11}	U_{21}	U_{31}	U_{41}
GIBBS****	-2.435	2.122	-2.429	1.271	2.331
ACTUAL	-1.797	0.936	-3.079	0.558	3.604
GIBBS S.D.*	1.189	0.792	0.647	0.829	1.399

	U_{51}	U_{61}	U_{71}	U_{81}	U_{91}
GIBBS****	-1.346	1.175	3.343	3.266	0.095
ACTUAL	-2.548	0.713	1.216	3.795	-0.468
GIBBS S.D.*	0.916	1.383	0.883	1.037	1.355

* Standard deviation of the marginal posterior distribution
** Standard error of the HLM estimate of the parameter
**** Posterior Mean

Table 6.14-A

Generated data: Model 3, θ = 0.20 and k = 40

Variance components estimates

Hyper-parameters

	θ		$\sigma_\circ^2$	τ^2
GIBBS****	0.313	GIBBS****	25.840	1.869
MOMENT EST	0.322	HLM	47.353**	2.340
ACTUAL	0.200	ACTUAL	25.000	2.250
GIBBS S.D.*	0.093	GIBBS S.D.*	3.601	0.940

CHI-SQUARE 405.029 D.F. = 39 P-value = 0.000

Within-groups residual variances

	σ_1^2	σ_5^2	σ_9^2	σ_{13}^2	σ_{17}^2
GIBBS****	14.152	35.146	52.942	16.728	28.825
ACTUAL	16.970	33.372	67.623	17.316	21.081
GIBBS S.D.*	6.476	13.091	13.311	4.566	10.611
C.V.***	0.458	0.372	0.251	0.273	0.368
n_j	12	17	34	29	16

	σ_{21}^2	σ_{25}^2	σ_{29}^2	σ_{33}^2	σ_{37}^2
GIBBS****	26.171	39.628	190.809	11.235	15.297
ACTUAL	40.242	64.957	184.955	12.052	10.423
GIBBS S.D.*	4.754	10.809	43.059	5.450	2.906
C.V.***	0.182	0.273	0.226	0.485	0.190
n_j	59	27	42	10	58

* Standard deviation of the marginal posterior distribution
** Pooled within-group variance estimate
*** Coefficient of variation
**** Posterior Mean

Table 6.14-B

Generated data: Model 3, θ = 0.20 and k = 40

Regression coefficients and random error U_j estimates

Regression coefficients

	γ_0	γ_1	γ_2	β_1	β_2	β_3
GIBBS****	7.977	3.932	2.780	1.002	3.490	0.746
HLM	8.112	3.952	2.831	1.000	3.476	0.750
ACTUAL	8.000	3.500	2.750	1.000	3.500	0.750
GIBBS S.D.*	0.285	0.124	0.136	0.009	0.009	0.007
HLM S.E.**	0.333	0.152	0.166	0.012	0.011	0.009

Random error U_j

	U_1	U_5	U_9	U_{13}	U_{17}
GIBBS****	-1.024	0.357	-1.175	1.889	0.171
ACTUAL	-2.831	0.945	-1.059	2.246	0.568
GIBBS S.D.*	1.019	0.950	0.993	0.826	0.973

	U_{21}	U_{25}	U_{29}	U_{33}	U_{37}
GIBBS****	1.289	0.533	0.453	0.529	-0.310
ACTUAL	1.755	2.368	-0.490	1.539	-0.381
GIBBS S.D.*	0.687	0.934	1.079	1.057	0.770

* Standard deviation of the marginal posterior distribution
** Standard error of the HLM estimate of the parameter
**** Posterior Mean

Table 6.15-A

Generated data: Model 3, θ = 0.20 and k = 15

Variance components estimates

Hyper-parameters

	θ		σ_o^2	τ^2
GIBBS****	0.435	GIBBS****	8.719	2.064
MOMENT EST	0.319	HLM	16.590**	2.061
ACTUAL	0.200	ACTUAL	10.000	2.250
GIBBS S.D.*	0.226	GIBBS S.D.*	2.309	1.209

CHI-SQUARE 106.600 D.F. = 14 P-value = 0.000

Within-groups residual variances

	σ_1^2	σ_3^2	σ_5^2	σ_7^2	σ_9^2
GIBBS****	3.796	10.678	3.858	37.938	12.871
ACTUAL	3.559	6.980	2.339	31.373	16.116
GIBBS S.D.*	1.834	2.302	1.603	9.166	3.378
C.V.***	0.483	0.216	0.416	0.242	0.262
n_j	12	42	17	37	34

	σ_{11}^2	σ_{13}^2	σ_{15}^2
GIBBS****	19.420	9.881	46.783
ACTUAL	19.165	11.071	49.120
GIBBS S.D.*	3.763	2.869	13.431
C.V.***	0.194	0.290	0.287
n_j	55	29	29

* Standard deviation of the marginal posterior distribution
** Pooled within-group variance estimate
*** Coefficient of variation
**** Posterior Mean

Table 6.15-B

Generated data: Model 3, θ = 0.20 and k = 15

Regression coefficients and random error U_j estimates

Regression coefficients

	γ_0	γ_1	γ_2	β_1	β_2	β_3
GIBBS****	7.774	3.512	2.947	1.004	3.488	0.743
HLM	7.566	3.525	2.975	1.015	3.484	0.744
ACTUAL	8.000	3.500	2.750	1.000	3.500	0.750
GIBBS S.D.*	0.490	0.115	0.218	0.013	0.011	0.009
HLM S.E.**	0.503	0.130	0.246	0.015	0.012	0.011

Random error U_j

	U_1	U_3	U_5	U_7	U_9
GIBBS****	0.875	0.486	0.557	-0.320	1.079
ACTUAL	0.149	0.915	0.478	-0.556	2.500
GIBBS S.D.*	1.095	1.140	0.714	0.911	0.755

	U_{11}	U_{13}	U_{15}
GIBBS****	-1.920	-0.911	1.137
ACTUAL	-2.992	-1.568	1.961
GIBBS S.D.*	0.813	1.006	1.040

* Standard deviation of the marginal posterior distribution
** Standard error of the HLM estimate of the parameter
**** Posterior Mean

Table 6.16-A

Generated data: Model 3, θ = 0.02 and k = 100

Variance components estimates

Hyper-parameters

	θ		σ_o^2	τ^2
GIBBS****	0.029	GIBBS****	30.205	3.361
MOMENT EST	0.054	HLM	32.231**	3.290
ACTUAL	0.020	ACTUAL	30.000	3.063
GIBBS S.D.*	0.010	GIBBS S.D.*	1.184	0.683

CHI-SQUARE 206.971 D.F. = 99 P-value = 0.000

Within-groups residual variances

	σ_1^2	σ_{11}^2	σ_{21}^2	σ_{31}^2	σ_{41}^2
GIBBS****	27.289	28.398	22.138	25.187	30.355
ACTUAL	27.242	26.377	26.066	25.761	29.086
GIBBS S.D.*	5.976	4.280	3.510	4.057	6.672
C.V.***	0.219	0.151	0.159	0.161	0.220
n_j	12	55	59	43	9

	σ_{51}^2	σ_{61}^2	σ_{71}^2	σ_{81}^2	σ_{91}^2
GIBBS****	31.380	28.421	32.066	37.254	33.112
ACTUAL	38.593	28.148	25.195	28.179	32.216
GIBBS S.D.*	5.334	6.337	4.800	6.000	6.805
C.V.***	0.170	0.223	0.150	0.161	0.206
n_j	36	9	49	42	14

* Standard deviation of the marginal posterior distribution
** Pooled within group-variance estimate
*** Coefficient of variation
**** Posterior Mean

Table 6.16-B

Generated data: Model 3, θ = 0.02 and k = 100

Regression coefficients and random error U_j estimates

Regression coefficients

	γ_0	γ_1	γ_2	β_1	β_2	β_3
GIBBS****	7.863	3.724	2.855	1.002	3.490	0.744
HLM	7.864	3.722	2.860	1.002	3.492	0.742
ACTUAL	8.000	3.500	2.750	1.000	3.500	0.750
GIBBS S.D.*	0.225	0.116	0.200	0.005	0.005	0.009
HLM S.E.**	0.219	0.113	0.193	0.005	0.005	0.009

Random error U_j

	U_1	U_{11}	U_{21}	U_{31}	U_{41}
GIBBS****	-1.423	1.450	-1.360	0.834	1.150
ACTUAL	-1.258	0.656	-2.155	0.390	2.523
GIBBS S.D.*	1.169	0.713	0.665	0.805	1.252

	U_{51}	U_{61}	U_{71}	U_{81}	U_{91}
GIBBS****	-1.080	0.618	2.432	2.268	0.134
ACTUAL	-1.783	0.499	0.851	2.656	-0.328
GIBBS S.D.*	0.937	1.353	0.805	0.875	1.238

* Standard deviation of the marginal posterior distribution
** Standard error of the HLM estimate of the parameter
**** Posterior Mean

Table 6.17-A

Generated data: Model 3, θ = 0.02 and k = 40

Variance components estimates

Hyper-parameters

	θ		$\sigma_\circ^2$	τ^2
GIBBS****	0.031	GIBBS****	15.104	5.922
MOMENT EST	0.020	HLM	16.084**	5.389
ACTUAL	0.020	ACTUAL	15.000	6.250
GIBBS S.D.*	0.017	GIBBS S.D.*	0.974	1.599

CHI-SQUARE 66.218 D.F. = 39 P-value = 0.003

Within-groups residual variances

	σ_1^2	σ_5^2	σ_9^2	σ_{13}^2	σ_{17}^2
GIBBS****	13.732	15.412	14.665	14.005	16.535
ACTUAL	13.445	14.186	17.284	13.484	13.841
GIBBS S.D.*	3.229	3.482	2.548	2.665	3.459
C.V.***	0.235	0.226	0.174	0.190	0.209
n_j	12	17	34	29	16

	σ_{21}^2	σ_{25}^2	σ_{29}^2	σ_{33}^2	σ_{37}^2
GIBBS****	13.157	12.930	18.131	13.852	18.624
ACTUAL	18.566	14.631	19.086	19.518	14.904
GIBBS S.D.*	2.113	2.569	3.093	3.288	2.892
C.V.***	0.161	0.199	0.171	0.237	0.155
n_j	59	27	42	10	58

* Standard deviation of the marginal posterior distribution
** Pooled within-group variance estimate
*** Coefficient of variation
**** Posterior Mean

Table 6.17-B

Generated data: Model 3, θ = 0.02 and k = 40

Regression coefficients and random error U_j estimates

Regression coefficients

	γ_0	γ_1	γ_2	β_1	β_2	β_3
GIBBS****	8.297	4.603	2.837	0.996	3.484	0.749
HLM	8.319	4.579	2.831	0.999	3.483	0.748
ACTUAL	8.000	3.500	2.750	1.000	3.500	0.750
GIBBS S.D.*	0.422	0.326	0.218	0.014	0.010	0.011
HLM S.E.**	0.397	0.289	0.218	0.014	0.010	0.011

Random error U_j

	U_1	U_5	U_9	U_{13}	U_{17}
GIBBS****	-2.482	1.192	-1.502	3.093	0.660
ACTUAL	-4.718	1.575	-1.765	3.744	0.946
GIBBS S.D.*	1.399	1.032	1.020	0.782	1.098

	U_{21}	U_{25}	U_{29}	U_{33}	U_{37}
GIBBS****	1.594	1.856	-1.293	0.369	-0.901
ACTUAL	2.925	3.947	-0.817	2.566	-0.634
GIBBS S.D.*	0.741	0.845	0.826	1.594	1.028

* Standard deviation of the marginal posterior distribution
** Standard error of the HLM estimate of the parameter
**** Posterior Mean

Table 6.18-A

Generated data: Model 3, θ = 0.02 and k = 15

Variance components estimates

Hyper-parameters

	θ		σ_o^2	τ^2
GIBBS****	0.067	GIBBS****	32.417	3.363
MOMENT EST	0.060	HLM	35.615**	3.104
ACTUAL	0.020	ACTUAL	30.000	4.000
GIBBS S.D.*	0.051	GIBBS S.D.*	4.231	2.253
CHI-SQUARE	29.843	D.F. = 14	P-value = 0.006	

Within-groups residual variances

	σ_1^2	σ_3^2	σ_5^2	σ_7^2	σ_9^2
GIBBS****	25.497	55.196	30.492	32.398	28.137
ACTUAL	20.225	42.615	22.852	25.591	30.396
GIBBS S.D.*	8.123	11.692	8.370	6.365	5.662
C.V.***	0.319	0.212	0.274	0.196	0.201
n_j	12	42	17	37	34

	σ_{11}^2	σ_{13}^2	σ_{15}^2
GIBBS****	33.894	37.273	30.061
ACTUAL	33.129	44.566	27.008
GIBBS S.D.*	5.721	8.680	6.977
C.V.***	0.169	0.233	0.232
n_j	55	29	29

* Standard deviation of the marginal posterior distribution
** Pooled within-group variance estimate
*** Coefficient of variation
**** Posterior Mean

Table 6.18-B

Generated data: Model 3, θ = 0.02 and k = 15

Regression coefficients and random error U_j estimates

Regression coefficients

	γ_0	γ_1	γ_2	β_1	β_2	β_3
GIBBS****	5.487	4.505	2.910	2.040	3.451	1.712
HLM	5.504	4.507	2.933	2.035	3.452	1.705
ACTUAL	6.000	4.500	2.500	2.000	3.500	1.750
GIBBS S.D.*	0.637	0.303	0.370	0.054	0.034	0.065
HLM S.E.**	0.622	0.281	0.346	0.052	0.033	0.067

Random error U_j

	U_1	U_3	U_5	U_7	U_9
GIBBS****	1.023	0.680	0.705	-0.583	1.266
ACTUAL	0.199	1.220	0.638	-0.742	3.333
GIBBS S.D.*	1.506	1.566	1.221	1.124	1.045

	U_{11}	U_{13}	U_{15}
GIBBS****	-2.152	-1.205	1.834
ACTUAL	-3.990	-2.091	2.615
GIBBS S.D.*	1.122	1.290	1.167

* Standard deviation of the marginal posterior distribution
** Standard error of the HLM estimate of the parameter
**** Posterior Mean

Table 6.19-A

HSB data: Model 1: MATHACH = BASE + SECTOR + MEANSES MINORITY + GENDER + SES

Variance components estimates

Hyper-parameters

	θ		$\sigma_\circ^2$	τ^2
GIBBS***	0.007	GIBBS***	35.502	1.824
MOMENT EST	0.010	HLM	35.908**	1.790
GIBBS S.D.*	0.004	GIBBS S.D.*	0.751	0.319
CHI-SQUARE	233.039	D.F. = 159	P-value = 0.000	

Within-groups residual variances

	σ_1^2	σ_{17}^2	σ_{33}^2	σ_{49}^2	σ_{65}^2
GIBBS***	39.826	35.557	35.339	33.639	36.259
GIBBS S.D.*	4.557	3.902	3.749	3.637	3.609
n_j	47	29	47	53	58

	σ_{81}^2	σ_{97}^2	σ_{113}^2	σ_{129}^2	σ_{145}^2
GIBBS***	38.854	36.921	35.065	34.116	35.909
GIBBS S.D.*	3.986	3.927	3.857	3.748	3.998
n_j	66	35	44	49	36

* Standard deviation of the marginal posterior distribution
** Pooled within group-variance estimate
*** Posterior mean

Table 6.19-B

HSB data: Model 1: MATHACH = BASE + SECTOR + MEANSES MINORITY + GENDER + SES

Regression coefficients and random error U_j estimates

Regression coefficients

	BASE γ_0	SECTOR γ_1	MEANSES γ_2	MINORITY β_1	GENDER β_2	SES β_3
GIBBS***	13.310	1.727	2.152	-2.826	-1.245	1.892
HLM	13.316	1.718	2.106	-2.838	-1.244	1.917
GIBBS S.D.*	0.212	0.287	0.357	0.211	0.161	0.115
HLM S.E.**	0.202	0.278	0.356	0.201	0.159	0.108

Random error U_j

	U_1	U_{17}	U_{33}	U_{49}	U_{65}
GIBBS***	-0.604	1.747	0.161	-0.425	-1.038
GIBBS S.D.*	0.732	0.857	0.769	0.687	0.755

	U_{81}	U_{97}	U_{113}	U_{129}	U_{145}
GIBBS***	-0.605	1.148	-1.742	0.230	1.229
GIBBS S.D.*	0.700	0.826	0.776	0.703	0.833

* Standard deviation of the marginal posterior distribution
** Standard error of the HLM estimate of the parameter
*** Posterior mean

Table 6.20-A

HSB data: Model 2: MATHACH = BASE + SECTOR + GENDER + SES

Variance components estimates

Hyper-parameters

	θ		$\sigma_\circ^2$	τ^2
GIBBS***	0.008	GIBBS***	36.402	3.550
MOMENT EST	0.008	HLM	36.818**	3.396
GIBBS S.D.*	0.004	GIBBS S.D.*	0.699	0.515
CHI-SQUARE	219.569	D.F. = 159	P-value = 0.000	

Within-groups residual variances

	σ_1^2	σ_{17}^2	σ_{33}^2	σ_{49}^2	σ_{65}^2
GIBBS***	41.255	36.181	37.185	34.213	36.626
GIBBS S.D.*	4.990	4.247	3.946	3.496	3.931
n_j	47	29	47	53	58

	σ_{81}^2	σ_{97}^2	σ_{113}^2	σ_{129}^2	σ_{145}^2
GIBBS***	39.494	37.160	35.774	34.270	37.458
GIBBS S.D.*	4.030	4.298	3.777	3.416	4.289
n_j	66	35	44	49	36

* Standard deviation of the marginal posterior distribution
** Pooled within group-variance estimate
*** Posterior mean

Table 6.20-B

HSB data: Model 2: MATHACH = BASE + SECTOR + GENDER + SES

Regression coefficients and random error U_j estimates

Regression coefficients

	BASE γ_0	SECTOR γ_1	GENDER β_1	SES β_2
GIBBS***	12.323	2.116	-1.207	2.309
HLM	12.347	2.100	-1.203	2.341
GIBBS S.D.*	0.229	0.336	0.161	0.112
HLM S.E.**	0.237	0.330	0.164	0.105

Random error U_j

	U_1	U_{17}	U_{33}	U_{49}	U_{65}
GIBBS***	-0.678	3.644	0.643	0.842	-3.512
GIBBS S.D.*	0.871	0.981	0.816	0.778	0.742

	U_{81}	U_{97}	U_{113}	U_{129}	U_{145}
GIBBS***	0.468	-0.151	-4.173	1.087	0.954
GIBBS S.D.*	0.717	0.917	0.901	0.802	0.944

* Standard deviation of the marginal posterior distribution
** Standard error of the HLM estimate of the parameter
*** Posterior mean

Table 6.21-A

HSB data: Model 3: MATHACH = BASE + MEANSES + GENDER + SES

Variance components estimates

Hyper-parameters

	θ		σ_o^2	τ^2
GIBBS***	0.008	GIBBS***	36.303	2.580
MOMENT EST	0.011	HLM	36.798**	2.496
GIBBS S.D.*	0.004	GIBBS S.D.*	0.712	0.395
CHI-SQUARE	238.342	D.F. = 159	P-value = 0.000	

Within-groups residual variances

	σ_1^2	σ_{17}^2	σ_{33}^2	σ_{49}^2	σ_{65}^2
GIBBS***	41.358	35.680	36.938	33.793	36.397
GIBBS S.D.*	5.063	4.028	3.858	3.765	3.979
n_j	47	29	47	53	58

	σ_{81}^2	σ_{97}^2	σ_{113}^2	σ_{129}^2	σ_{145}^2
GIBBS***	40.024	37.082	35.367	34.199	37.417
GIBBS S.D.*	4.498	4.074	3.812	3.530	4.204
n_j	66	35	44	49	36

* Standard deviation of the marginal posterior distribution
** Pooled within group-variance estimate
*** Posterior mean

Table 6.21-B

HSB data: Model 3: MATHACH = BASE + MEANSES + GENDER + SES

Regression coefficients and random error U_j estimates

Regression coefficients

	BASE γ_0	MEANSES γ_1	GENDER β_1	SES β_2
GIBBS***	13.278	3.668	-1.185	2.123
HLM	13.280	3.610	-1.183	2.153
GIBBS S.D.*	0.170	0.358	0.167	0.111
HLM S.E.**	0.168	0.368	0.163	0.108

Random error U_j

	U_1	U_{17}	U_{33}	U_{49}	U_{65}
GIBBS***	-0.294	1.026	-0.021	-0.114	-1.339
GIBBS S.D.*	0.869	0.888	0.761	0.717	0.715

	U_{81}	U_{97}	U_{113}	U_{129}	U_{145}
GIBBS***	0.236	0.459	-2.260	1.054	1.001
GIBBS S.D.*	0.737	0.896	0.783	0.759	0.898

* Standard deviation of the marginal posterior distribution
** Standard error of the HLM estimate of the parameter
*** Posterior mean

Table 6.22-A

HSB data: Model 4: MATHACH = BASE + SECTOR + SES

Variance components estimates

Hyper-parameters

	θ		$\sigma_\circ^2$	τ^2
GIBBS***	0.008	GIBBS***	36.564	3.799
MOMENT EST	0.008	HLM	37.037**	3.685
GIBBS S.D.*	0.004	GIBBS S.D.*	0.724	0.564
CHI-SQUARE	220.466	D.F. = 159	P-value = 0.000	

Within-groups residual variances

	σ_1^2	σ_{17}^2	σ_{33}^2	σ_{49}^2	σ_{65}^2
GIBBS***	41.847	36.353	37.309	34.250	36.847
GIBBS S.D.*	5.183	4.103	4.086	3.503	3.844
n_j	47	29	47	53	58

	σ_{81}^2	σ_{97}^2	σ_{113}^2	σ_{129}^2	σ_{145}^2
GIBBS***	39.946	37.522	35.941	34.406	37.989
GIBBS S.D.*	4.159	4.573	3.929	3.513	4.355
n_j	66	35	44	49	36

* Standard deviation of the marginal posterior distribution
** Pooled within group-variance estimate
*** Posterior mean

Table 6.22-B

HSB data: Model 4: MATHACH = BASE + SECTOR + SES

Regression coefficients and random error U_j estimates

Regression coefficients

	BASE γ_0	SECTOR γ_1	SES β_1
GIBBS***	11.703	2.103	2.350
HLM	11.719	2.101	2.375
GIBBS S.D.*	0.222	0.332	0.108
HLM S.E.**	0.228	0.341	0.105

Random error U_j

	U_1	U_{17}	U_{33}	U_{49}	U_{65}
GIBBS***	-0.801	3.555	1.095	0.877	-2.967
GIBBS S.D.*	0.884	0.970	0.818	0.735	0.776

	U_{81}	U_{97}	U_{113}	U_{129}	U_{145}
GIBBS***	0.561	-0.311	-4.171	1.593	0.884
GIBBS S.D.*	0.783	0.919	0.873	0.812	0.934

* Standard deviation of the marginal posterior distribution
** Standard error of the HLM estimate of the parameter
*** Posterior mean

Table 6.23-A

HSB data: Model 5: MATHACH = BASE + SECTOR + GENDER

Variance components estimates

Hyper-parameters

	θ		σ_o^2	τ^2
GIBBS***	0.016	GIBBS***	37.998	6.424
MOMENT EST	0.019	HLM	38.854**	6.242
GIBBS S.D.*	0.005	GIBBS S.D.*	0.851	0.861
CHI-SQUARE	291.573	D.F. = 159	P-value = 0.000	

Within-groups residual variances

	σ_1^2	σ_{17}^2	σ_{33}^2	σ_{49}^2	σ_{65}^2
GIBBS***	46.613	36.517	38.510	32.513	35.701
GIBBS S.D.*	6.900	5.451	5.469	4.703	4.778
n_j	47	29	47	53	58

	σ_{81}^2	σ_{97}^2	σ_{113}^2	σ_{129}^2	σ_{145}^2
GIBBS***	45.759	40.558	35.976	34.322	41.787
GIBBS S.D.*	5.768	6.166	4.827	4.446	5.936
n_j	66	35	44	49	36

* Standard deviation of the marginal posterior distribution
** Pooled within group-variance estimate
*** Posterior mean

Table 6.23-B

HSB data: Model 5: MATHACH = BASE + SECTOR + GENDER

Regression coefficients and random error U_j estimates

Regression coefficients

	BASE γ_0	SECTOR γ_1	GENDER β_1
GIBBS***	12.080	2.807	-1.337
HLM	12.113	2.791	-1.371
GIBBS S.D.*	0.312	0.448	0.171
HLM S.E.**	0.298	0.426	0.171

Random error U_j

	U_1	U_{17}	U_{33}	U_{49}	U_{65}
GIBBS***	-1.388	5.665	1.459	1.940	-4.969
GIBBS S.D.*	0.952	1.068	0.900	0.808	0.835

	U_{81}	U_{97}	U_{113}	U_{129}	U_{145}
GIBBS***	1.023	-0.900	-5.504	1.439	0.659
GIBBS S.D.*	0.879	1.023	0.948	0.828	1.013

* Standard deviation of the marginal posterior distribution
** Standard error of the HLM estimate of the parameter
*** Posterior mean

Table 6.24-A

HSB data: Model 6: MATHACH = BASE + MEANSES + SES

Variance components estimates

Hyper-parameters

	θ		σ_o^2	τ^2
GIBBS***	0.009	GIBBS***	36.533	2.753
MOMENT EST	0.012	HLM	37.019**	2.692
GIBBS S.D.*	0.005	GIBBS S.D.*	0.761	0.400
CHI-SQUARE	240.750	D.F. = 159	P-value = 0.000	

Within-groups residual variances

	σ_1^2	σ_{17}^2	σ_{33}^2	σ_{49}^2	σ_{65}^2
GIBBS***	42.197	35.813	37.292	34.027	36.388
GIBBS S.D.*	4.923	4.028	4.079	3.891	3.901
n_j	47	29	47	53	58

	σ_{81}^2	σ_{97}^2	σ_{113}^2	σ_{129}^2	σ_{145}^2
GIBBS***	40.194	37.526	35.623	34.302	38.145
GIBBS S.D.*	4.301	4.543	4.098	3.796	4.424
n_j	66	35	44	49	36

* Standard deviation of the marginal posterior distribution
** Pooled within group-variance estimate
*** Posterior mean

Table 6.24-B

HSB data: Model 6: MATHACH = BASE + MEANSES + SES

Regression coefficients and random error U_j estimates

Regression coefficients

	BASE γ_0	MEANSES γ_1	SES β_1
GIBBS***	12.651	3.740	2.160
HLM	12.661	3.675	2.191
GIBBS S.D.*	0.149	0.403	0.114
HLM S.E.**	0.149	0.378	0.109

Random error U_j

	U_1	U_{17}	U_{33}	U_{49}	U_{65}
GIBBS***	-0.315	0.957	0.362	-0.159	-0.817
GIBBS S.D.*	0.895	0.872	0.833	0.734	0.726

	U_{81}	U_{97}	U_{113}	U_{129}	U_{145}
GIBBS***	0.200	0.409	-2.202	1.508	1.028
GIBBS S.D.*	0.740	0.870	0.786	0.745	0.885

* Standard deviation of the marginal posterior distribution
** Standard error of the HLM estimate of the parameter
*** Posterior mean

Table 6.25-A

HSB data: Model 7: MATHACH = BASE + SECTOR

Variance components estimates

Hyper-parameters

	θ		$\sigma_\circ^2$	τ^2
GIBBS***	0.017	GIBBS***	38.243	6.871
MOMENT EST	0.020	HLM	39.151**	6.680
GIBBS S.D.*	0.005	GIBBS S.D.*	0.925	0.930
CHI-SQUARE	299.157	D.F. = 159	P-value = 0.000	

Within-groups residual variances

	σ_1^2	σ_{17}^2	σ_{33}^2	σ_{49}^2	σ_{65}^2
GIBBS***	47.627	36.521	38.249	32.763	36.508
GIBBS S.D.*	6.670	5.717	5.185	4.435	5.064
n_j	47	29	47	53	58

	σ_{81}^2	σ_{97}^2	σ_{113}^2	σ_{129}^2	σ_{145}^2
GIBBS***	46.460	40.872	36.236	34.949	43.323
GIBBS S.D.*	5.865	5.905	5.293	4.869	7.007
n_j	66	35	44	49	36

* Standard deviation of the marginal posterior distribution
** Pooled within group-variance estimate
*** Posterior mean

Table 6.25-B

HSB data: Model 7: MATHACH = BASE + SECTOR

Regression coefficients and random error U_j estimates

Regression coefficients

	BASE γ_0	SECTOR γ_1
GIBBS***	11.398	2.790
HLM	11.393	2.805
GIBBS S.D.*	0.294	0.442
HLM S.E.**	0.293	0.439

Random error U_j

	U_1	U_{17}	U_{33}	U_{49}	U_{65}
GIBBS***	-1.455	5.739	2.113	2.037	-4.376
GIBBS S.D.*	1.006	1.105	0.934	0.840	0.830

	U_{81}	U_{97}	U_{113}	U_{129}	U_{145}
GIBBS***	1.115	-1.163	-5.459	2.055	0.568
GIBBS S.D.*	0.877	1.043	0.907	0.847	1.076

* Standard deviation of the marginal posterior distribution
** Standard error of the HLM estimate of the parameter
*** Posterior mean

Table 6.26-A

HSB data: Model 8: MATHACH = BASE + SES

Variance components estimates

Hyper-parameters

	θ		$\sigma_\circ^2$	τ^2
GIBBS***	0.007	GIBBS***	36.621	4.944
MOMENT EST	0.007	HLM	37.034**	4.768
GIBBS S.D.*	0.003	GIBBS S.D.*	0.739	0.713
CHI-SQUARE	214.629	D.F. = 159	P-value = 0.000	

Within-groups residual variances

	σ_1^2	σ_{17}^2	σ_{33}^2	σ_{49}^2	σ_{65}^2
GIBBS***	41.826	36.284	37.445	34.474	36.657
GIBBS S.D.*	4.728	4.176	3.898	3.621	3.474
n_j	47	29	47	53	58

	σ_{81}^2	σ_{97}^2	σ_{113}^2	σ_{129}^2	σ_{145}^2
GIBBS***	39.659	37.365	36.274	34.900	37.725
GIBBS S.D.*	3.791	4.028	3.803	3.673	3.847
n_j	66	35	44	49	36

* Standard deviation of the marginal posterior distribution
** Pooled within group-variance estimate
*** Posterior mean

Table 6.26-B

HSB data: Model 8: MATHACH = BASE + SES

Regression coefficients and random error U_j estimates

Regression coefficients

	BASE γ_0	SES β_1
GIBBS***	12.640	2.358
HLM	12.657	2.390
GIBBS S.D.*	0.189	0.112
HLM S.E.**	0.188	0.106

Random error U_j

	U_1	U_{17}	U_{33}	U_{49}	U_{65}
GIBBS***	-1.600	3.137	2.160	1.972	-2.071
GIBBS S.D.*	0.904	1.007	0.810	0.807	0.755

	U_{81}	U_{97}	U_{113}	U_{129}	U_{145}
GIBBS***	1.586	-1.110	-3.379	2.693	0.163
GIBBS S.D.*	0.757	1.010	0.902	0.809	0.931

* Standard deviation of the marginal posterior distribution
** Standard errorthe HLM estimate of the parameter
*** Posterior mean

Table 6.27-A

HSB data: Model 9: MATHACH = BASE

Variance components estimates

Hyper-parameters

	θ		$\sigma_\circ^2$	τ^2
GIBBS***	0.016	GIBBS***	38.189	8.943
MOMENT EST	0.020	HLM	39.148**	8.615
GIBBS S.D.*	0.005	GIBBS S.D.*	0.868	1.146
CHI-SQUARE	299.157	D.F. = 159	P-value = 0.000	

Within-groups residual variances

	σ_1^2	σ_{17}^2	σ_{33}^2	σ_{49}^2	σ_{65}^2
GIBBS***	47.229	36.496	38.592	32.718	36.002
GIBBS S.D.*	7.184	5.733	5.258	4.268	4.959
n_j	47	29	47	53	58

	σ_{81}^2	σ_{97}^2	σ_{113}^2	σ_{129}^2	σ_{145}^2
GIBBS***	46.327	41.056	36.334	34.998	42.652
GIBBS S.D.*	5.972	6.164	5.137	5.077	6.185
n_j	66	35	44	49	36

* Standard deviation of the marginal posterior distribution
** Pooled within group-variance estimate
*** Posterior mean

Table 6.27-B

HSB data: Model 9: MATHACH = BASE

Regression coefficient and random error U_j estimates

Regression coefficient

	BASE γ_0
GIBBS***	12.638
HLM	12.637
GIBBS S.D.*	0.249
HLM S.E.**	0.244

Random error U_j

	U_1	U_{17}	U_{33}	U_{49}	U_{65}
GIBBS***	-2.625	4.817	3.499	3.509	-3.054
GIBBS S.D.*	0.967	1.062	0.842	0.769	0.810

	U_{81}	U_{97}	U_{113}	U_{129}	U_{145}
GIBBS***	2.568	-2.244	-4.114	3.496	-0.507
GIBBS S.D.*	0.797	1.115	0.943	0.875	1.033

* Standard deviation of the marginal posterior distribution
** Standard error of the HLM estimate of the parameter
*** Posterior mean

CHAPTER 7

Discussion

Discussion of the results is presented in this chapter. It is focused on comparing the application of Bayesian approach via Gibbs sampling in multi-level analysis to the empirical Bayes approach via HLM analysis when there is heterogeneity of variance in the first level. Suggestions and recommendations for future research are also presented in this chapter.

Bayes and Empirical Bayes Estimation

The main purpose of this empirical study was to apply the fully Bayesian approach to the analysis of multi-level data for the cases where the homogeneity of variance assumption can not be granted and when interest of the research is focused on making inferences on some or all of the groups variances. Available empirical Bayes methods for analyzing multi-level data often assume homogeneity of variance and concentrate on obtaining empirical Bayes estimates for the regression coefficients. The question here is: "How do Bayes estimates of variance components and regression coefficients behave when computed via Gibbs sampling?"

In general, the results presented in the previous chapter suggest that there are no substantial differences between the two approaches in the estimation and inferences about the regression coefficients. This finding suggests that HLM estimates of the regression coefficients, particularly for group effects are robust to the violation of homogeneity of variance. However, when it comes to the estimation of the variance components, HLM estimates of $\sigma_\circ^2$ are found to be positively biased especially when there exists clear evidence of heterogeneity of variance. A moderate heterogeneity of variance with large number of groups seems to have little effect on the HLM estimate of $\sigma_\circ^2$. This was demonstrated in all models used on the High School and Beyond data set.

Since empirical Bayes estimates of the regression coefficients are conditioned on estimating the variance components, it was anticipated that the regression coefficient estimates will be affected by the uncertainty in estimating those variance components, especially in the cases when the HLM estimate of $\sigma_\circ^2$ is quite different from its true value. However, when they are compared to the Bayes estimates and the actual values of the regression coefficients they were found to be within the same range of values with about the same random error of estimate. What appears to happen is that larger HLM estimates of $\sigma_\circ^2$ help compensate for the heterogeneity of variance, σ_j^2 when it comes to the estimation of regression coefficients, and their standard errors of estimate.

One advantage of applying the Bayesian approach to the analysis of multi-level data is the ability to obtain the full posterior distribution for each parameter involved in the model under study. A wide rage of statistics describing a particular parameter and its distribution such as mean, mode, percentile points and variance can be obtained from the posterior distribution of the parameter.

Furthermore the ability to obtain the marginal posterior distribution for each of the residual variances and the parameters θ and $\sigma_\circ^2$ of their prior distribution becomes quite important when research interest is focused on the residual variances themselves and their heterogeneity. To assess this heterogeneity, a coefficient of variation (C.V.) for the residual variance can be easily found as $\sqrt{\frac{2\theta}{1-4\theta}}$ (Lindley, 1965). Inferences about a particular group residual variance also can be made using its marginal posterior distribution.

The application of the Bayesian approach to the analysis of multi-level data involved the use of Gibbs sampling. One disadvantage of this procedure is the time and cost involved in the process. This problem is reflected in this study by having only one sample for each of the 18 conditions used in the study. Limiting the analysis to one sample prevented us from carefully studying the characteristics of the parameter estimates using their mean squared errors and statistical tests against their true values.

The idea of getting the marginal distribution of a variable from its conditional distribution on another variable is very appealing. Gibbs sampling and Data augmentation are two approaches which adopt that idea in many classical and Bayesian analyses. However, developments in their applications are limited to certain research problems. General algorithms and software that utilize these approaches are not widely available to practitioners for use in their practices.

Suggestions and Recommendations for Future Research

To investigate the differences between the fully Bayesian approach and the empirical Bayes approach more closely, it is recommended to use the empirical Bayes model that allows for heterogeneity of variance, $\{\sigma_j^2\}$, in a simulation study where many estimates of the same parameters in a given model are derived from several samples. Based on those many estimates of the parameters, a probability statement and mean squared errors can be used to compare the two approaches in parameter estimation.

Simulating several samples for a given model might seem straightforward in studying the hyper-parameters θ, $\sigma_\circ^2$, τ^2 and λ, where their values can be pre-specified for the data generation. However, it becomes more complex in studying the parameters $\{\sigma_j^2\}$ and $\{U_j\}$ where $j=1,\cdots,k$. That is because the true values for both sets of parameters, $\{\sigma_j^2\}$ and $\{U_j\}$ represent two random samples (see "Data creation" section in chapter 5) from their prior distributions with pre-specified values of the hyper-parameters. This means that we have no direct control on the actual true values of $\{\sigma_j^2\}$ and $\{U_j\}$. The only way that one can alter the true values of $\{\sigma_j^2\}$ and $\{U_j\}$ is by either respecifying the values of the hyper-parameters of their prior distributions or respecifying totally different parametric forms for their priors. More specifically, the true values for $\{\sigma_j^2\}$ represent a random sample drawn from an inverse gamma distribution with $\alpha=\frac{1}{2\theta}$ and $\beta=\frac{2\theta}{\sigma_\circ^2}$ (see equation 5.59). To alter these true values we need either to respecify the values of $\sigma_\circ^2$ and θ in the prior distribution of $\{\sigma_j^2\}$ in 5.59, or chose different parametric form for their prior.

Similarly, the true values for $\{U_j\}$ represent a random sample drawn from a normal distribution with mean equal to zero and variance equal to τ^2. To alter these true values we either respecify the value of τ^2 in the prior distribution of $\{U_j\}$ or chose different parametric form of the prior such as t-distribution (see Seltzer, 1993).

Realizations of the outcome variable Y and the predictors in Z are based on the generated values of $\{\sigma_j^2\}$ and $\{U_j\}$ in the above steps and the pre-specified values of λ. Thus, randomness in Y and Z, took place in two steps: in the generation of the true values of $\{\sigma_j^2\}$ and $\{U_j\}$, and in the generation of Y and Z which are based on the generated values of $\{\sigma_j^2\}$, $\{U_j\}$ in the first step and λ. Therefore, the posterior estimates of the hyper-parameters, θ, $\sigma_\circ^2$, τ^2 and λ based on a given sample of Y and Z can only be as "good" (close to the true values of θ, $\sigma_\circ^2$, τ^2 and λ) as those estimates of θ, $\sigma_\circ^2$, τ^2 and λ produced by the parameters, $\{\sigma_j^2\}$ and $\{U_j\}$. Another way of explaining this is that if we define θ^*, $\sigma_\circ^{2*}$, λ^* and τ^{2*} as estimates of θ, $\sigma_\circ^2$, λ and τ^2, which are based only on the true values of $\{\sigma_j^2\}$ and $\{U_j\}$; also we define θ', $\sigma_\circ^{2'}$, λ' and $\tau^{2'}$ as estimates for the same parameters which are based on the parameters $\{\sigma_j^2\}$ and $\{U_j\}$ and the data Y and Z, then the estimates θ', $\sigma_\circ^{2'}$, λ' and $\tau^{2'}$ cannot be as good as θ^*, $\sigma_\circ^{2*}$, λ^* and τ^{2*} in estimating the true values of θ, $\sigma_\circ^2$, λ and τ^2. That is because of the extra randomness in Y and Z added through the generation of $\{\sigma_j^2\}$ and $\{U_j\}$.

When several (say G) samples of $\{\sigma_j^2\}$ and $\{U_j\}$ are being generated for the same model, the chance of misrepresenting the true values of the hyper-parameters by one bad sample will be greatly reduced by the presence of the other samples of $\{\sigma_j^2\}$ and $\{U_j\}$. A mean squared error of estimates and a distribution of the estimates derived from these samples can be used to make inferences about the hyper-parameters.

If only one set of realization of Y and Z is generated for each of the G samples of the true values of $\{\sigma_j^2\}$ and $\{U_j\}$, then the problem of having only one set of estimates of $\{\sigma_j^2\}$

and $\{U_j\}$ will be the same as in the hyper-parameters. That is because for each one of these G samples of $\{\sigma_j^2\}$ and $\{U_j\}$ we only have one set of realizations for the variables Y and Z. Therefore, we can only obtain one set of estimates of $\{\sigma_j^2\}$ and $\{U_j\}$ for each of the G samples of Y and Z.

To overcome this problem we simply generate several (say J) samples of the variables Y and Z, for each of the G sample of the true values of $\{\sigma_j^2\}$ and $\{U_j\}$. This allows us to obtain a distribution of J estimates of $\{\sigma_j^2\}$ and $\{U_j\}$, which can be used to make inferences about the two sets of parameters $\{\sigma_j^2\}$ and $\{U_j\}$. Therefore, the total number of samples that need to be generated to estimate all the parameters in the model is equal to GJ samples.

One of the debatable issues in Bayesian analysis is the choice of the prior distributions (Deely and Lindley, 1981). In this study conjugate priors for the second stage (exchangeable) parameters were normal $(0\ ,\ \tau^2)$ for the $\{U_j\}$, and inverse gamma $\Gamma^{-1}\left(\frac{1}{2\theta}, \frac{2\theta}{\sigma_\circ^2}\right)$ for $\{\sigma_j^2\}$. Priors for the third stage parameters (hyper-parameters) were chosen to be proportional to constants to reflect noninformative priors. It would be extremely useful to know how inferences about all parameters of the model change when different priors are being used in both stages.

In this study, densities of the produced marginal posterior distributions were approximated using the kernel method. These estimates were used graphically in determining the stopping point for the iteration process of Gibbs sampling. Another method of estimating a density of a marginal posterior distribution for a particular parameter is by using equation 4.5 on the mixture of the conditional densities for a particular parameter. Zeger and Karim (1991) recommended the use of this method for getting better estimates for the tail of the distribution. Gelfand, Smith and Lee (1992) argued that using the mixture of the conditional density produces a more accurate representation of the density function than the kernel estimate. It would be very useful to know if overlay graphs of densities produced by (4.5) are more efficient in determining the stopping point for Gibbs sampling. This is important because of the large amount of computer resources required by Gibbs sampling.

This study attempted to model conditions often found in many education research projects where only intercepts from regression models of many groups were allowed to be random in multi-level analysis. Interactions between individual characteristics and group characteristics were not considered in this study. As a result of this layout, the variance-covariance matrix of the regression coefficients T reduced to a scaler τ^2 which represents the variance of the intercept. It is possible that as more elements added to this variance-covariance matrix (i.e., allowing some of the within-group effects to be random), the Bayesian estimates of these effects might then become different from their empirical Bayes estimates. Thus, it would be extremely informative if the analysis in this study were expanded to allow some of the within-group regression coefficients to be random with the possibility of them being related to the group level variables. This will change the scaler τ^2 to a full variance-covariance matrix, $\mathbf{T}$. Investigating such models provides some insight on what conditions the Gibbs sampling is superior and what conditions other statistical procedures are adequate.

Based on the result of this study, investigators who are only interested in the regression coefficients when analyzing data similar to the one in this study are advised to use empirical Bayes procedures, with assumption of homogeneity of variance, and for the cases with large number of groups, k. Note in table 6.9-B that when k=15 and θ=0.20, the HLM standard errors of the γ's are too small. However, when k=100 and θ=0.20, (table 6.13-B), the HLM standard errors of the γ's are nearly perfect. For these cases (large k) empirical Bayes estimates of the regression coefficients and their associated tests are not that different from those obtained by fully Bayesian procedures. Also, computer programs and application software that utilize empirical Bayes procedure are widely available. When there is a clear evidence of heterogeneity of variance and research interest is focused on studying group variances, applying the fully Bayesian procedure might become more informative than empirical Bayes procedure.

References

Aitkin, M., and Longford, N. (1986). Statistical Modeling Issues in School Effectiveness Studies. Journal of the Royal Statistical Society, Series A, 149, (1), 1-43.

Bandura A. (1977). Social learning theory. New Jersey, Englewood Cliffs: PRENTICE HALL.

Bartlett, M. S., and Kendall, D. G. (1946). The Statistical Analysis of Variance-Heterogeneity and the Logarithmic Transformation. Supp/Journal of Royal Statistical Society, 8, 128-138.

Bassiri, D. (1988). Large and Small Sample Properties of Maximum Likelihood Estimates for the Hierarchical Linear Model. Ddoctoral Dissertation. Michigan State University.

Berlin, K. (1984). Statistical Analysis of Mixed Linear Models with Inhomogeneous Variances. COMPSTAT Physica-Verlag, Vienna for IASC, International Association for Statistical Computing. 209-213.

Bretthorst, G. L. (1988). Bayesian Spectrum Analysis and Parameter Estimation. Lecture Notes in Statistics (48). New York, Heidelberg: Springer-Verlag.

Burstein, L. (1980). The Analysis of Multilevel Data in Educational Research and Evaluation. In D.C. Berliner (Ed.), Review of Research in Education, 8. Washington, DC: American Educationa Research Association.

Bloom, B. S. (1984). The 2-Sigma Problem: The Search for Methods of Group Instruction as Effective as One-to-One. Educational Researcher, 13, (6), 4-16.

Box, G. E. P., and Tiao, G. C. (1964). A Bayesian Approach to the Importance of Assumptions Applied to the Comparison of Variances. Biometrika, 51, (1) and (2), 153-167.

Box, G. E. P., and Tiao, G. C. (1973). Bayesian inference in statistical analysis. Reading, Massachusetts: Addison-Wesley.

Braun, H. I., Jones, D. H., Rubin, D. B., and Thayer, D. T. (1983). Empirical Bayes Estimation in the General Linear Model with Data of Deficient Rank. Psychometrika, 48, (2), 171-181.

Bryk, A. S., and Raudenbush, S. W. (1987). Application of Hierarchical Linear Models to Assessing Change. Psychological Bulletin 101, (1), 147-158.

Bryk, A. S., and Raudenbush, S. W. (1988). Heterogeneity of Variance in Experimental Studies: A Challenge to Conventional Interpretations. Psychological Bulletin, 104, (3), 396-404.

Casella, G., and George, E. I. (1992). Explaining the Gibbs Sampler. The American Statistician, 46, (3), 167-174.

Cook, R. D., and Weisberg, S. (1983). Diagnostics for Heteroscedasticity in Regression. Biometrika, 70, (1), 1-10.

Cooley, W. W., Bond, L., and Mao, B. (1981). Analyzing Multilevel Data. In Berk, R. A. (Ed.), Educational evaluation methodology. Baltimore: Johns Hopkins University press.

Cox, D. R., and Solomon, P. J. (1986). Analysis of Vriability with Large Numbers of Small Samples. Biometrika, 73, (3), 543-554.

Cronbach, L. J., and Webb, N. (1975). Between and Within Class Effects in Reported Aptitude-by-Treatment Interaction: Reanalysis of a Study by G. L. Anderson. Journal of Educational Psychology, 6, 717-724.

Deely, J. J., and Lindley, D. V. (1981). Bayes Empirical Bayes. Journal of the American Satistical Association, 76, (376), 833-841.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data Via the EM Algorithm. Royal Statistical Society Journal, Series. B, 39, 1-22.

Efron, B., and Morris, C. (1972). Empirical Bayes on Vector Observations. Biometrika, 59.

Efron, B., and Morris, C. (1973). Stein's Estimation Rule and Its Competitors-An Empirical Bayes Approach. Journal of the American Statistical Association, 68, 117-130.

Efron, B., and Morris, C. (1975). Data analysis Using Stein's estimator and Its Generalizations. Journal of the American Statistical Association, 72, 311-319.

Efron, B., and Morris, C. (1977). Comments. Journal of the American Statistical Association, 72, 91-93.

Fotiu, R. P. (1989). A Comparison of the EM and Data Augmentation Algorithms on Simulated Small Sample Hierarchical Data from Research on Education. Doctoral Dissertation, Michigan State University.

Gelfand, A. E., Hills S. E., Racine-Poon, A., and Smith, A. F. M. (1990). Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling. Journal of the American Statistical Association, 85, (412), 972-985.

Gelfand, A. E., and Smith, A. F. M. (1990). Sampling-Based Approaches to Calculating Marginal Densities. Journal of the American Statistical Association, 85, (410), 398-409.

Gelfand, A. E., Smith, A. F. M., and Lee, T. (1992). Bayesian Analysis of Constrained Parameter and Truncated Data Problems Using Gibbs Sampling. Journal of the American Statistical Association, 87, (418), 523-532.

Goldstein, H. (1986). Multilevel Mixed Linear Model Analysis Using Iterative Generalized Least Squares. Biometrika, 73, (1), 43-56.

Harville, D. A. (1977). Maximum Likelihood Approachs to Variance Component Estimation and to Related Problems. Journal of the American Statistical Association, 72, 320-338.

Hogg, R. V., and Tanis, E. A. (1977). Probability and statistical inference. New York: Macmillan Publishing Co.

Kasim, R. (1986, April). Examining Variances in Hierarchical Models. Paper presented at the annual meeting of the American Educational Research Association, San Fransisco CA.

Knapp, T. R. (1977). The Unit of Analysis Problem in Applications of Simple Correctional Research. Journal of Educational Statistics, 2, (3), 171-186.

Laird, N. M., and Ware, J. H. (1982). Random-Effect Models for Longitudinal Data. Biometrics, 38, 963-974.

Leonard, T. (1975). A Bayesian Approach to the Linear model with unequal variances. Technometrics, 17, (1), 95-102.

Lewis, P. A. W., and Orav, E. J. (1989). Simulation methodology for mtatisticians, operations analysts, and engineers. Volume 1. Pacific Grove, California: Wadsworth & Brooks/Cole, Statistics/Probability Series.

Lindley, D. V. (1965). Introduction to probability and statistics from a Bayesian viewpoint, part 2. Cambridge: Cambridge University Press.

Lindley, D. V. (1971). Foundations of statistical inference (ed. by V. P. Godmabe and D. A. Sprott). The Estimation of Many Parameters. Toronto: Holt, Rinehart, and Winston, 435-455.

Lindley, D. V., and Smith, A. F. M. (1972). Bayesian Estimates for the Linear Model. Journal of the Royal Statistical Society, Series B. 34, 1-41.

Linhart, H. (1965). Approximate Confidence Limits for the Coefficient of Variation of Gamma Distributions. Biometrics, 21, 733-738.

Longford, N. T. (1987). A Fast Scoring Algorithm for Maximum Likelihood Estimation in Unbalanced Mixed Models with Nested Effects. Biometrika, 74, 817-827.

Mason, W. M., Wong, G. Y., and Entwistle, B. (1984). Contextual Analysis through the Multilevel Linear Model. In S. Leinhardt (Ed.). Sociological Methodology, 72-103. San Francisco: Jossey-Bass.

Morris, C. N. (1987). Comment on Tanner and Wong (1987). Journal of the American Statistical Association, 82, 542-543.

Rao, C. R. (1970). Estimation of Heteroscedastic Variances in Linear Models. Journal of the American Statistical Association, 56, 161-172.

Rao, C. R. (1972). Estimation of Variance and Covariance Components in Linear Models. Journal of the American Statistical Association, 67, 112-115.

Raudenbush, S. W. (1984). Application of a Hierarchical Linear Model in Educational Research. Doctoral dissertation. Harvard University, Graduate School of Education.

Raudenbush, S. W. (1988). Educational Applications of Hierarchical Linear Models: A Review. Journal of Educational Statistics. 13, (2), 85-116.

Raudenbush, S. W., and Bryk, A. S. (1985). Empirical Bayes Meta-Analysis. Journal of Educational Statistics, 10, (2), 75-98.

Raudenbush, S. W., and Bryk, A. S. (1986). A Hierarchical Model for Studying School Effects. Sociology of Education, 59, 1-17.

Raudenbush, S. W., and Bryk, A. S. (1987). Examining Correlates of Diversity. Journal of Educational Statistics, 12, (3), 241-269.

Raudenbush, S. W., and Bryk, A. S. (1988). Methodological Advances in Analyzing the Effects of Schools and Classrooms on Student Learning. In E.Z. Rothkopf (Ed.), Review of Research in Education 423-475. Washington, DC: American Educational Research Association.

Robinson, W. S. (1950). Ecological Correlations and the Behavior of Individuals. American Sociological Review, 15, 351-357.

Rosenberg, B. (1973). Linear Regression with Randomly Dispersed Parameters. Biometrika, 60, 61-75.

Rubin, D. B. (1981). Estimation in Parallel Randomized Experiments. Journal of Educational Statistics, 6, (4), 377-400.

Rubin, D. B. (1983). Some Applications of Bayesian Statistics to educational data. The Statistician, 32, 55-68.

Rubin, D. B. (1987). Multiple Imputation for Nonresponse in surveys. New York: Wiley and Sons.

Searle, S. R. (1971). Topics in Variance Component Estimation. Biometrics, 27, 1-76.

Seltzer, M. H. (1988). The Use of Data Augmentation in Fitting Hierarchical Models to Educational Data. Doctoral Dissertation. The University of Chicago.

Seltzer, M. H. (1993). Sensitivity Analysis for Fix Effect in Hierarchical Model: A Gibbs Sampling Approach. Journal of Educational Statistics, 6, (4), 377-400.

Singh, B., and Sedransk, J. (1988). Variance Estimation in Stratified Sampling When Strata Sample Sizes are Small. Sankhya: The Indian Journal of Statistics, 50, Series B, Pt. 3, 382-393.

Smith, A. F. M. (1973). A General Bayesian Linear Model. Journal of the Royal Statistical Society, Series B. 35, 67-75.

Strenio, J. F., Weisberg, H. 1., and Bryk, A. S. (1983).

Empirical Bayes Estimation of Individual Growth-curve Parameters and Their Relationship to Covariates. Biometrics, 39, 71-86.

Tanner M. A. (1993). Tools for statistical inference. Observed data and data augmentation methods. Lecture Notes in Statistics 67. New York, Heidelberg: Springer-Verlag.

Tanner, M. A., and Wong, W. H. (1987). The Calculation of Posterior Distributions by Data Augmentation. Journal of the American Statistical Association, 82, (398), 528-540.

Winer, B. J. (1971). Statistical principles in experimental design. Second Edition. New York: McGraw-Hill Book Company.

Zeger, S. L., and Karim, M.R. (1991). Generalized Linear Moddels with Random Effects; A Gibbs Samplin Approach. Journal of the American Statistical association, 86, (413), 79-86.