Journal of Official Statistics, Vol. 16, No. 3, 2000, pp. 191–209

Model-Based Alternatives to Trimming Survey Weights

Michael R. Elliott[1] *and Roderick J.A. Little*[2]

In sample surveys with unequal probabilities of inclusion, units are often weighted by the inverse of the probability of inclusion to avoid biased estimates of population quantities such as means. Highly disproportional sample designs yield large weights, which can result in weighted estimates that have a high variance. Weight trimming reduces large weights to a fixed cutpoint value and adjusts weights below this value to maintain the untrimmed weight sum. This approach reduces variance at the cost of introducing some bias. An alternative approach uses random-effects models to induce shrinkage across weight strata. We compare these two approaches, and introduce extensions of each: a compound weight pooling model that allows Bayesian averaging over estimators based on different trimming points, and a weight smoothing model based on a nonparametric spline function for the underlying weight stratum means. The latter method performs well in simulations as compared with alternative estimators. Methods are also applied to estimates of depression using weighted data from the National Comorbidity Survey.

Key words: Sample surveys inference; sampling weights; unit nonresponse adjustments; random-effects models; nonparametric regression.

1. Introduction

This article concerns the analysis of sample surveys where units have differential probabilities of inclusion. We use the term ''inclusion'' here to encompass both selection into the sample and response given selection, so that a unit nonresponse is regarded as ''selected'' but not ''included.'' The probability of selection may vary in stratified designs or samples selected with probability proportional to size. The probability of inclusion also varies in equal probability designs with nonresponse that is differential across unit characteristics. In these settings, estimators like sample means that assign the included units the same weight are biased when there is a correlation between the probability of inclusion and the values of the sampled data. Unit i is usually weighted by the inverse of the probability of inclusion to remove this bias. For example, the Horvitz-Thompson estimator (Horvitz and Thompson 1952) of a population total $T = \sum_{i=1}^{N} y_i$ from a sample is given by $\hat{T} = \sum_{i \in s} w_i y_i$, where $w_i = 1/\pi_i$, π_i is the probability of inclusion and s is the subset of the population units sampled. The probabilities of inclusion may be known in advance for both sampled and nonsampled units of the population, as in stratified random

[1]Department of Biostatistics and Epidemiology, University of Pennsylvania, School of Medicine, 423 Guardian Drive, Philadelphia, PA 19104, U.S.A. E-mail: melliott@cceb.upenn.edu
[2]Department of Biostatistics, School of Public Health, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109, USA.
Acknowledgments: This research is supported by National Science Foundation Grant DMS-9803720, and by U.S. Bureau of the Census, Contract Number 50-YABC-7-66020.

sampling, or may be estimated from the sample, as when there is unit nonresponse and the probability of response is estimated using observed sample characteristics.

Weighting often increases an estimate's variance. This increase can overwhelm the reduction in bias, so that the mean squared error actually increases under a weighted analysis. This is particularly likely when the weights are highly variable, when the correlation between the probability of inclusion and the data is weak, or when the sample size is small. Perhaps the most common approach to dealing with this problem is *weight trimming* (Potter 1990; Kish 1992; Alexander et al. 1997; Little et al. 1997), in which weights larger than some value w_0 are fixed as w_0 and the remaining weights are adjusted upward by a constant so that the weighted sample size remains unchanged. This manipulation of the weights reflects a traditional "design"-based approach to survey inference, which treats the data values in the population y_i as fixed and the assignment of the sampling indicators I_i as random.

An alternative strategy is to apply the model-based approach to survey inference, which makes distributional assumptions about the y_i and uses the model to predict the non-sampled values of y. A useful way of providing protection against the effects of model misspecification is to formulate a model for the survey outcomes within strata defined by the probabilities of inclusion (Little 1983, 1991; Rubin 1983). Standard weighted estimates are then obtained when the stratum means of survey outcomes are treated as fixed effects, and smoothing of the weights is achieved by treating the stratum means as random effects (Holt and Smith 1979; Ghosh and Meeden 1986; Little 1991, 1993; Lazzeroni and Little 1998). We call these random-effects models "weight-smoothing models."

This article compares weight trimming and weight smoothing, and extends the earlier modelling work in two directions. Classic weight trimming effectively pools all units with weights higher than the trimming point into a stratum with a single (reduced) weight. We extend this approach with a more general compound weight pooling model consisting of a collection of simple weight pooling models with different pooling points. We then put a prior distribution over the pooling points to achieve Bayesian averaging over the models in the set. We also propose a weight smoothing model with a smooth non-parametric mean structure, which is designed to be resistant to model misspecification. Empirical study by simulation and by application to real survey data indicates that both proposed methods have attractive properties, with the nonparametric weight smoothing model performing best overall.

Background and extensions of weight trimming and weight smoothing methods are considered in Section 2. Section 3 examines the root mean squared error and nominal confidence interval coverage of estimators considered in Section 2 under a disproportionately stratified sample design. Section 4 applies these estimators to data from the National Comorbidity Survey (Kessler 1992), a U.S.-wide multi-stage, disproportionately stratified survey of persons aged 15–54 concerning psychiatric disorders. Section 5 summarizes our findings and suggests areas for future study. Throughout this article, we assume the primary quantity of interest is the finite population mean $\bar{Y}$. Our models assume normally-distributed data, although extensions to non-normal distributions are possible, and we consider both normal and non-normal outcomes in the simulations and real-world examples.

2. Weight Pooling and Weight Smoothing Models

2.1. Simple weight-stratum pooling models

Weight trimming effectively pools units with high weights by assigning them a common, trimmed weight. Suppose the population can be divided into H weight strata by the set of *ordered* distinct values of the weights w_h. Let n_h be the number of included units and N_h the population size in weight stratum h, so that $w_h = N_h/n_h$ for $h = 1, \ldots, H$. We assume here that N_h is known, as when the weight strata come from a stratified or post-stratified random sample. The untrimmed (design-based) weighted mean estimator is then $\bar{y}_w = \sum_h \sum_i w_h y_{hi} / \sum_h \sum_i w_h = \sum_h N_h/N_+ \bar{y}_h$. Weight trimming typically proceeds by establishing an *a priori* cutpoint, say 3 for the normalized weights, and multiplying the remaining weights by a normalizing constant $\gamma = (n - \sum \kappa_i w_0)/\sum(1 - \kappa_i)w_i$, where κ_i is an indicator variable for whether or not $w_i \geq w_0$. The trimmed mean estimator is thus given by

$$\begin{aligned} \bar{y}_{wt} &= \sum_{h=1}^{l-1} \gamma N_h/N_+ \bar{y}_h + \sum_{h=l}^{H} w_0 n_h/N_+ \bar{y}_h \\ &= \gamma \sum_{h=1}^{l-1} N_h/N_+ \bar{y}_h + \frac{w_0 \sum_{h=l}^{H} n_h}{N_+} \bar{y}^{(l)} \end{aligned} \tag{2.1}$$

where $\gamma = (N_+ - w_0 \sum_{h=1}^{H} n_h)/\sum_{h=1}^{l-1} N_h$ and $\bar{y}^{(l)} = (1/\sum_{h=1}^{H} n_h) \sum_{h=l}^{H} n_h \bar{y}_h$

The choice of cutpoint w_0 is often ad-hoc. Potter (1990) discusses systematic methods for choosing w_0, including weight distribution, National Assessment of Educational Progress (NAEP), and MSE trimming procedures. The weight distribution technique assumes that the weights follow an inverted and scaled beta distribution. The parameters of the inverse-beta distribution are estimated by method-of-moment estimators, and weights from the upper tail of the distribution, say where $1 - F(w_i) < .01$, are trimmed-to w_0 such that $1 - F(w_0) = .01$. The NAEP procedure (Benrud et al. 1978) trims all weights $w_i > c(1/n) \sum w_i^2$ for a fixed c, then iterates the procedure until all weights are below some factor of the mean of the squared sum. The MSE trimming procedure (Cox and McGrath 1981) estimates MSE at a variety of trimming levels and chooses the one at which $\widehat{MSE}$ is minimized. We prefer the latter procedure to the first two since it relates the choice of trimming point to the survey outcome, although this property leads to practical complications in a real survey setting with many survey variables.

To proceed, note that the choice of $w_0 = \sum_{h=l}^{H} N_h / \sum_{h=l}^{H} n_h$ yields $\gamma = 1$ and $\bar{y}_{wt} = \sum_{h=1}^{l-1}(N_h/N_+ \bar{y}_h) + (\sum_{h=l}^{H} N_h)/N_h)/N_+ \bar{y}^{(l)}$, which corresponds to the estimate for a model that assumes distinct stratum means for the smaller weight strata and a common mean for the large weight strata, that is:

$$\begin{aligned} &y_{hi}|\mu_h \sim N(\mu_h, \sigma^2) \quad h < l \\ &y_{hi}|\mu_l \sim N(\mu_l, \sigma^2) \quad h \geq l \\ &\mu_h,\ \mu_l \propto \text{constant} \end{aligned} \tag{2.2}$$

We call (2.2) the simple weight pooling model.

2.2. *Compound weight pooling models*

We extend (2.2) by treating the pooling level l as a realization of the random variable L with support $(1,\ldots,H)$. Assuming that the location of pooling level L is a *priori* equally likely across the H weight strata, we obtain the compound weight pooling model:

$$\begin{aligned} y_{hi}|\mu_h &\sim N(\mu_h, \sigma^2)\ h < l \\ y_{hi}|\mu_l &\sim N(\mu_l, \sigma^2)\ h \geq l \\ p(L = l) &= 1/H \end{aligned} \tag{2.3}$$

Then

$$\begin{aligned} E(\bar{Y}|\mathbf{y}) &= E(E\bar{Y}|\mathbf{y}, l)) \\ &= \sum_{l=1}^{H} \left(\sum_{h=1}^{l-1} \frac{N_h}{N_+} \bar{y}_h + \frac{\sum_{h=l}^{H} N_h}{N_+} \bar{y}_{(l)} \right) p(L = l|y) \end{aligned} \tag{2.4}$$

That is, pooling is conducted at every possible level and the weighted average computed, where the weighting is based on the posterior probability for the model that pools from the lth stratum onward.

To derive $p(L = l|\mathbf{y})$, we note that (2.3) is a special case of a Bayesian variable selection problem (Halpern 1973; Atkinson 1978; Spiegelhalter and Smith 1982) with $y|\beta_l, l, \sigma^2 \sim N(X_l\beta_l, \sigma^2 I)$, where X_l is an $n \times l$ matrix consisting of an intercept and dummy variables for each of the first $l-1$ weight strata. Utilizing priors of the form $p(\sigma^2|l) = (1/\sigma^2)^{l/2+1}$ (Dempster et al. 1977) and $p(\beta_l|l) = (2\pi)^{-l}$ (Halpern 1973) yields

$$p(L = l \mid \mathbf{y}) = \frac{\left[\prod_{h=1}^{l-1} n_h \left(\sum_{h=l}^{H} n_h \right) \right]^{-1/2} Q_l^{-n}}{\sum_{l=1}^{H} \left[\left[\prod_{h=1}^{l-1} n_h \left(\sum_{h=l}^{H} n_h \right) \right]^{-1/2} Q_l^{-n} \right]} \tag{2.5}$$

where $Q_l^2 = \sum_{h=1}^{l-1} \sum_i (y_{hi} - \bar{y}_h)^2$

Bayesian variable selection is a controversial topic, in part because of sensitivity of inferences to the choice of priors. We discuss our choices and possible alternatives in more detail in the Appendix.

2.3. *Weight smoothing models*

Instead of mimicking the idea of weight trimming, we can simply model the weight-stratum means directly as random effects. The general form of the weight smoothing models we consider is

$$\begin{aligned} y_{hi}|\mu_h &\overset{ind}{\sim} N(\mu_h, \sigma^2) \\ \boldsymbol{\mu} &\sim N_H(\boldsymbol{\phi}, D) \end{aligned} \tag{2.6}$$

where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_H), \boldsymbol{\phi} = (\phi_1, \ldots, \phi_H), \boldsymbol{\phi}, D$, and σ^2 all have non-informative priors, and h indexes the ''weight strata,'' with constant inclusion probabilities. Unlike the weight

pooling models, the weight strata do not need to be ordered by probability of inclusion; a more natural ordering may be used if available, e.g., if the weight strata represent a disproportionately stratified sample by age. Under the model (2.6),

$$E(\bar{Y}|\mathbf{y}) = \sum_h [n_h \bar{y}_h + (N_h - n_h)\hat{\mu}_h]/N_+ \tag{2.7}$$

where $\hat{\mu}_h = E(\bar{Y}_h \,|\, \mathbf{y}) = E(\mu_h \,|\, \mathbf{y})$

The unweighted and fully weighted means are obtained as estimators of $E(\bar{Y}|\mathbf{y})$ as $D \to 0$ and $D \to \infty$, respectively. We consider the following special cases of the model:

Exchangeable random effects (XRE): (Holt and Smith 1979; Ghosh and Meeden 1986; Little 1991; Lazzaroni and Little 1998)

$$\phi_h = \mu \text{ for all } h, \; D = \tau^2 I_H \tag{2.8}$$

Autoregressive (AR1): (Lazzaroni and Little 1998)

$$\phi_h = \mu \text{ for all } h, \; D = r^2\{\rho^{|i-j|}\} \tag{2.9}$$

Linear (LIN): (Lazzaroni and Little 1998)

$$\phi_h = \alpha + \beta h, \; D = \tau^2 I_H \tag{2.10}$$

Nonparametric (NPAR):

$$\phi_h = f(h), \; D = 0 \tag{2.11}$$

where $f(h)$ is a twice differentiable smooth function of h,

$$\left\{ f : f^v \text{ absolutely continuous}, \; v = 0, 1, \int (f^{(2)}(u))^2 du < \infty \right\}$$

and $f(h)$ minimizes the residual sum of squares plus a roughness penalty parameterized by λ:

$$\sum_h \sum_i (y_{hi} - f(h))^2 + \lambda \int (f^{(2)}(u))^2 du \tag{2.12}$$

(Wahba 1978; Hastie and Tibshirani 1990).

All of these models can be written in the mixed-effect form (Laird and Ware 1982)

$$\mathbf{y} = NX\beta + NZ\mathbf{u} + \epsilon \tag{2.13}$$

where N is an $n \times H$ "incidence" matrix relating the distinct weight strata to the data ($n_{jk} = 1$ if y_j is in stratum k and 0 otherwise), $\boldsymbol{X}$ is an $H \times p$ fixed-effect design matrix, β is a $p \times 1$ vector of fixed-effect parameters, Z is an $H \times q$ random-effect design matrix, $\boldsymbol{u} \sim N_q(0, G)$, and $\epsilon \sim N(0, \sigma^2 I_n)$. This formulation yields the following replacements for (2.13):

XRE:

$$X = \mathbf{1}_H, Z = I_H, G = \tau^2 I_H \tag{2.14}$$

AR1:

$$X = \mathbf{1}_H, Z = I_H, G = \tau^2\{\rho^{|i-j|}\} \tag{2.15}$$

LIN:

$$X = \begin{pmatrix} 1 & 1 \\ \vdots & \vdots \\ 1 & H \end{pmatrix}, Z = I_H, G = \tau^2 I_H \tag{2.16}$$

NPAR:

$$X = \begin{pmatrix} 1 & 1 \\ \vdots & \vdots \\ 1 & H \end{pmatrix}, Z_{H\times(H-1)} \text{ such that } ZZ' = \Omega, G = (\sigma^2/H\lambda) I_{H-1} \tag{2.17}$$

where
$\Omega_{hk} = \int_0^1 ((h-1)/(H-1) - t)_+ ((k-1)/(H-1) - t)_+ dt, (x)_+ = x$ if $x \geq 0$ and $(x)_+ = 0$ if $x < 0$, $h, k = 1, \ldots, H$, and λ is the penalty parameter in (2.16) (Speed, in discussion of Robinson 1991; Wang 1998).

Under these formulations,

$$\hat{\boldsymbol{\mu}} = X\hat{\beta} + Z\hat{\boldsymbol{\mu}} \tag{2.18}$$

where $\hat{\beta} = (X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}\bar{\boldsymbol{y}}$ and $\hat{\boldsymbol{u}} = \hat{G}Z'\hat{V}^{-1}(\bar{\boldsymbol{y}} - X\hat{\beta})$. Here $V = ZGZ' + \sigma^2\Sigma$, where $\Sigma = diag(1/n_h)$, and $\bar{\boldsymbol{y}} = (\bar{\boldsymbol{y}}_1, \ldots, \bar{\boldsymbol{y}}_H)'$. In the case of XRE, AR1, and LIN, estimates of G and σ^2, and thus of β and $\boldsymbol{u}$, may be obtained by maximum likelihood (ML) or restricted maximum likelihood (REML) methods. In NPAR the REML likelihood given by Model (2.17) and the likelihood given by (2.11) differ by only a constant when $\hat{f}(h) = \boldsymbol{X}_h'\hat{\beta} + \boldsymbol{Z}_h'\hat{\boldsymbol{u}}$ is a natural cubic spline with knots at $(1, \ldots, H)$ (Wahba 1985; Green 1987; Wang 1998); hence we utilize ML estimates for (2.14)-(2.16) and REML estimates for (2.17) to obtain mean and variance component estimates and thus $\hat{\beta}$ and $\hat{\boldsymbol{u}}$.

These weight smoothing models allow compromises between weighted and unweighted estimates. As an example, note that, under the XRE model, $\widehat{\mu}_h = w_h\bar{y}_h + (1 - w_h)\tilde{y}$, where $w_h = \tau^2 n_h/\tau^2 n_h + \sigma^2$ and $\tilde{y}$ is an overall weighted mean given by $(\sum_h n_h/(n_h\tau^2 + \sigma^2))^{-1} \sum_h n_h/(n_h\tau^2 + \sigma^2)\, \bar{y}_h$. As $\tau^2 \to \infty$, $w_h \to 1$ so that $\widehat{\mu}_h \to \bar{y}_h$. Thus a flat prior for μ_h recovers the fully-weighted estimator, which can then be viewed as a fixed effects ANOVA model. On the other hand, as $\tau^2 \to 0$, $w_h \to 0$ so that $\widehat{\mu}_h \to \tilde{y}|_{\tau^2=0} = \tilde{y}$, which estimates the excluded units at the pooled mean since the model now assumes that y_{hi} are drawn from a common mean. Similarly for NPAR, $\lambda \to 0$ implies $\bar{y}_{NPAR} \to \bar{y}_w$ and $\lambda \to \infty$ implies $\bar{y}_{NPAR} \to \bar{y}_{linear}|_{\tau^2=0}$, so NPAR should be somewhat less effective than LIN when the means are linear, but should have reduced bias when the mean structure is nonlinear.

3. Simulation Study

A simulation similar to that in Little (1991) was constructed to test these methods in a controlled environment. Two populations of $N = 36,000$ were constructed consisting of 10 strata:

Stratum h	1	2	3	4	5	6	7	8	9	10
N_h	800	1,000	1,200	1,500	2,000	3,000	4,000	5,000	7,500	10,000

The values of Y_{hi} were generated as

$$y_{hi} = \delta_h + \epsilon_{hi}$$

where

$$\boldsymbol{\delta}^C = (22.5, 14.4, 9.0, 4.8, 1.8, -1.2, -1.8, -2.16, -1.92, -1.8)$$
$$\boldsymbol{\delta}^D = (-1.8, -1.92, -2.16, -1.8, -1.2, 1.8, 4.8, 9.0, 14.4, 22.5)$$
$$\boldsymbol{\delta}^E = (10.88, 10.88, 10.88, 10.88, 10.88, 10.88, 10.88, 10.88, 10.88, 10.88)$$
$$\boldsymbol{\delta}^L = (-12.09, -8.64, -5.20, -1.75, 1.70, 5.14, 8.59, 12.03, 15.48, 18.93)$$

and

$$\epsilon_{hi} \overset{ind}{\sim} \begin{cases} N(0, \sigma^2) & \text{with probability a) 1 or b).9} \\ \text{log normal}(-log(\sigma^2+1)/2, log(\sigma^2+1)) - 1 & \text{with probability a) 0 or b).1} \end{cases}$$

Disproportional samples of size 500 (90, 80, 70, 60, 50, 50, 40, 30, 20, 10) and size 100 (18, 16, 14, 12, 10, 10, 8, 6, 4, 2) were then drawn (maximum normalized weight = 13.9). To save space, we concentrate on the $n = 500$ results here, while commenting on results for $n = 100$ that are qualitatively different. Note that the mean structures $\boldsymbol{\delta}^C$ and $\boldsymbol{\delta}^D$ are best- and worst-case scenarios for weight trimming, with $C =$ close and $D =$ distant means in the high-weight strata. The $E =$ equal mean structure $\boldsymbol{\delta}^E$ provides the best-case scenario for the XRE and AR1 models, and the $L =$ linear structure $\boldsymbol{\delta}^L$ provides the best-case scenario for the LIN model; parameters are chosen so that $E(\bar{Y}|\boldsymbol{\delta}^D) = E(\bar{Y}|\boldsymbol{\delta}^E) = E(\bar{Y}|\boldsymbol{\delta}^L)$. The log-normal contamination is chosen so that the mean and variance equals the uncontaminated normal distribution for a given value of σ^2, but yields skewed and outlying values in data otherwise assumed to be normally distributed, thus giving a simple check of the robustness of the various weighting procedures to non-normality.

Two hundred stratified random samples were drawn as described from each population. The two primary outcomes of interest are root mean squared error (RMSE) and coverage of nominal 90% confidence intervals. RMSE is estimated as $\sqrt{(1/200\sum_{i=1}^{200}(\hat{\theta}_i - \theta)^2)}$, where θ is the population mean and $\hat{\theta}_i$ is the estimate from the ith of the 200 samples.

3.1. Stratum pooling methods

A 2×2 design was used to compare the stratum pooling methods: mean structures $\boldsymbol{\delta}^C$ and $\boldsymbol{\delta}^D$, and normal and contaminated normal error distributions. For each design, we considered estimates from five weighting schemes: unweighted (WT = 1); fully weighted, $\bar{y}_w$ (FWT); ''crude'' trimming, $\bar{y}_{wt}$, where the maximum normalized weight is 3 (WT < 3); the MSE trimming method (MSET) where the trimming points are selected from all possible pooling values (in these simulations described, 10 distinct cutpoints are possible); and the posterior mean from the compound weight pooling model given by (2.4) and (2.5) (CWP). The variance estimates used to calculate 90% confidence intervals are given in Table 1. Note that the second term in the variance estimator of $\bar{y}_{cwp}$ includes the variance added for estimating L.

3.1.1. Performance when stratum means favor trimming

The top left panel of Figure 1 compares the RMSE ratio of the unweighted and three other trimmed estimators to the fully weighted estimator, for normal data with a stratum mean configuration $\boldsymbol{\delta}^C$ that favors pooling of the high-weight strata. The unweighted

Table 1. Variance estimates of weight pooling estimators

Mean Estimator	Variance Estimator
$\bar{y}$	$(1-f)\,s^2/n$
$\bar{y}_w$	$\left(s^2_{p_H}\sum_h (1-f_h)n_h w_h^2 / \left(\sum_h n_h w_h\right)^2\right) = s^2_{p_H}\sum_h P_h^2(1-f_h)/n_h$
$\bar{y}_{wt}$, $\bar{y}_{MSET}$	$\left(\dfrac{s^2_{p_l}\sum_{h=1}^{l-1}(1-f_h)n_h w_h^2 + s_l^2(1-f_l)n_l w_0^2}{\left(\sum_{h=1}^{l-1} n_h w_h + w_0 n_l\right)^2}\right)$ $= s^2_{p_l}\sum_{h=1}^{l-1} P_h^2(1-f_h)/n_h + P_l^2(1-f_l)s_l^2/n_l$
$\bar{y}_{CWP}$	$\sum_{l=1}^{H} Var(\bar{y}_{wt}\|L=l)P(L=l\|y) + \left(\sum_{l=1}^{H}\bar{y}_l^2 P(L=l\|y) - \bar{y}^2_{CWP}\right)$

Notation:

$n_l = \sum_{h=l}^{H} n_h;\ N_l = \sum_{h=l}^{H} N_h;\ P_l = \left(\sum_{h=l}^{H} N_h\right)/N$

$f = n/N; f_h = n_h/N_h; f_l = n_l/N_l$

$\bar{y}, \bar{y}_h$ = overall and stratum mean; $\bar{y}^{(l)} = 1/n_l \sum_{h=l}^{H}\sum_i y_{hi}$

s^2, s_h^2 = overall and stratum variance; $s_l^2 = 1/(n_l - 1)\sum_{h=l}^{H}\sum_i (y_{hi} - \overline{y^{(l)}})^2$

$s^2_{p_l} = \left(\sum_{h=1}^{l-1} n_h - l\right)^{-1}\sum_{h=1}^{l-1}\sum_{i=1}^{n_h}(y_{hi} - \bar{y}_h)^2)$

$P(L = l|y)$ is given by (2.5)

estimator (WT = 1) is poor when $\log \sigma < 2$ and some form of weighting is needed to counteract bias. ''Crude'' weight trimming (WT < 3) is an improvement, but fails badly when $\log \sigma < 0$. The compound weight pooling estimator (CWP) works well in this setting: when σ is small and weighting is needed to counteract bias, CWP mimics the fully weighted estimator, and when σ is large and variance is of primary concern, CWP behaves more like an unweighted estimator. Figure 2 provides insight on how the estimator works by plotting the posterior probabilities of pooling the L largest weight strata as a function of σ, for a single normal sample with stratum means $\boldsymbol{\delta}^C$. When $\sigma \approx 0$, the differences among the means are distinguished and pooling is prevented. As σ increases, the small differences among the large strata are ignored and pooling proceeds. Eventually all but the well-distinguished and most heavily sampled strata are pooled. The MSET estimator behaves like CWP when $\log \sigma$ is less than 2, but is inferior to CWP when σ is relatively

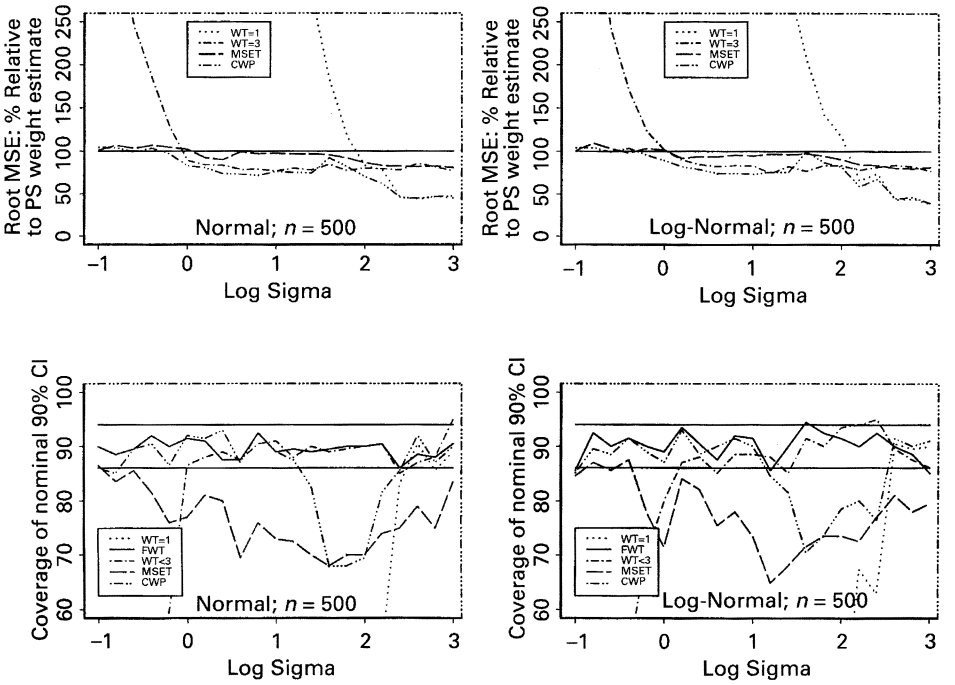

Fig. 1. Average RMSE relative to unweighted estimator and nominal 90% coverage (200 simulations) of unweighted (WT = 1), crude trimming (WT < 3), minimum MSE (MSET), and compound weight pooling (CWP) estimators, when stratum means favor trimming

large. The MSE savings of trimming are obtained at smaller values of σ and are somewhat larger when $n = 100$.

The bottom left panel of Figure 1 shows the coverage of the nominal 90% confidence intervals of each method over 200 samples, for the favorable mean configuration. The coverage of the crude trimming estimator (WT < 3) is very poor when $\log \sigma < 0$, and the method has unacceptable bias. For the smaller sample size, both FWT and CWP have good coverage; for the larger sample size, CWP suffers coverage problems when the between- and within-variances are approximately equal. The coverage of MSET is markedly below the nominal level, because $Var(\bar{y}_{wt})$ in Table 1 does not account for the variability in estimating the cutpoint. This variability is difficult to incorporate except perhaps by bootstrapping or jackknifing the entire procedure for selecting the cutpoint.

The right panels of Figure 1 show results under the favorable mean configuration for data generated using the contaminated normal model. As expected, CWP and MSET estimators still yielded reduced mean squared error, although gains were somewhat less than for normal data. Coverage for these methods is somewhat poorer than for normal data, as might be expected.

3.1.2. Performance when stratum means do not favor trimming

The mean structure δ^D is an unfavorable scenario for trimming, since the highest weight stratum has a mean substantially different from the other strata. In this case pooling of the

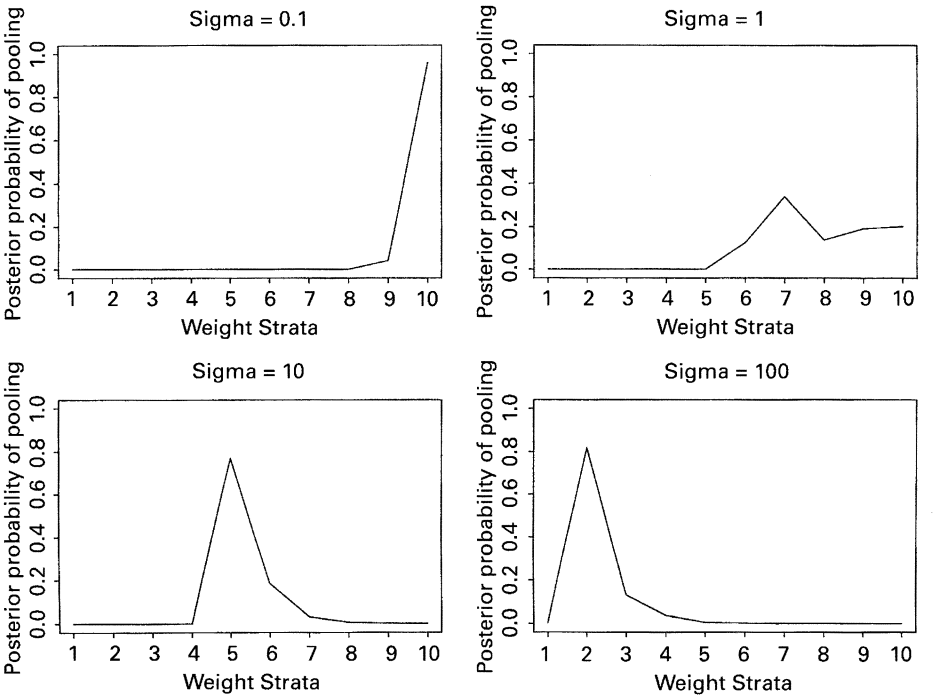

Fig. 2. Posterior probability of pooling the l largest weight strata as a function of variance when stratum means favor trimming

strata is inappropriate, and will lead to biased estimators. This will increase RMSE unless σ is very large relative to the between-strata mean differences and the sample size.

The top left panel of Figure 3 gives the RMSE relative to the fully weighted estimator under the assumption of normality, for this unfavorable case. The unweighted and crude trimming methods perform very poorly unless σ is large. The CWP estimator behaves well for small or large values of the variance. However, it is less satisfactory for intermediate values of σ, tending to overpool. The MSET estimator is more protective of overpooling, although it is less effective than CWP at reducing RMSE for large values of σ.

The CWP interval estimates have below nominal coverage when σ is moderate (bottom left panel of Figure 3). The MSET estimator displays coverage rates closer to nominal levels, but also has poorer coverage than the FWT estimator when variance is moderate. Both the MSE and coverage of the CWP estimator are improved for the small sample case.

The right panels of Figure 3 show results for contaminated normal data with an unfavorable mean configuration. The weight trimming estimators generally perform poorly relative to the fully-weighted estimator in large sample sizes, and the problems of overpooling for CWP are somewhat intensified.

3.2. *Weight smoothing methods*

For weight smoothing models, mean structures $\boldsymbol{\delta}^E$, $\boldsymbol{\delta}^L$, and $\boldsymbol{\delta}^D$ were considered with normal and contaminated normal errors. Five weighting schemes were considered: the

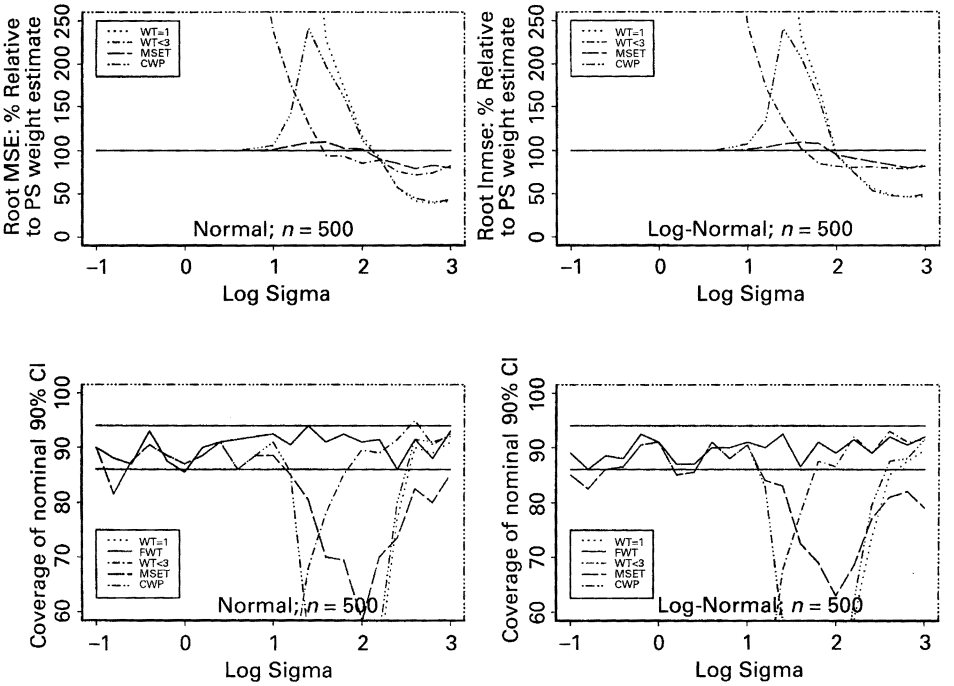

Fig. 3. Average RMSE relative to unweighted estimator and nominal 90% coverage (200 simulations) of unweighted (WT = 1), crude trimming (WT < 3), minimum MSE (MSET), and compound weight pooling (CWP) estimators, when stratum means do ***not*** *favor trimming*

fully-weighted estimator $\bar{y}_w$ (FWT), and posterior estimate of $\bar{Y}$ from (2.7) where $\hat{\mu}_h$ is obtained under the XRE, AR1, LIN, and NPAR models. Results for the $\boldsymbol{\delta}^C$ mean structure were similar to those for $\boldsymbol{\delta}^D$ for the weight smoothing methods considered, and hence are omitted.

With regard to variance estimation, a standard empirical Bayes analysis based on (2.14) through (2.17) yields

$$\begin{aligned} Var(\bar{Y}|\mathbf{y}) &= (\boldsymbol{N}-\boldsymbol{n})' Var(\hat{\boldsymbol{\mu}} - \bar{\boldsymbol{Y}}^*)(\boldsymbol{N}-\boldsymbol{n})/N_+^2 \\ &= (\boldsymbol{N}-\boldsymbol{n})'(\sigma^2\Lambda + ZGZ' + A(\sigma^2\Sigma + ZGZ')A' - 2AZGZ')(\boldsymbol{N}-\boldsymbol{n})/N_+^2 \end{aligned} \quad (3.1)$$

(Holt and Smith 1979; Lazzaroni and Little 1998), where Y^* is the set of unobserved values of Y, $(\boldsymbol{N}-\boldsymbol{n})$ is an $H \times 1$ vector of counts of the unobserved population $(N_h - n_h)$, $\Lambda = diag(N_h - n_n)^{-1}$, and $\hat{\boldsymbol{\mu}} = \hat{A}\bar{\mathbf{y}}$ is given by replacing $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{u}}$ in (2.18) when G and σ^2 are known. Specifically,

$$A = (I - ZGZ'V^{-1})X(X'V^{-1}X)^{-1}X'V^{-1} + ZGZ'V^{-1} \quad (3.2)$$

The estimator of $Var(\bar{Y}|\mathbf{y})$ given by replacing G and σ^2 in (3.1) and (3.2) with $\hat{G}$ and $\hat{\sigma}^2$ will be biased downward, since it ignores uncertainty in the estimates of G and σ^2 as fixed. A fully Bayesian analysis accounts for this uncertainty and should yield better estimates of $Var(\bar{Y}|\mathbf{y})$ and consequently better coverage properties of the posterior estimator of $\bar{Y}$. (See also Pfeffermann et al. 1998.) The t-type corrections suggested by Lazzaroni

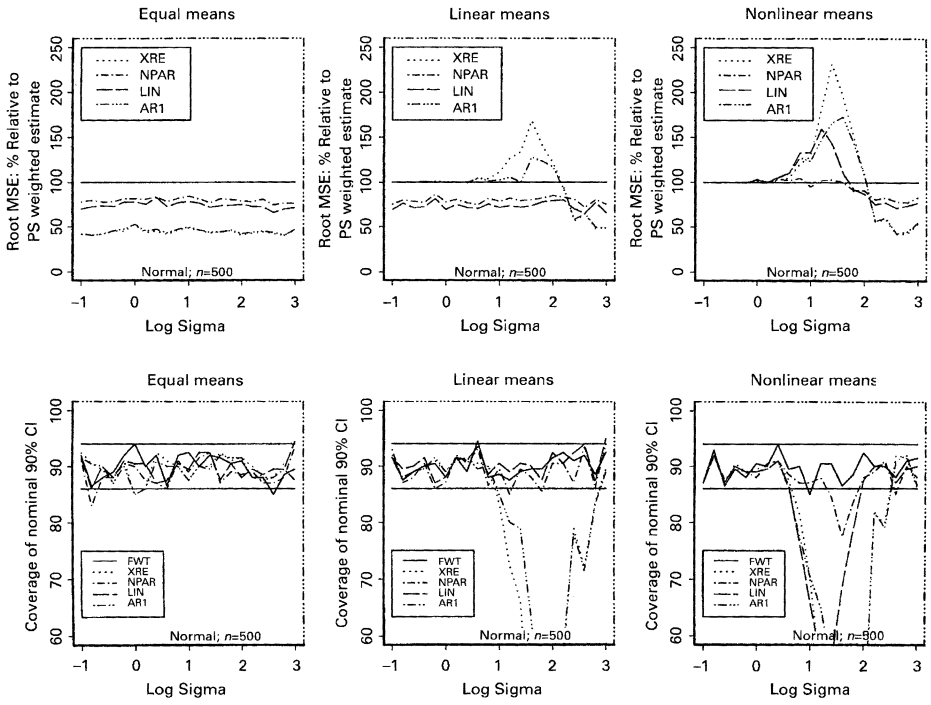

Fig. 4. Average RMSE relative to unweighted estimator and nominal 90% coverage (200 simulations) of exchangeable random effects (XRE), autoregressive (AR1), linear (LIN), and nonparametric (NPAR) weight smoothing estimators. Normally-distributed error terms

and Little (1998) to provide an approximate adjustment for variability in $\hat{G}$ and $\hat{\sigma}^2$ are overly conservative when H is small and are not considered here.

Figure 4 compares the RMSE for $\bar{y}_{XRE}$, $\bar{y}_{AR1}$, $\bar{y}_{LIN}$, and $\bar{y}_{NPAR}$ relative to $\bar{y}_w$ when the errors are normally-distributed. As expected, the XRE and AR1 estimators do well when the means are equal. Both these estimators perform poorly relative to $\bar{y}_w$ when the means are unequal, although as expected the AR1 estimator is more robust than the XRE estimator. The LIN estimator works well for both equal and linear mean structures, although as expected it is less efficient than XRE or AR1 when the stratum means are equal. For the nonlinear mean structure $\boldsymbol{\delta}^D$, LIN performs poorly relative to FWT for moderate σ^2. Since $\bar{y}_{NPAR} \rightarrow \bar{y}_w$ as the roughness penalty $\lambda \rightarrow 0$ and $\bar{y}_{NPAR} \rightarrow \bar{y}_{LIN}$ as $\lambda \rightarrow \infty$, NPAR can be viewed as a compromise between LIN and FWT; the simulations suggest that this compromise works well. Specifically, NPAR performs nearly as well as LIN for equal and linear mean structures. When the stratum means are nonlinear, NPAR mimics FWT for small to moderate values of σ^2, and mimics LIN as σ^2 increases and the RMSE of LIN is lower than the RMSE of FWT.

Figure 4 shows the coverage of the various weight smoothing estimators for different mean structures and variances. All estimators have good coverage properties when the true superpopulation means are equal, since all models allow equal means. The XRE and AR1 models yield intervals with poor coverage when the means follow a linear trend and variance is moderate. The LIN model has moderate coverage problems when the trend

is nonlinear. The NPAR and FWT procedures are close to nominal levels for all mean structures and all values of σ considered.

Examining of RMSE and coverage when outliers are present and the mean structure is nonlinear shows that results are similar to those obtained under the normal distribution, with low variance yielding design-type estimators and large variances yielding trimmed or unweighted estimators. For moderate variance, the XRE, AR1 and LIN models perform somewhat more poorly with respect to coverage than for uncontaminated normal data, while the NPAR estimator is more robust to contamination.

4. An Application: The National Comorbidity Survey

The National Comorbidity Survey (NCS) (Kessler 1992), conducted in 1990–1992, was a national face-to-face survey of the noninstitutionalized U.S. population (excluding Alaska and Hawaii) aged 15–54 regarding the prevalence, risk factors, and consequences of psychiatric morbidity and comorbidity. The 8,098 respondents were selected using probability methods from 1,205 block-level segments, with a response rate of 82.4%. The sample design included an oversample of persons aged 15–24, and a complex case weighting scheme to adjust for this oversample of young persons, together with adjustments for unequal household size, nonresponse, new construction, subsampling of difficult-to-reach respondents (''holdback''), and poststratification to National Health Interview Survey estimates in eight categories defined by various combinations of education, number of persons in household, place of residence, and ethnicity. The final weight, after normalization to sum to the sample size, ranged from 0.10 to 5.67, with a standard deviation of 0.97. More details on the construction of NCS weights are provided in Little et al. (1997).

Data from the NCS was utilized to determine which subjects had experienced depression as defined by the American Psychiatric Association (1987); 18.0% of the subjects were found to fit the criteria. We applied nine weighting methods to the estimated mean of the depression indicator for the entire sample of subjects and for the subset of African-American subjects. Estimates and associated estimated bias and mean squared error are displayed in Table 2. The estimated RMSE was calculated under the assumption that the fully weighted estimator $\bar{y}_w$ is unbiased. The true population mean is of course not known here, and the assumption that the fully weighted estimator is unbiased and the use of estimated RMSE as a criterion tends to favor the FWT and MSET methods. To account for the effects of the multi-stage sample design and the fact that the population strata sample sizes $\{N_h\}$ were estimated, the variances of the mean estimators were calculated using a jackknife repeated replications method (Wolter 1985). The estimated squared bias of a mean estimator $\bar{y}^*$ is given by $\max(\hat{B}^2 - \hat{V}_{01}, 0)$ where $\hat{B} = \bar{y}^* - \bar{y}_w$ and $\hat{V}_{01}$ is the variance of $\hat{B}$, estimated using the jackknife (Little et al. 1997).

The number of weight strata H varied from 492 to 3,944, with $1 \leq n_h \leq 447$, depending on the sample under consideration. Because the stratum pooling estimators require $n_h \geq 2$, some ''pre-pooling'' of weight strata was necessary to fit the stratum pooling models. While not required, this prepooling was also done for the weight smoothing estimators, both to allow direct comparisons of the methods, and to speed use of the jackknife estimator. We present results where strata were prepooled at the fifth and first percentiles, creating approximately 20 and 100 weight strata respectively.

Table 2. Estimated proportion of U.S. and African-American population aged 15–54 ever reporting symptoms of depression, with associated bias and root mean squared error for various weighted estimators

Mean Estimator	Est. Mean ($\times 10^{-3}$)		Est. Bias ($\times 10^{-4}$)		Est. RMSE ($\times 10^{-4}$)	
	$H = 20$	$H = 99$	$H = 20$	$H = 99$	$H = 20$	$H = 99$
U.S.-Wide						
$\bar{y}$	180	180	91	91	100	100
$\bar{y}_w$	171	171	0	0	67	67
$\bar{y}_{wt<3}$	170	170	−3	−3	67	67
$\bar{y}_{MSET}$	173	172	21	11	80	65
$\bar{y}_{CWP}$	173	176	21	51	53	73
$\bar{y}_{XRE}$	173	174	25	30	62	65
$\bar{y}_{AR1}$	173	174	26	30	71	68
$\bar{y}_{LIN}$	168	169	23	−20	64	68
$\bar{y}_{NPAR}$	170	170	−8	−6	65	66
African-American						
$\bar{y}$	128	128	84	84	108	108
$\bar{y}_w$	119	119	0	0	147	147
$\bar{y}_{wt<3}$	111	111	−85	−85	141	141
$\bar{y}_{MSET}$	120	121	9	21	138	130
$\bar{y}_{CWP}$	125	122	57	28	128	121
$\bar{y}_{XRE}$	128	125	84	57	117	108
$\bar{y}_{AR1}$	126	125	72	57	105	107
$\bar{y}_{LIN}$	114	116	−48	−34	126	131
$\bar{y}_{NPAR}$	114	115	−48	−47	126	128

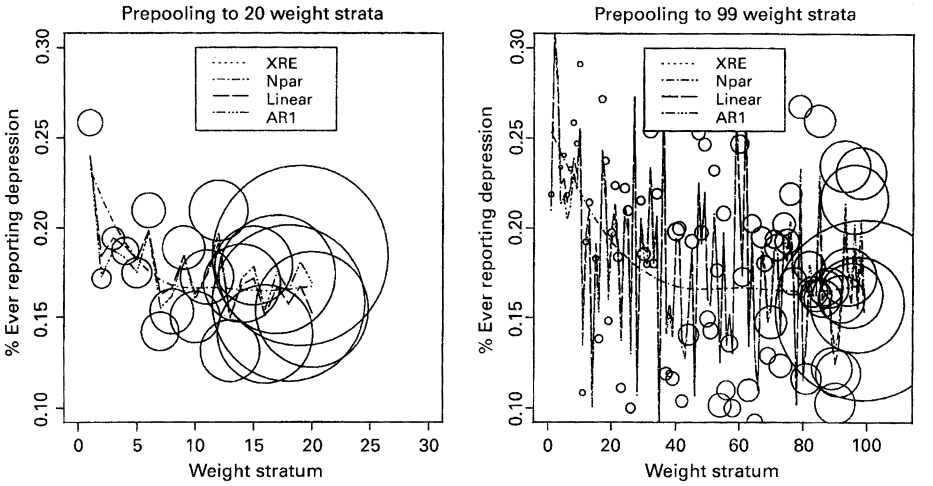

Fig. 5. Percent ever depressed versus inverse of probability of selection. Circles represent relative size of estimated population within each weight stratum, centered at sampled stratum mean. Lines represent posterior estimates of weight stratum mean

The proportion of the population N_h/N_+ in each weight stratum was unknown, and estimated from the fully-weighted sample by $\sum_i w_{hi}/n_+$ for the normalized weights w_{hi}. Also, because the sampling fraction was extremely small, the finite population correction was negligible and was ignored.

4.1. Depression indicator

The relationship between ever reporting depressive symptoms and probability of inclusion in Figure 5 indicates that pooling of the high-weight strata may be reasonable. Table 2 indicates that the CWP estimator of this quantity is quite effective in reducing RMSE. The crude trimming and MSET estimators are similar to the FWT estimator. Note that although the CWP and MSET estimators give similar point estimates when 20 weight strata were utilized, the jackknife-estimated variance of the MSET method is larger, since only a fixed trimming point could be selected for each jackknife sample. This difference diminishes when 99 weight strata are utilized.

Table 2 shows that the smaller sample size and relatively weak correlation between the weights and the depression indicator among African-Americans 15–54 gives MSE advantage to the unweighted estimator, which the CWP estimator approximates better than the other stratum pooling estimators. The approximate exchangeability of the means gives preference to the XRE and AR1 models, with the smaller sample size enhancing their MSE performance in the African-American subsample. The NPAR estimator approximates the LIN estimator as conditions are favorable for shrinkage, yielding substantial savings over the FWT estimator when the sample size is smaller.

5. Discussion

Survey weights are generally trimmed in an ad-hoc manner, with little attention given to the optimum degree of trimming. We have considered a number of methods that use the data to determine adjustments of the weights that involve appropriate bias-variance tradeoffs. One approach is to obtain an estimate of root mean squared error and then choose a trimming point that minimizes this estimate (Cox and McGrath 1981; Potter 1990). This method performed reasonably well in our simulations, although model-based methods were more efficient for some problems, and confidence intervals that fail to reflect uncertainty in the trimming point did not achieve nominal levels of coverage.

Our model-based procedures are divided into two classes, weight pooling models and weight smoothing models. The compound weight pooling model is proposed as a model-based analogue of weight trimming that allows Bayesian averaging over estimates based on different trimming points. Bayesian methodology also allows the uncertainty about the choice of trimming point to be included in the inference. In our empirical studies, this model did well in terms of RMSE when the mean configuration was favorable towards trimming, but tended to over-pool for some regions of the parameter space when the mean configuration was not favorable towards trimming. The over-pooling is even more problematic when confidence coverage, rather than MSE, is of interest, since the resulting bias results in intervals that are systematically shifted away from the population mean. An immediate extension of the CWP model would allow pooling of all possible combinations of the weight strata, or all possible pairs of adjacent strata. The former

involves considering

$$\sum_{l=0}^{H-2} \binom{H}{H-l}$$

pooling possibilities, and the latter 2^{H-1} possibilities, rather than the $H-1$ possibilities of our approach here. Since the number of possible pooling options quickly becomes large as H increases, a Monte Carlo method may be needed to handle the computations.

There is no compelling need to mimic trimming methods, and we also consider weight smoothing models that treat the unknown weight stratum means as random variables with their own mean and covariance structure (Holt and Smith 1979; Little 1991; Lazzaroni and Little 1998). Choosing between models in this class involves trade-offs between robustness and efficiency. In particular, assuming exchangeable means and including between-stratum variance components to induce shrinkage, as in the XRE and AR1 models, yields estimators that have good properties when the sample design is highly disproportional and the data highly variable, but that are vulnerable to model misspecification when the between-stratum and within-stratum variances are approximately equal. In contrast, adding parameters to the mean structure, as in the LIN and NPAR models, reduces the problem of misspecification at some cost in efficiency. The NPAR model has the advantage of being more ''believable'' when the strata are nominal rather than ordinal, so there is less reason to believe a linear trend exists in the data. It yields estimates that behave somewhat like the MSET estimator, but with more efficiency and better confidence coverage properties. Indeed, this model was nearly as robust to alternative mean configurations as the fully-weighted estimator in simulations, yet approximates the efficiency of the LIN model estimator when variance overwhelms bias. It also has the advantage of being more stable than other weight smoothing model estimators when weight strata are combined to speed model fitting.

Our discussion has focused on the finite population mean and models with Gaussian distributional assumptions. However simplistic, these assumptions nonetheless encompass a very substantial amount of survey analysis. The Central Limit Theorem renders the normality assumption for the stratum sample means tenable if the number of elements in each weight stratum is moderate or large. Otherwise, extensions of weight smoothing models to exponential family distributions via generalized linear mixed models (Zhang et al. 1998) can also be envisioned. The use of weight smoothing models to estimate parameters other than the mean, such as population regression coefficients, is another important area for further research.

Finally, these methods are computationally more complex than the standard estimates with full or crudely trimmed weights, and some practitioners may feel they are too complex for survey practice. However, modern computing power has made these more complex methods practicable on a production basis, and we feel that a robust model-based procedure such as NPAR might be safely applied on a routine basis, with better inferences resulting.

Appendix

One difficulty with Bayesian variable selection models is that the posterior probability that model l is correct is a Bayes factor transformed from an odds scale to a probability scale.

Bayes factors are ill-defined if the prior distributions are weakly or non-informative, and often have strange behavior if continuous alternative hypotheses exist, as in our model. Spike and slab prior distributions for the regression coefficients ($P(\beta_{lj} = 0|l) > 0$) (Mitchell and Beauchamp 1988), a "natural" choice for variable selection models in the regression context, require choice of an unknown latent hyperparameter, which is an undesirable feature in our context. Assuming instead continuous proper priors for $\beta_l|l$ and $\sigma^2|l$ of the form $p(\beta_l|l) = N(0, \sigma^2 \sum_l)$ and $p(\sigma^2|l) = \text{Inv-}\chi^2(\nu_l, v_l)$, it can be shown (Halpern 1973) that

$$p(L = l|y) = \frac{(\nu_1 v_1)/2)^{\nu_l/2}\Gamma(\nu_l'/2)|\Sigma_l'|^{1/2}}{(\nu_l' v_l')/2)^{\nu_l'/2}\Gamma(\nu_l/2)|\Sigma_l|^{1/2}}$$

where

$$\Sigma_l' = (\Sigma_l^{-1} + X_l^T X_l)^{-1}$$

$$\nu_l' = \nu_l + n$$

$$v_l' = 1/\nu_l'(\nu_l v_l + y^T y - \mu_l^T(\Sigma_l')^{-1}\mu_l)$$

$$\mu_1 = \Sigma_l' X_l^T y$$

Letting the prior degree-of-freedom parameters ν_l tend to 0 yields a relatively uninformative prior for the $\sigma^2|l$. However, letting Σ_l^{-1} tend to 0 to create a relatively uninformative prior for $\beta_l|l$ can yield unstable estimates of the posterior probabilities due to the $|\Sigma_l|^{1/2}$ term in the denominator. Using standard noninformative priors $p(\sigma^2|l) \propto \sigma^{-1}$ and $\rho(\beta_l|l) \propto 1$ yields a posterior estimator with bizarre degree-of-freedom behaviour (Atkinson 1978). In particular, if $H = 2$ and $E(y_{1_i}) = E(y_{2i}) = \beta_0$, the posterior probability of pooling will approach 0 as $\sigma^2 \to \infty$ for fixed n. That is, unless $\beta_1 = 0$ *and* there is sufficient evidence to allow pooling in the form of a small variance relative to the sample size, pooling will always be rejected in favor of the (incorrect) larger model. To counter the tendency of the posterior probability to overly favor the larger model, we considered Schwarz's likelihood (Schwarz 1978). However, we found that in practice use of this Bayesian Information Criterion (BIC) penalized likelihood overcorrected, yielding overpooled estimates.

Halpern (1973) provides an argument for $p(\beta_l|l) = (2\pi)^{-l}$ as a consequence of formulating improper priors with different dimensions. Dempster (1977) includes without comment $p(\sigma^2|l) = (1/\sigma^2)^{l/2+1}$ as one of the "57 varieties" of regression discussed. We suggest its use on practical grounds: it removes the sensitivity of the posterior pooling probability to σ^2 when the means are equal, and tends to favor pooling when the variances are large and there is little ability to distinguish between the means. Halpern (1973) argues that it is inappropriate to make the variances of different models different, but we think the residual variance σ^2 might be expected to be smaller when the number of different means is known to be larger. Yet another prior that might bear further inspection is that of Spiegelhalter and Smith (1982), who suggest inflating the probability of the lth model by a factor of $((H - l + 2)/2)^{1/2}$ based on a "minimum training sample" argument. Given the lack of consensus in the literature, we believe our choice of prior is at least as reasonable as others that have been suggested.

6. References

Alexander, C.H., Dahl, S., and Weidman, L. (1997). Making Estimates From the American Community Survey. Paper presented at the 1997 Joint Statistical Meetings, Anaheim, CA.

American Psychiatric Association (1987). Diagnostic and Statistical Manual of Mental Disorders DSM-III-R, Third Edition, Revised. Washington, DC: American Psychiatric Association.

Atkinson, A.C. (1978). Posterior Probabilities for Choosing Regression Models. Biometrika, 65, 39–48.

Berund, C.H. et al. (1978). Final Report on National Assessment of Education Progress: Sampling and Weighting Activities for Assessment Year 08. Research Triangle Park, North Carolina: National Assessment of Education Progress.

Cox, B.G. and McGrath, D.S. (1981). An Examination of the Effect of Sample Weight Truncation on the Mean Squared Error of Survey Estimates. Paper presented at the 1981 Biometric Society ENAR meeting, Richmond, VA.

Dempster, A.P., Schatzoff, M., and Wermuth, N. (1977). A Simulation Study of Alternatives to Ordinary Least Squares (with discussion). Journal of the American Statistical Association, 72, 77–106.

Ghosh, M. and Meeden, G. (1986). Empirical Bayes Estimation of Means from Stratified Samples. Journal of the American Statistical Association, 81, 1058–1062.

Green, P.J. (1987). Penalized Likelihood for General Semi-Parametric Regression Models. International Statistical Review, 55, 245–260.

Halpern, E.F. (1973). Polynomial Regression from a Bayesian Approach. Journal of the American Statistical Association, 68, 137–143.

Hastie, T.J. and Tibshirani, R.J. (1990). Generalized Additive Models, London: Chapman and Hall.

Holt, D. and Smith, T.M.F. (1979). Poststratification. Journal of the Royal Statistical Society, A142, 33–46.

Horvitz, D.G. and Thompson, D.J. (1952). A Generalization of Sampling Without Replacement from a Finite Universe. Journal of the American Statistical Association, 47, 663–685.

Kessler, R. (1992). National Comorbidity Survey. Ann Arbor, MI: Survey Research Center, Institute for Social Research.

Kish, L. (1992). Weighting for Unequal P_i. Journal of Official Statistics, 8, 183–200.

Laird, N.M. and Ware, J.H. (1982). Random Effects Models for Longitudinal Data. Biometrics, 38, 963–974.

Lazzeroni, L.C. and Little, R.J.A. (1998). Random Effects Models for Smoothing Post-Stratification Weights. Journal of Official Statistics, 14, 61–78.

Little, R.J.A. (1983) Estimating a Finite Population Mean from Unequal Probability Samples. Journal of the American Statistical Association, 78, 596–604.

Little, R.J.A. (1991). Inference with Survey Weights. Journal of Official Statistics, 7, 405–424.

Little, R.J.A. (1993). Poststratification: A Modeler's Perspective. Journal of the American Statistical Association, 88, 1001–1012.

Little, R.J.A., Lewitzky, S., Heeringa, S., Lepkowski, J., and Kessler, R.C. (1997). Assessment of Weighting Methodology for the National Comorbidity Survey. American Journal of Epidemiology, 146, 439–449.

Mitchell, T.J. and Beauchamp, J.J. (1988). Bayesian Variable Selection in Linear Regression. Journal of the American Statistical Association, 83, 1023–1032.

Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H., and Rasbash, J. (1998). Weighting for Unequal Selection Probabilities in Multilevel Models (with discussion). Journal of the Royal Statistical Society, B60, 23–56.

Potter, F. (1990). A Study of Procedures to Identify and Trim Extreme Sample Weights. Proceedings of the Survey Research Methods Section, American Statistical Association, 225–230.

Robinson, G.K. (1991). That BLUP is a Good Thing: the Estimation of Random Effects (with discussion). Statistical Science, 6, 15–51.

Rubin, D.B. (1983). Comment: Probabilities of Selection and Their Role for Bayesian Modeling in Sample Surveys. Journal of the American Statistical Association, 78, 803–805.

Schwarz, G. (1978). Estimating the Dimension of a Model. The Annals of Statistics, 6, 461–464.

Spiegelhalter, D.J. and Smith, A.F.M. (1982). Bayes Factors for Linear and Log-Linear Models with Vague Prior Information. Journal of the Royal Statistical Society, B44, 377–387.

Wahba, G. (1978). Improper Priors Spline Smoothing, and the Problem of Guarding Against Model Errors in Regression. Journal of the Royal Statistical Society, B40, 364–372.

Wahba, G. (1985). A Comparison of GCV and GML for Choosing the Smoothing Parameters in the Generalized Spline Smoothing Problem. The Annals of Statistics, 4, 1378–1402.

Wang, Y. (1998). Smoothing Spline Models with Correlated Random Errors. Journal of the American Statistical Association, 93, 341–348.

Wolter, K. (1985). Introduction to Variance Estimation. New York: Springer-Verlag.

Zhang, D., Lin, X., Raz, J., and Sowers, M. (1998). Semiparametric Stochastic Mixed Models for Longitudinal Data. Journal of the American Statistical Association, 93, 710–719.

Received September 1999
Revised May 2000