Statistica Sinica **9**(1999), 345-359

METHODOLOGY FOR POOLING SUBPOPULATION REGRESSIONS WHEN SAMPLE SIZES ARE SMALL AND THERE IS UNCERTAINTY ABOUT WHICH SUBPOPULATIONS ARE SIMILAR

Richard Evans and J. Sedransk

Menninger Clinic and Case Western Reserve University

Abstract: Inference for parameters associated with small geographical areas or domains of study requires considerable care because the subpopulation sample sizes are usually very small. Since sample survey data are usually clustered, hierarchical models are often appropriate. However, the customary hierarchical models may specify more exchangeability than is warranted. Thus, we propose an alternative model that is more flexible. We consider the case of a set of multiple linear regressions, one for each subpopulation. The objective is to make inference about one or more regression coefficients, $\underline{\beta}_i$. We derive the posterior mean and variance of $\underline{\beta}_i$, and obtain simplified versions of these moments by using reference-type prior distributions. We use a set of numerical examples to contrast our method with the more conventional hierarchical analysis, and to exhibit the large gains in precision that are possible.

Key words and phrases: Hierarchical model, meta analysis.

1. Introduction

Even in large sample surveys such as the U.S. National Health Interview Survey (NHIS), the emphasis is on the provision of national estimates. However, there is also a need for estimates for small geographical areas or domains of study (e.g., age/race/sex/education classes). When sample sizes in such subpopulations are small, it is generally accepted that the customary randomization-based estimates will not be satisfactory for inference for subpopulation parameters. Thus, investigators have proposed alternative estimators based on realistic models. For example, in the NHIS, the principal variables are binary, and Malec, Sedransk, Moriarity, and LeClere (1997) have investigated models of the following type.

Assume that each individual in the population is assigned to one of K mutually exclusive and exhaustive classes based on the individual's socioeconomic/demographic status. Let Y_{ikj} denote a binary random variable for individual j in class k, cluster i where $i = 1, \ldots, L$, $k = 1, \ldots, B$, and $j = 1, \ldots, N_{ik}$. Within cluster i and class k, and conditional on p_{ik}, the Y_{ikj} are assumed to be independent Bernoulli random variables with $Pr(Y_{ikj} = 1 \mid p_{ik}) = p_{ik}$. A column

vector of M covariates, $\underline{X}_k = (X_{k1}, \ldots, X_{kM})'$, is assumed to be the same for each individual j in class k, cluster i. Given $\underline{X}_k$ and a column vector of regression coefficients, $\underline{\beta}_i = (\beta_{i1}, \ldots, \beta_{iM})'$,

$$logit(p_{ik}) = \underline{X}_k' \underline{\beta}_i. \tag{1.1}$$

Conditional on $\underline{\eta}$ and Γ, the $\underline{\beta}_i$ are assumed to be independently distributed with

$$\underline{\beta}_i \sim N(\underline{\eta}, \Gamma) \tag{1.2}$$

where $\underline{\eta}$ is a vector of regression coefficients, and Γ is an $M \times M$ positive definite matrix. Finally, reference prior distributions are assigned to $\underline{\eta}$ and Γ; i.e.,

$$p(\underline{\eta}, \Gamma) \propto \text{constant}. \tag{1.3}$$

While the NHIS is a multistage personal interview survey and the specification in (1.1)-(1.3) is relatively simple, this type of model is often concordant with the observed data (see Malec, Sedransk, Moriarity and LeClere (1997), Section 5.3).

One may wish to make a direct inference about the $\underline{\beta}_i$, or predictive inference for a finite population quantity such as the total, $\sum_{i \in I} \sum_{k \in K} \sum_{j=1}^{N_{ik}} Y_{ikj}$, where I is the collection of clusters that define the small area, K is the collection of classes that define the domain of study, and N_{ik} is the total number of individuals in cluster i, class k. In either case, inference about the $\underline{\beta}_i$ is central (for details about the predictive inference, see Malec, Sedransk, Moriarity, and LeClere (1997)).

The assumption in (1.2) and (1.3) that the $\underline{\beta}_i$ are exchangeable may be questioned. To make inference about a specific $\underline{\beta}_i$, we propose an alternative to (1.2) and (1.3) that provides a way that the *observed data* can be used to determine the weights to be assigned to the estimates from the set of clusters $\{1, \ldots, L\}$. If one uses (1.2) and (1.3) without modification, the amount of pooling is dictated solely by the prior parameters. For example, let $L = 6$ and assume that $(\underline{\beta}_1, \underline{\beta}_2, \underline{\beta}_3)$ and $(\underline{\beta}_4, \underline{\beta}_5, \underline{\beta}_6)$ are two subsets with similar values of the $\underline{\beta}_i$ within each subset and a sharp separation between the two subsets. Using (1.2) and (1.3) may produce a point estimator of $\underline{\beta}_1$ that has a large, inappropriate, contribution from the data from regressions $4, 5$ and 6.

Because of the complexity of (1.1)-(1.3), we investigate a somewhat simpler specification: multiple linear regression.

The remainder of this paper is organized as follows. The notation and model are presented in Section 2, while the methodology for posterior inference is described in Section 3. The results of a small numerical study of the properties of our method are in Section 4. Section 5 has a brief summary.

2. Notation and Model

Given a column vector of parameters, $\underline{\beta}_i = (\beta_{i1}, \ldots, \beta_{im}, \ldots, \beta_{iM})'$, for regressions $i = 1, \ldots, L$ (corresponding to L subpopulations) and a column vector of covariates $\underline{X}_{ij} = (x_{ij1}, \ldots, x_{ijm}, \ldots, x_{ijM})'$, an observation j from regression i is

$$Y_{ij} = \underline{X}'_{ij}\underline{\beta}_i + \epsilon_{ij}, \quad i = 1, \ldots, L, \quad j = 1, \ldots, n_i.$$

Using the approach in Malec and Sedransk (1992), a class of prior distributions for the vectors $\underline{\beta}_i$, $i = 1, \ldots, L$, is given to reflect the beliefs that (a) for each $m = 1, \ldots, M$ there are subsets of $\{\beta_{1m}, \ldots, \beta_{Lm}\}$, such that the β_{im} within each subset are similar, and (b) there is uncertainty about the composition of such subsets of $\{\beta_{1m}, \ldots, \beta_{Lm}\}$. We assume that there is independence from one subset to another. We also postulate that there is independence across the regression parameter index m. First, modelling the relationships among dissimilar regression parameters is difficult. Second, pooling parameters across the index m would give estimates that have no clear interpretation.

Given the vector $\underline{\beta} = (\underline{\beta}_1^{'}, \ldots, \underline{\beta}_L^{'})'$ the sampling distribution for the data $\underline{y}$ is

$$\underline{y} \mid \underline{\beta} \sim N\left(X\underline{\beta}, \Sigma_1\right) \tag{2.1}$$

where

$$\begin{aligned} \underline{y} &= (y_{11}, \ldots, y_{1n_1}, \ldots, y_{Ln_L})', \\ X &= \text{block diagonal}(X_i), \quad X_i' = (\underline{X}_{i1} \vdots \cdots \vdots \underline{X}_{in_i}), \\ \Sigma_1 &= \text{block diagonal}(\sigma_i^2 I_{n_i \times n_i}), \text{ assumed known.} \end{aligned}$$

For the prior distribution, define G as the set of all partitions of the set of experiment labels $B = \{1, \ldots, L\}$, and $G^* = G \times G \times \cdots \times G$, the Cartesian product of M copies of G. An element $\underline{g}$ of G^* is then a vector of M partitions of the set B. The mth element, g_m, of $\underline{g}$ is a partition that dictates the grouping of the set $\{\beta_{1m}, \ldots, \beta_{Lm}\}$. Denote by $S_k(g_m)$ the set of experiment labels in subset k, partition g_m, $k = 1, \ldots, d(g_m)$, where $d(g_m)$ is the total number of subsets in partition g_m. Finally, define $d(\underline{g}) = \sum_m d(g_m)$ as the total number of subsets in partition $\underline{g}$.

To specify the prior distribution for $\underline{\beta}$ first condition on $\underline{g}$. One may represent the desired similarity of the β_{im} by assuming (a) there is independence between the elements of $\underline{\beta}$ belonging to $S_k(g_m)$ and $S_{k'}(g_{m'})$ for $k \neq k'$ or $m \neq m'$, and (b) for sets $\{\beta_{1m}, \ldots, \beta_{Lm}\}$, and conditional on $\underline{\nu}(g_m)$, the elements of the vector $\underline{\beta}$ are independent with

$$\underline{\beta} \mid \underline{\nu}(\underline{g}) \sim N(A(\underline{g})\underline{\nu}(\underline{g}), \Sigma_2(\underline{g})). \tag{2.2}$$

The matrix $A(\underline{g})$ is $ML \times d(\underline{g})$, with the $[(i-1)M+m]$th row corresponding to β_{im}. This row has all 0's except for a 1 in the zth column; if $\beta_{im} \in S_k(g_m)$, $z = (\sum_{i=1}^{m-1} d(g_i)) + k$. Also

$$\begin{aligned}
\underline{\nu}(\underline{g}) &= (\underline{\nu}(g_1), \ldots, \underline{\nu}(g_M))', \;\; \underline{\nu}(g_m) = (\nu_1(g_m), \ldots, \nu_{d(g_m)}(g_m))', \\
\Sigma_2(\underline{g}) &= \text{block diagonal}\Big(\Sigma_2^*(g_m)\Big), \; m = 1, \ldots M, \\
\Sigma_2^*(g_m) &= \text{block diagonal}\Big(\delta_k^2(g_m) I_{|S_k(g_m)| \times |S_k(g_m)|}\Big), \; k = 1, \ldots, d(g_m),
\end{aligned}$$

where $|S_k(g_m)|$ is the size of the set $S_k(g_m)$ and $I_{|S_k(g_m)| \times |S_k(g_m)|}$ is the identity matrix with the specified dimensions. The effect of this specification is that all β_{im} in $S_k(g_m)$ are independent and identically distributed with mean $\nu_k(g_m)$ and variance $\delta_k^2(g_m)$. The second stage is

$$\underline{\nu}(\underline{g}) \mid \underline{\theta}(\underline{g}) \sim N\left(\underline{\theta}(\underline{g}), \Sigma_3(\underline{g})\right), \tag{2.3}$$

where

$$\begin{aligned}
\underline{\theta}(\underline{g}) &= (\underline{\theta}(g_1), \ldots, \underline{\theta}(g_M))', \;\; \underline{\theta}(g_m) = (\theta_1(g_m), \ldots, \theta_{d(g_m)}(g_m))', \\
\Sigma_3(\underline{g}) &= \text{block diagonal}(\Sigma_3(g_m)), \; m = 1, \ldots, M, \\
\Sigma_3(g_m) &= \text{diag}\Big(\gamma_k^2(g_m)\Big), \; k = 1, \ldots, d(g_m).
\end{aligned}$$

Finally, we assign the prior probability of a partition vector $\underline{g}$ to be $p(\underline{g})$. To simplify the presentation of the theory in Section 3 we assume that the covariance matrices are known. A small numerical example, presented in Section 4, illustrates the methodology when the variance components are unknown. A special case of practical importance is to select the $\Sigma_3(\underline{g})$ to represent little prior, relative to sample, information about the $\underline{\nu}(\underline{g})$; see Section 3.

3. Posterior Inference

We summarize the posterior distribution of $\underline{\beta}$ with the first two moments; i.e.,

$$E(\underline{\beta} \mid \underline{y}) = \sum_{\underline{g}} E(\underline{\beta} \mid \underline{y}, \; \underline{g}) p(\underline{g} \mid \underline{y}), \tag{3.1}$$

and

$$\text{Cov}\,(\underline{\beta} \mid \underline{y}) = \sum_{\underline{g}} \text{Cov}\,(\underline{\beta} \mid \underline{y}, \; \underline{g}) p(\underline{g} \mid \underline{y}) + \text{Cov}_{\,\underline{g}|\underline{y}}(E(\underline{\beta} \mid \underline{y}, \; \underline{g})). \tag{3.2}$$

Theorems 3.1-3.3, given below, are the ones needed to evaluate (3.1) and (3.2). The proofs are straightforward, but require a substantial amount of algebraic manipulation; see Evans ((1997), Chapter 6) for details.

Theorem 3.1. *The expectation of $\underline{\beta}$, conditional on $\underline{g}$ and under the specification* (2.1) *to* (2.3)*, is*

$$E(\underline{\beta} \mid \underline{y},\ \underline{g}) = \lambda(\underline{g})\hat{\underline{\beta}} + (I - \lambda(\underline{g}))A(\underline{g})\phi(\underline{g})\hat{\underline{\nu}}(\underline{g}) + (I - \lambda(\underline{g}))A(\underline{g})(I - \phi(\underline{g}))\underline{\theta}(\underline{g}), \tag{3.3}$$

where

$$\begin{aligned}
\lambda(\underline{g}) &= (X'\Sigma_1^{-1}X + \Sigma_2^{-1}(\underline{g}))^{-1}X'\Sigma_1^{-1}X,\\
\phi(\underline{g}) &= (A(\underline{g})'\Sigma_*^{-1}(\underline{g})A(\underline{g}) + \Sigma_3^{-1}(\underline{g}))^{-1}A(\underline{g})'\Sigma_*^{-1}(\underline{g})A(\underline{g}),\\
\hat{\underline{\nu}}(\underline{g}) &= (A(\underline{g})'\Sigma_*^{-1}(\underline{g})A(\underline{g}))^{-1}A(\underline{g})'\Sigma_*^{-1}(\underline{g})\hat{\underline{\beta}},\\
\Sigma_*^{-1}(\underline{g}) &= (X'\Sigma_1^{-1}X)(X'\Sigma_1^{-1}X + \Sigma_2^{-1}(\underline{g}))^{-1}\Sigma_2^{-1}(\underline{g}),\\
\hat{\underline{\beta}} &= \left(X'\Sigma_1^{-1}X\right)^{-1}X'\Sigma_1^{-1}\underline{y}.
\end{aligned}$$

Theorem 3.2. *The covariance of $\underline{\beta}$, conditional on $\underline{g}$ and under the specification* (2.1) *to* (2.3)*, is*

$$\begin{aligned}
&\mathrm{Cov}\,(\underline{\beta} \mid \underline{y},\ \underline{g})\\
&= (X'\Sigma_1^{-1}X + \Sigma_2^{-1}(\underline{g}))^{-1} + W(\underline{g})\left[A(\underline{g})'\Sigma_*^{-1}(\underline{g})A(\underline{g}) + \Sigma_3^{-1}(\underline{g})\right]^{-1}W(\underline{g})',
\end{aligned} \tag{3.4}$$

where

$$W(\underline{g}) = (X'\Sigma_1^{-1}X + \Sigma_2^{-1}(\underline{g}))^{-1}\Sigma_2^{-1}(\underline{g})A(\underline{g}).$$

Theorem 3.3. *The posterior probability of $\underline{g}$, under the specification* (2.1) *to* (2.3)*, is*

$$p(\underline{g} \mid \underline{y}) \propto p(\underline{g})Z_0(\underline{g})Z_1(\underline{g})Q(\underline{g}) \tag{3.5}$$

where

$$\begin{aligned}
Z_0(\underline{g}) &= \mid \Sigma_2(\underline{g}) \mid^{-\frac{1}{2}} \mid X'\Sigma_1^{-1}X + \Sigma_2(\underline{g})^{-1} \mid^{-\frac{1}{2}},\\
Z_1(\underline{g}) &= \mid \Sigma_3(\underline{g}) \mid^{-\frac{1}{2}} \mid A(\underline{g})'\Sigma_*^{-1}(\underline{g})A(\underline{g}) + \Sigma_3(\underline{g})^{-1} \mid^{-\frac{1}{2}},\\
\Sigma_{**}^{-1}(\underline{g}) &= A(\underline{g})'\Sigma_*^{-1}(\underline{g})A(\underline{g})(A(\underline{g})'\Sigma_*^{-1}(\underline{g})A(\underline{g}) + \Sigma_3^{-1}(\underline{g}))^{-1}\Sigma_3^{-1}(\underline{g}),\\
Q(\underline{g}) &= \exp\left\{-\tfrac{1}{2}(\hat{\underline{\beta}} - A(\underline{g})\hat{\underline{\nu}}(\underline{g}))'\Sigma_*^{-1}(\underline{g})(\hat{\underline{\beta}} - A(\underline{g})\hat{\underline{\nu}}(\underline{g}))\right\}\\
&\quad \times \exp\left\{-\tfrac{1}{2}(\hat{\underline{\nu}}(\underline{g}) - \underline{\theta}(\underline{g}))'\Sigma_{**}^{-1}(\underline{g})(\hat{\underline{\nu}}(\underline{g}) - \underline{\theta}(\underline{g}))\right\}.
\end{aligned}$$

Formula (3.3) can be expressed as a weighted average of the sampling precision of $\hat{\underline{\beta}}$, $X'\Sigma_1^{-1}X$, prior precision of $\underline{\beta}$, $\Sigma_2^{-1}(\underline{g})$, sampling precision of $\hat{\underline{\nu}}(\underline{g})$, $A(\underline{g})'\Sigma_*^{-1}(\underline{g})A(\underline{g})$, and prior precision of $\underline{\theta}(\underline{g}), \Sigma_3^{-1}(\underline{g})$; i.e.,

$$E(\underline{\beta} \mid \underline{y},\ \underline{g}) = (B_1 + B_2)^{-1}\left[B_1\hat{\underline{\beta}} + B_2A(\underline{g})\left\{(A_1 + A_2)^{-1}[A_1\hat{\underline{\nu}}(\underline{g}) + A_2\underline{\theta}(\underline{g})]\right\}\right],$$

where $B_1 = X'\Sigma_1 X$, $B_2 = \Sigma_2^{-1}(\underline{g})$, $A_1 = A(\underline{g})'\Sigma_*^{-1}A(\underline{g})$, and $A_2 = \Sigma_3^{-1}(\underline{g})$.

Formula (3.5) has the appealing property that $p(\underline{g} \mid \underline{y})$ increases as $(\underline{\hat{\nu}}(\underline{g}) - \underline{\theta}(\underline{g}))'\Sigma_{**}^{-1}(\underline{g})(\underline{\hat{\nu}}(\underline{g}) - \underline{\theta}(\underline{g}))$ decreases and as the "within subset" distance, $(\underline{\hat{\beta}} - A(\underline{g})\underline{\hat{\nu}}(\underline{g}))'\Sigma_*^{-1}(\underline{g})(\underline{\hat{\beta}} - A(\underline{g})\underline{\hat{\nu}}(\underline{g}))$, decreases.

The second stage prior distribution in (2.3) is too unwieldy to be useful. We first find the consequences of letting $\Sigma_3^{-1}(\underline{g}) \to 0$, $\forall \underline{g} \in G^*$. Writing $\Sigma_3^{-1}(\underline{g}) \to 0$, $\forall \underline{g} \in G^*$, as $\Sigma_3^{-1} \to 0$, $\lim_{\Sigma_3^{-1}\to 0} E(\underline{\beta}|\underline{y}) = \sum_{\underline{g}} \lim_{\Sigma_3^{-1}\to 0}\{E(\underline{\beta}|\underline{y},\underline{g})p(\underline{g}|\underline{y})\}$ and $\lim_{\Sigma_3^{-1}\to 0} \mathrm{Cov}\,(\underline{\beta}|\underline{y}) = \sum_{\underline{g}} \lim_{\Sigma_3^{-1}\to 0}\{\mathrm{Cov}\,(\underline{\beta}|\underline{y},\underline{g})p(\underline{g}|\underline{y})\} + \lim_{\Sigma_3^{-1}\to 0} \mathrm{Cov}_{\,\underline{g}|\underline{y}}(E(\underline{\beta}|\underline{y},\underline{g}))$.

Using (3.1),

$$\lim_{\Sigma_3^{-1}\to 0} E(\underline{\beta} \mid \underline{y},\underline{g}) = \lambda(\underline{g})\hat{\underline{\beta}} + (I - \lambda(\underline{g}))A(\underline{g})\underline{\hat{\nu}}(\underline{g}), \tag{3.6}$$

and, using (3.4),

$$\lim_{\Sigma_3^{-1}\to 0} \mathrm{Cov}\,(\underline{\beta} \mid \underline{y},\underline{g}) = (X'\Sigma_1^{-1}X + \Sigma_2^{-1}(\underline{g}))^{-1} + W(\underline{g})\left\{A(\underline{g})'\Sigma_*^{-1}(\underline{g})A(\underline{g})\right\}^{-1} W(\underline{g})'. \tag{3.7}$$

To find $\lim_{\Sigma_3^{-1}\to 0} p(\underline{g}|\underline{y})$, we use (3.5) to write

$$\begin{aligned} p(\underline{g} \mid \underline{y}) &= \frac{p(\underline{g})Z_0(\underline{g})Z_1(\underline{g})Q(\underline{g})}{\sum_{\underline{g}'} p(\underline{g}')Z_0(\underline{g}')Z_1(\underline{g}')Q(\underline{g}')} \\ &= \left[\sum_{\underline{g}'} \frac{p(\underline{g}')}{p(\underline{g})}\frac{Z_0(\underline{g}')}{Z_0(\underline{g})}\frac{Z_1(\underline{g}')}{Z_1(\underline{g})}\frac{Q(\underline{g}')}{Q(\underline{g})}\right]^{-1}. \end{aligned} \tag{3.8}$$

From (3.8),

$$\lim_{\Sigma_3^{-1}\to 0} p(\underline{g} \mid \underline{y}) = \Big[\sum_{\underline{g}'} \frac{p(\underline{g}')}{p(\underline{g})}\frac{Z_0(\underline{g}')}{Z_0(\underline{g})} \lim_{\Sigma_3^{-1}\to 0}\frac{Z_1(\underline{g}')}{Z_1(\underline{g})} \lim_{\Sigma_3^{-1}\to 0}\frac{Q(\underline{g}')}{Q(\underline{g})}\Big]^{-1}.$$

Using the definition of $\Sigma_{**}^{-1}(\underline{g})$ in (3.5),

$$\lim_{\Sigma_3^{-1}\to 0} Q\Big(\underline{g}\Big) = \exp\left\{-\tfrac{1}{2}(\hat{\underline{\beta}} - A(\underline{g})\underline{\hat{\nu}}(\underline{g}))'\Sigma_*^{-1}(\underline{g})(\hat{\underline{\beta}} - A(\underline{g})\underline{\hat{\nu}}(\underline{g}))\right\}. \tag{3.9}$$

The remaining term is

$$\lim_{\Sigma_3^{-1}\to 0}\frac{Z_1(\underline{g}')}{Z_1(\underline{g})} = \lim_{\Sigma_3^{-1}\to 0}\frac{|\,\Sigma_3(\underline{g}')\,|^{-\frac{1}{2}}|\,A(\underline{g}')'\Sigma_*^{-1}(\underline{g}')A(\underline{g}') + \Sigma_3(\underline{g}')^{-1}\,|^{-\frac{1}{2}}}{|\,\Sigma_3(\underline{g})\,|^{-\frac{1}{2}}|\,A(\underline{g})'\Sigma_*^{-1}(\underline{g})A(\underline{g}) + \Sigma_3(\underline{g})^{-1}\,|^{-\frac{1}{2}}}, \tag{3.10}$$

and the limit in (3.10) depends on the rates at which $\Sigma_3^{-1}(\underline{g}) \to 0$ and $\Sigma_3^{-1}(\underline{g}') \to 0$.

For example, let $\gamma_k^{-2}(g_m) = a \times \gamma^{-2}$. Then for $a \in R^+$, it can be shown that

$$\lim_{\Sigma_3^{-1} \to 0} p(\underline{g} \mid \underline{y}) = \tilde{p}(\underline{g} \mid \underline{y}) \text{ if } d(\underline{g}) = \min\{d(\underline{g}') : \underline{g}' \epsilon G^*\}$$

and is 0 otherwise, where

$$\tilde{p}(\underline{g} \mid \underline{y}) \propto p(\underline{g}) Z_0(\underline{g}) \mid A(\underline{g})' \Sigma_*^{-1}(\underline{g}) A(\underline{g}) \mid^{-\frac{1}{2}} \times$$
$$\exp\left\{-\tfrac{1}{2}(\hat{\underline{\beta}} - A(\underline{g})\hat{\underline{\nu}}(\underline{g}))' \Sigma_*^{-1}(\underline{g})(\hat{\underline{\beta}} - A(\underline{g})\hat{\underline{\nu}}(\underline{g}))\right\}.$$

In other words, $\lim_{\Sigma_3^{-1} \to 0} p(\underline{g} \mid \underline{y})$ is non-zero only for those partitions having the lowest dimension; i.e., the minimal value of $d(\underline{g})$. If there is only one partition, $\underline{g}_0$, with the lowest dimension, $\lim_{\Sigma_3^{-1} \to 0} p(\underline{g}_0 \mid \underline{y}) = 1$ (e.g., $\underline{g}_0 = (g_0, \ldots, g_0)$ where g_0 is the partition of $\{\beta_{1m}, \ldots, \beta_{Lm}\}$ that assigns all L members to a single subset; i.e., $d(g_m) = 1$).

An alternative is to let $\Sigma_3^{-1} \to 0$ subject to a constraint on the last stage generalized variance, $\lim_{\Sigma_3^{-1} \to 0} \mid \Sigma_3^{-1}(\underline{g}) \mid / \mid \Sigma_3^{-1}(\underline{g}') \mid = c^*,\ c^* \in R^+,\ \underline{g} \neq \underline{g}'$, which gives

$$\lim_{\Sigma_3^{-1} \to 0} Z_1(\underline{g}')/Z_1(\underline{g}) = \mid A(\underline{g}')' \Sigma_*^{-1}(\underline{g}') A(\underline{g}') \mid^{-\frac{1}{2}} \mid A(\underline{g})' \Sigma_*^{-1}(\underline{g}) A(\underline{g}) \mid^{\frac{1}{2}} \times (c^*)^{-\frac{1}{2}}. \tag{3.11}$$

Substituting (3.9) and (3.11) into (3.8) it can be shown that $\lim_{\Sigma_3^{-1} \to 0} p(\underline{g} \mid \underline{y})$ is not invariant to changes in the scale of the data.

Our solution to determine the rate at which $\Sigma_3^{-1} \to 0$ is to assume that the Kullback-Leibler information for discriminating between the posterior distribution, $p(\underline{\nu}(\underline{g}) \mid \underline{y},\ \underline{g})$, and prior distribution, $p(\underline{\nu}(\underline{g}) \mid \underline{g})$, is constant over $\underline{g}$. This information quantity, $I(\underline{g})$, is given (for a general normal linear model) by Goel and DeGroot (1981). For the specification (2.1), (2.2), and (2.3), it can be shown that

$$I(\underline{g}) = \tfrac{1}{2} \ln \mid \Sigma_3(\underline{g}) H^{-1}(\underline{g}) \mid + \tfrac{1}{2} tr[H(\underline{g}) \Sigma_3(\underline{g})^{-1} - I] + \tfrac{1}{2} tr[\Sigma_3(\underline{g})^{-1} \underline{a}(\underline{g}) \underline{a}(\underline{g})'], \tag{3.12}$$

where

$$H^{-1}(\underline{g}) = \Sigma_3(\underline{g})^{-1} + (XA(\underline{g}))'[\Sigma_1 + X\Sigma_2(\underline{g})X']^{-1}(XA(\underline{g})),$$
$$\underline{h}(\underline{g}) = (A(\underline{g})X)'[\Sigma_1 + X\Sigma_2(\underline{g})X']^{-1}\underline{y} + \underline{\theta}(\underline{g}),$$
$$\underline{a}(\underline{g}) = H(\underline{g})\underline{h}(\underline{g}) - \underline{\theta}(\underline{g}).$$

The expression for $p(\underline{g} \mid \underline{y})$ that we propose is given by Theorem 3.4.

Theorem 3.4. *Under* (2.1)-(2.3), *if* $I(\underline{g}) = \xi$, $\xi \in R^+$, $\forall \underline{g} \in G^*$, *then*

$$\lim_{\Sigma_3^{-1} \to 0} p(\underline{g} \mid \underline{y}) \propto p(\underline{g}) \times \mid \Sigma_2(\underline{g}) \mid^{-\frac{1}{2}} \mid X'\Sigma_1^{-1}X + \Sigma_2(\underline{g})^{-1} \mid^{-\frac{1}{2}} \times \exp\{-d(\underline{g})/2\}$$
$$\times \exp\left\{-\tfrac{1}{2}(\hat{\underline{\beta}} - A(\underline{g})\hat{\underline{\nu}}(\underline{g}))'\Sigma_*^{-1}(\underline{g})(\hat{\underline{\beta}} - A(\underline{g})\hat{\underline{\nu}}(\underline{g}))\right\}. \quad (3.13)$$

Outline of the Proof.

The first step in the proof is to use the condition $I(\underline{g}) = \xi$ to solve for $\mid \Sigma_3(\underline{g})H^{-1}(\underline{g}) \mid$ in (3.12). This gives

$$\mid \Sigma_3(\underline{g})H^{-1}(\underline{g}) \mid^{-\frac{1}{2}} = \exp\left\{-\xi + \tfrac{1}{2}tr[H(\underline{g})\Sigma_3(\underline{g})^{-1} - I] + \tfrac{1}{2}tr[\Sigma_3(\underline{g})^{-1}\underline{a}(\underline{g})\underline{a}(\underline{g})']\right\}. \quad (3.14)$$

The second step is to verify that

$$\mid \Sigma_3(\underline{g})H^{-1}(\underline{g}) \mid^{-\frac{1}{2}} = \mid \Sigma_3(\underline{g}) \mid^{-\frac{1}{2}} \mid A(\underline{g})'\Sigma_*^{-1}(\underline{g})A(\underline{g}) + \Sigma_3(\underline{g})^{-1} \mid^{-\frac{1}{2}}. \quad (3.15)$$

The third step is to combine (3.14) and (3.15) to replace, in (3.5), the multiplier $\mid \Sigma_3(\underline{g}) \mid^{-\frac{1}{2}} \mid A(\underline{g})'\Sigma_*^{-1}(\underline{g})A(\underline{g}) + \Sigma_3(\underline{g})^{-1} \mid^{-\frac{1}{2}}$ by $\exp\{-\xi + \frac{1}{2}tr[H(\underline{g})\Sigma_3(\underline{g})^{-1} - I] + \frac{1}{2}tr[\Sigma_3(\underline{g})^{-1}\underline{a}(\underline{g})\underline{a}(\underline{g})']\}$. The final step of the proof is to let $\Sigma_3^{-1} \to 0$ in (3.5).

Note that the term $\exp(-d(\underline{g})/2)$ in (3.13) penalizes partitions $\underline{g}$ that have a large number of subsets.

The model, (2.1)−(2.3), specified independence across the regression parameter index m, the purpose being to avoid combining dissimilar regression coefficients. Nevertheless, an implicit pooling across m can occur; i.e., for some $\underline{g} \in G^*$ and some i, $i' = 1, \ldots, L$, $\exists m \neq m'$ such that $\mathrm{Cov}\left(\beta_{im}, \beta_{i'm'} \mid \underline{y}, \underline{g}\right) \neq 0$. However, Evans ((1997), Chapter 6) has shown that if $X'\Sigma_1^{-1}X$ is a diagonal matrix, $\mathrm{Cov}\left(\beta_{im}, \beta_{i'm'} \mid \underline{y}, \underline{g}\right) = 0$ for $m \neq m'$. One may obtain this zero covariance by using the Gram - Schmidt orthogonalization process to find a matrix Q such that $X^* = \Sigma_1^{-\frac{1}{2}}XQ$ is orthogonal. Defining $\underline{y}^* = \Sigma_1^{-\frac{1}{2}}\underline{y}$, $\underline{\beta}^* = Q^{-1}\underline{\beta}$ and $\underline{\eta} = \Sigma_1^{-\frac{1}{2}}\underline{\epsilon}$, (2.1) can be rewritten as

$$\underline{y}^* = X^*\underline{\beta}^* + \underline{\eta} \quad (3.16)$$

where $\underline{\eta} \sim N(0, I)$ and $X^{*'}X^*$ is a diagonal matrix. Then, using (2.2) and the definition of $\underline{\beta}^*$,

$$\underline{\beta}^* \mid \underline{\nu}(\underline{g}) \sim N\left(Q^{-1}A(\underline{g})\underline{\nu}(\underline{g}), Q^{-1}\Sigma_2(\underline{g})Q^{-1\prime}\right). \quad (3.17)$$

Thus, one uses the specification (3.16), (3.17) and (2.3) rather than (2.1), (2.2) and (2.3).

4. Properties

To illustrate properties of the methodology described in Section 3 we have carried out two small numerical investigations. The first study assumes known variances while the second allows these variances to be unknown. Throughout the first study there are $L = 6$ simple linear regressions. For simplicity, we take $\sigma_i^2 = \sigma^2$ and $\delta_k^2(g_m) = \delta^2$. We assume that there is sufficient prior information that the only partitions $\underline{g}$ assigned positive prior probability are those having at least two members in each subset of the constituent partitions, g_1 and g_2. Conservatively, we assigned $p(\underline{g})$ to be constant for all such partitions which have the form $P \times P$. Here, P defines the 41 possible partitions of the $L = 6$ intercepts or slopes. Each member of P has the form $\{(123456)\}, \{(i_1, i_2), (i_3, i_4), (i_5, i_6)\}$ or $\{(i_1, i_2, i_3), (i_4, i_5, i_6)\}$ where $i_j \epsilon \{123456\}$ and $i_j \neq i_k$ for $j \neq k$.

The overall posterior mean, $E(\underline{\beta} \mid \underline{y})$, and the posterior covariance matrix, $\mathrm{Cov}\,(\underline{\beta} \mid \underline{y})$, are obtained from (3.1), (3.2), (3.6), (3.7), and (3.13).

The data that we have used to illustrate our method are a modification of a data set in Moore and McCabe (1993) where, for state i, X_i is the average teacher's salary and Y_i is the average expenditure per pupil (note that Moore and McCabe regress X on Y). We use six simple linear regressions, one for each of six Census Divisions (e.g., the east north central division includes Ohio, Indiana, Illinois, Michigan, and Wisconsin). The values of X are from the Moore and McCabe data set. We then selected values of the intercepts, $\beta_{11}, \ldots, \beta_{61}$, and slopes, $\beta_{12}, \ldots, \beta_{62}$, and variance, σ^2, to achieve a desired amount of separation of the six slopes into two subsets $(\beta_{12}, \beta_{22}, \beta_{32})$ and $(\beta_{42}, \beta_{52}, \beta_{62})$. Finally, we obtained values of the Y_{ij} from

$$Y_{ij} = \beta_{i1} + \beta_{i2} X_{ij} + \epsilon_{ij}, \quad i = 1, \ldots, 6, \ j = 1, \ldots, n_i, \tag{4.1}$$

where the ϵ_{ij} are independent with $\epsilon_{ij} \sim N(0, \sigma^2)$. We also considered a range of values of δ^2.

To avoid the "implicit pooling problem" described in Section 3, we first center the X values for each regression so that $\sum_{j=1}^{n_i} X_{ij} = 0$. Then, from (3.6),

$$\lim_{\Sigma_3^{-1} \to 0} E(\beta_{i2} \mid \underline{y}, \ \underline{g}) = \lambda_i \hat{\beta}_{i2} + (1 - \lambda_i) A_{6+i}(g_2) \underline{\hat{\nu}}(g_2), \tag{4.2}$$

where $\hat{\beta}_{i2}$ is the least squares estimator of β_{i2}, $\lambda_i = \delta^2 \{\delta^2 + (\sigma^2 / \sum_{j=1}^{n_i} X_{ij}^2)\}^{-1}$, $A_{6+i}(g_2)$ is the row vector of length $d(g_2)$ obtained from the $(6+i)th$ row of $A(\underline{g})$ by deleting the first $d(g_1)$ columns, and $\underline{\hat{\nu}}(g_2)$ is the column vector obtained from $\underline{\hat{\nu}}(\underline{g})$ (formula (3.6)) by deleting the first $d(g_1)$ rows. For simplicity, we refer to $\lim_{\Sigma_3^{-1} \to 0} E(\beta_{i2} \mid \underline{y}, \ \underline{g})$ as $E(\beta_{i2} \mid \underline{y}, \ \underline{g})$.

Taking $\sigma^2 = 0.0043$, we present in Part a of Table 1 the value of λ_i and three estimators of β_{i2} for $i = 1, \ldots, 6$. Three different values of δ^2 are used. The three estimators are

a. $\hat{\beta}_{i2}$; i.e., $E(\beta_{i2} \mid \underline{y}, \underline{g})$ with $\lambda_i = 1$,
b. $E(\beta_{i2} \mid \underline{y}, g_2^*)$ where g_2^* is the conventional "pool all" partition that includes $\beta_{12}, \ldots, \beta_{62}$ in a *single subset* (i.e., $d(g_2^*) = 1$), and
c. the unconditional posterior mean, $E(\beta_{i2} \mid \underline{y})$ from (3.1) using (4.2) and $p(\underline{g} \mid \underline{y})$ from (3.13).

Table 1. Values of the alternative estimators of slope, β_{i2}, for different choices of δ^2.

	Regression					
	1	2	3	4	5	6
a. Example 1.						
$\hat{\beta}_{i2}$	.0189	.0416	.0958	.2340	.2370	.2884
δ^2=.0025						
λ	.93	.86	.98	.96	.88	.99
$E(\beta_{i2} \mid \underline{y}, g_2^*)$	.0284	.0568	.0971	.2307	.2269	.2872
$E(\beta_{i2} \mid \underline{y})$	.0210	.0428	.0957	.2342	.2372	.2881
δ^2=.0005						
λ	.73	.56	.90	.82	.59	.96
$E(\beta_{i2} \mid \underline{y}, g_2^*)$	.0574	.0937	.1024	.2207	.2055	.2827
$E(\beta_{i2} \mid \underline{y})$	.0290	.0479	.0919	.2381	.2448	.2870
δ^2=.0001						
λ	.35	.20	.64	.47	.22	.81
$E(\beta_{i2} \mid \underline{y}, g_2^*)$	.1210	.1483	.1248	.2034	.1893	.2670
$E(\beta_{i2} \mid \underline{y})$	.0483	.0594	.0843	.2496	.2577	.2837

	Regression					
	1	2	3	4	5	6
b. Example 2.						
$\hat{\beta}_{i2}$	.1404	.1624	.1995	.2228	.2559	.2873
δ^2=.0010						
λ	.84	.72	.95	.74	.90	.98
$E(\beta_{i2} \mid \underline{y}, g_2^*)$	.1520	.1770	.2003	.2207	.2517	.2856
$E(\beta_{i2} \mid \underline{y})$	.1441	.1641	.1996	.2218	.2560	.2867
δ^2=.0005						
λ	.73	.56	.90	.59	.82	.96
$E(\beta_{i2} \mid \underline{y}, g_2^*)$	.1612	.1863	.2013	.2204	.2488	.2841
$E(\beta_{i2} \mid \underline{y})$	.1452	.1615	.1997	.2193	.2578	.2864
δ^2=.0001						
λ	.35	.20	.64	.22	.47	.81
$E(\beta_{i2} \mid \underline{y}, g_2^*)$	.1970	.2140	.2096	.2263	.2408	.2759
$E(\beta_{i2} \mid \underline{y})$	.1461	.1520	.2015	.2093	.2662	.2850

With a large value of δ^2 such as $\delta^2 = 0.0025$, the λ_i are close to 1 and all three estimators of β_{i2} have similar values. As the value of δ^2 decreases it becomes clear that the "pool all" estimator, $E(\beta_{i2} \mid \underline{y},\ g_2^*)$, is inappropriate. For example, for $\delta^2 = 0.0005$, $\hat{\beta}_{22} = 0.0416$. However, $E(\beta_{22} \mid \underline{y},\ g_2^*) = 0.0937$ because the latter includes contributions from $\hat{\beta}_{42}$, $\hat{\beta}_{52}$, and $\hat{\beta}_{62}$, each of which exceeds 0.23. Conversely, $E(\beta_{22} \mid \underline{y}) = 0.0479$ essentially includes contributions only from $\hat{\beta}_{12}$, $\hat{\beta}_{22}$, and $\hat{\beta}_{32}$ since the posterior probability associated with the partition $\{(\beta_{12}, \beta_{22}, \beta_{32}),\ (\beta_{42}, \beta_{52}, \beta_{62})\}$ is 0.9892. When $\delta^2 = 0.0001$ the contrast between $E(\beta_{i2} \mid \underline{y},\ g_2^*)$ and $E(\beta_{i2} \mid \underline{y})$ is even clearer. Here, for $i = 1, 2, 3$, $E(\beta_{i2} \mid \underline{y})$ includes contributions *only* from these three regressions, and there is substantial shrinkage as well; e.g., $\hat{\beta}_{12} = 0.0189$ while $E(\beta_{12} \mid \underline{y}) = 0.0483$. However, $E(\beta_{12} \mid \underline{y},\ g_2^*) = 0.1210$ because it includes very large contributions from $\hat{\beta}_{42}$, $\hat{\beta}_{52}$, and $\hat{\beta}_{62}$.

A second way to contrast $E(\beta_{i2} \mid \underline{y})$ and $E(\beta_{i2} \mid \underline{y},\ g_2^*)$ is to define the separations $M_b = \mid \max_{i=1,2,3} E(\beta_{i2} \mid \underline{y},\ g_2^*) - \min_{i=4,5,6} E(\beta_{i2} \mid \underline{y},\ g_2^*) \mid$ and $M_c = \mid \max_{i=1,2,3} E(\beta_{i2} \mid \underline{y}) - \min_{i=4,5,6} E(\beta_{i2} \mid \underline{y}) \mid$. When $\delta^2 = 0.0005$, $M_b = 0.1$ and $M_c = 0.15$ while for $\delta^2 = 0.0001$, $M_b = 0.04$ and $M_c = 0.17$. In each case, the "pool all" estimator inappropriately includes data from regressions that are different from the one for which inferences are desired; this is well illustrated for $\delta^2 = 0.0001$ where M_b is very small.

Table 2. Ratio, R, of posterior variance of β_{i2} with uniform (reference) prior to unconditional posterior variance of β_{i2}, Var $(\beta_{i2} \mid \underline{y})$.

	Regression					
δ^2	1	2	3	4	5	6
a. Example 1.						
.0025	1.044	1.093	1.014	1.027	1.085	1.006
.0005	1.221	1.482	1.064	1.134	1.445	1.027
.0001	1.854	2.938	1.203	1.562	2.963	1.096
b. Example 2.						
.0010	1.094	1.204	1.030	1.193	1.060	1.012
.0005	1.156	1.360	1.050	1.351	1.111	1.022
.0001	1.323	2.017	1.105	2.354	1.495	1.075

Note: See Table 1 for additional information about the two examples.

If we do not pool the data from the relevant regressions, the increase in posterior variance is substantial. We present in Part a of Table 2 the ratio, R_i, $i = 1, \ldots, 6$, of $\sigma^2 / \sum_{j=1}^{n_i} X_{ij}^2$ to the unconditional posterior variance of β_{i2},

$\mathrm{Var}\,(\beta_{i2} \mid \underline{y})$. Here $\sigma^2/\sum_{j=1}^{n_i} X_{ij}^2$ is the posterior variance of β_{i2} when a uniform prior distribution is assigned to β_{i2}. When δ^2 is large, the relative increases, $R_i - 1$, are modest. However, for $\delta^2 = 0.0005$, the values of $100(R_i - 1)$ range from 2.7% to 48.2%. When $\delta^2 = 0.0001$, $100(R_i - 1)$ ranges from 9.6% to 196.3%.

Our second example in the first study has six regressions with the same values of the X_{ij}, and $\sigma^2 = 0.0043$. However, the slopes were chosen to be concordant with the partition $\{(\beta_{12}, \beta_{22}), (\beta_{32}, \beta_{42}), (\beta_{52}, \beta_{62})\}$ rather than $\{(\beta_{12}, \beta_{22}, \beta_{32}), (\beta_{42}, \beta_{52}, \beta_{62})\}$. In addition, the separations between adjacent subsets are smaller than in the first example. Thus, as one would expect, the unconditional means, $E(\beta_{i2} \mid \underline{y})$, are more similar to the "pool all" means, $E(\beta_{i2} \mid \underline{y},\ g_2^*)$, in Example 2 than in Example 1 (compare Parts a and b of Table 1). This shows the versatility of the unconditional estimator, $E(\beta_{i2} \mid \underline{y})$. Moreover, the increases in posterior variance (Table 2, Part b) are still large.

Finally, all of the posterior correlation coefficients of the β_{i2} are small. The maximal value is 0.38 for Example 1 and 0.56 for Example 2. Of the 90 posterior correlation coefficients (two examples, three δ^2, fifteen correlations per example and value of δ^2), only eleven exceed 0.10. The correlation coefficients are essentially zero for all pairs of slopes in different subsets.

The second study has $L = 4$ simple linear regressions with the specification as in (4.1). For each regression, $n = 3$ and $X_{ij} \epsilon \{-1, 0, 1\}$. We then selected values of the intercepts, $\beta_{11}, \ldots, \beta_{41}$, slopes, $\beta_{12}, \ldots, \beta_{42}$, and variance, σ^2, to achieve separation of the four intercepts into two subsets (β_{11}, β_{21}) and (β_{31}, β_{41}). The values of the Y_{ij} were then obtained from (4.1) with $\sigma^2 = 42$. (In the first study reported in this section we took $\sigma^2 = .0043$; in the second investigation we found it was convenient to rescale the values of Y so that they are 10^2 larger.) The only partitions, $\underline{g}$, assigned positive prior probability are the 16 having at least two members in each subset of the constituent partitions, g_1 and g_2. These partitions, assigned equal probability, are defined by the Cartesian product, $P \times P$, where $P = \{[(12), (34)], [(13), (24)], [(14), (23)], (1234)\}$ defines the four possible partitions of the $L = 4$ intercepts or slopes. Using (2.2) with $M = 2$ we take $\delta_k^2(g_1) = \delta_1^2$ and $\delta_k^2(g_2) = \delta_2^2$. That is, there is a common value, δ_1^2, of the variances of the intercepts corresponding to all partitions and subsets. Independent, inverse gamma prior distributions are assigned to σ^2, δ_1^2, and δ_2^2. The prior for σ^2 has mean 42 and variance 100 while each of the δ_i^2 has mean 1 and variance 2.

We use the Metropolis algorithm to sample $(g, \sigma^2, \delta_1^2, \delta_2^2)$. Using (2.1) and (3.13) the posterior distribution of g, $\sum_1$ and $\sum_2$, $f(g, \sum_1, \sum_2 | \underline{y})$, is proportional to the product of (3.13), $|\sum_1|^{-\frac{1}{2}}$, and the independent prior densities of $\sum_1$ and $\sum_2$. The Metropolis algorithm generates random walk Markov chains for $(\sigma^2, \delta_1^2, \delta_2^2)$ and an independent Markov chain for g. The convergence of the

Markov chain may be slow unless the data provide at least moderate support for all of the partitions. This occurs because a near-absorbing state is produced by the failure of the chain to move over the sample space of $\underline{g}, G^*$. Using the approach in Carlin and Chib (1995), extended to accommodate more than two "models," we adjust $p(\underline{g})$ to correct for the imbalance. That is, we increase the value of $p(\underline{g})$ for partitions that are rarely observed in the chain. (Adjusting $p(\underline{g})$ is a trial-and-error process, performed before the iterates are saved.) The realized Markov chain is corrected (by re-weighting) to account for the adjusted $p(\underline{g})$; i.e., to represent chains sampled using a uniform prior distribution on $\underline{g}$. We use the sampled $(g, \sigma^2, \delta_1^2, \delta_2^2)$ together with $E(\underline{\beta}|\underline{y}, \underline{g}, \sigma^2, \delta_1^2, \delta_2^2)$ from (3.6), $\mathrm{Cov}\,(\underline{\beta}|\underline{y}, \underline{g}, \sigma^2, \delta_1^2, \delta_2^2)$ from (3.7) and $p(\underline{g}|\underline{y}, \sigma^2, \delta_1^2, \delta_2^2)$ from (3.13)to obtain $E(\underline{\beta}|\underline{y})$ and $\mathrm{Cov}\,(\underline{\beta}|\underline{y})$.

Of the 16 partitions assigned positive prior probability, the four presented in Table 3 account for .9999 of the total posterior probability. The first line of Table 4a gives the least squares estimates of the intercepts, $\hat{\beta}_{i1}$, while the first line of Table 4b gives the corresponding estimates of the slopes, $\hat{\beta}_{i2}$. These estimates suggest the results in Table 3; i.e., for the intercepts there is overwhelming support for the partition $\{(12), (34)\}$ while for the slopes the support is distributed among four partitions.

Table 3. Values of the posterior probabilities, $p(\underline{g}|\underline{y})$, for the second study (variances are unknown)

Partition	Probability, $p(\underline{g}\|\underline{y})$
$\{(12), (34)\}, \{(12), (34)\}$	.1880
$\{(12), (34)\}, \{(13), (24)\}$	.2905
$\{(12), (34)\}, \{(14), (23)\}$	.1789
$\{(12), (34)\}, \{1234\}$	.3425
Others	.0001

Note: The left-most entry in each row describes the partition for the intercept; the second entry describes the partition for the slope.

Table 4, organized in the same way as Table 1, contrasts the three estimators: least squares, "pool all," $E(\beta_{ij}|\underline{y}, g_j^*)$, and unconditional posterior mean, $E(\beta_{ij}|\underline{y})$. It is clear from Table 4a that the conventional "pool all" estimator incorrectly pools the data from all four regressions, while the proposed estimator, $E(\beta_{i1}|\underline{y})$, pools the data from only the relevant regressions. For example, $E(\beta_{11}|\underline{y})$, and $E(\beta_{21}|\underline{y})$ use data only from the first two regressions. The situation for the slope is quite different. There is no obvious grouping of the four slopes, and the "pool all" estimator and proposed estimator have similar values. This example again illustrates the versatility of the proposed estimator.

Table 4. Values of the alternative estimators of the intercept, β_{i1}, and slope, β_{i2}, for the second study (variances are unknown)

	Regression			
	1	2	3	4
a. Intercept				
$\hat{\beta}_{i1}$	3.02	6.48	20.26	25.01
$E(\beta_{i1} \mid \underline{y}, g_1^*)$	13.24	13.39	13.97	14.18
$E(\beta_{i1} \mid \underline{y})$	4.66	4.85	22.50	22.77
b. Slope				
$\hat{\beta}_{i2}$	12.05	16.27	13.61	18.20
$E(\beta_{i2} \mid \underline{y}, g_2^*)$	14.96	15.06	15.00	15.11
$E(\beta_{i2} \mid \underline{y})$	14.15	15.53	14.50	15.95

If we do not pool the data from the relevant regressions, the increase in posterior variance is substantial. We start with (2.1), applied to only *one regression*, and assign a locally uniform prior distribution to $\underline{\beta}$. We assign to σ^2 exactly the same prior distribution (inverse gamma with mean 42, variance 100) as used throughout the second study. Define the variance of β_{ij} under this specification by $\mathrm{Var}\,(\beta_{ij}|\underline{y}_i)$ – to emphasize that inference depends only on data from the ith regression. Then define $R_{ij} = \mathrm{Var}\,(\beta_{ij}|\underline{y}_i)/\mathrm{Var}\,(\beta_{ij}|\underline{y})$ where the denominator is the unconditional (with respect to $\sigma^2, \delta_1^2, \delta_2^2, g$) posterior variance of β_{ij}. The eight values of R_{ij} (slope and intercept for each of four regressions) range from 3.38 to 3.63, a *very large* increase in variance.

Finally, the empirical density of σ^2 is positively skewed with mean 38 and variance 59. The empirical densities of δ_1^2 and δ_2^2 are mixture distributions (with no gaps) with means 0.73 and 0.71 and variances .13 and .08.

5. Discussion

The model that we have proposed is more flexible than the conventional hierarchical model which would start with (2.1) but then require that for each m, $(\beta_{1m}, \ldots, \beta_{Lm})$ are conditionally independent and identically distributed. With the conventional model, inference about β_{im} may use data from subpopulations that have characteristics substantially different from the ith. The numerical results suggest that our methodology assigns the L regressions to appropriate subsets, leading to sensible unconditional inferences, $E(\underline{\beta} \mid \underline{y})$ and $\mathrm{Cov}\,(\underline{\beta} \mid \underline{y})$ (see (3.1) and (3.2)). Moreover, $\mathrm{Cov}\,(\underline{\beta} \mid \underline{y})$ properly accounts for the uncertainty about $\underline{g}$.

We have shown how to accommodate unknown variances such as σ^2, δ_1^2 and δ_2^2. It is important to develop more efficient computational methods, and to investigate how the choice of prior distribution affects posterior inference.

To accommodate additional features of the survey design, the specification in (2.1) and (2.2) may have to be modified. In practical situations it should be possible to limit substantially the number of partitions that are assigned positive prior probability.

Acknowledgement

The authors are grateful to a referee for helpful suggestions.

References

Carlin, B. and Chib, S. (1995). Bayesian model choice via Markov chain Monte Carlo methods. *J. Roy. Statist. Soc. Ser. B* **57**, 473-484.

Evans, R. (1997). Bayesian inference when the pooling of data is uncertain. Ph.D. dissertation, State University of New York at Albany.

Goel, P. K. and DeGroot, M. H. (1981). Information about hyperparameters in hierarchical models. *J. Amer. Statist. Assoc.* **76**, 140-147.

Malec, D. and Sedransk, J. (1992). Bayesian methodology for combining the results from different experiments when the specifications for pooling are uncertain. *Biometrika* **79**, 593-601.

Malec, D., Sedransk, J., Moriarity, C. and LeClere, F. (1997). Small area inference for binary variables in the National Health Interview Survey. *J. Amer. Statist. Assoc.* **92**, 815-826.

Moore, D. and McCabe, D. (1993). *Introduction to the Practice of Statistics*. 2nd edition. W. H. Freeman, New York.

The Menninger Clinic, 5800 SW Sixth Avenue, Topeka KS 66601, U.S.A.

E-mail: evansrb@menninger.edu

Department of Statistics, Case Western Reserve University, 10900 Euclid Avenue, Cleveland, OH 44106-7054, U.S.A.

E-mail: jxs123@po.cwru.edu

(Received March 1997; accepted September 1998)