

## Model Averaging Methods for Weight Trimming

Michael R. Elliott<sup>1</sup>

In sample surveys where sampled units have unequal probabilities of inclusion, associations between the inclusion probabilities and the statistic of interest can induce bias. Weights equal to the inverse of the probability of inclusion are often used to counteract this bias. Highly disproportional sample designs have highly variable weights, which can introduce undesirable variability in statistics such as the population mean or linear regression estimates. Weight trimming reduces large weights to a fixed maximum value, reducing variability but introducing bias. Most standard approaches are ad-hoc in that they do not use the data to optimize bias-variance tradeoffs. This manuscript develops variable selection models, termed “weight pooling” models, that extend weight trimming procedures in a Bayesian model averaging framework to produce “data driven” weight trimming estimators. We develop robust yet efficient models that approximate fully-weighted estimators when bias correction is of greatest importance, and approximate unweighted estimators when variance reduction is critical.

*Key words:* Sample survey; sampling weights; Bayesian population inference; weight pooling; variable selection; fractional Bayes Factors.

### 1. Introduction

Analyses of data from samples designed to have differential probabilities of inclusion typically use case weights equal to the inverse of the probability of inclusion to reduce bias in the estimators of population quantities of interest. An example is the Horvitz-Thompson estimator (Horvitz and Thompson 1952) of a population mean  $\bar{Y} = N^{-1} \sum_{i=1}^N y_i$  given by  $\hat{Y} = N^{-1} \sum_{i \in s} w_i y_i$ , where  $w_i = 1/\pi_i$ ,  $\pi_i$  is the probability of inclusion and  $s$  is the subset of the population units sampled. This fully-weighted estimator is unbiased for the population mean. For the wide class of nonlinear estimators such as ratio estimators or linear regression slopes that are functions of linear statistics, bias can be reduced and consistent estimates of population values obtained by replacing implicit means or totals with their weighted equivalents (Binder 1983).

It is generally accepted that sampling weights should be utilized when considering descriptive statistics such as means and totals, although even here, highly variable probabilities of selection can give rise to bias-variance tradeoffs and the desire to employ

<sup>1</sup> Department of Biostatistics, School of Public Health, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109, USA., and Survey Methodology Program, Institute for Social Research, 426 Thompson St., Ann Arbor, MI 48106, U.S.A. Email: mreliott@umich.edu

**Acknowledgments:** This research was supported by National Institute of Heart, Lung, and Blood grant R01-HL-068987-01. The author acknowledges Jack Chen for his assistance with programming, and Dr. Andrew Tershakovec for his assistance with the Barker’s Hypothesis analysis, as well as the Editor, Associate Editor, and three anonymous reviewers whose comments improved the manuscript.

weight trimming (Little et al. 1997). However, when estimating “analytical” models (Cochran 1977, p.4) that focus on associations between, e.g., risk factors and health outcomes estimated via linear and generalized linear models, the decision to use sampling weights is less definitive (cf. Korn and Gaubard 1999, pp.180–182). Consider a population generated from

$$Y_i = A + BX_i + CX_i^2 + \epsilon_i, X_i \sim U(0, 1), \epsilon_i \stackrel{iid}{\sim} N(0, 1) \quad (1.1)$$

where the superpopulation model of interest is the conditional distribution of  $Y_i$  given  $X_i$ , modeled by

$$Y_i = \alpha + \beta X_i + \epsilon_i, \epsilon_i \stackrel{iid}{\sim} N(0, \sigma^2) \quad (1.2)$$

The superpopulation model is correctly specified when  $C = 0$  and misspecified when  $C \neq 0$ . We consider two sampling schemes: an ignorable sampling scheme that oversamples large values of  $X_i$ , and a nonignorable scheme that oversamples large values of  $Y_i$  at a given value of  $X_i$ . The sampling scheme is ignorable in the regression context when the sampling probability is a function of  $X_i$  only and thus the inclusion indicator  $I_i$  is independent of  $Y_i|X_i$  because our goal is to determine the distribution of  $Y|X$ ; nonignorable designs in the regression setting retain an association between  $Y_i$  and  $I_i$ , even conditional on  $X_i$ . Of course, designs in which  $I_i$  depends on  $X_i$  are nonignorable for parameters that describe the marginal distribution of  $Y_i$ , unless  $Y_i \perp X_i$  (see Section 2). We assume that the goal of the modeler is to describe the association between  $Y$  and  $X$  using the regression slope  $\beta$  from the superpopulation model. If the superpopulation model is correctly specified, the target quantity of interest could be either the superpopulation slope or the population slope defined by  $B = \sum_{i=1}^N A_i(Y_i - \bar{Y})$ , where  $A_i = (X_i - \bar{X})/\sum_{i=1}^N (X_i - \bar{X})^2$ ,  $\bar{Y} = N^{-1}\sum_{i=1}^N Y_i$ ,  $\bar{X} = N^{-1}\sum_{i=1}^N X_i$  (the “corresponding descriptive population quantity” in Pfeffermann (1993)). If the superpopulation model is misspecified, then only the population slope makes sense as a target quantity. The unweighted ordinary least squares (OLS) estimator and (case-)weighted least squares (WLS) estimator of  $\alpha$  and  $\beta$  respectively are given by

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = (X^T S X)^{-1} X^T S Y$$

$$\begin{pmatrix} \hat{\alpha}^w \\ \hat{\beta}^w \end{pmatrix} = (X^T S^w X)^{-1} X^T S^w Y$$

where the  $i$ th row of  $X$ ,  $X_i^T$ , is given by  $(1 X_i)^T$ ,  $S$  is a diagonal matrix of sample inclusion indicators  $S_i$ , and  $S^w$  replaces  $S_i$  in  $S$  with  $S_i/\pi_i$ . Thus, the WLS estimator replaces the means and totals in the unweighted estimator with the Horvitz-Thompson equivalents.

Table 1 shows the results from 500 simulations for equivalent populations of  $N = 10,000$ , under correctly specified and misspecified models and ignorable and non-ignorable sample designs, for sample sizes of  $n = 50$  and  $n = 500$ . When the sample design is ignorable (probability of selection depends only on  $X$ ) and the mean model correctly specified, both the unweighted and fully-weighted estimators are essentially

Table 1. % Bias (MSE in parentheses) for population slope for population generated under  $Y_i|X_i \sim N(A + BX_i + CX_i^2, 1)$ ,  $i = 1, \dots, 10,000$ , and superpopulation model is given by  $Y_i|X_i \sim N(\alpha + \beta X_i, \sigma^2)$ : correctly specified ( $A = 0, B = 2, C = 0$ ), misspecified ( $A = 0, B = 2, C = -1$ ). Sampling design ignorable for population slope ( $P(S_i|Y_i, X_i) \propto X_i^{.75}$ ) or nonignorable for population slope ( $P(S_i|Y_i, X_i) \propto \exp(.5Y_i/(X_i + .25) - 1)$ ). Results from 500 simulations

	Sampling ignorable?			
	Yes		No	
	Superpopulation model correctly specified?			
	Yes	No	Yes	No
$n = 50$				
$\hat{\beta}$	-0.2(.168)	-15.4(.203)	-30.8(.511)	-63.8(.518)
$\hat{\beta}^w$	0.7(.264)	-3.8(.274)	-5.1(.239)	-10.5(.303)
$n = 500$				
$\hat{\beta}$	-0.5(.015)	-16.3(.040)	-28.6(.339)	-55.7(.321)
$\hat{\beta}^w$	-0.8(.033)	-0.3(.032)	-.7(.039)	0.0(.041)

unbiased, and the larger variance of the weighted estimator results in a larger mean squared error (MSE). When the sampling is ignorable but the mean model incorrectly specified (linear instead of quadratic), the weighted estimator provides protection against model misspecification, but can introduce large variability into the estimator (note the larger MSEs for the weighted estimator when  $n = 50$ ). When the sample design is nonignorable for the population slope, the weighted population slope estimator  $\hat{\beta}^w$  accounts for the underrepresentation of smaller values of  $Y$  when  $X$  is small, reducing the negative bias in the slope; in these simulations this bias in the unweighted estimator was a greater contributor to MSE than the variance from the weighted estimator.

The fully-weighted estimators  $(\hat{\alpha}^w \hat{\beta}^w)^T$  are sometimes termed “pseudo-maximum likelihood” estimators (PMLEs) (Binder 1983; Pfeiffermann 1993) because they are “design consistent” for the MLEs that would solve the score equations under the sampling model defined in (1.2) if we had observed data for the entire population:

$$U(\beta) = \sum_{i=1}^N \frac{d}{d\beta} \log f(y_i|\beta) = \sum_{i=1}^N X_i(X_i^T \beta - y_i) = 0 \tag{1.3}$$

In brief, design consistency implies that the difference between the population target quantity and the estimate derived from the sample tends to zero as the sample size and population size jointly increase, or that this difference will on average tend to 0 from repeated sampling of the population, where samples are selected in an identical fashion from  $t \rightarrow \infty$  replicates of the population: see Särndal (1980) or Isaki and Fuller (1982).

### 1.1. Weight Trimming

While PMLEs are popular in practice for the reasons discussed above, their bias reduction typically comes at the cost of increased variance. This increase can overwhelm the reduction in bias, so that the mean squared error (MSE) actually increases under a weighted analysis, as in the example in Table 1. Even in cases where disproportional sample designs do reduce

variance, as in “optimal allocation” where strata with more variable outcomes are oversampled (Kish 1965), designs that are optimal for one outcome may not be optimal for another, or for examination of associations (e.g., regression models).

Perhaps the most common approach to dealing with this problem is *weight trimming* or *winsorization* (Potter 1990; Kish 1992; Alexander et al. 1997), in which weights larger than some value  $w_0$  are fixed as  $w_0$ . Thus bias is introduced to reduce variance, with the goal of an overall reduction in MSE. This manipulation of the weights reflects a traditional design-based approach to survey inference.

Other design-based methods have been considered in the literature. Potter (1990) discusses systematic methods for choosing  $w_0$ , including the weight distribution 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 MSE trimming procedure (Cox and McGrath 1981) determines the empirical MSE at a variety of trimming levels  $t = 1, \dots, T$  under the assumption that the true population mean is given by the fully weighted estimate:  $\hat{MSE}_t = (\hat{\theta}_t - \hat{\theta}_Y)^2 + \hat{V}(\hat{\theta}_t)$ , where  $t = 1$  corresponds to the unweighted data and  $t = T$  to the fully-weighted data, and  $\hat{\theta}_t$  is the value of the statistic using the trimmed weights at level  $t$ . The trimming level is then given by the level  $l$  minimized  $\hat{MSE}_t$  over  $t$ .

In addition to adjusting for unequal probabilities of selection, case weights are also used to calibrate sample elements to known control totals in the population (Deville and Särndal 1992), either jointly (poststratification weights) or marginally (raking weights). In the calibration literature, techniques have been developed that allow generalized poststratification or raking adjustments to be bounded to prevent the construction of extreme weights (Folsom and Singh 2000). Beaumont and Alavi (2004) extend this idea to develop estimators that focus on trimming large weights of highly influential or outlying observations. While these bounds trim extreme weights to a fixed outpoint value, the choice of this cutpoint remains arbitrary. Another approach is to consider robust regression estimates (Hampel et al. 1986) that downweight highly influential observations, although applications which consider downweighting influence statistics as an alternative to weight trimming in the context of survey designs are limited (Zaslavsky et al. 2001 considered their use with ratio estimators).

This article develops an alternative approach to weight trimming that considers the case weights as stratifying variables within strata defined by the probability of inclusion. These “inclusion strata” may correspond to formal strata from a disproportional stratified sample design, or may be “pseudo-strata” based on collapsed or pooled weights derived from selection, poststratification, and/or nonresponse adjustments. Ordering these weight strata by the inverse of the probability of selection and collapsing together the largest valued strata mimics weight trimming by assuming the underlying data from these combined strata are exchangeable (conditional on any covariates of interest). In a regression setting, this model can be posed as a variable selection problem, where dummy variables for the inclusion strata interact with the regression parameters; subtracting from or adding to the inclusion strata design matrix allows for a larger or lesser degree of weight trimming. By averaging over all possible of these “weight pooling” models, we can compute an

estimator of the population parameter of interest whose bias-variance tradeoff is data-driven. By allowing for all contiguous inclusion strata (strata whose weights are closest in value) to be considered for pooling, we induce a high degree of robustness into our model, protecting against the over-pooling that simpler models suffered from (Elliott and Little 2000). We embed this model in a Bayesian framework, as we believe it provides a natural setting for model averaging, as well as a proper framework for population inference.

Section 2 reviews Bayesian finite population inference. Section 3 develops the weight pooling models for linear regression models in a fully Bayesian setting. Section 4 provides simulation results to determine the repeated sampling properties of the weight pooling estimators of linear regression parameters in a disproportionately stratified sample design and compares them with standard design-based estimators. Section 5 illustrates the use of the weight-pooling estimator using data from the National Health and Nutrition Examination Survey to consider evidence for “Barker’s Hypothesis” that low birth weight babies are at greater risk for cardiovascular disease later in life. Section 6 summarizes the results of the simulations and considers extensions to generalized linear models.

## 2. Bayesian Finite Population Inference

Let the population data for a population with  $i = 1, \dots, N$  units be given by  $Y = (y_1, \dots, y_N)$ , with associated covariate vectors  $X = (\mathbf{x}_1, \dots, \mathbf{x}_N)$  and sampling indicator variable  $I = (I_1, \dots, I_N)$ , where  $I_i = 1$  if the  $i$ th element is sampled and 0 otherwise. Similar to design-based population inference, Bayesian population inference focuses on population quantities of interest  $Q(Y)$ , such as population means  $Q(Y) = \bar{Y}$ . In contrast to design-based inference, however, one posits a model for the population data  $Y$  as a function of parameters  $\boldsymbol{\theta} : Y \sim f(Y|\boldsymbol{\theta})$ . Inference about  $Q(Y)$  is made based on the posterior predictive distribution of  $p(Y_{nob}|Y_{obs}, I)$ , where  $Y_{nob}$  consists of the elements of  $Y_i$  for which  $I_i = 0$ :

$$p(Y_{nob}|Y_{obs}, I) = \frac{\int \int p(Y_{nob}|Y_{obs}, \boldsymbol{\theta}, \phi) p(I|Y, \boldsymbol{\theta}, \phi) p(Y_{obs}|\boldsymbol{\theta}) p(\boldsymbol{\theta}, \phi) d\boldsymbol{\theta} d\phi}{\int \int p(Y_{nob}|Y_{obs}, \boldsymbol{\theta}, \phi) p(I|Y, \boldsymbol{\theta}, \phi) p(Y_{obs}|\boldsymbol{\theta}) p(\boldsymbol{\theta}, \phi) d\boldsymbol{\theta} d\phi} \quad (2.1)$$

where  $\phi$  models the inclusion indicator. If we assume that  $\phi$  and  $\boldsymbol{\theta}$  are a priori independent and if the distribution of sampling indicator  $I$  is independent of  $Y$ , the sampling design is said to be “unconfounded” or “noninformative”; if the distribution of  $I$  depends only on  $Y_{obs}$ , then the sampling mechanism is said to be “ignorable” (Rubin 1987), equivalent to the standard missing data terminology (the unobserved elements of the population can be thought of as missing by design).

Under ignorable sampling designs  $p(\boldsymbol{\theta}, \phi) = p(\boldsymbol{\theta})p(\phi)$  and  $p(I|Y, \boldsymbol{\theta}, \phi) = p(I|Y_{obs}, \phi)$ , and thus (2.1) reduces to

$$\frac{\int p(Y_{nob}|Y_{obs}, \boldsymbol{\theta}) p(Y_{obs}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int \int p(Y_{nob}|Y_{obs}, \boldsymbol{\theta}) p(Y_{obs}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} dY_{obs}} = p(Y_{nob}|Y_{obs})$$

allowing inference about  $Q(Y)$  to be made without explicitly modeling the sampling inclusion parameter  $I$  (Ericson 1969; Holt and Smith 1979; Little 1993; Rubin 1987; Skinner et al. 1989). In the regression setting, where inference is desired about parameters

that govern the distribution of  $Y$  conditional on fixed and known covariates  $X$ , (2.1) becomes

$$p(Y_{nob}|Y_{obs}, X, I) = \frac{\int \int p(Y_{nob}|Y_{obs}, X, \boldsymbol{\theta}, \phi) p(I|Y, X, \boldsymbol{\theta}, \phi) p(Y_{obs}|X, \boldsymbol{\theta}) p(\boldsymbol{\theta}, \phi) d\boldsymbol{\theta} d\phi}{\int \int \int p(Y_{nob}|Y_{obs}, X, \boldsymbol{\theta}, \phi) p(I|Y, X, \boldsymbol{\theta}, \phi) p(Y_{obs}|X, \boldsymbol{\theta}) p(\boldsymbol{\theta}, \phi) d\boldsymbol{\theta} d\phi dY_{nobs}}$$

which reduces to

$$p(Y_{nob}|Y_{obs}, X) = \frac{\int p(Y_{nob}|Y_{obs}, X, \boldsymbol{\theta}) p(Y_{obs}|X, \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int \int p(Y_{nob}|Y_{obs}, X, \boldsymbol{\theta}) p(Y_{obs}|X, \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta} dY_{nobs}}$$

if and only if  $I$  depends only on  $(Y_{obs}, X)$ , of which dependence on  $X$  only is a special case. Thus if inference is desired about a regression parameter  $Q(Y, X)|X$ , then a noninformative or more generally ignorable sample design can allow inclusion to be a function of the fixed covariates.

### 2.1. Accommodating Unequal Probabilities of Selection

Maintaining the ignorability assumption for the sampling mechanism often requires accounting for the sample design in both the likelihood and the prior model structure. In the case of the disproportional probability-of-inclusion sample designs, this can be accomplished by developing an index  $h = 1, \dots, H$  of the probability of inclusion (Little 1983, 1991); this could either be a one-to-one mapping of the case weight order statistics to their rankings, or a preliminary “pooling” of the case weights using, e.g., the  $100/H$  percentiles of the case weights. 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  $h = 1, \dots, H$ . We assume here that  $N_h$  is known, as when the weight strata come from a stratified random sample. (If  $N_h$  is unknown, as would be the case when the weights are constructed from estimated probabilities of inclusion via calibration or nonresponse adjustments, it can be replaced with  $\hat{N}_h = n_h w_h$ .  $\hat{N}_h$  can be treated as known, or if the underlying within-stratum samples are small, uncertainty in  $\hat{N}_h$  can be incorporated into the model by treating  $n_1, \dots, n_H$  as a multinomial distribution of size  $n$  parameterized by unknown inclusion stratum probabilities  $q_1, \dots, q_H$  with, e.g., a Dirichlet prior (Lu and Gelman 2003). Draws of  $N_h^*$  could then be obtained as  $N w_h q_h^*/n$ , where  $q_h^*$  is drawn from the Dirichlet posterior for  $q$ . If the weights within a stratum are not all equal, then  $w_h$  can be approximated by the inverse of the mean probability of inclusion with the stratum given by  $n_h / \sum_{i \in h} w_{hi}^{-1}$ ). The data are then modeled by

$$y_{hi} | \boldsymbol{\theta}_h \sim f(y_{hi}; \boldsymbol{\theta}_h), i = 1, \dots, N_h$$

for all elements in the  $h$ th inclusion stratum, where  $\boldsymbol{\theta}_h$  allows for an interaction between the model parameter(s)  $\boldsymbol{\theta}$  and the inclusion stratum  $h$ . Putting a noninformative prior distribution on  $\boldsymbol{\theta}_h$  then reproduces a fully-weighted analysis with respect to the expectation of the posterior predictive distribution of  $Q(Y)$ .

### 3. Weight Pooling Models

Weight trimming effectively pools units with large weights by assigning them a common, trimmed weight. The untrimmed (design-based) weighted mean estimator in a disproportionally stratified design is then  $\bar{y}_w = \sum_h \sum_i w_h y_{hi} / \sum_h \sum_i w_h = \sum_h (N_h / N_+) \bar{y}_h$ , where  $N_+ = \sum_h N_h$ , the total population. 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  $\gamma = (N_+ - \sum \kappa_i w_o) / \sum (1 - \kappa_i) w_i$ , where  $\kappa_i$  is an indicator variable for whether or not  $w_i \geq w_o$ . The trimmed mean estimator is thus given by

$$\bar{y}_{wt} = \sum_{h=1}^{l-1} \gamma \frac{N_h}{N_+} \bar{y}_h + \sum_{h=l}^H \frac{w_0 n_h}{N_+} \bar{y}_h = \gamma \sum_{h=1}^{l-1} \frac{N_h}{N_+} \bar{y}_h + \frac{w_0 \sum_{h=l}^H n_h}{N_+} \bar{y}^{(l)}$$

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

$$y_{hi} | \mu_h \sim N(\mu_h, \sigma^2) \quad h < l \tag{3.1}$$

$$y_{hi} | \mu_l \sim N(\mu_l, \sigma^2) \quad h \geq l$$

$$\mu_h, \mu_l, \log \sigma \propto \text{const.}$$

Elliott and Little (2000) considered an extension of this model where we no longer assume the cutpoint  $l$  is known:

$$y_{hi} | \mu_h \sim N(\mu_h, \sigma^2) \quad h < l$$

$$y_{hi} | \mu_l \sim N(\mu_l, \sigma^2) \quad h \geq l$$

$$p(L = l) = 1/H$$

$$p(\sigma^2 | L = l) = \sigma^{-(l+1/2)}$$

$$p(\beta | \sigma^2, L = l) = (2\pi)^{-l}$$

where  $\mu_1 = \beta_0, \dots, \mu_l = \beta_0 + \beta_{l-1}$ . This “weight pooling” model averages the estimators obtained from all possible weight trimming cutpoints, where each estimator contributes to the final average based on the probability that the cutpoint is “correct.” This posterior probability is determined via Bayesian variable selection models that determine the posterior probability of each cutpoint model conditional on the observed data.

#### 3.1. Weight Pooling Models for Linear Regression

This manuscript extends Elliott and Little (2000) in two ways. First, we consider the linear regression of  $Y_i$  on fixed covariates  $\mathbf{x}_i$ . Thus the most general model must allow for interactions between the probability of selection and the linear regression slopes; the full interaction model (a different slope within each probability-of-selection stratum,

equivalent to no pooling) approximately reproduces the fully-weighted estimator, while the minimal model (a single slope across all probability-of-selection strata, equivalent to full pooling) approximately reproduces the unweighted estimator. Pooling of some, but not all, of the strata, reproduces the trimmed estimator where the degree of trimming is determined by the degree to which the data suggest that distinct probability-of-selection strata have similar linear regression slopes. Second, we allow for the pooling of all conterminous inclusion strata. This increases the robustness of the model, by permitting the lowest probability-of-selection strata to interact with the linear regression slopes even when higher probability-of-selection strata are pooled. Thus

$$y_{hi} | \mathbf{x}_{hi}, \boldsymbol{\beta}_l, \sigma^2, L = l \stackrel{\text{ind}}{\sim} N(\mathbf{Z}_{li}^T \boldsymbol{\beta}_l, \sigma^2) \quad (3.2)$$

$$\boldsymbol{\beta}_l | \sigma^2, L = l \sim N(\boldsymbol{\beta}_0, \sigma^2 \boldsymbol{\Sigma}_0)$$

$$\sigma^2 | L = l \sim \text{Inv} - \chi^2(a, s^2)$$

$$p(L = l) = 2^{-(H-1)}$$

where  $\mathbf{Z}_{li} = D_{hl} \otimes \mathbf{x}_{hi}$  and  $D_{hl}$  is a vector of dummy variables that pool the appropriate conterminous inclusion strata based on the  $l$ th pooling pattern.

Table 2 shows the set of pooling patterns when  $H = 4$ . Under weak or noninformative priors, the first four pooling strata mimic standard weight trimming estimators, with  $L = 1$

Table 2. The set of  $\{D_{hl}\}$  when four weight strata are present: all patterns of pooling coterminous strata

Pooling pattern index	Dummy variable pattern	Number of pooled strata
$L = 1$ (complete pooling)	$D_{hl} = (1)$ for all $h$	$H^* = 1$
$L = 2$ (pool highest three weight strata)	$D_{hl} = (1\ 0)$ for all $h = 1$ $D_{hl} = (0\ 1)$ for $h \geq 2$	$H^* = 2$
$L = 3$ (pool highest two weight strata)	$D_{hl} = (1\ 0\ 0)$ for $h = 1$ $D_{hl} = (0\ 1\ 0)$ for $h = 2$ $D_{hl} = (0\ 0\ 1)$ for $h \geq 3$	$H^* = 3$
$L = 4$ (no pooling)	$D_{hl} = (1\ 0\ 0\ 0)$ for $h = 1$ $D_{hl} = (0\ 1\ 0\ 0)$ for $h = 2$ $D_{hl} = (0\ 0\ 1\ 0)$ for $h = 3$ $D_{hl} = (0\ 0\ 0\ 1)$ for $h = 4$	$H^* = 4$
$L = 5$ (pool all but highest weight stratum)	$D_{hl} = (1\ 0)$ for $h \leq 3$ $D_{hl} = (0\ 1)$ for $h = 4$	$H^* = 2$
$L = 6$ (pool first and last two weight strata)	$D_{hl} = (1\ 0)$ for $h \leq 2$ $D_{hl} = (0\ 1)$ for $h \geq 3$	$H^* = 2$
$L = 7$ (pool middle two strata)	$D_{hl} = (1\ 0\ 0)$ for $h = 1$ $D_{hl} = (0\ 1\ 0)$ for $h = 2, 3$ $D_{hl} = (0\ 0\ 1)$ for $h = 4$	$H^* = 3$
$L = 8$ (pool lowest two weight strata)	$D_{hl} = (1\ 0\ 0)$ for $h \leq 2$ $D_{hl} = (0\ 1\ 0)$ for $h = 3$ $D_{hl} = (0\ 0\ 1)$ for $h = 4$	$H^* = 3$

corresponding to an unweighted analysis and  $L = 4$  corresponding to a fully-weighted analysis.

Our population quantity of interest  $\mathbf{B} = (B_1, \dots, B_p)^T$  is the slope that solves the population score Equation (1.3) where

$$U_N(\mathbf{B}) = \sum_{i=1}^N \frac{\partial}{\partial \mathbf{B}} \log f(y_i; \mathbf{B}) = \sum_{i=1}^N -\frac{1}{\sigma^2} (y_i - \mathbf{x}_i^T \mathbf{B}) \mathbf{x}_i = 0$$

or

$$\mathbf{B} = \left( \sum_{i=1}^N \mathbf{x}_i \mathbf{x}_i^T \right)^{-1} \left( \sum_{i=1}^N \mathbf{x}_i y_i \right)$$

Note that the quantity  $\mathbf{B}$  such that  $U(\mathbf{B}) = 0$  is always a meaningful population quantity of interest even if the model is misspecified (i.e.,  $y_i$  is not exactly linear with respect to the covariates), since it is the linear approximation of  $\mathbf{x}_i$  to  $E(Y_i|\mathbf{x}_i)$ .

The posterior predictive distribution of  $\mathbf{B}$  is then given by

$$p(\mathbf{B}|y, X) = \sum_l \int p(\mathbf{B}|y, X, \boldsymbol{\theta}_l) p(\boldsymbol{\theta}_l|y, X) d\boldsymbol{\theta}_l$$

for  $\boldsymbol{\theta}_l = (\boldsymbol{\beta}_l, \sigma^2, L = l)$ . Simulations from  $p(\mathbf{B}|y, X)$  can be obtained by first obtaining a draw from  $p(\boldsymbol{\theta}_l|y, X)$ , and then computing  $\mathbf{B} = \left[ \sum_{h=1}^H W_h \sum_{i=1}^{n_h} \mathbf{Z}_{li} \mathbf{Z}_{li}^T \right]^{-1} \times \left[ \sum_{h=1}^H W_h \left( \sum_{i=1}^{n_h} \mathbf{Z}_{li} \mathbf{Z}_{li}^T \right) \boldsymbol{\beta}_l \right]$ , where  $W_h = N_h/n_h$  for the population size  $N_h$  and sample size  $n_h$  is the  $h$ th inclusion stratum. Note that this preserves the distribution of the covariates under the sample design while allowing the slopes to still be fully-modeled.

A direct draw from  $p(\boldsymbol{\theta}_l|y, X) = p(\boldsymbol{\beta}_l|\sigma^2, L = l, y, X) p(\sigma^2|L = l, y, X) p(L = l|y, X)$  is possible if  $H$  is of modest size; otherwise a Metropolis step can be run to obtain an approximation to the marginal posterior of  $p(L = l|y, X)$ , and direct draws obtained accordingly. Details are provided in the Appendix.

### 3.2. Fractional Bayes Factors

In the absence of strong prior information to define  $p(\boldsymbol{\theta}_l)$ , the Bayes Factors (BF) comparing weight pooling model  $l$  with weight pooling model  $l'$

$$BF(y, X) = \frac{p(L = l|y, X)}{p(L = l'|y, X)} = \frac{p(y|L = l, X) p(L = l)}{p(y|L = l', X) p(L = l')} = \frac{\int f(y|\boldsymbol{\theta}_l) p(\boldsymbol{\theta}_l) d\boldsymbol{\theta}_l p(L = l)}{\int f(y|\boldsymbol{\theta}_{l'}) p(\boldsymbol{\theta}_{l'}) d\boldsymbol{\theta}_{l'} p(L = l')}$$

can be quite sensitive to the choice of  $p(\boldsymbol{\theta}_l)$  (Kass and Raftery 1995). We have a similar issue in our weight pooling model, since our marginal pooling probabilities are simply Bayes Factors converted from the odds to the probability scale. To counter this, we consider the “fractional Bayes factor” approach proposed in O’Hagan (1995). The concept extends the training-sample idea first proposed in Spiegelhalter and Smith (1982). A fraction  $b$  of the sample is set aside as to provide a data-based proper prior for  $\boldsymbol{\theta}_l$ . O’Hagan (1995) shows that the resulting Bayes factor for comparing model  $l$  with model  $l'$  using the data-based prior, which he terms a fractional Bayes factor (FBF), is of the form

$BF_b(y, X) = q_l(f, y, X)P(L = l)/q_{l'}(f, y, X)P(L = l')$ , where

$$q_l(f, y, X) = \frac{\int p(\boldsymbol{\theta}_l)f(y|\boldsymbol{\theta}_l)d\boldsymbol{\theta}_l}{\int p(\boldsymbol{\theta}_l)f(y|\boldsymbol{\theta}_l)^b d\boldsymbol{\theta}_l}$$

Small values of  $b$  should be most efficient at choosing correct models, while larger values of  $b$  are protective against outliers (data generated under a model not in the classes considered). O'Hagan proposed  $n^{-1}\log n$  and  $n^{-1/2}$  as increasingly "robust" choices of  $b$ . O'Hagan assumes a noninformative prior  $h(\boldsymbol{\theta}_l)$  in contrast to our proper prior, but very weakly informative priors, as we use in simulations and examples below, can be used as well. The Appendix provides details describing the use of FBF in the weight pooling application.

#### 4. Simulation Results

##### 4.1. Mean Models

We considered the repeated sampling properties of our proposed models for estimating population means given by  $\bar{Y} = N^{-1}\sum_{i=1}^N Y_i$  (i.e.,  $\mathbf{x}_i = 1$  for all  $i$ ). We generated data under the following model:

$$Y_i|Y_i \in h, \boldsymbol{\mu}, \sigma^2 \sim N(\boldsymbol{\mu}_h, \sigma^2)$$

The population size of the  $H = 10$  selection strata were as follows:

$$N = (800, 1,000, 1,200, 1,500, 2,000, 3,000, 5,000, 7,500, 10,000)$$

from which disproportional samples of size 500 and 100 were drawn:

$$\mathbf{n} = (90, 80, 70, 60, 50, 50, 40, 30, 20, 10)$$

$$\mathbf{n} = (18, 16, 14, 12, 10, 10, 8, 6, 4, 2)$$

(maximum normalized weight = 13.9).

We considered two patterns for the means across 10 inclusion strata:

$$1. \quad \boldsymbol{\mu}_C = (22.5, 14.4, 9.0, 4.8, 1.8, -1.2, -1.8, -2.16, -1.92, -1.8)'$$

$$2. \quad \boldsymbol{\mu}_D = (-1.8, -1.92, -2.16, -1.8, -1.2, 1.8, 4.8, 9.0, 14.4, 22.5)'$$

and considered values of  $\sigma^2 = 10^l$ ,  $l = -1, 0, \dots, 3$ ; 200 simulations were generated for each value of  $\sigma^2$ . The mean pattern  $\boldsymbol{\mu}_C$  would generally be favorable for weight trimming, since the means for the low probability-of-selection weight strata are approximately equal;  $\boldsymbol{\mu}_D$  would generally be unfavorable for weight trimming, since the means for the low probability-of-selection weight strata differ substantially. Generally, weight trimming should be more favorable as  $\sigma^2 \rightarrow \infty$  and the effect of the bias correction is minimized; the fully-weighted estimator will generally be favored as  $\sigma^2 \rightarrow 0$ , and bias correction is paramount.

For priors, we considered  $\boldsymbol{\mu}_0 = \hat{\boldsymbol{\mu}} = (\bar{y}_1, \dots, \bar{y}_H)'$ ,  $\boldsymbol{\Sigma}_0 = c\sum_{h=1}^H \sum_{i=1}^{n_h} (y_{hi} - \bar{y}_h)^2$  and  $a = s = 10^{-8}$  (see (3.2)). This is a "data-based" prior that centers all the inclusion means

at their unweighted sample values, with a variance scaled by the sample size  $n$  so that it is equivalent to a variance estimate based on a single observation. We further scale this prior by a factor  $c \geq 1$  to allow for reduced informativeness; we consider  $c = 1,000$  in the simulations below, making the prior effectively noninformative. We term the estimator of  $\bar{Y}$  obtained under this model PWT. We also consider the Fractional Bayes Factor data-based prior as well; PWTF1, which uses a training fraction of  $\log n/n$ , and PWTF2, which uses a larger training fraction, of  $n^{-1/2}$ . O'Hagan suggests that PWTF1 will be more efficient at choosing the correct model when the true model is among the models considered, whereas PWTF2 will be more robust (have better repeated sampling properties when the true model is not among the models considered).

In addition to these three weight pooling models, we consider the standard designed-based (fully weighted) estimator (FWT), as well as two trimmed weight (TWT3, TWT7) and unweighted (UNWT) estimators. The TWT3 estimator is obtained by replacing the weights  $w_{hi}$  with trimmed values  $w_{hi}^t$  that set the maximum normalized value to 3:  $w_{hi}^t = N\tilde{w}_{hi}^t / \sum_{h=1}^H n_h \tilde{w}_h^t$ , where  $\tilde{w}_{hi}^t = \min(w_{hi}, 3N/n)$ ; this approximately corresponds to the weight pooling model (3.1) with  $l = 6$ . The TWT7 estimator uses trimmed values that set the maximized values to 7, approximately corresponds to the weight pooling model (3.1) with  $l = 8$ . The UNWT estimator obtained by fixing  $w_{hi} = N/n$  for all  $h, i$ . We estimate their variance using the Taylor Series (linearization) approximation (Binder 1983) that accounts for weighting and stratification.

Table 3 shows the root mean squared error (RMSE) relative to the fully-weighted estimator and nominal 95% coverage for the three design-based and three model-based estimators of the population mean, as a function of the variance  $\sigma^2$ , under  $\mu_C$ , the structure

Table 3. Squared root of mean squared error (RMSE) relative to RMSE of fullt-weighted estimator, and true coverage of the 95% CI or PPI of population mean estimator under the model  $\mu_C$  that is consistent with weight trimming

	RMSE relative to FWT					True coverage				
	Variance $\log_{10}$					Variance $\log_{10}$				
$n = 100$										
Estimator	-2	0	2	4	6	-2	0	2	4	6
UNWT	358.35	37.72	2.90	0.54	0.42	0	0	0	93	98
FWT	1	1	1	1	1	93	92	85	92	88
TWT3	36.80	3.92	0.65	0.70	0.65	0	0	94	96	96
TWT7	3.65	1.62	0.88	0.84	0.84	6	77	93	92	96
PWT	1.01	0.94	0.92	0.91	0.89	95	98	90	96	95
PWTF1	1.06	0.89	0.83	0.81	0.80	94	97	91	96	95
PWTF2	1.04	0.89	0.86	0.85	0.83	94	97	92	96	94
$n = 500$										
UNWT	739.48	72.52	7.78	0.69	0.39	0	0	0	80	98
FWT	1	1	1	1	1	98	96	98	96	96
TWT3	76.14	7.56	1.02	0.64	0.63	0	0	88	97	98
TWT7	8.96	2.63	1.15	0.82	0.80	0	29	97	99	100
PWT	1.03	0.91	0.84	0.83	0.81	94	94	98	92	96
PWTF1	1.09	0.85	0.76	0.76	0.72	92	96	99	95	96
PWTF2	1.06	0.87	0.82	0.82	0.79	94	95	98	95	96

that favors weight trimming. Table 4 shows the equivalent measures under  $\mu_D$ , the structure that is not consistent with weight trimming.

Even when the mean structure is favorable for weight trimming, the unweighted estimator (UNWT) and crude trimming estimators (TWT3, TWT7) behave poorly when  $\sigma^2$  is small, but have better MSE properties than the fully weighted estimator and conservative coverage when the within-stratum variance is considerably larger than the between-stratum variance. The trimmed estimator requires a smaller residual variance to have better MSE properties than the fully weighted estimator, but the unweighted estimator has the best MSE properties for the largest residual variance. The fully weighted estimator is design-unbiased; coverage is approximately correct for  $n = 500$ , but anti-conservative when  $n = 100$  due to the poor asymptotic approximation. The pooled weight estimator under the flat prior nearly dominates the fully-weighted estimator with respect to MSE and has approximately correct coverage when  $n = 100$ , since asymptotic assumptions are not necessary for the Bayesian estimator. Similar results are found for the pooled weight estimator using the fractional Bayes Factor priors, except that the increase in efficiency is greater for larger  $\sigma^2$  (RMSE reductions of nearly 30%).

When the mean structure is not favorable for weight trimming, the UNWT and TWT estimators both have larger MSE than the FWT estimator and very poor coverage except for very large  $\sigma^2$ . The pooled weight estimators are fairly robust, with slightly increased MSE relative to the fully weighted estimator for intermediate values of  $\sigma^2$ , and improved MSE relative to the fully weighted estimator for large values of  $\sigma^2$ . The true coverage of the pooled weight estimator is somewhat less than the nominal coverage when  $n = 100$  but is still better than that of the fully weighted estimator, again reflecting the lack of need for asymptotic assumptions in the Bayesian paradigm.

Table 4. Squared root of mean squared error (RMSE) relative to RMSE of full-weighted estimator, and true coverage of the 95% CI or PPI of population mean estimator under the model  $\mu_D$  that is consistent with weight trimming.

Estimator	RMSE relative to FWT					True coverage				
	Variance $\log_{10}$					Variance $\log_{10}$				
$n = 100$										
UNWT	471.23	37.73	4.37	0.56	0.39	0	0	0	94	98
FWT	1	1	1	1	1	94	88	88	93	90
TWT3	192.64	15.45	2.04	0.68	0.65	0	0	28	98	96
TWT7	23.49	7.83	2.74	1.09	0.86	0	0	30	82	94
PWT	1.00	1.01	1.08	0.90	0.91	96	88	90	94	96
PWTF1	1.00	0.99	1.17	0.82	0.81	96	88	88	96	96
PWTF2	1.00	1.00	1.13	0.85	0.84	95	90	89	94	96
$n = 500$										
UNWT	1015.9	92.02	9.81	1.01	0.45	0	0	0	58	98
FWT	1	1	1	1	1	96	96	98	95	98
TWT3	415.58	37.64	4.02	0.75	0.66	0	0	0	95	98
TWT7	56.42	17.71	5.79	2.03	0.91	0	0	0	70	96
PWT	1.01	1.00	1.08	0.91	0.80	96	96	94	95	94
PWTF1	1.00	1.01	1.15	0.88	0.70	95	94	92	94	96
PWTF2	1.00	1.00	1.11	0.89	0.80	96	94	93	94	96

#### 4.2. Linear Regression Models

For the linear regression model, we generated population data under a linear spline as follows:

$$Y_i|X_i, \boldsymbol{\beta}, \sigma^2 \sim N(\beta_0 + \sum_{h=1}^{10} \beta_h(X_i - h)_+, \sigma^2)$$

$$X_i \sim \text{UNI}(0, 10), i = 1, \dots, N = 20,000.$$

where  $(x)_+ = x$  if  $x \geq 0$  and  $(x)_+ = 0$  if  $x < 0$ . A noninformative, disproportionately stratified sampling scheme sampled elements as a function of  $X_i$  ( $I_i$  equals 1 if sampled and 0 otherwise):

$$H_i = [X_i]$$

$$P(I_i = 1|H_i) = \pi_{h_i} \propto (1 + H_i/2.5)H_i$$

This created 10 strata, defined by the integer portions of the  $X_i$  values. A total of  $n = 1,000$  elements were sampled without replacement for each simulation (maximum normalized weight  $N \approx 14.9$ ). The object of the analysis is to obtain the population slope  $\beta_1 = \sum_{i=1}^N (Y_i - \bar{Y})(X_i - \bar{X}) / \sum_{i=1}^N (X_i - \bar{X})^2$ .

We considered three patterns for  $\boldsymbol{\beta}$ :

1.  $\boldsymbol{\beta}_C = (0, 0, 0, 0, .5, .5, 1, 1, 2, 2, 4)'$
2.  $\boldsymbol{\beta}_D = (0, 11, -4, -2, -2, -1, -1, -.5, -.5, 0, 0)'$
3.  $\boldsymbol{\beta}_E = (0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0)'$ .

and considered values of  $\sigma^2 = 10^l, l = 1, \dots, 5$ ; 200 simulations were generated for each value of  $\sigma^2$ . The effect of model misspecification increases as  $\sigma^2 \rightarrow 0$  as the bias of the estimators becomes larger relative to the variance, and conversely decreases as  $\sigma^2 \rightarrow \infty$ . Under  $\boldsymbol{\beta}_C$ , weight trimming is likely to be a productive strategy under smaller values of  $\sigma^2$  than under  $\boldsymbol{\beta}_D$ , since the low probability-of-selection slopes are equal. Under  $\boldsymbol{\beta}_E$ , the linear regression model for the population is correctly specified, and the unweighted estimator should be most efficient.

We use priors equivalent to the “data-based” priors we used for population means, extended to population slopes:  $\boldsymbol{\beta}_0 = \hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \mathbf{y}$ ,  $\Sigma_0 = cn \text{Var}(\hat{\boldsymbol{\beta}})$  for  $\text{Var}(\hat{\boldsymbol{\beta}}) = \hat{\tau}^2 (X^T X)^{-1}$ ,  $\hat{\tau}^2 = (n - p)^{-1} (\mathbf{y} - X \hat{\boldsymbol{\beta}})^T (\mathbf{y} - X \hat{\boldsymbol{\beta}})$ ,  $a = s = 10^{-8}$ , and  $c = 1,000$ . We again consider Fractional Bayes Factor with training fraction of  $\log n/n$  and  $n^{-1/2}$ .

As in the population mean evaluation, we consider the FWT, TWT3, TWT7 and UNWT estimators, again estimating their variance using the Taylor Series (linearization) approximation that accounts for weighting and stratification. As in the mean model TWT3 approximately corresponds to the weight pooling model (3.1) with  $l = 6$ , and TWT7 approximately corresponds to the weight pooling model (3.1) with  $l = 8$ .

Table 5 shows the root mean squared error (RMSE) relative to the fully-weighted estimator and nominal 95% coverage for the three design-based and three model-based estimators of the population slope (second component of  $\hat{\mathbf{B}}$ ) as a function of the

Table 5. Squared root of mean squared error (RMSE) relative to RMSE of full-weighted estimator, and true coverage of the 95% CI or PPI of population mean estimator under the model  $\beta_C$  that is consistent with weight trimming.

Estimator	RMSE relative to FWT					True coverage				
	Variance $\log_{10}$					Variance $\log_{10}$				
	1	2	3	4	5	1	2	3	4	5
UNWT	15.27	4.68	1.75	0.61	0.57	0	0	16	87	96
FWT	1	1	1	1	1	96	92	94	95	94
TWT3	5.45	1.83	0.80	0.61	0.57	0	22	95	98	98
TWT7	2.72	1.18	0.73	0.74	0.66	19	82	98	96	98
PWT	0.99	0.98	0.97	0.93	0.93	96	94	96	96	96
PWTF1	1.00	1.01	1.00	0.73	0.55	90	88	91	92	97
PWTF2	0.94	0.90	0.84	0.72	0.70	96	94	96	96	99

variance  $\sigma^2$ , under  $\beta_C$ , the structure that favors weight trimming for smaller values of  $\sigma^2$ . Tables 6 and 7 show the equivalent measures under  $\beta_D$  and  $\beta_E$ , the structures that respectively favor weight trimming for only larger values of  $\sigma^2$ , and the correctly specified linear model. Under all three models, the nominal coverage of the 95% CI of the fully weighted estimator is approximately correct.

The unweighted and trimmed estimators are always biased because of model misspecification, although the reduction in variance overwhelms bias correction for large  $\sigma^2$ , yielding approximately correct nominal 95% CI coverage and smaller MSEs relative to the fully weighted estimator. When the model is correctly specified, the unweighted and trimmed estimators reduce RMSE by 35–45%, and the Nominal 95% CI coverage is correct.

The weight pooling estimator with noninformative prior generally tracks the fully weighted estimator in the presence of model misspecification, although for large  $\sigma^2$  there is a 10% reduction in RMSE. Nominal 95% coverage is correct except for small values of  $\sigma^2$  under  $\beta_D$ , the model least favorable to weight trimming. Under the correctly specified

Table 6. Squared root of mean squared error (RMSE) relative to RMSE of full-weighted estimator, and true coverage of the 95% CI or PPI of population mean estimator under the model  $\beta_D$  that is consistent with weight trimming.

Estimator	RMSE relative to FWT					True coverage				
	Variance $\log_{10}$					Variance $\log_{10}$				
	1	2	3	4	5	1	2	3	4	5
UNWT	9.69	3.68	1.52	0.57	0.63	0	0	25	93	96
FWT	1	1	1	1	1	92	91	96	94	96
TWT3	5.40	2.22	1.00	0.65	0.68	0	7	88	98	99
TWT7	3.24	1.43	0.89	0.85	0.75	6	69	93	95	98
PWT	1.00	1.00	1.01	0.93	0.90	84	92	93	96	98
PWTF1	1.02	1.04	1.11	0.60	0.53	85	92	90	96	98
PWTF2	1.03	1.03	0.96	0.74	0.70	88	93	94	98	96

Table 7. Squared root of mean squared error (RMSE) relative to RMSE of full-weighted estimator, and true coverage of the 95% CI or PPI of population mean estimator under the correctly specified model  $\beta_E$ .

Estimator	RMSE relative to FWT					True coverage				
	Variance $\log_{10}$					Variance $\log_{10}$				
	1	2	3	4	5	1	2	3	4	5
UNWT	0.55	0.46	0.55	0.50	0.49	94	96	94	96	96
FWT	1	1	1	1	1	96	95	96	94	96
TWT3	0.64	0.54	0.66	0.60	0.59	96	100	98	98	98
TWT7	0.75	0.70	0.68	0.71	0.72	98	97	98	98	98
PWT	0.93	0.91	0.93	0.93	0.93	94	98	98	94	96
PWTF1	0.62	0.56	0.63	0.61	0.59	98	98	94	96	95
PWTF2	0.69	0.70	0.72	0.71	0.68	97	97	97	98	98

model, the weight pooling estimator with noninformative prior has a 5–10% reduction in RMSE, with correct nominal 95% PPI coverage.

The weight pooling estimator with the smaller training fraction FBF prior (PWTF1) has equivalent RMSE to the fully-weighted estimator when  $\sigma^2$  is small under  $\beta_C$  and weight trimming is not warranted, but has equivalent RMSE to the unweighted estimator when  $\sigma^2$  is large and weight trimming is appropriate. A similar pattern is seen under  $\beta_D$  except that PWTF1 “overpools” somewhat for intermediate levels of  $\sigma^2$ , leading to slightly larger RMSE than the fully-weighted estimator. Under the correctly specified model  $\beta_E$ , PWTF1 has RMSE properties similar to that of TWT3, with a 35–45% reduction in RMSE. There is modest undercoverage of the nominal 95% PPI when  $\sigma^2$  is small and the model is misspecified.

The weight pooling estimator with the larger training fraction FBF prior (PWTF2) is more robust than PWTF1, with little increase in RMSE over the fully-weighted estimator even when the model is misspecified and  $\sigma^2$  is small, but retaining substantial RMSE reductions (over 30%) when bias correction is unimportant or the model is correctly specified. Coverage properties of the 95% PPI are correct, except for modest undercoverage under the “worst case” model ( $\beta_D$  with small  $\sigma^2$ ).

## 5. Application: Consideration of the Barker Hypothesis Using NHANES Data

Barker et al. (1993) described an association between on the one hand low birth weight, on the other adult cardiovascular disease and type 2 diabetes. It was postulated that in the face of a nutritionally stressed fetal environment, the fetus adapts in a manner which predisposes to the development of insulin resistance and increased CVD risk factors in later life. This hypothesis has been evaluated by others (Curhan et al. 1996; Rich-Edwards et al. 1997, among many), but usually in convenience samples. A few analyses have considered whether evidence of the “Barker Hypothesis” exists in children (Forrester et al. 1996; Matthes et al. 1994), again with convenience samples and limited ethnic diversity.

To evaluate the Barker hypothesis in children using a population-based sample, we use the National Health and Examination Nutrition Survey III (NHANES III). NHANES III (U.S. Department of Health and Human Services 1997) is a U.S.-wide survey designed to collect information about the diet and health status of the U.S. population. The survey was conducted between 1988 and 1994 with 33,394 subjects, drawn from a probability sample of the U.S. population with a complex sample design construction. The primary sample units (PSUs) consisting of standard metropolitan statistical areas (SMSAs), counties, or groups of counties were collapsed into strata. Strata containing a single large SMSA had that PSU selected with probability 1. Two PSUs were selected from the remaining strata using a “controlled selection” process that selected PSUs proportional to population while assuring balance on key covariates such as region, socioeconomic status, etc. Within each PSU clusters of dwelling units were sampled using controlled selection as well, and a systematic sample of addresses were then selected from each cluster. Oversamples of minorities (African- and Mexican-Americans) and the young (<6) and old (60 + ) were also obtained. The NHANES III sampling weights are highly variable:  $215 < w_i < 79,382$ , where 8% of the weights have normalized values larger than 3. The weights include a nonresponse adjustment as well as a post-stratification adjustment to known Census age-sex-geographic-ethnicity (non-Hispanic Caucasian, non-Hispanic African-American, Mexican-American, and other) totals that also account for the age-ethnicity oversampling, and included crude trimming adjustments at each step. (No detail is provided about the weight trimming procedures except that fewer than 1% of the cases have trimmed weights (Mohadjer et al. 1996). In the analysis below, the weights were grouped into 10 strata for the weight-pooling model.

To evaluate this hypothesis using the population-based estimates in NHANES, we regress non-HDL cholesterol on birth weight and birth weight<sup>2</sup> among 4–12 year olds, unadjusted and adjusting for age, gender, age x gender, and current body-mass index (BMI). Table 8 shows the unweighted, fully-weighted, weight trimming (to a maximum normalized value of 3), pooled weight, and fractional Bayes factor pooled weight estimators along with estimates of bias and mean squared error under the assumption that the fully weighted estimator is unbiased for both the unadjusted and adjusted models. Because a fully-weighted regression estimator  $\hat{\beta}_w$  is unbiased only in expectation, the estimated squared bias of a regression estimator  $\hat{\beta}^*$  is given by  $\max((\hat{\beta}^* - \hat{\beta}_w)^2 - \hat{V}_{01}, 0)$  where  $\hat{V}_{01} = \widehat{Var}(\hat{\beta}^*) + \widehat{Var}(\hat{\beta}_w) - 2\widehat{Cov}(\hat{\beta}^*, \hat{\beta}_w)$  (Kish 1992). To account for the effects of clustering and stratification in the multi-stage sample design, the variances of the regression estimators were calculated using a bootstrap (Davidson and Hinckley 1997, pp.92–102) where PSUs were resampled with replacement within strata. For each resampled dataset, the unweighted and fully-weighted estimates were computed as  $B^u = (X'X)^{-1}X'y$  and  $B^w = (X'WX)^{-1}X'Wy$  respectively, where  $W$  is the  $n \times n$  diagonal case weight matrix. Point estimates under the weight pooling method were computed as  $B^p = \sum_{l=1}^{H^*} \hat{B}_l P(L=l|y, X)$ , where  $\hat{B}_l = [\sum_{h=1}^H W_h \sum_{i=1}^{n_h} \mathbf{Z}_{li} \mathbf{Z}_{li}^T]^{-1} [\sum_{h=1}^H W_h (\sum_{i=1}^{n_h} \mathbf{Z}_{li} \mathbf{Z}_{li}^T) \hat{\beta}_l]$  for  $\hat{\beta}_l = (Z_l' Z_l)^{-1} Z_l' y$  where  $\mathbf{Z}_l$  consists of the stacked vectors of  $\mathbf{Z}_{li}$ . To compute  $P(L=l|y, X)$  we used a Fractional Bayes Factor data-based prior with a training fraction of  $n^{-1/2}$  (PWTF2).

In this example, the unweighted estimator appears to have better RMSE properties than the fully-weighted estimator, particularly for the linear term; the unweighted and

Table 8. Change in non-HDL cholesterol (mg/dL) associated with each 1 lb. change in birth weight, among U.S. 4–12 year-olds, using unweighted (UNWT), fully-weighted (FWT), trimmed weight (TWT3), pooled weight (PWT), and fractional Bayes factor pooled weight (PWTF2) estimators; unadjusted and adjusted for age, gender, and age x gender interactions. Point estimates for PWT and PWTF2 models from posterior median; 95% CI or PPI in subscript. RMSE = estimated root mean squared error, treating fully-weighted estimator as unbiased in expectation. Data from National Health and Nutrition Examination Survey III

	Unadjusted		Adjusted	
	Birth weight	Birth weight <sup>2</sup>	Birth weight	Birth weight <sup>2</sup>
<b>UNWT</b>				
Est. <sub>95%CI</sub>	0.25 <sub>-0.41,0.85</sub>	-0.19 <sub>-0.42,0.03</sub>	-0.08 <sub>-0.84,0.44</sub>	0.15 <sub>-0.45,-0.01</sub>
Bias	0.38	0.02	0.42	0.06
RMSE	0.32	0.12	0.36	0.11
<b>FWT</b>				
Est. <sub>95%CI</sub>	-0.13 <sub>-1.39,1.04</sub>	-0.21 <sub>-0.58,0.08</sub>	-0.51 <sub>-1.75,0.40</sub>	-0.21 <sub>-0.55,0.00</sub>
Bias	0	0	0	0
RMSE	0.61	0.15	0.57	0.13
<b>TWT3</b>				
Est. <sub>95%CI</sub>	-0.06 <sub>-1.00,0.81</sub>	-0.19 <sub>-0.48,0.05</sub>	-0.35 <sub>-1.06,0.32</sub>	-0.24 <sub>-0.63,0.07</sub>
Bias	0.07	0.02	0.16	-0.03
RMSE	0.46	0.12	0.34	0.18
<b>PWT</b>				
Est. <sub>95%CI</sub>	0.25 <sub>-0.46,0.81</sub>	-0.19 <sub>-0.40,0.01</sub>	-0.12 <sub>-0.75,0.39</sub>	-0.28 <sub>-0.45,-0.00</sub>
Bias	0.38	0.02	0.40	-0.07
RMSE	0.31	0.11	0.29	0.11
<b>PWTF2</b>				
Est. <sub>95%CI</sub>	0.18 <sub>-1.24,0.92</sub>	-0.19 <sub>-.46,0.03</sub>	-0.13 <sub>-1.20,0.39</sub>	-0.21 <sub>-0.52,0.02</sub>
Bias	0.31	0.02	0.38	0.00
RMSE	0.50	0.13	0.40	0.12

weighted quadratic terms are approximately equal under both models. The weight pooling estimator compromises between the unweighted and fully-weighted estimator for the unadjusted linear term, but tracks the unweighted estimator in the adjusted model. The weight pooling estimator tracks the unweighted estimator in the unadjusted model and compromises between the weighted and unweighted estimator in the adjusted model; the fractional Bayes factor weight pooling estimator compromises between the unweighted and fully-weighted estimator for the unadjusted linear term, but tracks the unweighted estimator in the adjusted model. The weight pooling estimator has the best MSE properties, somewhat smaller than those of the unweighted estimator; the variance of the fractional Bayes factor weight pooling estimator is somewhat larger than that of the unweighted estimator. The crude trimming estimator has the next-best MSE properties, with the fully-weighted estimator having the maximum MSE for both the unadjusted and adjusted models.

Both the unadjusted and adjusted estimates suggest that a quadratic effect might be present, with extremely underweight, and normal and above-normal weight children having lower levels of non-HDL cholesterol than moderately underweight children. However, the trends were not jointly significant using a Wald test with 2 degrees of freedom using either the unweighted, fully-weighted, or weight-pooling estimators.

## 6. Discussion

In this manuscript we have developed a “weight smoothing” methodology that allows the data to make a principled tradeoff between bias and variance – approximating the fully-weighted estimator when bias is of great importance, but moving toward the unweighted estimator when variance overwhelms the squared bias correction factor. This model generalizes the work of Elliott and Little (2000), where population inference was restricted to population means using a weight pooling model that mimicked weight trimming. A shortcoming of the previous model was lack of robustness: by considering submodels that pooled only the largest weight strata, data structures that favored fully-weighted estimators were “overpooled” and the resulting bias yielded MSEs that were larger than the fully-weighted estimators’ MSEs. Here we consider a model that allows for the pooling of all conterminous inclusion strata. This yields weight pooling estimators that are protected against overpooling, but have limited efficiency gains over fully-weighted estimators. By considering the “Fractional Bayes Factors” of O’Hagan (1995), in which a fraction  $b$  of the sample is set aside as to provide a data-based proper prior, we showed that our resulting estimators retained their robustness properties while gaining considerable efficiencies over standard fully-weighted estimators. This article also extends the weight pooling method to consider population linear regression slopes as well as population means.

We also applied the methods to assess “Barker’s Hypothesis,” an association between on the one hand low birth weight, on the other adult cardiovascular disease and Type 2 diabetes (Barker et al. 1993), using the nationally representative National Health and Examination Nutrition Survey III. In this situation, the unweighted estimates of the quadratic effect of birth weight on non-HDL cholesterol generally had the best RMSE properties; however the weight pooling estimators outperformed the fully-weighted estimators.

When sampling weights are used to account for misspecification of the mean in a regression setting, it could be argued that the correct approach is to correctly specify the mean to eliminate discrepancies between the fully-weighted and unweighted estimates of the regression parameters. However, perfect specification is an unattainable goal, and even good approximations might be highly biased if case weights are ignored if the sampling probabilities are highly variable, even if the sampling itself is noninformative. In the informative sampling setting, it may be impossible to determine whether discrepancies between weighted and unweighted estimates are due to model misspecification or to the sample design itself. Finally, even misspecified regression models have the attractive feature in the finite population setting of yielding a unique target population quantity. Consequently accounting for the probability of inclusion in linear model settings continues to be advised, and methods that balance between a low-bias, high variance fully-weighted analysis and a large bias, low variance unweighted analysis remain useful.

The next logical extension of the weight pooling methods is into the generalized linear model setting. The situation is complicated here by the lack of a closed form solution for  $p(y|L=l, X)$  outside of the Gaussian special case, making it difficult to compute a Fraction Bayes Factor to enhance efficiency. One possibility is to utilize Laplace

approximations (Tierney and Kadane 1986). In general, we have  $p(L = l|y) = C_l / \sum_l C_l$  where  $C_l = \int f(y|\theta_l)p(\theta_l)l\theta_l$ . By approximating the posterior with a normal distribution, we estimate  $C_l$  with  $(2\pi)^{l/2} |\sum_l^{-1}|^{1/2} f(y|\hat{\theta}_l)p(\hat{\theta}_l)$  where  $\hat{\theta}_l$  is a value with high posterior probability (a median or mode). DiCiccio et al. (1997) discuss improvements on this approximation that may be utilized as well.

**7. Appendix**

From (3.2), we obtain a direct draw from the posterior of  $p(\beta_l, \sigma^2, L = l|y, X)$  as follows:

- (1)  $p(L=l|y, X) = p(y|L=l, X)P(L=l) / \sum_l p(y|L=l, X)P(L=l)$ , where  $p(y|L=l, X) \propto |\Psi_l|^{1/2} [\Delta_l - \theta_l^T \Psi_l \theta_l]^{-(n+a)/2}$  for  $\Psi_l = ((Z_l^T Z_l) + \Sigma_0)^{-1}$ ,  $\theta_l = (Z_l^T Z_l)b + \sum_0 \beta_0$ ,  $\Delta_l = b^T (Z_l^T Z_l)b + \beta_0^T \sum_0^{-1} \beta_0 + Q_l^2 + as^2$ ,  $b = (Z_l^T Z_l)^{-1} Z_l^T y$ , and  $Q_l^2 = y^T (I_{pH^*} - H_l)y$ ,  $H_l = Z_l (Z_l^T Z_l)^{-1} Z_l^T$ .
- (2)  $\sigma^2 | L = l, y, X \sim Inv - \chi^2(n + a, \Delta_l - \theta_l^T \Psi_l \theta_l)$
- (3)  $\beta_l | \sigma^2, L = l, y, X \sim N(\Gamma_l A_l, \sigma^2 \Gamma_l)$ ,  $A_l = Z_l^T y + \Sigma_0^{-1} \beta_0$ ,  $\Gamma_l = [\Sigma_0^{-1} + (Z_l^T Z_l)]^{-1}$

We derive these marginal and conditional distributions in reverse order to simplify computation and notation.

(3) is derived by noting that

$$\begin{aligned}
 p(\beta_l | \sigma^2, L = l|y, X) &\propto f(y|X, \beta_l, \sigma^2, L = l)p(\beta_l | \sigma^2, L = l) \\
 &\propto \exp\left(-\frac{1}{2\sigma^2} [(b - \beta_l)^T (Z_l^T Z_l)(b - \beta_l) + (\beta_l - \beta_0)^T \Sigma_0^{-1} (\beta_l - \beta_0)]\right) \\
 &\propto f(b|\beta_l, \sigma^2, L = l)f(\beta_l | \sigma^2, L = l)
 \end{aligned}$$

for

$$\begin{aligned}
 b|\beta_l, \sigma^2, L = l &\sim N(\beta_l, \sigma^2(Z_l^T Z_l)^{-1}) \\
 \beta_l|\sigma^2, L = l &\sim N(\beta_0, \sigma^2 \Sigma_0)
 \end{aligned}$$

and thus by standard results (Gelman et al. 2004, pp. 85–86)

$$\beta_l | b, \sigma^2, L = l \sim N(\tilde{\beta}, \tilde{\Sigma})$$

where

$$\begin{aligned}
 \tilde{\beta} &= [(\sigma^2 \Sigma_0)^{-1} + (\sigma^2 (Z_l^T Z_l)^{-1})^{-1}]^{-1} [(\sigma^2 (Z_l^T Z_l)^{-1})^{-1} b + (\sigma^2 \Sigma_0)^{-1} \beta_0] \\
 &= [\Sigma_0^{-1} + Z_l^T Z_l]^{-1} [Z_l^T y + \Sigma_0^{-1} \beta_0]
 \end{aligned}$$

and  $\tilde{\Sigma} = \sigma^2 [\Sigma_0^{-1} + Z_l^T Z_l]^{-1}$ .

(2) is derived by

$$\begin{aligned} p(\sigma^2|y, X, L = l) &\propto \int_{-\infty}^{\infty} f(y|\beta_l, \sigma^2, L = l, X)p(\beta_l|\sigma^2, L = l)p(\sigma^2|L = l)d\beta_l \\ &\propto (2\pi)^{-((n+pH^*)/2)}(\sigma^2)^{-((n+pH^*+a)/2+1)} \\ &\quad \times \int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}[(\beta_l - b)^T(Z_l^T Z_l)(\beta_l - b) \right. \\ &\quad \left. + (\beta_l - \beta_0)^T \Sigma_0^{-1}(\beta_l - \beta_0) + Q_l^2 + as^2]\right)d\beta_l \end{aligned}$$

Now

$$\begin{aligned} &(\beta_l - b)^T(Z_l^T Z_l)(\beta_l - b) + (\beta_l - \beta_0)^T \Sigma_0^{-1}(\beta_l - \beta_0) + Q_l^2 + as^2 \\ &= \beta_l^T(Z_l^T Z_l + \Sigma_0^{-1})\beta_l - 2\beta_l^T[(Z_l^T Z_l)b + \Sigma_0^{-1}\beta_0] + b^T(Z_l^T Z_l)b \\ &\quad + \beta_0^T \Sigma_0^{-1}\beta_0 + Q_l^2 + as^2 \\ &= (\beta_l - \Psi_l \theta)^T \Psi_l^{-1}(\beta_l - \Psi_l \theta) + \Delta_l - \theta_l^T \Psi_l \theta_l \end{aligned}$$

Thus

$$\begin{aligned} &\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}[(\beta_l - b)^T(Z_l^T Z_l)(\beta_l - b) + (\beta_l - \beta_0)^T \Sigma_0^{-1}(\beta_l - \beta_0) + Q_l^2 + as^2]\right)d\beta_l \\ &= (2\pi\sigma^2)^{(pH^*/2)}|\Psi_l|^{1/2} \exp\left(-\frac{1}{2\sigma^2}[\Delta_l - \theta_l^T \Psi_l \theta_l]\right) \end{aligned}$$

from the normalizing constant for  $N(\mu, \Sigma)$  distribution, and thus

$$p(\sigma^2|L = l, y, X) \propto (2\pi)^{-(n/2)}(\sigma^2)^{-((n+a/2)+1)}|\Psi_l|^{1/2} \exp\left(-\frac{1}{2\sigma^2}[\Delta_l - \theta_l^T \Psi_l \theta_l]\right)$$

which is the kernel of a scaled inverse chi-square distribution with  $n + a$  degrees of freedom and scaling factor  $\Delta_l - \theta_l^T \Psi_l \theta_l$ .

(1) then follows from (2):

$$p(L = l|y, X) \propto p(y|L = l, X)p(L = l)$$

where

$$\begin{aligned}
 p(y|L=l, X) &= \int_0^\infty \int_{-\infty}^\infty f(y|\beta_l, \sigma^2, L=l, X) p(\beta_l|\sigma^2, L=l) p(\sigma^2|L=l) d\beta_l d\sigma^2 \\
 &\propto \int_0^\infty (2\pi)^{-(n/2)} (\sigma^2)^{-((n+a/2)+1)} |\Psi_l|^{1/2} \exp\left(-\frac{1}{2\sigma^2} [\Delta_l - \theta_l^T \Psi_l \theta_l]\right) d\sigma^2 \\
 &\propto (2\pi)^{-(n/2)} |\Psi_l|^{1/2} \Gamma\left(\frac{n+a}{2}\right) \left(\frac{n+a}{2}\right)^{-(n+a)/2} \left[\frac{\Delta_l - \theta_l^T \Psi_l \theta_l}{n+a}\right]^{-(n+a)/2} \\
 &\propto |\Psi_l|^{1/2} [\Delta_l - \theta_l^T \Psi_l \theta_l]^{-(n+a)/2}
 \end{aligned}$$

from the normalizing constant for the  $Inv - \chi^2(n, s^2)$  distribution.

### 7.1. Fractional Bayes Factors

To implement O’Hagan’s (1995) Fractional Bayes Factors for the marginal weight pooling selection probability, we replaced

$$p(L=l|y, X) \propto p(L=l) \int_0^\infty \int_{-\infty}^\infty f(y|\beta_l, \sigma^2, L=l, X) p(\beta_l|\sigma^2, L=l) p(\sigma^2|L=l) d\beta_l d\sigma^2$$

with

$$p(L=l|y, X) \propto p(L=l) \frac{\int_0^\infty \int_{-\infty}^\infty f(y|\beta_l, \sigma^2, L=l, X) p(\beta_l|\sigma^2, L=l) p(\sigma^2|L=l) d\beta_l d\sigma^2}{\int_0^\infty \int_{-\infty}^\infty f(y|\beta_l, \sigma^2, L=l, X)^b p(\beta_l|\sigma^2, L=l) p(\sigma^2|L=l) d\beta_l d\sigma^2}$$

where  $0 < b < 1$  represents a “training fraction” of the data set aside to provide prior information for the parameters for the  $l$ th pooling model. From the derivation of (1) above we have

$$\begin{aligned}
 &\int_0^\infty \int_{-\infty}^\infty f(y|\beta_l, \sigma^2, L=l, X)^b p(\beta_l|\sigma^2, L=l) p(\sigma^2|L=l) d\beta_l d\sigma^2 \\
 &\propto |\Psi_{bl}|^{1/2} [\Delta_{bl} - \theta_{bl}^T \Psi_{bl} \theta_{bl}]^{-(bn+a)/2}
 \end{aligned}$$

for

$$\begin{aligned}
 \Psi_{bl} &= ((bZ_l^T Z_l) + \Sigma_0)^{-1}, \theta_{bl} = b(Z_l^T Z_l)b + \Sigma_0 \beta_0 \\
 \Delta_{bl} &= b[b^T(Z_l^T Z_l)b + Q_l^2] + \beta_0^T \Sigma_0^{-1} \beta_0 + as^2
 \end{aligned}$$

Thus using FBF, we have

$$p(L=l|y, X) \propto p(L=l) \frac{[\Delta_{bl} - \theta_{bl}^T \Psi_{bl} \theta_{bl}]^{(bn+a)/2} |\Psi_l|^{1/2}}{[\Delta_l - \theta_l^T \Psi_l \theta_l]^{(n+a)/2} |\Psi_{bl}|^{1/2}}$$

## 8. References

- Alexander, C.H., Dahl, S., and Weidman, L. (1997). Making Estimates from the American Community Survey. *Proceedings of the American Statistical Association, Social Statistics Section*, 88–97.
- Barker, D.J.P., Gluckman, P.D., Godfrey, K.M., Harding, J.E., Owens, J.A., and Robinson, J.S. (1993). Fetal Nutrition and Cardiovascular Disease in Adult Life. *Lancet*, 341, 938–941.
- Beaumont, J.F. and Alavi, A. (2004). Robust Generalized Regression Estimation. *Survey Methodology*, 30, 195–208.
- Binder, D.A. (1983). On the Variances of Asymptotically Normal Estimators from Complex Surveys. *International Statistical Review*, 51, 279–292.
- Cochran, W.G. (1977). *Sampling Techniques*. New York: Wiley.
- Cox, B.G. and McGrath, D.S. (1981). An Examination of the Effect of Sample Weight Truncation on the Mean Square Error of Survey Estimates. Paper presented at the 1981 Biometric Society ENAR Meeting, Richmond, VA.
- Curhan, G.C., Willett, W.C., Rimm, E.B., Spiegelman, D., Ascherio, A.L., and Stampfer, M.J. (1996). Birth Weight and Adult Hypertension. *Diabetes Mellitus and Obesity in U.S. Men. Circulation*, 94, 3246–3250.
- Davidson, A.C. and Hinckley, D.V. (1997). *Bootstrap Methods and Their Applications*. Cambridge: Cambridge Press.
- Deville, J.C. and Särndal, C-E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, 87, 376–382.
- DiCiccio, T.J., Kass, R.E., Raftery, A., and Wasserman, L. (1997). Computing Bayes Factors by Combining Simulation and Asymptotic Approximations. *Journal of the American Statistical Association*, 92, 903–915.
- Elliott, M.R. and Little, R.J.A. (2000). Model-based Approaches to Weight Trimming. *Journal of Official Statistics*, 16, 191–210.
- Ericson, W.A. (1969). Subjective Bayesian Modeling in Sampling Finite Populations. *Journal of the Royal Statistical Society, Series B*, 31, 195–234.
- Folsom, R.E. and Singh, A.C. (2000). The Generalized Exponential Model for Sampling Weight Calibration for Extreme Values, Nonresponse, and Poststratification. *Proceedings of the American Statistical Association, Survey Research Methods Section*, 598–603.
- Forrester, T.E., Wilks, R.J., Bennett, F.I., Simeon, D., Osmond, C., Allen, M., Chung, A.P., and Scott, P. (1996). Fetal Growth and Cardiovascular Risk Factors in Jamaican Schoolchildren. *British Medical Journal*, 312, 156–160.
- Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2004). *Bayesian Data Analysis*, (2nd edn.). Boca Raton, FL: Chapman and Hall/CRC.
- Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., and Stahel, W.A. (1986). *Robust Statistics: The Approach Based on Influence Function*. New York: Wiley.
- Holt, D. and Smith, T.M.F. (1979). Poststratification. *Journal of the Royal Statistical Society, Series A*, 142, 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.
- Isaki, C.T. and Fuller, W.A. (1982). Survey Design under a Regression Superpopulation Model. *Journal of the American Statistical Association*, 77, 89–96.
- Kass, R.E. and Raftery, A.E. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90, 773–795.
- Kish, L. (1965). *Survey Sampling*. New York: John Wiley and Sons.
- Kish, L. (1992). Weighting for Unequal  $P_i$ . *Journal of Official Statistics*, 8, 183–200.
- Korn, E.L. and Graubard, B.I. (1999). *Analysis of Health Surveys*. New York: Wiley.
- 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). Post-stratification: 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.
- Lu, H. and Gelman, A. (2003). A Method for Estimating Design-based Sampling Variances for Surveys with Weighting, Poststratification, and Raking. *Journal of Official Statistics*, 19, 133–152.
- Matthes, J.W., Lewis, P.A., Davies, D.P. and Bethel, J.A. (1994). Relation between Birth Weight at Term and Systolic Blood Pressure in Adolescence. *British Medical Journal*, 308, 1074–1077.
- Mohadjer, L., Montaquila J., Waksberg, J., Bell, B., James, P., Flores-Cervantes, I., and Montes, M. (1996). National Health and Nutrition Examination Survey III: Weighting and Estimation Methodology, Executive Summary. Prepared by Westat, Inc., for the National Center for Health Statistics, Hyattsville, MD. At <http://www.cdc.gov/nchs/data/nhanes/nhanes3/cdrom/NGHS/MANUALS/WGT E XEC.PDF>
- O'Hagan, A. (1995). Fraction Bayes Factors for Model Comparison. *Journal of the Royal Statistical Society, Series B*, 57, 99–138.
- Potter, F.A. (1990). Study of Procedures to Identify and Trim Extreme Sample Weights. *Proceedings of the American Statistical Association, Survey Research Methods Section*. 225–230.
- Pfeffermann, D. (1993). The Role of Sampling Weights when Modeling Survey Data. *International Statistical Review*, 61, 317–337.
- Rich-Edwards, J.W., Stampfer, M.J., Manson, J.E., Rosner, B., Hankinson, S.E., Colditz, G.A., Willett, W.C., and Hennekens, C.H. (1997). Birth Weight and Risk of Cardiovascular Disease in a Cohort of Women Followed Up Since 1976. *British Medical Journal*, 315, 396–400.
- Rubin, D.B. (1987). *Multiple Imputation for Non-Response in Surveys*. New York: Wiley.
- Särndal, C.E. (1980). On  $\pi$ -inverse Weighting Versus Best Linear Unbiased Weighting in Probability Sampling. *Biometrika*, 67, 639–650.

- Skinner, C.J., Holt, D., and Smith, T.M.F. (1989). *Analysis of Complex Surveys*. New York: Wiley.
- 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, Series B*, 44, 377–387.
- Tierney, L.J. and Kadane, J. (1986). Accurate Approximations for Posterior Moments and Marginal Densities. *Journal of the American Statistical Association*, 81, 82–86.
- Zaslavsky, A.M., Schenker, N., and Belin, T.R. (2001). Downweighting Influential Clusters in Surveys: Application to the 1990 Post Enumeration Survey. *Journal of the American Statistical Association*, 96, 858–869.

Received August 2006

Revised April 2008