Using Hierarchical Models to Estimate a Weighted Average of Stratum-specific Parameters

Babette A. Brumback[1*], Larry H. Winner[2], George Casella[2], Malay Ghosh[2], Allyson Hall[3], and Paul Duncan[3]

[1]Department of Epidemiology and Biostatistics
[3]Department of Health Services Research, Management, and Policy
College of Public Health and Health Professions
[2]Department of Statistics
College of Liberal Arts and Sciences
University of Florida
Gainesville, FL 32611, USA

[*]corresponding author
e-mail: brumback@phhp.ufl.edu
phone: 352-273-5366
fax: 352-273-5365

February 23, 2007

SUMMARY

Many applications of statistics involve estimating a weighted average of stratum-specific parameters. We consider the case of known weights, motivated by a stratified survey sample of Medicaid participants in which the parameters are population stratum means and the known weights are determined by the population sampling frame. Assuming heterogeneous parameters, it is common to estimate the weighted average with the weighted sum of sample stratum means; under homogeneity, one ignores the known weights in favor of precision weighting. We focus on a general class of estimators, based on hierarchical models, that encompasses these two methods but which also includes adaptive approaches. One basic adaptive approach corresponds to using the DerSimonian and Laird (1986) model for the parameters. We compare this with a novel alternative, which models the variances of the parameters as inversely proportional to the known weights. For two strata, the two approaches coincide, but for three or more strata, they differ. We also present computational details and apply the methods to the Medicaid data for illustration and comparison in terms of mean squared error.

KEY WORDS: hierarchical model; weighted average; stratified sample; adaptive estimation; heterogeneity; shrinkage; survey sampling

1. Introduction

Many applications of statistics involve estimating a weighted average of stratum-specific parameters, representable as

$$\theta^s = \Sigma_{i=1}^n w_i \theta_i,$$

where n is the number of strata, θ_i, $i = 1, \ldots, n$ are the stratum-specific parameters and w_i, $i = 1, \ldots, n$ are the weights, which are positive and sum to one. To motivate this article, we consider one such example, which arises from a stratified sampling design to survey Medicaid participants in two Florida counties, Duval and Broward, for satisfaction with medical care before and after Medicaid Reform. The website `ahca.myflorida.com/Medicaid` provides information about the reform. Surveyed participants each belong to one of five Medicaid health care plans in Duval County or to one of fourteen plans in Broward County. In order to maximize power for plan-to-plan comparisons, equal sample sizes of 315 participants will be drawn per plan; however, county-level summaries are also of interest. For illustration, we will focus on using sample data to estimate for each of the two counties the mean age as of June 1, 2006 of Medicaid participants under 85 years old who were continously enrolled for 6 months and who have valid phone numbers; because participant age is known for the entire population, we will be able to compare our results with the "truth". The stratum-specific parameters are the mean ages for each of the health care plans within each county. The weights are the relative sizes of the health care plans in each of the counties, respectively. Population and sample data are presented in Table 1. Using the population data, we can compute θ^s, the "true" mean age, for each county; for Duval, $\theta^s = 16.0634$, and for Broward, $\theta^s = 15.7181$. For expository purposes, we will also focus on estimating the "true" mean age for plans 1,2, and 10 (the three plans with the most adults) combined in Duval County, $\theta^s = 16.3354$, as well as for plans 1 and 2 in Duval County, $\theta^s = 17.2049$, and finally for plans 4 and 6 in Duval County, $\theta^s = 13.5983$.

Probably the most common estimator of θ^s is the weighted sum of the stratum-specific sample means Y_i, $i = 1, \ldots, n$:

$$\hat{\theta}_a^s = \Sigma_{i=1}^n w_i Y_i,$$

which is the best linear unbiased estimator (BLUE) of θ^s assuming the model

$$Y_i = \theta_i + e_i, \tag{1}$$

where the θ_i are fixed effects, and the e_i are independent, approximately normal, mean zero error terms with variances σ_i^2. However, if the stratum-specific population means are nearly homogeneous ($\theta_i = \theta_0$, $i = 1, \ldots, n$), then an estimator with lower mean-squared error weights the stratum-specific sample means proportionally to their inverse-variances $1/\sigma_i^2$, $i = 1, \ldots, n$:

$$\hat{\theta}_b^s = \frac{\Sigma_{i=1}^n \frac{Y_i}{\sigma_i^2}}{\Sigma_{i=1}^n \frac{1}{\sigma_i^2}},$$

County	Plan	Size	Population Mean	Sample Mean	Standard Error
Duval	1	2364	17.6142	16.4624	0.84383
Duval	2	22126	17.1612	17.3253	0.96718
Duval	4	655	9.4875	9.7463	0.30147
Duval	6	5005	14.1363	14.0505	0.68854
Duval	10	26812	15.5412	15.4785	0.81467
Broward	1	1635	20.1854	19.9434	1.02065
Broward	2	11888	19.6333	20.9653	1.15109
Broward	3	3807	23.3295	25.3604	1.22483
Broward	4	1592	9.6735	9.3911	0.31058
Broward	5	4203	19.4929	18.8638	0.96248
Broward	6	3622	19.7760	20.7541	1.05755
Broward	7	9803	8.9737	9.0365	0.27134
Broward	8	6079	15.0898	14.8424	0.87780
Broward	9	4915	13.7887	12.8639	0.74514
Broward	10	6916	13.4814	14.0492	0.76740
Broward	11	10050	17.1009	16.9052	0.98841
Broward	12	1310	20.4160	21.6371	1.04670
Broward	13	13538	13.0447	13.6365	0.73828
Broward	14	3370	17.8735	18.0644	0.95388

Table 1: Population and Sample Data for the Two Counties

where in practice, we will substitute the sample standard errors s_i for σ_i, $i = 1, \ldots, n$. This estimator is the BLUE of θ^s assuming the model

$$Y_i = \theta_0 + e_i,$$

where θ_0 is a fixed effect, and the e_i are as before. In practice, homogeneity is usually not known *a priori*, and thus one strategy is to first test the null hypothesis of homogeneity using a one-way analysis of variance, and then to select either $\hat{\theta}^s_a$ or $\hat{\theta}^s_b$ depending on whether the test rejects or not.

For the Medicaid data of Duval and Broward Counties considered separately, the tests of homogeneity reject with p-values less than 0.001. However, when just plans 1,2, and 10 are evaluated in Duval County, the test does not reject, based on a p-value of 0.331. Following the strategy we have just outlined, then, we would choose $\hat{\theta}^s_a$ for each county-level summary, and $\hat{\theta}^s_b$ for the 3-plan summary of Duval County.

In this article, we focus on a general class of estimators based on hierarchical models. These encompass $\hat{\theta}^s_a$ and $\hat{\theta}^s_b$, but also include adaptive approaches resulting from modeling the θ_i as random and selecting the variance parameters to minimize mean squared error. One basic adaptive approach corresponds to using the DerSimonian and Laird (1986) model for the parameters. We compare this with a novel alternative, which models the variances of the parameters as inversely proportional to the w_i. We will evaluate competing estimators of θ^s under a sampling model that conditions on the θ_i but not on the e_i, $i = 1, \ldots, n$. Thus, bias, variance, and mean-squared error will be computed under a model that resamples the e_i but leaves the θ_i fixed. Whereas in this article, our inferences are confined to θ^s, the weighted mean of θ_i for our sample

of n strata, one might instead be interested in the weighted mean of θ_i for a larger population of N strata (in which case, the strata are usually called clusters), that is, in

$$\theta^p = \Sigma_{i=1}^{N} w_i^p \theta_i,$$

where the w_i^p are the weights normalized to the population rather than to the sample of n clusters. In this case, the sampling distribution of competing estimators would be based on resampling both the θ_i and the e_i.

In the earlier literature, DuMouchel and Duncan (1983) considered estimation of θ^s in a more general context; reduced to our problem, their results answer the question of when to use $\hat{\theta}_a^s$ or $\hat{\theta}_b^s$. They developed a test based on the difference statistic $\hat{\theta}_a^s - \hat{\theta}_b^s$, an approach that goes beyond the simple testing of homogeneity outlined above by incorporating the weights w_i into the decision. Pfeffermann and Nathan (1981) focused on estimating θ^p in a more general context; reduced to our problem, their method corresponds to using the DerSimonian and Laird (1986) model for the θ_i and estimating θ^p with $\Sigma_{i=1}^{n} w_i \hat{\theta}_i$, where the $\hat{\theta}_i$ are posterior means.

Following Fay and Herriot (1979), who adapted the James-Stein estimator (James and Stein 1961) and applied it to simultaneously estimate income in several areas with populations less than 1,000, the literature on small-area estimation is rife with hierarchical models for the θ_i; see, for example, the review article by Ghosh and Rao (1994) and the references therein. However, the focus in that literature is on simultaneous estimation of the θ_i rather than on their weighted sum.

Another related literature is that of multilevel modeling of complex survey data, the title of a recent paper by Rabe-Hesketh and Skrondal (2006). Following Pfeffermann et al. (1998), this literature concentrates on estimating θ^p but in a more general context; when reduced to our problem, the basic idea is to model the distribution of the θ_i in the entire population using a multilevel model that incorporates θ^p as a parameter. The next step is to relate the data, generated under informative sampling, to the census model using either weighting likelihoods (i.e. pseudolikelihoods, as in Rabe-Hesketh and Skrondal 2006) or weighted iterative generalized least squares algorithms (as in Pfeffermann et al. 1998). In the present paper, we apply multilevel (i.e. hierarchical) models for the sample (rather than the census) of θ_i, but rather than including θ^s (or, more generally, θ^p) as a direct parameter in the models, we estimate it indirectly as $\Sigma_{i=1}^{n} w_i \hat{\theta}_i$, where the $\hat{\theta}_i$ are posterior means. However, as we show in section 2, our novel hierarchical model, which models the variances of the θ_i as inversely proportional to the w_i, effectively incorporates θ^s as a direct parameter.

The last literature we will consider is that of generalized estimating equations (GEE, Liang and Zeger 1986), where the focus is on estimating the "population average" of clustered data. In the present paper, we reduce each of the clusters to a single statistic Y_i, which associates with a weight w_i or w_i^p, depending on whether θ^s or θ^p is the target of inference. In a recent paper, Williamson, Datta, and Satten (2003) address the problem of GEE estimation with so-called informative cluster size; when reduced to our setting with the

θ_i specified as cluster means rather than as general parameters (such as odds ratios, for example), what they propose is to estimate θ^p for the special case of all $w_i^p = \frac{1}{N}$, by using $\hat{\theta}_a^s$ with $w_i = \frac{1}{n}$ and the Y_i encoding the sample means of the clusters. However, their method is much different from the methods we consider when the θ_i are general parameters, such as odds ratios, because their model makes incomprehensible assumptions about the relationship between the θ_i and θ^p.

The present article is organized as follows. In section 2, we present the general hierarchical model and consider three special cases. In section 3, we focus on optimal estimation with only two strata and show that the adaptive approach arising from the DerSimonian and Laird (1986) model coincides with our novel approach when $n = 2$. In section 4, we show that the two approaches differ when there are three or more strata. In section 5, we discuss the details of estimation, including estimation of sampling distributions and mean squared error (MSE). In section 6, we apply the resulting methods to the Medicaid data for illustration and a comparison study of the estimators in terms of their estimated MSEs. Section 7 concludes with a summary, a discussion of related issues, and proposals for future work.

2. Adaptive Estimators Based on Hierarchical Models

We focus on the general hierarchical model

$$\begin{aligned} Y|\theta,\mu,\delta &\sim N(\theta,\Sigma) \\ \theta|\mu,\delta &= Z_W\mu+\delta \\ \delta &\sim N(h_W,T_W) \\ \mu &\sim U^p(-\infty,\infty), \end{aligned} \tag{2}$$

with $Y = (Y_1,\ldots,Y_n)^T$; $W = (w_1,\ldots,w_n)^T$; $\theta = (\theta_1,\ldots,\theta_n)^T$; Σ a symmetric positive definite known matrix; Z_W an $n \times p$ matrix, possibly dependent upon W; the vector h_W and the symmetric positive definite matrix T_W functions of W and of hyperparameters; $U^p(-\infty,\infty)$ the improper uniform distribution for a vector of p independent parameters; and δ and μ independent of one another given W. For the remainder of the paper we will suppress the subscripts on Z_W, h_W, and T_W. Additionally, we wish to emphasize that all distributions are conditional on the known weights constituting W.

Under mild restrictions on Z, T, and Σ (Appendix 1), the conditional distribution of $(\mu^T,\delta^T)^T$ given Y is multivariate normal with mean $(\hat{\mu}^T,\hat{\delta}^T)^T$ solving

$$\begin{aligned} Z^T\Sigma^{-1}Z\mu + Z^T\Sigma^{-1}\delta &= Z^T\Sigma^{-1}Y \\ \Sigma^{-1}Z\mu + (\Sigma^{-1}+T^{-1})\delta &= \Sigma^{-1}Y + T^{-1}h. \end{aligned} \tag{3}$$

Multiplying the second of these equations by Z^T and comparing to the first, we find that $\hat{\delta}$ satisfies the p-dimensional linear constraint

$$Z^TT^{-1}(\hat{\delta}-h) = 0. \tag{4}$$

This follows naturally from our setup, which uses the $n + p$ parameters in μ and δ to represent the n-dimensional parameter θ. We will later observe that this constraint influences our interpretation for μ, analogous to how constraints in models with nonrandom μ and δ would do.

The equations at (3) can be re-expressed to solve for $\hat{\delta}$ independently of $\hat{\mu}$ as follows:

$$\begin{aligned} \left(I - (\Sigma^{-1} + T^{-1})^{-1}\Sigma^{-1}P_Z^{\Sigma}\right)\hat{\delta} &= (\Sigma^{-1} + T^{-1})^{-1}\Sigma^{-1}\left((I - P_Z^{\Sigma})Y - \Sigma T^{-1}h\right), \\ Z^T T^{-1}\hat{\delta} &= Z^T T^{-1}h \end{aligned} \tag{5}$$

with $P_Z^{\Sigma} = Z(Z^T\Sigma^{-1}Z)^{-1}Z^T\Sigma^{-1}$. In Appendix 2 we show that, again under mild restrictions on Z, T, and Σ, the solution to these $n + p$ equations in n unknowns exists uniquely as

$$\begin{aligned} \hat{\delta} &= (A + T^{-1})^{-1}(AY + T^{-1}h), \\ \text{with } A &= (I - P_Z^{\Sigma})^T\Sigma^{-1}(I - P_Z^{\Sigma}) = \Sigma^{-1}(I - P_Z^{\Sigma}). \end{aligned} \tag{6}$$

It follows from (3) that $\hat{\mu}$ equals

$$\hat{\mu} = (Z^T\Sigma^{-1}Z)^{-1}(Z^T\Sigma^{-1})(Y - \hat{\delta}). \tag{7}$$

The posterior distribution of θ^s given Y has mean

$$\begin{aligned} \hat{\theta}^s &= W^T(Z\hat{\mu} + \hat{\delta}) \\ &= W^T\left((I - P_Z^{\Sigma})\hat{\delta} + P_Z^{\Sigma}Y\right), \end{aligned} \tag{8}$$

which can also be expressed as

$$\hat{\theta}^s = W^T\left((I - B)Y + BP_Z^{\Sigma+T}(Y - h) + Bh\right), \tag{9}$$

where $B = \Sigma(\Sigma + T)^{-1}$ (Appendix 3).

The posterior mean $\hat{\theta}^s$ is an admissible estimator of θ^s under squared-error loss for any particular specification of the general hierarchical model (2) (Lehmann and Casella, 1998, chapter 5). We can furthermore turn $\hat{\theta}^s$ into an adaptive estimator by letting the data guide our selection of the hyperparameters implicit in h and T. In this paper, we focus on T (we will later set $h = 0$); we specify $T = \tau^2 T_0$ with T_0 a known matrix (possibly dependent upon W), so that τ^2 tunes the eigenvalues of T towards 0 or ∞. When the eigenvalues of T tend towards ∞, the first equation at (5) tends towards

$$(I - P_Z^{\Sigma})\hat{\delta} = (I - P_Z^{\Sigma})Y,$$

which implies that $Z\hat{\mu} + \hat{\delta} \to Y$ (see equation (8)), so that

$$\hat{\theta}^s \to W^T Y = \hat{\theta}_a^s; \tag{10}$$

whereas, when the eigenvalues tend towards 0, $\hat{\delta} \to h$ (see the first equation at (5), in which $(\Sigma^{-1} + T^{-1}) \approx T^{-1}$), so that

$$\hat{\theta}^s \to W^T\left(P_Z^{\Sigma}Y + (I - P_Z^{\Sigma})h\right), \tag{11}$$

which equals $\hat{\theta}_b^s$ when (a) $Z = 1_p$ (the vector of p ones), (b) Σ equals the diagonal matrix with entries $\sigma_1^2, \ldots, \sigma_n^2$ (denoted henceforth as Σ_d), and (c) $h = 0$. If we choose τ^2 to minimize the estimated MSE of $\hat{\theta}^s$, we will adapt towards (10) or (11) depending on which limit optimizes the bias-variance trade-off relative to the other. We observe that $\hat{\theta}_a^s$ is minimax and that the right-hand side of (11) is admissible under squared-error loss (Lehmann and Casella, 1998, chapter 5.)

We next compare and contrast three particular specifications of the general hierarchical model (2).

Model 1

Model 1 sets $T_0 = I_n$, the $n \times n$ identity matrix; $h = 0$; $Z = 1_n$; and $\Sigma = \Sigma_d$. This leads to

$$\hat{\mu} = \frac{\Sigma_{i=1}^n \frac{Y_i}{\tau^2+\sigma_i^2}}{\Sigma_{i=1}^n \frac{1}{\tau^2+\sigma_i^2}},$$

and

$$\hat{\delta}_i = \frac{\tau^2}{\tau^2 + \sigma_i^2}(Y_i - \hat{\mu}).$$

Furthermore, $\hat{\delta}$ satisfies the constraint $\Sigma_{i=1}^n \hat{\delta}_i = 0$. Model 1 corresponds to the basic one-way mixed effects ANOVA model used by DerSimonian and Laird (1986) for meta-analysis, in which W is typically $(1/n)1_n$ to reflect the equally weighted mean of study-specific effects. With that choice of W, $\hat{\theta}^s = \hat{\mu}$. But for general W,

$$\begin{aligned} \hat{\theta}_1^s &= \Sigma_{i=1}^n \eta_i Y_i, \quad \text{where} \\ \eta_i &= \gamma_i \left(\tau^2 w_i + \frac{\Sigma_{j=1}^n \gamma_j \sigma_j^2 w_j}{\Sigma_{j=1}^n \gamma_j} \right) \quad \text{and} \\ \gamma_i &= \frac{1}{\tau^2 + \sigma_i^2}; \end{aligned} \tag{12}$$

we observe that $\Sigma_{i=1}^n \eta_i = 1$. Note that $\hat{\theta}_1^s$ approaches $\hat{\theta}_a$ or $\hat{\theta}_b$ as $\tau^2 \to \infty$ or $\tau^2 \to 0$.

Model 2

Model 2 sets T_0 equal to a diagonal matrix with entries equal to $1/w_i$, $i = 1, \ldots, n$ and keeps h,Z, and Σ the same as in Model 1. This leads to

$$\begin{aligned} \hat{\mu} &= \Sigma_{i=1}^n \eta_i Y_i, \text{ with} \\ \eta_i &= \frac{\frac{1}{\frac{\tau^2}{w_i}+\sigma_i^2}}{\Sigma_{i=1}^n \frac{1}{\frac{\tau^2}{w_i}+\sigma_i^2}}, \end{aligned} \tag{13}$$

and

$$\hat{\delta}_i = \left(\frac{1}{\sigma_i^2} + \frac{w_i}{\tau^2} \right)^{-1} \left(\frac{Y_i - \hat{\mu}}{\sigma_i^2} \right).$$

Note that $\Sigma_{i=1}^n w_i \hat{\delta}_i = 0$. Model 2 was briefly discussed by Brumback and Brumback (2005) in the present context but with estimated weights, and by Ghosh and Maiti (1998) in the context of multilevel modeling

under informative sampling. For general W, we find that the posterior mean of θ^s under Model 2, $\hat{\theta}_2^s$, equals the $\hat{\mu}$ of (13), and that it, like $\hat{\theta}_1^s$, also approaches $\hat{\theta}_a$ or $\hat{\theta}_b$ as $\tau^2 \to \infty$ or $\tau^2 \to 0$.

The Role of μ in Model 2 versus Model 1

That $\hat{\theta}_2^s$ equals the $\hat{\mu}$ of (13) parallels the role of μ in Model 2. The distribution specified for δ implies that the θ_i corresponding to larger w_i will be closer to μ; this information together with the constraint on $\hat{\delta}$ induces us to interpret μ as θ^s and δ_i as $\theta_i - \theta^s$ in Model 2. In contrast, the μ of Model 1 represents $\bar{\theta} = (1/n)\Sigma_i\theta_i$, whereas the δ_i represent $\theta_i - \bar{\theta}$. On first thought, this makes Model 2 ideal for our goal of estimating θ_s. However, one might use the data to negotiate between Model 1 versus Model 2; as we show in Appendix 4, if at least one w_i is distinct from the rest, and if we condition on μ, a "supermodel" that encompasses Model 1 and Model 2 is identifiable. In Sections 3 and 4, we refine this result for our purpose of estimating θ^s. In section 3, we show that for $n = 2$, $\hat{\theta}_1^s$ and $\hat{\theta}_2^s$ are identical when we select the τ^2 of Model 2 to equal the τ^2 of Model 1 multiplied by $2w_1(1 - w_1)$. In section 4, we show that $\hat{\theta}_1^s$ and $\hat{\theta}_2^s$ diverge when $n \geq 3$.

Model 3

We consider one more particular specification before moving on. Model 3 sets T_0 equal to I_n, sets h equal to a parametric function of W, e.g. $(W - \bar{W})\beta$, where $\bar{W} \equiv ((1/n)\Sigma_{i=1}^n w_i)\, 1_n$ and β is a scalar parameter, and keeps Z and Σ the same as in Models 1 and 2. This leads to

$$\hat{\mu} = \frac{\Sigma_{i=1}^n \frac{Y_i - h_i}{\sigma_i^2 + \tau^2}}{\Sigma_{i=1}^n \frac{1}{\sigma_i^2 + \tau^2}},$$

where h_i is the i^{th} element of h, and

$$\hat{\delta}_i = \left(\frac{1}{\sigma_i^2} + \frac{1}{\tau^2}\right)^{-1} \left(\frac{Y_i - \hat{\mu}}{\sigma_i^2} + \frac{h_i}{\tau^2}\right).$$

Thus,

$$\begin{aligned}
\hat{\theta}_3^s &= \Sigma_{i=1}^n \left(\eta_i Y_i + w_i \sigma_i^2 \gamma_i (h_i - \frac{\Sigma_{j=1}^n \gamma_j h_j}{\Sigma_{j=1}^n \gamma_j})\right), \text{ where} \\
\eta_i &= \gamma_i \left(\tau^2 w_i + \frac{\Sigma_{j=1}^n \gamma_j \sigma_j^2 w_j}{\Sigma_{j=1}^n \gamma_j}\right), \text{ and} \\
\gamma_i &= \frac{1}{\tau^2 + \sigma_i^2}. \qquad (14)
\end{aligned}$$

We observe that this estimator is not a simple weighted average of the Y_i except when $h_i \equiv 0$. In practice, one would often choose to let $Z\mu$ subsume h; however, it is worth recalling that $\hat{\theta}_3^s$ is admissible under squared-error loss no matter what value we choose for h. We note that $\hat{\theta}_3^s \to \hat{\theta}_a^s$ as $\tau^2 \to \infty$, and that it approaches

$$\hat{\theta}_b^s + \Sigma_{i=1}^n w_i \left(h_i - \frac{\Sigma_{i=1}^n \frac{h_i}{\sigma_i^2}}{\Sigma_{i=1}^n \frac{1}{\sigma_i^2}}\right),$$

as $\tau^2 \to 0$. Due to space constraints, we leave the development of this approach for future work.

3. "Optimal" Estimation with Only Two Strata

A natural question is whether we can use our adaptive estimators with $h = 0$ to approximately minimize mean-squared error (MSE) over all estimators of θ^s of the form $\Sigma_{i=1}^n \nu_i Y_i$, where the ν_i are positive weights that sum to one. In this paper, we investigate the question for $n = 2$. The generalization to $n > 2$ is much more complicated, and we leave it as a topic for future research. We consider the MSE of $\nu Y_1 + (1-\nu)Y_2$ for known ν, computed based on the fixed effects model (1):

$$\nu^2\left((\theta_1-\theta_2)^2+(\sigma_1^2+\sigma_2^2)\right)-2\nu\left(w_1(\theta_1-\theta_2)^2+\sigma_2^2\right)+\left(w_1^2(\theta_1-\theta_2)^2+\sigma_2^2\right).$$

This function of ν is minimized when

$$\nu=\nu^*=\frac{w_1(\theta_1-\theta_2)^2+\sigma_2^2}{(\theta_1-\theta_2)^2+(\sigma_1^2+\sigma_2^2)}=\frac{2w_1s_\theta^2+\sigma_2^2}{2s_\theta^2+(\sigma_1^2+\sigma_2^2)}, \tag{15}$$

where $s_\theta^2=\frac{1}{n-1}\Sigma_{i=1}^n(\theta_i-\bar{\theta})^2$, which equals $(\theta_1-\theta_2)^2/2$ when $n=2$.

We observe that the weight $\hat{\theta}_1^s$ associates with Y_1 is

$$\nu'=\frac{2w_1\tau^2+\sigma_2^2}{2\tau^2+\sigma_1^2+\sigma_2^2},$$

which equals ν^* when $\tau^2 = s_\theta^2$. According to Model 1, conditionally on μ, $\text{var}(\theta_i) = \tau^2$; thus, when $n = 2$, the risk-minimizing choice of τ^2 is a logical choice based on Model 1 and coincides with substituting s_θ^2 for $\text{var}(\theta_i)$.

Next we consider Model 2. We observe that the weight $\hat{\theta}_2^s$ associates with Y_1 is

$$\nu'=\frac{w_1\tau^2+w_1(1-w_1)\sigma_2^2}{\tau^2+w_1(1-w_1)(\sigma_1^2+\sigma_2^2)},$$

which equals ν^* when $\tau^2 = 2s_\theta^2 w_1(1-w_1)$. According to Model 2, conditionally on μ,

$$\frac{1}{n}\Sigma_{i=1}^n\text{var}(\theta_i)=\frac{\tau^2}{n}\Sigma_{i=1}^n\frac{1}{w_i}. \tag{16}$$

Approximating $\frac{1}{n}\Sigma_{i=1}^n\text{var}(\theta_i)$ by s_θ^2 and noting that $\frac{1}{2}\Sigma_{i=1}^2\frac{1}{w_i} = 1/(2w_1(1-w_1))$ leads us again to $\tau^2 = 2s_\theta^2 w_1(1-w_1)$. Thus, when $n = 2$, the risk-minimizing choice of τ^2 is also a logical choice based on Model 2.

In practice, s_θ^2 must be estimated from the data, and the resulting "risk-minimizing" estimator loses its optimality; by its equivalence to $\hat{\theta}_1^s$ or $\hat{\theta}_2^s$ for particular choices of τ^2 we know that it is admissible, but because we know that $\hat{\theta}_1^s$ and $\hat{\theta}_2^s$ are also admissible for other choices of τ^2, the "risk-minimizing" choice does not lead to a dominant estimator.

In the above investigation, we have incidentally discovered that, for $n = 2$, $\hat{\theta}_1^s = \hat{\theta}_2^s$ when the choice of τ^2 for Model 2 equals the choice of τ^2 for Model 1 multiplied by $2w_1(1-w_1)$. In the next section we show that, for $n = 3$, the two estimators are not identical for any data-based choices of the two τ^2.

4. A Brief Comparison of Models 1 and 2 for Three Strata

When there are three strata, $\hat{\theta}_1^s$ and $\hat{\theta}_2^s$ are quite distinct. To demonstrate this, we present a counterexample to the claim that, when $n = 3$, one can always make a data-based choice of τ_1^2 for Model 1 and of τ_2^2 for Model 2 such that the resulting estimators are identical.

Counterexample Dataset We let $(w_1, w_2, w_3) = (1/2, 1/3, 1/6)$, all $\sigma_i^2 = 1$, and $(Y_1, Y_2, Y_3) = (2.35, 0, 2.35)$.

Figure 1 graphs the resulting (η_1, η_2, η_3) as a function of τ^2 for each of the two estimators, and shows us the reason we can not find τ_1^2 and τ_2^2 to make the estimators identical: the graph for η_2 associated with $\hat{\theta}_2^s$ lies strictly above that for η_2 associated with $\hat{\theta}_1^s$; this results in $\hat{\theta}_2^s < \hat{\theta}_1^s$ for all choices of τ_1^2 and τ_2^2 – see Figure 2.

Although the graphs are convincing, some algebra shows conclusively what the graphs lead us to believe. For the counterexample dataset, we find that for Model 1, $\hat{\mu} = \frac{1}{3}\Sigma_{i=1}^3 Y_i = \bar{Y}$ for all $\tau^2 \in (0, \infty)$; $\hat{\delta}_i = \frac{\tau^2}{\tau^2+1}(Y_i - \bar{Y})$; and $\hat{\delta}_1 = \hat{\delta}_3 = -\hat{\delta}_2/2$. Therefore, $\hat{\theta}_1^s = \bar{Y}$ for all τ^2. For Model 2, on the other hand, we show in Appendix 5 that the η_2 associated with $Y_2 = 0$ is greater than $1/3$ for all τ^2, and thus $\hat{\theta}_2^s < \bar{Y}$ for all τ^2.

5. Details of Estimation and Performance Evaluation

We develop and present techniques for estimation and performance evaluation assuming $h = 0$ and $T = \tau^2 T_0$ in the hierarchical model at (2). We base our performance evaluation on the MSE criterion computed under the fixed effects sampling model $Y|\theta, \mu, \delta \sim N(\theta, \Sigma)$. In practice, the MSE generally depends on θ and τ^2; therefore, we need to use an estimated version. For the Medicaid example, we are able to compare the estimated version with a simulated version based on the true θ because we have data on patient age for the entire sampling frame.

The results of Section 2 lead to estimators for θ and θ^s, assuming known τ^2, as follows:

$$\begin{aligned} \hat{\theta} &= HY \text{ and} \\ \hat{\theta}^s &= W^T HY, \text{ where} \\ H &= P_Z^\Sigma + (I - P_Z^\Sigma)\left(\Sigma^{-1}(I - P_Z^\Sigma) + \frac{1}{\tau^2}T_0^{-1}\right)^{-1}\Sigma^{-1}(I - P_Z^\Sigma). \end{aligned}$$

We note that H depends on τ^2 and on the particular specification for the hierarchical model at (2). We propose to select τ^2 to minimize the estimated MSE of $\hat{\theta}^s$. The actual bias and variance are given by

$$\begin{aligned} \text{bias}_a &= W^T(H - I)\theta \text{ and} \\ \text{var}_a &= W^T H \Sigma H^T W. \end{aligned} \tag{17}$$

We estimate the bias as

$$\hat{\text{bias}} = W^T(H - I)\hat{\theta} = W^T(H - I)HY, \tag{18}$$

and the MSE as $\hat{\text{mse}} = \hat{\text{bias}}^2 + \text{var}_a$. Because $\hat{\text{mse}}$ depends upon τ^2 in a complicated way, we use the `optimize()` function within Version 2.3.0 of the statistical programming language `R` (2006) to minimize $\hat{\text{mse}}$ as a function of τ^2; in turn, we use the minimizing $\hat{\tau}^2$ to evaluate $\hat{\text{mse}}$.

Performance Evaluation

In section 6, we compare the performance of $\hat{\theta}_a^s$, $\hat{\theta}_b^s$, $\hat{\theta}_1^s$, and $\hat{\theta}_2^s$ in terms of their estimated MSEs, as follows. For $\hat{\theta}_1^s$ and $\hat{\theta}_2^s$, we use $\hat{\text{mse}}$ as described above. For $\hat{\theta}_a^s$ and $\hat{\theta}_b^s$, τ^2 does not depend upon the data. Instead,

$$\hat{\theta}_a^s = W^T Y = \lim_{\tau^2 \to \infty} \hat{\theta}_i^s \quad \text{and} \quad \hat{\theta}_b^s = W^T P_Z^\Sigma Y = \lim_{\tau^2 \to 0} \hat{\theta}_i^s, \tag{19}$$

for $i = 1$ or 2. For $\hat{\theta}_a^s$, $\lim_{\tau^2 \to \infty} H$ is undefined, and thus we compute the bias and variance directly from (19) as zero and $W^T \Sigma W$, respectively. For $\hat{\theta}_b^s$, $\lim_{\tau^2 \to 0} H = P_Z^\Sigma$, and the variance can be computed as the limit of (17). The estimation of bias is more complicated, because $\lim_{\tau^2 \to 0}(H - I)H = 0$, leading to $\hat{\text{bias}} = 0$, an unreasonable estimate. We thus propose to estimate the bias of $\hat{\theta}_b^s$ in two ways, using $W^T(H - I)H_i Y$ for $i = 1, 2$, where H_1 is the H of Model 1 and H_2 is the H of Model 2. For H_1, we select τ^2 as for $\hat{\theta}_1^s$, and for H_2, we select τ^2 as for $\hat{\theta}_2^s$.

A difficulty with our MSEs is that they are derived assuming that τ^2 is known, when in fact we are selecting it based on the data. This difficulty propogates into our estimated MSEs as well. For the Medicaid data, however, because we know θ and θ^s, we can compare our estimated MSEs to simulated 'true' MSEs. Specifically, we generate 10,000 datasets with Y simulated according to the fixed effects sampling model $Y|\theta, \mu, \delta \sim N(\theta, \Sigma)$. For each dataset, we estimate $\hat{\theta}_a^s$, $\hat{\theta}_b^s$, $\hat{\theta}_1^s$, and $\hat{\theta}_2^s$ as above. The 'true' MSE is then reported as the mean of the 10,000 squared errors, where one error is the estimate minus the true θ^s.

In usual practice, θ and θ^s will not be known. Thus, we would modify the above procedure, replacing θ and θ^s by one of their estimated versions for the simulation. Additionally, whereas throughout this paper we assume Σ is known, this is not the case in practice. The estimators of θ^s and MSE will be modified to use a consistent estimator of Σ generated prior to the data reduction to Y, W, Σ; the simulations should be modified accordingly, so that Σ is reestimated for each simulated dataset.

6. Application to Medicaid Data

Table 2 compares $\hat{\theta}_a^s$, $\hat{\theta}_b^s$, $\hat{\theta}_1^s$, and $\hat{\theta}_2^s$ as applied to the Medicaid data of Table 1. The column labeled **Approx MSE** presents the estimated, i.e. approximated, version of MSE, whereas the column labeled **Simulated MSE** presents the simulated 'true' version, computed as described in the preceding section. The rows present five different analyses, corresponding to different target summaries of plans within counties.

The adaptive nature of $\hat{\theta}_1^s$ and $\hat{\theta}_2^s$ is clearly evident in these results. For summaries 1,3, and 5, both estimators tend towards $\hat{\theta}_a^s$ due to heterogeneity of the plans. For summary 2, both tend towards $\hat{\theta}_b^s$, reflecting approximate homogeneity. For summary 4, all four estimators are nearly equal by a coincidence,

in which precision weighting is effectively the same as weighting based on plan size. By adapting to $\hat{\theta}_a^s$ or $\hat{\theta}_b^s$ according to estimated MSE, $\hat{\theta}_1^s$ and $\hat{\theta}_2^s$ avoid the large MSEs seen with $\hat{\theta}_a^s$ for summary 2 and with $\hat{\theta}_b^s$ for summaries 1,3, and 5.

Observe that the simulated MSEs are always larger than the estimated MSEs for $\hat{\theta}_1^s$ and $\hat{\theta}_2^s$; extra error appears when we account for the data-based selection of τ^2. The discrepancy between estimated and simulated MSEs for $\hat{\theta}_b^s$ results from using estimated versus 'true' values of θ in the computations. As expected, estimated and simulated MSEs are approximately equal for $\hat{\theta}_a$.

Perhaps most interesting is a direct comparison of $\hat{\theta}_1^s$ and $\hat{\theta}_2^s$. These estimators are identical for the two-plan summaries of Duval County, consistent with our results from section 3. For summary 4, the two estimators are nearly identical (differing only in simulated MSEs) because $\hat{\theta}_a^s$ and $\hat{\theta}_b^s$ are so close together. Basing a choice between $\hat{\theta}_1^s$ and $\hat{\theta}_2^s$ on estimated rather than simulated MSE would lead to the wrong selection for summaries 1 and 5, the only two summaries for which this choice could matter. Overall, however, we see that, when applied to the Medicaid data, $\hat{\theta}_1^s$ and $\hat{\theta}_2^s$ are essentially the same for all practical purposes.

7. Discussion

We have developed and compared adaptive estimators of a weighted average of stratum-specific parameters based on varying specificiations of a general hierarchical model. The estimators are especially useful when prior knowledge indicates that the parameters might be relatively homogeneous. Two estimators that we have studied in depth correspond to using either the DerSimonian and Laird (1986) model (Model 1) for the parameters or the novel model briefly touched on by Brumback and Brumback (2005) and Ghosh and Maiti (1998) (Model 2). In our investigation, we have uncovered the importance that the constraints satisfied by posterior means have for interpretability of the direct parameters of hierarchical models. We have also discovered that Model 1 and Model 2 lead to distinctly different estimators of θ^s when the number of strata exceeds two. However, these two estimators performed very similarly when applied to the Medicaid data. Further research is warranted to ascertain whether these two estimators perform similarly in general, or only under certain circumstances.

In this paper, we have considered only the case of known w_i. In several related applications, the w_i are unknown and need to be estimated from the sample. Examples include adjustment for nonresponse bias, missing data, or confounding. Further research could study the effects of applying our estimators with estimated w_i. A related topic is that of estimating summary odds ratios under heterogeneity. Greenland (1982) discusses relative advantages of the Miettinen (1972) and Mantel-Haenszel (1959) summary odds ratio estimators when one cannot assume homogeneity; these summary measures can each be expressed as a weighted sum of stratum-specific odds ratios, with data dependent weights.

We have only briefly discussed estimation of θ^p, but it would be worthwhile to use general hierarchical models to extend the early ideas of Pfeffermann and Nathan (1981). A comparison of these extensions with

Summary	Estimator	Estimate	Approx MSE	Simulated MSE
1. Duval	θ^s (truth)	16.0634	NA	NA
	$\hat{\theta}^s_a$	16.0453	0.2931	0.2909
	$\hat{\theta}^s_b$	11.8086	16.5058, 17.2316*	18.7206
	$\hat{\theta}^s_1$	15.8641	0.2688	0.3542
	$\hat{\theta}^s_2$	15.9527	0.2763	0.3168
2. Duval plans 1,2	θ^s (truth)	17.2049	NA	NA
	$\hat{\theta}^s_a$	17.2420	0.7702	0.7690
	$\hat{\theta}^s_b$	16.8354	0.4043,0.4043	0.4445
	$\hat{\theta}^s_1$	16.8354	0.4043	0.5045
	$\hat{\theta}^s_2$	16.8354	0.4043	0.5045
3. Duval plans 4,6	θ^s (truth)	13.5983	NA	NA
	$\hat{\theta}^s_a$	13.5524	0.3719	0.3714
	$\hat{\theta}^s_b$	10.4387	9.1705, 9.1705	11.4143
	$\hat{\theta}^s_1$	13.4544	0.3626	0.3997
	$\hat{\theta}^s_2$	13.4544	0.3626	0.3997
4. Duval plans 1,2,10	θ^s (truth)	16.3354	NA	NA
	$\hat{\theta}^s_a$	16.3203	0.3568	0.3592
	$\hat{\theta}^s_b$	16.3217	0.2512, 0.2512	0.3889
	$\hat{\theta}^s_1$	16.3217	0.2512	0.3791
	$\hat{\theta}^s_2$	16.3217	0.2512	0.3821
5. Broward	θ^s (truth)	15.7181	NA	NA
	$\hat{\theta}^s_a$	16.0856	0.0776	0.0763
	$\hat{\theta}^s_b$	12.0907	15.7914, 15.7534	13.4604
	$\hat{\theta}^s_1$	16.0613	0.0770	0.0790
	$\hat{\theta}^s_2$	16.0565	0.0767	0.0796

Table 2: Results of Applying the Methods to the Medicaid Data. *Approximate MSEs for $\hat{\theta}^s_2$ were computed with bias estimated first replacing each θ_i with $\hat{\theta}_{1,i}$ and second with $\hat{\theta}_{2,i}$.

the latter work of Pfeffermann et al. (1998) and others currently active in multilevel modeling of complex survey data would be quite interesting.

Acknowledgements

George Casella was supported by NSF-DEB-0540745. We thank the Florida Agency for Health Care Administration for providing the Medicaid data.

Appendix

Appendix 1.

Here we verify (3). Let $\beta = (\mu^T, \delta^T)^T$, $X = [Z|I_n]$, I_n be the $n \times n$ identity matrix, C be the block diagonal matrix with first block the $p \times p$ matrix of zeroes and second block equal to T^{-1}, and m be the vector $(0_p^T, h^T)^T$, with 0_p as the vector of p zeroes. Suppose β has prior given by (2), and that Y given β is

MVN with mean $X\beta$ and variance Σ. Then minus two times the log of the posterior distribution of β given Y equals (up to a constant)

$$(Y - X\beta)^T\Sigma^{-1}(Y - X\beta) + (\beta - m)^T C(\beta - m), \tag{20}$$

which in turn equals (up to a constant)

$$\beta^T(X^T\Sigma^{-1}X + C)\beta - 2\beta^T(X^T\Sigma^{-1}Y + Cm).$$

Assuming that $(X^T\Sigma^{-1}X + C)$ is invertible (this implies mild restrictions on Z, T, and Σ), we can complete the square with

$$(X^T\Sigma^{-1}Y + Cm)^T(X^T\Sigma^{-1}X + C)^{-1}(X^T\Sigma^{-1}Y + Cm),$$

so that minus two times the log of the posterior equals (up to a constant)

$$(\beta - b)^T V^{-1}(\beta - b),$$

with

$$\begin{aligned} b &= (X^T\Sigma^{-1}X + C)^{-1}(X^T\Sigma^{-1}Y + Cm) \text{ and} \\ V &= (X^T\Sigma^{-1}X + C)^{-1}. \end{aligned}$$

Thus the posterior of $\beta = (\mu^T, \delta^T)^T$ is multivariate normal with mean b solving (3).

Appendix 2.

Here we show that the posterior mean of δ given Y is the solution to (6). By the properties of the multivariate normal distribution, it must also equal the $\hat{\delta}$ of (3); therefore, we know that it solves (5).

Minus two times the log of the joint distribution of Y, δ, μ is (up to a constant)

$$(Y - Z\mu - \delta)^T\Sigma^{-1}(Y - Z\mu - \delta) + (\delta - h)^T T^{-1}(\delta - h).$$

Recall that $P_Z^\Sigma = Z(Z^T\Sigma^{-1}Z)^{-1}Z^T\Sigma^{-1}$, and note that $P_Z^\Sigma Z = Z$. Write

$$\begin{aligned} &(Y - Z\mu - \delta)^T\Sigma^{-1}(Y - Z\mu - \delta) = \\ &\left([P_Z^\Sigma(Y - \delta) - P_Z^\Sigma Z\mu] + [(I - P_Z^\Sigma)(Y - \delta)]\right)^T \Sigma^{-1} \left([P_Z^\Sigma(Y - \delta) - P_Z^\Sigma Z\mu] + [(I - P_Z^\Sigma)(Y - \delta)]\right). \end{aligned}$$

Expand the quadratic form according to the terms in square brackets, and note that the cross term is zero since $(I - P_Z^\Sigma)^T\Sigma^{-1}P_Z^\Sigma = 0$. So this part of the log is

$$(P_Z^\Sigma Z\mu - P_Z^\Sigma(Y - \delta))^T\Sigma^{-1}(P_Z^\Sigma Z\mu - P_Z^\Sigma(Y - \delta)) + ((I - P_Z^\Sigma)(Y - \delta))^T\Sigma^{-1}((I - P_Z^\Sigma)(Y - \delta)).$$

Only the first term has μ, and if we examine it and use the fact that $P_Z^\Sigma Z = Z$ again, we have

$$(P_Z^\Sigma Z\mu - P_Z^\Sigma(Y - \delta))^T\Sigma^{-1}(P_Z^\Sigma Z\mu - P_Z^\Sigma(Y - \delta)) = (\mu - \hat{\mu}_\delta)^T(Z^T\Sigma^{-1}Z)(\mu - \hat{\mu}_\delta),$$

where $\hat{\mu}_\delta = (Z^T\Sigma^{-1}Z)^{-1}Z^T\Sigma^{-1}(Y-\delta)$. Thus we have that the conditional distribution of $\mu|Y,\delta$ is

$$\mu|Y,\delta \sim N(\hat{\mu}_\delta, (Z^T\Sigma^{-1}Z)^{-1}).$$

In this form, we can now integrate out μ from the joint distribution to get the marginal distribution of Y,δ. This has minus two times the log equal to (up to a constant)

$$((I-P_Z^\Sigma)(Y-\delta))^T\Sigma^{-1}((I-P_Z^\Sigma)(Y-\delta)) + (\delta-h)^TT^{-1}(\delta-h).$$

Define $A = (I-P_Z^\Sigma)^T\Sigma^{-1}(I-P_Z^\Sigma)$. Minus two times the log is then (up to a constant)

$$(\delta-Y)^TA(\delta-Y) + (\delta-h)^TT^{-1}(\delta-h).$$

Factor this as

$$(\delta-\hat{\delta})^T(A+T^{-1})(\delta-\hat{\delta}) + Y^TAY + h^TT^{-1}h - \hat{\delta}^T(A+T^{-1})\hat{\delta},$$

where

$$\hat{\delta} = (A+T^{-1})^{-1}(AY+T^{-1}h),$$

as in (6). Thus, as we wished to show, the conditional distribution of δ given Y is

$$\delta|Y \sim N(\hat{\delta}, (A+T^{-1})^{-1}).$$

Appendix 3.

From the identity

$$\begin{aligned}&(Y-\theta)^T\Sigma^{-1}(Y-\theta) + (\theta-Z\mu-h)^TT^{-1}(\theta-Z\mu-h)\\ &= \theta^T(\Sigma^{-1}+T^{-1})\theta - 2\theta^T\left(\Sigma^{-1}Y + T^{-1}(Z\mu+h)\right) + Y^T\Sigma^{-1}Y + (Z\mu+h)^TT^{-1}(Z\mu+h),\end{aligned}$$

it follows that

$$E(\theta|Y,\mu) = T(\Sigma+T)^{-1}Y + \Sigma(\Sigma+T)^{-1}(Z\mu+h).$$

Noting that $Y|\mu \sim N(Z\mu+h, \Sigma+T)$, from the identity

$$(Y-Z\mu-h)^T(\Sigma+T)^{-1}(Y-Z\mu-h) = \mu^T\left(Z^T(\Sigma+T)^{-1}Z\right)\mu - 2\mu^TZ^T(\Sigma+T)^{-1}(Y-h) + (Y-h)^T(\Sigma+T)^{-1}(Y-h),$$

it follows that

$$E(\mu|Y) = \left(Z^T(\Sigma+T)^{-1}Z\right)^{-1}Z^T(\Sigma+T)^{-1}(Y-h).$$

Hence

$$E(\theta|Y) = (I-B)Y + BP_Z^{\Sigma+T}(Y-h) + Bh,$$

where $B = \Sigma(\Sigma+T)^{-1}$. The equation at (9) follows.

Appendix 4.

Here we show that if at least one w_i is distinct from the rest, and if we condition on μ, a "supermodel" that encompasses Model 1 and Model 2 is identifiable.

Specifically, under either Model 1 or 2 we consider the density of $Y|\mu, \tau^2$, and we introduce a new parameter ψ that equals 1 or 2 for Model 1 or 2. This leads to a supermodel, indexed by parameters (ψ, μ, τ^2), such that for ψ equal to either 1 or 2, the distribution P_{ψ,μ,τ^2} of $Y_1, \ldots, Y_n$ is that of n independent normal random variables with mean μ; for $\psi = 1$ the variances are $\sigma_i^2 + \tau^2$, but for $\psi = 2$, the variances are $\sigma_i^2 + \tau^2/w_i$. To show that the supermodel is identifiable, we suppose it is not and derive a contradiction. We suppose that there exists $(\psi_1, \mu_1, \tau_1^2)$ unequal to $(\psi_2, \mu_2, \tau_2^2)$ such that $P_{\psi_1,\mu_1,\tau_1^2} = P_{\psi_2,\mu_2,\tau_2^2}$. If $\psi_1 = \psi_2$, then (μ_1, τ_1^2) must equal (μ_2, τ_2^2) because the mean and variance parameters of the corresponding multivariate normal distribution are identifiable. It thus suffices to consider the case of $\psi_1 = 1$ and $\psi_2 = 2$. Nonidentifiability implies that $\mu_1 = \mu_2$, because the marginal normal mean of Y_1 when $\psi = 1$ must be the same as that when $\psi = 2$. We furthermore learn that $\sigma_i^2 + \tau_1^2$ must equal $\sigma_i^2 + \tau_2^2/w_i$ for $i = 1, \ldots, n$, because the marginal normal standard deviations of the Y_i must also be the same for $\psi = 1$ as for $\psi = 2$. Consequently,

$$\tau_1^2 = \tau_2^2/w_i \text{ for all } i.$$

Because T is positive definite, τ_1^2 and τ_2^2 must be nonzero. Thus, unless all w_i are equal, we have a contradiction.

Appendix 5.

For Model 2 and the counterexample dataset,

$$\eta_2 = \frac{\frac{1}{3\tau^2+1}}{\frac{1}{2\tau^2+1} + \frac{1}{3\tau^2+1} + \frac{1}{6\tau^2+1}}.$$

We need to show that $\eta_2 > 1/3$ for all $\tau^2 \in (0, \infty)$. Multiplying numerator and demoninator by $3\tau^2 + 1$, it suffices to show that

$$\frac{3\tau^2+1}{2\tau^2+1} + \frac{3\tau^2+1}{6\tau^2+1} = 1 + \frac{\tau^2}{2\tau^2+1} + 1 - \frac{3\tau^2}{6\tau^2+1} < 2.$$

Assuming $\tau^2 \in (0, \infty)$, it is equivalent to show that

$$\frac{1}{2\tau^2+1} < \frac{3}{6\tau^2+1},$$

which is true because $6\tau^2 + 1 < 6\tau^2 + 3$.

References

Brumback, B.A. and Brumback, L.C. (2005). Comment on "Semiparametric Estimation of a Treatment Effect in a Pretest-Posttest Study with Missing Data," by Davidian M, Tsiatis A, and Leon S, *Statistical Science*, 20, 284-289.

DerSimonian, R. and Laird, N. (1986). Meta-Analysis in Clinical Trials. *Controlled Clinical Trials*, 7, 177-188.

DuMouchel, W.H. and Duncan, G.J. (1983). Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples. *Journal of the American Statistical Association*, 78, 535-543.

Fay, R.E. and Herriot, R.A. (1979). Estimates of Income for Small Places: An Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association*, 74, 269-277.

Ghosh, M. and Maiti, T. (1998). Discussion on the Papers by Firth and Bennett and Pfeffermann et al. *Journal of the Royal Statistical Society, Series B*, 60, 41-56.

Ghosh, M. and Rao, J.N.K. (1994). Small Area Estimation: An Appraisal (with Discussion). *Statistical Science*, 9, 55-93.

Greenland, S. (1982). Interpretation and Estimation of Summary Ratios Under Heterogeneity. *Statistics in Medicine*, 1, 217-227.

James, W. and Stein, Charles (1961). Estimation with Quadratic Loss. In *Proceedings of the Fourth Berkeley Symposium of Mathematical Statistics and Probability, Vol. 1*, Berkeley: University of California Press, 361-379.

Lehmann, E.L. and Casella, G. (1998). *Theory of Point Estimation*. New York: Springer-Verlag.

Liang, K.Y. and Zeger, S.L. (1986). Longitudinal Data Analysis Using Generalized Linear Models. *Biometrika*, 73, 13-22.

Mantel, N. and Haenszel, W.H. (1959). Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease. *Journal of the National Cancer Institute*, 22, 719-748.

Miettinen, O.S. (1972). Standardization of Risk Ratios. *American Journal of Epidemiology*, 96, 383-388.

Rabe-Hesketh, S. and Skrondal, A. (2006). Multilevel Modelling of Complex Survey Data. *Journal of the Royal Statistical Society, Series A*, 169, 805-827.

The R Development Core Team (2006). R 2.3.0 - A Language and Environment. `http://www.r-project.org/`.

Pfeffermann, D. and Nathan, G. (1981). Regression Analysis of Data from a Cluster Sample. *Journal of the American Statistical Association*, 76, 681-689.

Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H., and Rasbash, J. (1998). Weighting for Unequal Selection Probabilities in Multilevel Models. *Journal of the Royal Statistical Society, Series B*, 60, 23-40.

Williamson, J.M., Datta, S., and Satten, G.A. (2003). Marginal Analyses of Clustered Data When Cluster Size is Informative. *Biometrics*, 59, 36-42.

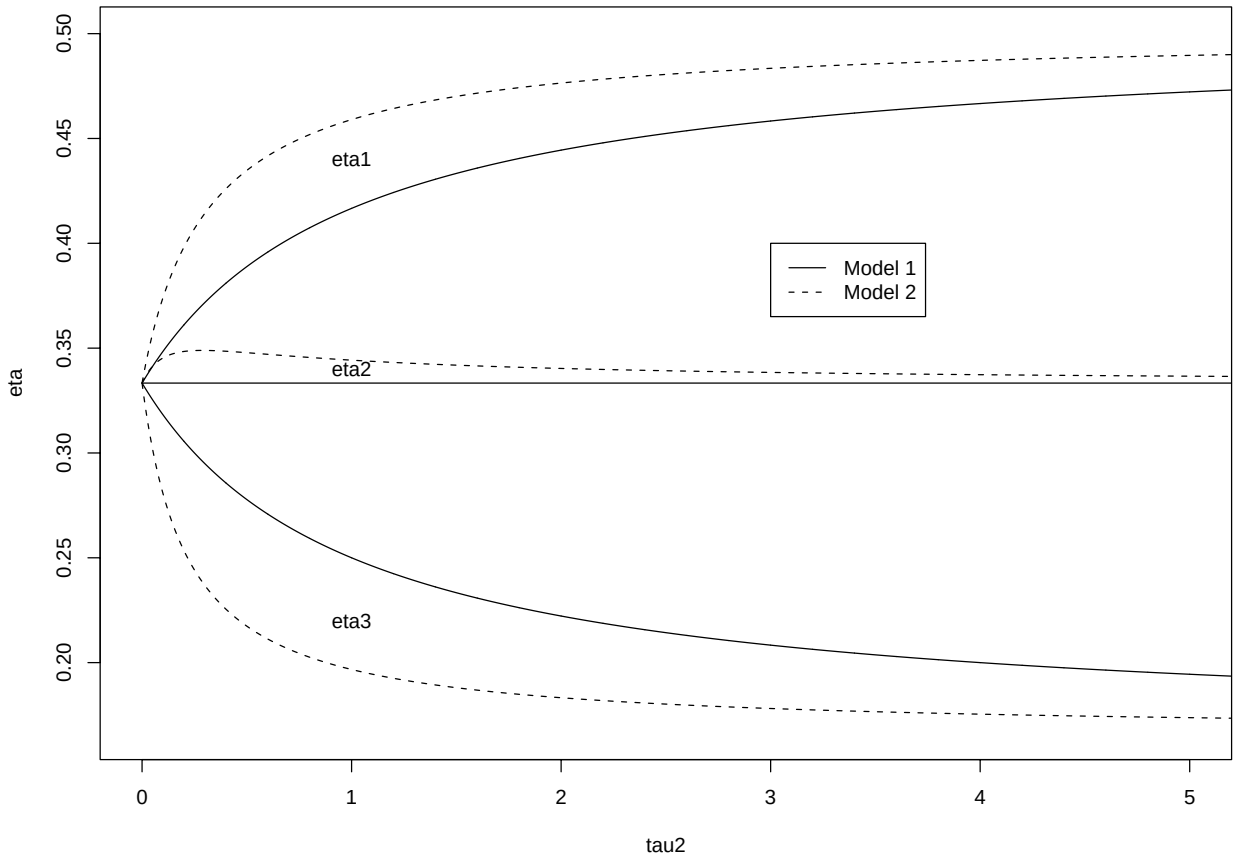

Figure 1: The η_i for Models 1 and 2 for the $n = 3$ Example Dataset of Section 4, as a function of τ^2.

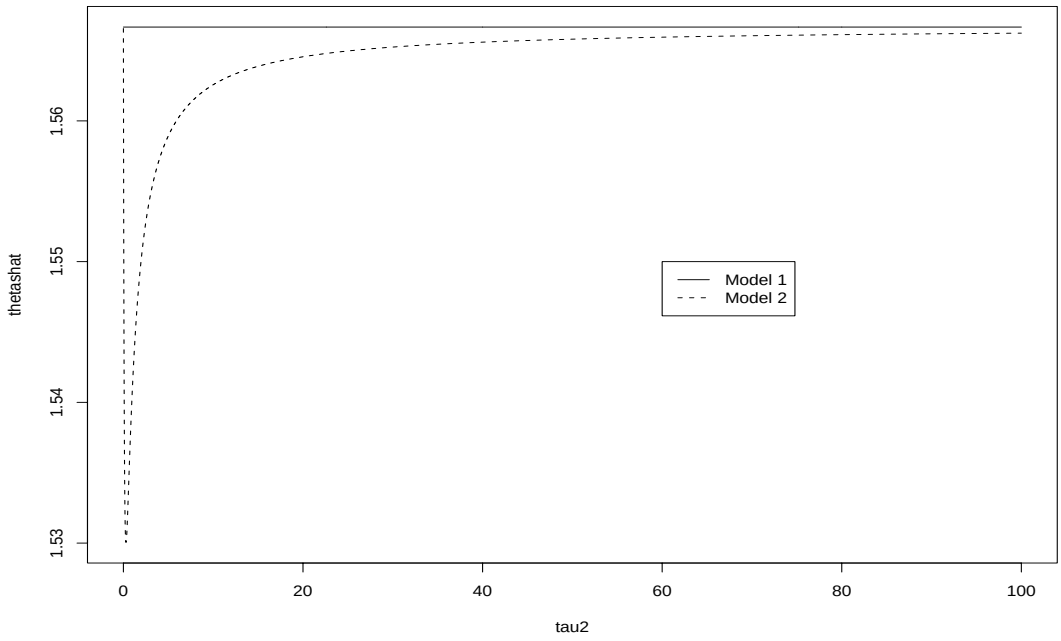

Figure 2: Estimated $\hat{\theta}_1^s$ and $\hat{\theta}_2^s$ as a function of τ^2 for the $n = 3$ Example Dataset of Section 4.