Model Averaging Methods for Weight Trimming in Generalized Linear Regression Models

Michael R. Elliott[1]

Acknowledgements: This research was supported in by National Institute of Heart, Lung, and Blood grant R01-HL-068987-01. The author acknowledges Jack Chen for his assistance with programming. The author also thanks Drs. Dennis Durbin and Flaura Winston of the Partners for Child Passenger Safety project for their assistance, as well as State Farm Insurance Companies for their support of the Partners for Child Passenger Safety project.

Abstract

In sample surveys where units have unequal probabilities of inclusion, associations between the inclusion probability and the statistic of interest can induce bias. This is true even in regression models, where the estimates of the population slope may be biased if the underlying mean model is misspecified or the sampling is non-ignorable. 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; weight trimming reduces large weights to a 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 uses Bayesian model averaging to create "data driven" weight trimming estimators. We develop robust 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, weight Winsorization, Bayesian population inference, weight pooling, variable selection, fractional Bayes Factors.

1 Introduction

Population-based samples with differential probabilities of inclusion typically use case weights equal to the inverse of the probability of inclusion to provide reduce bias in the estimators of population quantities of interest (Horvitz and Thompson 1952). By replacing unweighted sums in statistics with their weighted equivalents, bias can be removed from linear estimators and reduced in non-linear estimators (Binder 1983).

This bias reduction typically comes at the cost of increased variance. This increase can overwhelm the reduction in bias, so that the mean square error (MSE) actually increases under a weighted analysis. This is particularly likely if a) the sample size is small, b) the difference in the probability of inclusion is large, or c) the association between the probability of inclusion and the data (which drives the bias) is weak. This manuscript develops an alternative approach to weight trimming that considering 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 non-response 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

[1]Department of Biostatistics, School of Public Health, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109, USA.
Survey Methodology Program, Institute for Social Research, 426 Thompson St., Ann Arbor, MI 48106, USA.
E-mail: mrelliot@umich.edu.

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; substracting from or adding to the inclusion strata design matrix allows for a greater 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 continguous inclusion strata to be considered for pooling, we induce a high degree of robustness into our model, protecting against "over-pooling" from which models that crudely mimicked weight trimming suffered (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 an proper framework for population inference. In particular, we consider an alternative Bayesian modeling approach that focuses on population quantities of interest $Q(Y)$, such as populations means $Q(Y) = \overline{Y}$ or population least-squares regression slopes $Q(Y_1, Y_2) = \min_{B_0,B_1} \sum_{i=1}^{N} (Y_{i1} - B_0 - B_1 Y_{i2})^2$. Inference is made about $Q(Y)$ by considering the marginal posterior predictive distribution (Ericson 1969, Holt and Smith 1979, Skinner et al. 1989, Little 1993):

$$p(Q(Y) \mid y) = \int f(Q(Y) \mid \theta) p(\theta \mid y) d\theta = \frac{\int f(Q(Y) \mid \theta) f(y \mid \theta) p(\theta) d\theta}{\int f(y \mid \theta) p(\theta) d\theta}. \quad (1)$$

If the sampling indicator I is independent of Y, as is the case in probability sampling design, then the sampling mechanism is said to be unconfounded or non-informative (Rubin 1987, Little 2004), and inference about $Q(Y)$ can be made using $p(Q(Y) \mid y)$ alone. However, sensible models in (1) still need to account for the sample design in both the likelihood and prior model structure. For more detail about Bayesian survey inference in the context of regression models, see Elliott (2006).

Section 2 briefly reviews standard weight trimming methods. Section 3 develops our weight pooling models for generalized linear regression models. Section 4 provides simulation results to consider the repeated sampling properties of the weight pooling estimators of logistic regression parameters in a disproportional-stratified sample design and compares them with standard design-based estimators. Section 5 summarizes the results of the simulations and considers extensions to more complex sample designs.

2 Standard Weight Trimming Procedures

Standard weight trimming approach pick a single cutpoint w_0 at which all weights greater than this value are to be fixed, with the remaining weights are adjusted upward by a constant so that the trimmed and untrimmed weighted sample sizes are equal. Typically w_0 is chosen in an ad-hoc manner – say 3 times or 6 times the mean weight – without regard to whether the chosen cutpoint is optimal with respect to mean square error. 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 $t = 1, \ldots, T$ under the assumption that the true population mean is given by the fully weighted estimate: $\hat{MSE}_t = (\hat{\theta}_t - \hat{\theta}_T)^2 + \hat{V}(\theta_T)$, 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. More recently, the calibration literature has developed methods for adjusting design weights so that the

adjusted weights equal known population totals under a variety of minimizing distance constraints between the unadjusted and adjusted weights, thus generalizing poststratification and raking procedures (Deville and Sarndal 1992). Techniques have been developed that allow these adjustments to be bounded to prevent the construction of extreme weights (Deville and Sarndal 1992, Folsom and Singh 2000), but these bounds involving the winsorizing of extreme weights to a fixed cutpoint value, with the choice of this cutpoint remaining arbitrary.

3 Weight 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, although this assumption can be relaxed (Lu and Gelman 2003). The untrimmed (design-based) weighted mean estimator is then $\overline{y}_w = \frac{\sum_h \sum_i w_h y_{hi}}{\sum_h \sum_i w_h} = \sum_h \frac{N_h}{N_+} \overline{y}_h$. 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 $\gamma = (n - \sum \kappa_i w_o)/\sum(1-\kappa_i)w_i$, where κ_i is an indicator variable for whether or not $w_i \geq w_0$. The trimmed mean estimator is thus given by

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

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

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

$$y_{hi} \mid \mu_h \sim N(\mu_h, \sigma^2) \;\; h < l$$
$$y_{hi} \mid \mu_l \sim N(\mu_l, \sigma^2) \;\; h \geq l$$
$$\mu_h, \mu_l, \log \sigma \propto \text{const.}$$

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

$$y_{hi} \mid \mu_h \sim N(\mu_h, \sigma^2) \;\; h < l$$
$$y_{hi} \mid \mu_l \sim N(\mu_l, \sigma^2) \;\; h \geq l$$
$$p(L = l) = 1/H$$
$$p(\sigma^2 \mid L = l) = \sigma^{-(l+1/2)}$$
$$p(\boldsymbol{\beta} \mid \sigma^2, L = l) = (2\pi)^{-l}$$

where $\mu_1 = \beta_0 + \beta_1, \ldots, \mu_l = \beta_0 + \beta_{l-1}$. 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 and Generalized Linear Regression

This manuscript extends Elliott and Little (2000) in three ways. First, we consider the regression of Y_i on fixed covariates $\boldsymbol{x}_i$. Second, we allow for the pooling of all conterminous inclusion strata. Third, we allow for general exponential family outcomes, not just normally-distributed outcomes.

Generalized linear regression models (Nelder and Wedderburn 1972) postulate a likelihood for y_i of the form

$$f(y_i;\theta_i,\phi) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i,\phi)\right]$$

where $a_i(\phi)$ involves a known constant and a (nuisance) scale parameter ϕ, and the mean of y_i is related to a linear combination of fixed covariates $\boldsymbol{x}_i$ through a link function $g(\cdot)$: $E(y_i \mid \theta_i) = \mu_i$, where $g(\mu_i) = g(b'(\theta_i)) = \eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$. We also have $\mathrm{Var}(y_i \mid \theta_i) = a_i(\phi)V(\mu_i)$, where $V(\mu_i) = b''(\theta_i)$. The link is canonical if $\theta_i = \eta_i$, in which case $g'(\mu_i) = V^{-1}(\mu_i)$.

Indexing the inclusion stratum by h and allow for the pooling of all conterminous inclusion strata, we have

$$g(E[y_{hi} \mid \boldsymbol{\beta}_l, L = l]) = \boldsymbol{Z}_{li}^T\boldsymbol{\beta}_l$$

where $\boldsymbol{Z}_{li} = D_{hl} \otimes \boldsymbol{x}_{hi}$ where D_{hl} is a vector of dummy variables that pool the appropriate conterminous inclusion strata based on the lth pooling pattern. We assume priors of the form

$$\boldsymbol{\beta}_l \mid L = l \sim N(\boldsymbol{\beta}_0, \Sigma_0)$$

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

For normally distributed outcomes where $a_i(\phi) = \sigma^2$, we assume

$$\sigma^2 \mid L = l \sim Inv - \chi^2(a, s^2)$$

Our population quantity of interest $\boldsymbol{B}$ is the slope that solves the population score equation $U_N(\boldsymbol{B}) = 0$ where

$$U_N(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{\partial}{\partial\boldsymbol{\beta}} \log f(y_i;\boldsymbol{\beta}) =$$

$$\sum_{h=1}^{H}\sum_{i=1}^{N} \frac{(y_i - g^{-1}(\mu_i(\boldsymbol{\beta})))\boldsymbol{x}_i}{V(\mu_i(\boldsymbol{\beta}))g'(\mu_i(\boldsymbol{\beta}))}.$$

A closed form solution for $\boldsymbol{B}$ is available when the y_i are normally distributed: $\boldsymbol{B} = \left(\sum_{i=1}^N \boldsymbol{x}_i\boldsymbol{x}_i^T\right)^{-1}\left(\sum_{i=1}^N \boldsymbol{x}_i y_i\right)$; otherwise an iterative procedure such as iterative reweighted least squares must be used. Note that the quantity $\boldsymbol{B}$ such that $U(\boldsymbol{B}) = 0$ 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 $\boldsymbol{x}_i$ to $g(E(Y_i \mid \boldsymbol{x}_i))$.

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

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

for $\boldsymbol{\theta}_l = (\boldsymbol{\beta}_l, \phi, L = l)$. Simulations from $p(\boldsymbol{B} \mid y, X)$ can be obtained by first obtaining a draw from $p(\boldsymbol{\theta}_l \mid y, X)$, and then computing

$$\sum_{h=1}^{H} W_h \sum_{i=1}^{n_h} \frac{(\hat{y}_{hi} - g^{-1}(\mu_i(\hat{\boldsymbol{B}})))\boldsymbol{x}_{hi}}{V(\mu_{hi}(\hat{\boldsymbol{B}}))g'(\mu_{hi}(\hat{\boldsymbol{B}}))} = 0$$

where $W_h = N_h/n_h$ and $\hat{y}_{hi} = g^{-1}(\boldsymbol{Z}_{li}^T \boldsymbol{\beta}_l)$. Thus, in the example of logistic regression, where $V(\mu_i) = \mu_i(1-\mu_i)$ and $g'(\mu_i) = \mu_i^{-1}(1-\mu_i)^{-1}$, a posterior draw of $\boldsymbol{B}$ can be computed by solving for B_j, $j = 1, \ldots, p$

$$\sum_{h=1}^{H} W_h \sum_{i=1}^{n_h} x_{hij} \frac{\exp(x_{hij} B_j)}{1 + \exp(x_{hij} B_j)} = \sum_{h=1}^{H} W_h \sum_{i=1}^{n_h} x_{hij} \frac{\exp(x_{hij} \beta_{hj})}{1 + \exp(x_{hij} \beta_{hj})}$$

where β_{hj} corresponds to the jth value of the $\boldsymbol{\beta}_l$ parameter for the hth inclusion stratum as a function of the lth pooling pattern. When the y_i are normally distributed, a closed form solution is directly available: $\hat{\boldsymbol{B}} = \left[\sum_{h=1}^{H} W_h \sum_{i=1}^{n_h} \boldsymbol{Z}_{li} \boldsymbol{Z}_{li}^T\right]^{-1} \left[\sum_{h=1}^{H} W_h \left(\sum_{i=1}^{n_h} \boldsymbol{Z}_{li} \boldsymbol{Z}_{li}^T\right) \boldsymbol{\beta}_l\right]$ where $W_h = N_h/n_h$ for the population size N_h and sample size n_h is the hth 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 $p(\boldsymbol{\theta}_l \mid y, X) = p(\boldsymbol{\beta}_l \mid \sigma^2, L = l, y, X) p(\sigma^2 \mid L = l, y, X) p(L = l \mid y, X)$ is possible in the Gaussian setting if H is of modest size; otherwise a Metropolis step can be run to obtain an approximation to the marginal posterior of $p(L = l \mid y, X)$, and direct draws obtained accordingly. Details for the Gaussian model are provided in Section 6.1 of the Appendix. In the non-Gaussian setting, we approximately a direct draw by using a Laplace approximation to obtain a draw from $p(L = l \mid y, X)$ and a Metropolis step to obtain a draw from $p(\boldsymbol{\beta}_l \mid L = l, y, X)$; alternatively a Metropolis step may be used to obtain draws from $p(L = l \mid \boldsymbol{\beta} y, X)$ and an Markov Chain Monte Carlo algorithmn implemented instead. Details for the non-Gaussian model are provided in Section 6.2 of the Appendix.

3.2 Fractional Bayes Factors

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

$$BF(y, X) = \frac{p(L = l \mid y, X)}{p(L = l' \mid y, X)} = \frac{p(y \mid L = l, X) p(L = l)}{p(y \mid L = l', X) p(L = l')} = \frac{\int\int p(y \mid \boldsymbol{\beta}_l, \sigma^2 L = l, X) d\boldsymbol{\beta}_l d\sigma^2 p(L = l)}{\int\int p(y \mid \boldsymbol{\beta}_{l'}, \sigma^2 L = l', X) d\boldsymbol{\beta}_{l'} d\sigma^2 p(L = l')}$$

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

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

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 non-informative 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 Linear Regression

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

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

$$X_i \sim UNI(0, 10),\ i = 1, \ldots, N = 20000.$$

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

$$H_i = \lceil X_i \rceil$$

$$P(I_i = 1 \mid H_i) = \pi_h \propto (1 + H_i) H_i$$

This created 10 strata, defined by the integer portions of the X_i values A total of $n = 1000$ elements were sampled without replacement for each simulation (maximum normalized weight ≈ 11.0). The object of the analysis is to obtain the population slope $B_1 = \frac{\sum_{i=1}^{N}(Y_i - \overline{Y})(X_i - \overline{X})}{\sum_{i=1}^{N}(X_i - \overline{X})^2}$.

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. $\boldsymbol{\beta}_E = (0, 2, 0, 0, 0, 0, 0, 0, 0, 0,)'$.

and considered values of $\sigma^2 = 10^l$, $l = 1, \ldots, 5$; 200 simulations were generated for each value of σ^2. The effect of model misspecification increases as $\sigma^2 \to 0$ as the bias of the estimators becomes larger relative to the variance, and conversely decreases as $\sigma^2 \to \infty$. Under $\boldsymbol{\beta}_C$, weight trimming is likely to be a productive strategy under smaller values of σ^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: $\boldsymbol{\beta}_0 = \hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{y}$, $\Sigma_0 = cn\mathrm{Var}(\hat{\boldsymbol{\beta}})$ for $\mathrm{Var}(\hat{\boldsymbol{\beta}}) = \hat{\tau}^2 (X^T X)^{-1}$, $\hat{\tau}^2 = (n-p)^{-1}(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})^T(\boldsymbol{y} - X\hat{\boldsymbol{\beta}})$, $a = s = 10^{-8}$, and $c = 1000$. We also consider Fractional Bayes Factor with training fraction of $\log n/n$ and $n^{-1/2}$. 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).

As in the population mean evaluation, we consider the FWT, TWT, and UNWT estimators, again estimating their variance using the Taylor Series (linearization) approximation that accounts for weighting and stratification.

Table 1 shows the root mean square 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 $\hat{\boldsymbol{B}}$) as a function of the variance σ^2, under $\boldsymbol{\beta}_C$, the structure that favors weight trimming for smaller values of σ^2; Tables 2 and 3 show the equivalent measures

	RMSE relative to FWT					True Coverage				
	Variance $\log_{10}$					Variance $\log_{10}$				
Estimator	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
TWT	5.45	1.83	0.80	0.61	0.57	0	22	95	98	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

Table 1: Relative bias (%), square root of mean square error (RMSE) relative to RMSE of fully-weighted estimator, and true coverage of the 95% CI or PPI of population linear regression slope estimator under the misspecified model $\boldsymbol{\beta}_C$ that favors weight trimming.

under $\boldsymbol{\beta}_D$ and $\boldsymbol{\beta}_E$, the structures that respectively favor weight trimming for only larger values of σ^2, and the correctly specified linear model.

Under all three models, the nominal coverage of the 95% CI of 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 σ^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 nominal 95% CI coverage is correct.

The weight pooling estimator with non-informative prior generally tracks the fully weighted estimator in the presence of model misspecification, although for large σ^2 there is a 10% reduction in RMSE. Nominal 95% coverage is correct except for small values of σ^2 under $\boldsymbol{\beta}_D$, the model least favorable to weight trimming. Under the correctly specified model, the weight pooling estimator with non-informative 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 σ^2 is small under $\boldsymbol{\beta}_C$ and weight trimming is not warranted, but has equivalent RMSE to the unweighted estimator when σ^2 is large and weight trimming is appropriate. A similar pattern is seen under $\boldsymbol{\beta}_D$, except that PWTF1 "overpools" somewhat for intermediate levels of σ^2, leading to slightly higher RMSE that the fully-weighted estimator. Under the correctly specified model $\boldsymbol{\beta}_E$, PWTF1 has RMSE properties similar to that of TWT, with a 35-45% reduction in RMSE. There is modest undercoverage of the nominal 95% PPI when σ^2 is small and the model is misspecified.

The weight pooling estimator with the larger training fraction FBF prior (PWTF2) is more robust that PWTF1, with little increase in RMSE over the fully-weighted estimator even when the model is misspecified and σ^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 ($\boldsymbol{\beta}_D$ with small σ^2).

4.2 Logistic Regression

We consider logistic regression under a correctly specified and then under an increasingly misspecified model. We generated population data as follows:

$$P(Y_i = 1 \mid X_i) \sim BER(\text{expit}(2 - .4 * X_i + C * X_i^2))$$

	RMSE relative to FWT					True Coverage				
	Variance $\log_{10}$					Variance $\log_{10}$				
Estimator	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
TWT	5.40	2.22	1.00	0.65	0.68	0	7	88	98	99
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 2: Relative bias (%), square root of mean square error (RMSE) relative to RMSE of fully-weighted estimator, and true coverage of the 95% CI or PPI of population linear regression slope estimator under the misspecified model $\boldsymbol{\beta}_D$ that discourages weight trimming.

	RMSE relative to FWT					True Coverage				
	Variance $\log_{10}$					Variance $\log_{10}$				
Estimator	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
TWT	0.64	0.54	0.66	0.60	0.59	96	100	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

Table 3: Relative bias (%), square root of mean square error (RMSE) relative to RMSE of fully-weighted estimator, and true coverage of the 95% CI or PPI of population linear regression slope estimator under the correctly specified model $\boldsymbol{\beta}_E$.

$$X_i \sim UNI(0, 10),\ i = 1, \ldots, N = 20000.$$

where expit$(\cdot) = exp(\cdot)/(1 + exp(\cdot))$. The object of the analysis is to obtain the logistic population regression slope, defined as the value B_1 in the equation $\sum_i^N (y_i - \text{expit}(B_0 + B_1 x_i)) \begin{pmatrix} 1 \\ x_i \end{pmatrix} = 0$. A disproportional sampling scheme was implemented as described in the linear regression simulations. We consider values of $C = 0, .0158, .0273, .0368, .0454$, corresponding to curvature measures of $K = 0, .02, .04, .06$ at the midpoint 5 of the support for X, where $K(X; C) = | \frac{2C}{[1+(2CX-.75)^2]^{3/2}} |$; 200 simulations were generated for each value of C. A noninformative, disproportionally stratified sampling scheme sampled elements as a function of X_i (I_i equals 1 if sampled and 0 otherwise):

$$H_i = \lceil X_i \rceil$$

$$P(I_i = 1 \mid H_i) = \pi_h \propto (1 + H_i/2.5)H_i$$

A total of $n = 1000$ elements were sampled for each simulation (maximum normalized weight ≈ 7.5).

For priors, we considered a nearly non-informative prior of the form $\boldsymbol{\beta}_l \mid L = l \sim N(0, 225I)$, which assumes that the logistic regression parameters lie between -30 and 30 with 95% probability. We term the estimator of B_1 obtained under this model PWT. We again consider the Factional Bayes Factor data-based prior as well; PWTF1, which uses a training fraction of $n^{-1/2}$, and PWTF2, which uses a larger training fraction of 0.1.

In addition to these two weight pooling models, we consider the standard designed-based (fully weighted) estimator (FWT), as well as trimmed weight (TWT) and unweighted (UNWT) estimators. The TWT estimator is obtained by replacing the weights w_{hi} with trimmed values w_{hi}^t that set the maximum normalized value to 3: $w_{hi}^t = \frac{N\tilde{w}_{hi}^t}{\sum_{h=1}^H n_h \tilde{w}_h^t}$, where $\tilde{w}_{hi}^t = \min(w_{hi}, 3N/n)$, and 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 4 shows the relative bias, RMSE relative to the RMSE of the fully-weighted estimator, and true coverage of the nominal 95% CIs or PPIs associated with each of the six estimators of the population slope (B) for different values of curvature K, corresponding to increased degrees of misspecification.

The undersampling of small values of X means that the maximum likelihood estimator of B in the model misspecification setting will be unbiased for $K = 0$ and biased downward for $K = .02, .04, .06$ unless the sample design is accounted for. The trimmed estimator's bias is intermediate between the unweighted and fully weighted estimator. The weight pooling estimator with a non-informative prior, similar to the fully weighted estimator, showed little bias. The weight pooling FBF estimator with the smaller training fraction (PWTF1) had bias similar to the unweighted estimator, while the weight pooling FBF estimator with the larger training fraction (PWTF2) had bias similar to the trimmed weight estimator.

The unweighted estimator had substantially improved MSE (40% reduction) when the linear slope model was approximately correctly specified, but was highly biased with moderate to large degree of misspecification. The trimmed weight estimator and the weight pooling estimator with a non-informative prior both dominated the standard fully-weighted estimator over the range of simulations considered. The crude trimming estimator yielded up to 35% reduction in MSE, while the weight pooling estimator with non-informative priors yielded 10% reduction in MSE. The weight pooling estimators with the fractional Bayes factors had MSE reductions up to nearly 40% when the linear slope model was approximately correctly specified, but only PWTF2 was robust against model misspecification.

	Relative bias (%)				RMSE relative to FWT				True Coverage			
	Curvature K				Curvature K				Curvature K			
Estimator	0	.02	.04	.06	0	.02	.04	.06	0	.02	.04	.06
UNWT	-0.2	-12.8	-59.2	-217.9	.59	.87	1.51	1.72	96	86	32	22
FWT	-0.1	3.9	-0.1	16.6	1	1	1	1	86	96	92	86
TWT	0.3	-3.4	-21.4	-83.9	.65	.71	.90	.91	92	96	92	88
PWT	2.4	1.6	-2.8	-0.0	.96	.92	1.02	.91	87	92	93	89
PWTF1	-0.0	-7.7	-42.2	-281.1	.61	.65	1.08	1.47	97	98	90	90
PWTF2	-0.0	-2.8	-25.2	-96.7	.63	.61	.95	.97	98	99	95	97

Table 4: Relative bias (%), square root of mean square error (RMSE) relative to RMSE of fully-weighted estimator, and true coverage of the 95% CI or PPI of population logistic regression slope estimator under model misspecification.

The unweighted estimator had poor coverage except when the linear slope model was correctly specified, or nearly so. The fully-weighted, trimmed weight, and the weight pooling estimator with non-informative priors generally had approximately correct coverage, except for somewhat below nominal coverage when the linear model was badly misspecified. The fractional Bayes factor estimators had somewhat conservative coverage when the model is correctly specified, with the coverage dropping somewhat below nominal for model misspecification only for PWTF1.

5 Discussion

The model discussed in this manuscript generalizes the work Elliott and Little (2000), where population inference was restricted to population means using a weight pooling model that mimicked weight trimming. We consider a model that allows for the pooling of all conterminous inclusion strata, as well as utilizing data-based "fractional Bayes Factors" of O'Hagan (1995), we obtained robust estimators that can still gain considerable efficiencies over standard fully-weighted estimators. This manuscript also extended the weight pooling method to consider population regression slopes under the linear and generalized linear model, allowing for regression models for both coninuous and binary or count outcomes.

More generally, the methods discussed in this manuscript show the promise of adapting model-based methods to attack problems in survey data analysis. However, because these models rely on stratifying the data by probability of selection as a prelude to using pooling or shrinkage techniques to induce data-driven weight trimming, there is a natural correspondence between this methodology and (post)stratified sample designs in which strata correspond to disproportional probabilities of inclusion. Developing methods that accommodate a more general class of complex sample designs that include single or multi-stage cluster samples and/or strata that "cross" the weight strata remains an area for future work.

6 Appendix

6.1 Simulations from linear weight pooling models

We obtain a direct draw from the posterior of $p(\boldsymbol{\beta}_l, \sigma^2, L = l \mid y, X)$ as follows:

1. $p(L = l \mid y, X) = \frac{p(y|L=l,X)P(L=l)}{\sum_l p(y|L=l,X)P(L=l)}$, where $p(y \mid L = l, X) \propto |\Psi_l|^{1/2} \left[\Delta_l - \boldsymbol{\theta}_l^T \Psi_l \boldsymbol{\theta}_l\right]^{-(n+a)/2}$ for $\Psi_l = ((Z_l^T Z_l) + \Sigma_0)^{-1}$, $\boldsymbol{\theta}_l = (Z_l^T Z_l)\boldsymbol{b} + \Sigma_0 \boldsymbol{\beta}_0$, $\Delta_l = \boldsymbol{b}^T (Z_l^T Z_l)\boldsymbol{b} + \boldsymbol{\beta}_0^T \Sigma_0^{-1} \boldsymbol{\beta}_0 + Q_l^2 + as^2$, $\boldsymbol{b} = (Z_l^T Z_l)^{-1} Z_l^T \boldsymbol{y}$, and $Q_l^2 = \boldsymbol{y}^T (I_{pH^*} - H_l)\boldsymbol{y}$, $H_l = Z_l (Z_l^T Z_l)^{-1} Z_l^T$.
2. $\sigma^2 \mid L = l, y, X \sim Inv - \chi^2(n + a, \Delta_l - \boldsymbol{\theta}_l^T \Psi_l \boldsymbol{\theta}_l)$
3. $\boldsymbol{\beta}_l \mid \sigma^2, L = l, y, X \sim N(\Gamma_l A_l, \sigma^2 \Gamma_l)$, $A_l = Z_l^T \boldsymbol{y} + \Sigma_0^{-1} \boldsymbol{\beta}_0$, $\Gamma_l = [\Sigma_0^{-1} + (Z_l^T Z_l)]^{-1}$

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

3. is derived by noting that

$$p(\boldsymbol{\beta}_l \mid \sigma^2, L = l \mid y, X) \propto f(y \mid X, \boldsymbol{\beta}_l, \sigma^2, L = l) p(\boldsymbol{\beta}_l \mid \sigma^2, L = l) \propto$$

$$\exp\left(-\frac{1}{2\sigma^2}[(\boldsymbol{b} - \boldsymbol{\beta}_l)^T (Z_l^T Z_l)(\boldsymbol{b}\text{-}\boldsymbol{\beta}_l) + (\boldsymbol{\beta}_l - \boldsymbol{\beta}_0)^T \Sigma_0^{-1} (\boldsymbol{\beta}_l - \boldsymbol{\beta}_0)]\right) \propto$$

$$f_{\boldsymbol{b}}(\boldsymbol{b} \mid \boldsymbol{\beta}_l, \sigma^2, L = l) f_{\boldsymbol{\beta}_l}(\boldsymbol{\beta}_l \mid \sigma^2, L = l)$$

for

$$\boldsymbol{b} \mid \boldsymbol{\beta}_l, \sigma^2, L = l \sim N(\boldsymbol{\beta}_l, \sigma^2 (Z_l^T Z_l)^{-1})$$

$$\boldsymbol{\beta}_l \mid \sigma^2, L = l \sim N(\boldsymbol{\beta}_0, \sigma^2 \Sigma_0)$$

and thus by standard results (Gelman et al., 2004, p. 85-86)

$$\boldsymbol{\beta}_l \mid \boldsymbol{b}, \sigma^2, L = l \sim N(\tilde{\boldsymbol{\beta}}, \tilde{\Sigma})$$

where
$\tilde{\boldsymbol{\beta}} = [(\sigma^2 \Sigma_0)^{-1} + (\sigma^2 (Z_l^T Z_l)^{-1})^{-1}]^{-1} [(\sigma^2 (Z_l^T Z_l)^{-1})^{-1} \boldsymbol{b} + (\sigma^2 \Sigma_0)^{-1} \boldsymbol{\beta}_0] = [\Sigma_0^{-1} + Z_l^T Z_l]^{-1} [Z_l^T \boldsymbol{y} + \Sigma_0^{-1} \boldsymbol{\beta}_0]$
and $\tilde{\boldsymbol{\Sigma}} = \sigma^2 [\Sigma_0^{-1} + Z_l^T Z_l]^{-1}$.

2. is derived by

$$p(\sigma^2, \mid y, X, L = l) \propto \int_{-\infty}^{\infty} f(y \mid \boldsymbol{\beta}_l, \sigma^2, L = l, X) p(\boldsymbol{\beta}_l \mid \sigma^2, L = l) p(\sigma^2 \mid L = l) d\boldsymbol{\beta}_l \propto$$

$$(2\pi)^{-\frac{n+pH^*}{2}} (\sigma^2)^{-(\frac{n+pH^*+a}{2}+1)} \times$$

$$\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}[(\boldsymbol{\beta}_l - \boldsymbol{b})^T (Z_l^T Z_l)(\boldsymbol{\beta}_l - \boldsymbol{b}) + (\boldsymbol{\beta}_l - \boldsymbol{\beta}_0)^T \Sigma_0^{-1} (\boldsymbol{\beta}_l - \boldsymbol{\beta}_0) + Q_l^2 + as^2)]\right) d\boldsymbol{\beta}_l.$$

Now

$$(\boldsymbol{\beta}_l - \boldsymbol{b})^T (Z_l^T Z_l)(\boldsymbol{\beta}_l - \boldsymbol{b}) + (\boldsymbol{\beta}_l - \boldsymbol{\beta}_0)^T \Sigma_0^{-1} (\boldsymbol{\beta}_l - \boldsymbol{\beta}_0) + Q_l^2 + as^2 =$$

$$\boldsymbol{\beta}_l^T (Z_l^T Z_l + \Sigma_0^{-1}) \boldsymbol{\beta}_l - 2\boldsymbol{\beta}_l^T [(Z_l^T Z_l)\boldsymbol{b} + \Sigma_0^{-1} \boldsymbol{\beta}_0] + \boldsymbol{b}^T (Z_l^T Z_l)\boldsymbol{b} + \boldsymbol{\beta}_0^T \Sigma_0^{-1} \boldsymbol{\beta}_0 + Q_l^2 + as^2 =$$

$$(\boldsymbol{\beta}_l - \Psi_l \boldsymbol{\theta})^T \Psi_l^{-1} (\boldsymbol{\beta}_l - \Psi_l \boldsymbol{\theta}) + \Delta_l - \boldsymbol{\theta}_l^T \Psi_l \boldsymbol{\theta}_l$$

Thus

$$\int_{-\infty}^{\infty} \exp\left(-\frac{1}{2\sigma^2}[(\boldsymbol{\beta}_l - \boldsymbol{b})^T(Z_l^T Z_l)(\boldsymbol{\beta}_l - \boldsymbol{b}) + (\boldsymbol{\beta}_l - \boldsymbol{\beta}_0)^T\Sigma_0^{-1}(\boldsymbol{\beta}_l - \boldsymbol{\beta}_0 + Q_l^2 + as^2)]\right) d\boldsymbol{\beta}_l =$$

$$(2\pi\sigma^2)^{\frac{pH^*}{2}}|\Psi_l|^{1/2}\exp\left(-\frac{1}{2\sigma^2}[\Delta_l - \boldsymbol{\theta}_l^T\Psi_l\boldsymbol{\theta}_l]\right)$$

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

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

which is the kernel of a scaled inverse chi-sqaure distribution with $n + a$ degrees of freedom and scaling factor $\Delta_l - \boldsymbol{\theta}_l^T\Psi_l\boldsymbol{\theta}_l$.

1. then follows:

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

where

$$p(y \mid L = l, X) = \int_0^{\infty}\int_{-\infty}^{\infty} f(y \mid \boldsymbol{\beta}_l, \sigma^2, L = l, X)p(\boldsymbol{\beta}_l \mid \sigma^2, L = l)p(\sigma^2 \mid L = l)d\boldsymbol{\beta}_l d\sigma^2 \propto$$

$$\int_0^{\infty}(2\pi)^{-\frac{n}{2}}(\sigma^2)^{-(\frac{n+a}{2}+1)}|\Psi_l|^{1/2}\exp\left(-\frac{1}{2\sigma^2}[\Delta_l - \boldsymbol{\theta}_l^T\Psi_l\boldsymbol{\theta}_l]\right)d\sigma^2 \propto$$

$$(2\pi)^{-\frac{n}{2}}|\Psi_l|^{1/2}\Gamma\left(\frac{n+a}{2}\right)\left(\frac{n+a}{2}\right)^{-(n+a)/2}\left[\frac{\Delta_l - \boldsymbol{\theta}_l^T\Psi_l\boldsymbol{\theta}_l}{n+a}\right]^{-(n+a)/2} \propto$$

$$|\Psi_l|^{1/2}\left[\Delta_l - \boldsymbol{\theta}_l^T\Psi_l\boldsymbol{\theta}_l\right]^{-(n+a)/2}$$

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

6.1.1 Fractional Bayes Factors

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

$$p(L = l \mid y, X) \propto p(L = l)\int_0^{\infty}\int_{-\infty}^{\infty} f(y \mid \boldsymbol{\beta}_l, \sigma^2, L = l, X)p(\boldsymbol{\beta}_l \mid \sigma^2, L = l)p(\sigma^2 \mid L = l)d\boldsymbol{\beta}_l d\sigma^2$$

with

$$p(L = l \mid y, X) \propto p(L = l)\frac{\int_0^{\infty}\int_{-\infty}^{\infty} f(y \mid \boldsymbol{\beta}_l, \sigma^2, L = l, X)p(\boldsymbol{\beta}_l \mid \sigma^2, L = l)p(\sigma^2 \mid L = l)d\boldsymbol{\beta}_l d\sigma^2}{\int_0^{\infty}\int_{-\infty}^{\infty} f(y \mid \boldsymbol{\beta}_l, \sigma^2, L = l, X)^b p(\boldsymbol{\beta}_l \mid \sigma^2, L = l)p(\sigma^2 \mid L = l)d\boldsymbol{\beta}_l d\sigma^2}.$$

where $0 < b < 1$ represents a "training fraction" of the data set aside to provide prior information for the parameters for the lth pooling model. From the derivation of 1. above we have

$$\int_0^{\infty}\int_{-\infty}^{\infty} f(y \mid \boldsymbol{\beta}_l, \sigma^2, L = l, X)^b p(\boldsymbol{\beta}_l \mid \sigma^2, L = l)p(\sigma^2 \mid L = l)d\boldsymbol{\beta}_l d\sigma^2 \propto$$

$$|\Psi_{bl}|^{1/2}\left[\Delta_{bl} - \boldsymbol{\theta}_{bl}^T\Psi_{bl}\boldsymbol{\theta}_{bl}\right]^{-(bn+a)/2}$$

for for $\Psi_{bl} = ((bZ_l^T Z_l) + \Sigma_0)^{-1}$, $\boldsymbol{\theta}_{bl} = b(Z_l^T Z_l)\boldsymbol{b} + \Sigma_0\boldsymbol{\beta}_0$,
$\Delta_{bl} = b\left[\boldsymbol{b}^T(Z_l^T Z_l)\boldsymbol{b} + Q_l^2\right] + \boldsymbol{\beta}_0^T\Sigma_0^{-1}\boldsymbol{\beta}_0 + as^2$. Thus using FBF, we have

$$p(L = l \mid y, X) \propto p(L = l)\frac{\left[\Delta_{bl} - \boldsymbol{\theta}_{bl}^T\Psi_{bl}\boldsymbol{\theta}_{bl}\right]^{(bn+a)/2}|\Psi_l|^{1/2}}{\left[\Delta_l - \boldsymbol{\theta}_l^T\Psi_l\boldsymbol{\theta}_l\right]^{(n+a)/2}|\Psi_{bl}|^{1/2}}$$

6.2 Simulations from the generalized weight pooling model

6.2.1 Simulations from the generalized weight pooling model using direct draws

Draws from $p(\boldsymbol{\beta}_l, L = l \mid \boldsymbol{y}, X) = p(\boldsymbol{\beta}_l \mid L = l, \boldsymbol{y}, X)p(L = l \mid \boldsymbol{y}, X)$ can be made by drawing first from $p(L = l \mid \boldsymbol{y}, X)$ using a Laplace approximation (Tierney and Kadane 1986) to obtain $f(\boldsymbol{y} \mid L = l, X)$ and then a Metropolis step for $p(\boldsymbol{\beta}_l \mid L = l, \boldsymbol{y}, X)$.

Note that

$$p(L = l \mid \boldsymbol{y}, X) = \frac{f(\boldsymbol{y} \mid L = l, X)}{\sum_l f(\boldsymbol{y} \mid L = l, X)} \tag{2}$$

where

$$f(\boldsymbol{y} \mid X, L = l) = \int f(\boldsymbol{y} \mid X, \boldsymbol{\beta}_l, L = l)p(\boldsymbol{\beta}_l \mid L = l)d\boldsymbol{\beta}_l \approx$$

$$\int f(\boldsymbol{y} \mid X, \boldsymbol{\beta}_l, L = l)d\boldsymbol{\beta}_l \approx (2\pi)^{(pH^*)/2} \mid \hat{\Sigma}_{\hat{\boldsymbol{\beta}}_l} \mid^{1/2} f(\boldsymbol{y} \mid X, \hat{\boldsymbol{\beta}}_l, L = l)$$

where $\hat{\boldsymbol{\beta}}_l$ is the MLE of a GLM regressing $\boldsymbol{y}$ on Z_l, where Z_l consists of the stacked row vectors of Z_{li}^T, and $\hat{\Sigma}_{\hat{\boldsymbol{\beta}}_l}$ is the associated covariance matrix estimate for $\hat{\boldsymbol{\beta}}_l$ given by the inverse of the expected information matrix. The first approximation follows from assuming a non-informative or nearly non-informative prior on $\boldsymbol{\beta}_l \mid L = l$, and the second from the Laplace approximation to the true marginal distribution of $\boldsymbol{y}$.

Draws from $p(\boldsymbol{\beta}_l \mid L = l, \boldsymbol{y}, X)$ are made by running a Metropolis algorithm using a $N(0, k\hat{\Sigma}_{\hat{\boldsymbol{\beta}}_l})$ jumping distribution, where k is a tuning factor designed to obtain an acceptance rate of 20-30%. The algorithm starts at $\boldsymbol{\beta}_l^{(0)} = \hat{\boldsymbol{\beta}}_l$, and a proposal draw $\boldsymbol{\beta}_l^{prop} = \hat{\boldsymbol{\beta}}_l + \boldsymbol{e}$, $\boldsymbol{e} \sim N(0, k\hat{\Sigma}_{\hat{\boldsymbol{\beta}}_l})$ is made; $\boldsymbol{\beta}_l^{(1)} = (1 - u)\boldsymbol{\beta}_l^{(0)} + u\boldsymbol{\beta}_l^{prop}$, where u is a Bernoulli random variable with probability $\min(1, \frac{f(\boldsymbol{y}|X,\{\boldsymbol{\beta}_l^{prop}, L=l)}{f(\boldsymbol{y}|X,\{\boldsymbol{\beta}_l^{(0)}, L=l)})$. The algorithm proceeds until a sufficient number of draws T have been made to approximate the posterior distribution. In general $k = .1$ and $T = 200$ provided reasonable acceptance rates and sufficient coverage of the posterior interval.

Fractional Bayes Factors

When using the FBF prior, we replace $f(\boldsymbol{y} \mid L = l, X)$ in (2) with

$$f^*(\boldsymbol{y} \mid L = l, X) = \frac{\int f(\boldsymbol{y} \mid X, \boldsymbol{\beta}_l, L = l)p(\boldsymbol{\beta}_l \mid L = l)d\boldsymbol{\beta}_l}{\int f(\boldsymbol{y} \mid X, \boldsymbol{\beta}_l, L = l)^b p(\boldsymbol{\beta}_l \mid L = l)d\boldsymbol{\beta}_l}$$

for $0 < b < 1$. Under a nearly non-informative prior, we have, using the Laplace approximation,

$$\int f(\boldsymbol{y} \mid X, \boldsymbol{\beta}_l, L = l)^b d\boldsymbol{\beta}_l \approx$$

$$(2\pi)^{(pH^*)/2} \mid b^{-1}\hat{\Sigma}_{\hat{\boldsymbol{\beta}}_l} \mid^{1/2} f(\boldsymbol{y} \mid X, \hat{\boldsymbol{\beta}}_l, L = l)^b$$

so that

$$f^*(\boldsymbol{y} \mid L = l, X) \approx b^{(pH^*)/2} f(\boldsymbol{y} \mid X, \hat{\boldsymbol{\beta}}_l, L = l)^{(1-b)}.$$

6.2.2 Simulations from the generalized linear weight pooling model using an MCMC algorithm

Draws from the posterior distribution of $(\boldsymbol{\beta}_l, L = l)$ are obtained via the product space series method of Carlin and Chib (1995). This approach assumes that $\boldsymbol{y}$ is independent of $\{\boldsymbol{\beta}_{k\neq l}\}$ given that $L = l$. Assuming also that $\{\boldsymbol{\beta}_l\}$ are independent for $l = 1, \ldots, L$, we have that

$$p(\boldsymbol{y} \mid X, L = l) = \int f(\boldsymbol{y} \mid X, \boldsymbol{\beta}, L = l) p(\boldsymbol{\beta} \mid L = l) d\boldsymbol{\beta}$$

$$= \int f(\boldsymbol{y} \mid X, \boldsymbol{\beta}_l, L = l) p(\boldsymbol{\beta}_l \mid L = l) d\boldsymbol{\beta}_l$$

The form given to the "pseudoprior" $p(\boldsymbol{\beta}_{k\neq l} \mid L = l)$ is irrelevant, as it is chosen only to completely define the joint model specification:

$$p(\boldsymbol{y}, \boldsymbol{\beta}, L = l \mid X) = f(\boldsymbol{y} \mid X, \boldsymbol{\beta}_l, L = l) \prod_{j=1}^{2^{H-1}} \left\{p(\boldsymbol{\beta}_j \mid L = j)\right\} P(L = l)$$

We can then develop a Gibbs sampler that draws from $p(\boldsymbol{\beta}_l \mid L = l, \boldsymbol{\beta}_{k\neq l} \boldsymbol{y}, X)$ and then from $p(L = l \mid \boldsymbol{\beta}, \boldsymbol{y}, X)$.

With the model fixed at $L = l$, we obtain a draw of

$$p(\boldsymbol{\beta}_l \mid L = l, \boldsymbol{\beta}_{k\neq l}, \boldsymbol{y}, X) = p(\boldsymbol{\beta}_l \mid L = l, \boldsymbol{y}, X)$$

using the Metropolis step described in 7.1.

The full conditional $p(L \mid \boldsymbol{\beta}, \boldsymbol{y}, X)$ is given by

$$p(L = l \mid \boldsymbol{\beta}, \boldsymbol{y}, X) = \frac{f(\boldsymbol{y} \mid X, \boldsymbol{\beta}_l, L = l) \prod_{j=1}^{2^{H-1}} \left\{p(\boldsymbol{\beta}_j \mid L = j)\right\} P(L = l)}{\sum_{j=1}^{2^{H-1}} f(\boldsymbol{y} \mid X, \boldsymbol{\beta}_j, L = j) \prod_{i=1}^{2^{H-1}} \left\{p(\boldsymbol{\beta}_i \mid L = j)\right\} P(L = j)}$$

Because computing $\prod_{j=1}^{2^{H-1}} \left\{p(\boldsymbol{\beta}_j \mid L = j)\right\}$ is prohibitive except when H is small, we instead used a Metropolis step suggested by Dellaportas, Foster, and Ntzoufras (1998) to obtain a draw from $L \mid \boldsymbol{\beta}, \boldsymbol{y}, X$.

1. Propose new model l' with probability $h(l, l')$.
2. Generate $\boldsymbol{\beta}_{l'}$ from the pseudoprior $p(\boldsymbol{\beta}_{l'} \mid L \neq l')$.
3. Accept the new model l' with probability $\min\left\{1, \frac{f(\boldsymbol{y}|X, \boldsymbol{\beta}_{l'}, L=l') p(\boldsymbol{\beta}_{l'}|L=l') p(\boldsymbol{\beta}_l|L=l') P(L=l') h(l', l)}{f(\boldsymbol{y}|X, \boldsymbol{\beta}_l, L=l) p(\boldsymbol{\beta}_l|L=l) p(\boldsymbol{\beta}_{l'}|L=l) P(L=l) h(l, l')}\right\}$.

Carlin and Chib note that poor choices for pseudo-priors $p(\boldsymbol{\beta}_{k\neq l} \mid L = l)$ can yield slow convergence, and suggest matching them as closely as possible to the true model-specific posteriors. Because of the large number of models to be considered, we simply set the pseudo prior to be multivariate normal with mean $\hat{\boldsymbol{\beta}}_k$ given by the MLE of a GLM regressing $\boldsymbol{y}$ on Z_l, and covariance $\Sigma_{\hat{\boldsymbol{\beta}}_l}$ given by the inverse of the expected information matrix. Jumping probabilities to the l' models that exclude L were always given by the uniform dicrete distribution with probability $(2^{H-1} - 1)^{-1}$.

References

Association for the Advancement of Automotive Medicine. (1990). The Abbreviated Injury Scale, 1990 Revision. Association for the Advancement of Automotive Medicine, Des Plaines, Illinois.

Binder, D.A. (1983). On the variances of asymptotically normal estimators from complex surveys. International Statistical Review, 51, 279-292.

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.

Deville, J.C., Sarndal C-E. (1992). Calibration estimators in survey sampling. Journal of the American Statistical Association, 87, 376-382.

Durbin D.R., Bhatia E., Holmes J.H.. Shaw K.N., Werner J.V., Sorenson W., Winston F.K. (2001). Partners for Child Passenger Safety: A Unique Child- Specific Crash Surveillance System. Accident Analysis and Prevention, 33, 407-412.

Elliott, M.R. (2006). Model Averaging Methods for Weight Trimming. Unpublished manuscript.

Elliott, M.R., 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, B31, 195-234.

Folsom, R.E., Singh, A.C. (2000). The generalized exponential model for sampling weight calibration for extreme values, nonresponse, and poststratification. Proceedings of the Survey Research Methods Section, American Statistical Association, 2000, 598-603.

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

Horvitz, D.G., 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., Fuller, W.A. (1982). Survey design under a regression superpopulation model. Journal of the American Statistical Association, 77, 89-96.

Kass, R.E., Raftery, A.E. (1995). Bayes Factors. Journal of the American Statistical Association, 90, 773-795.

Korn, E.L., Graubard, B.I. (1999). Analysis of Health Surveys. Wiley: New York.

Lazzaroni, 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. (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., Kessler, R.C. (1997). Assessment of weighting methodology for the National Comorbidity Survey. American Journal of Epidemiology, 146, 439-449.

Little, R.J.A. (2004). To model or not model? Competing modes of inference for finite

population sampling. Journal of the American Statistical Association, 99, 546-556.

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, 133-151.

O'Hagan A. (1995). Fraction Bayes Factors for model comparison. Journal of the Royal Statistical Society, B57, 99-138.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models Journal of the Royal Statistical Society(A), 135, 370-384.

Potter, F. (1990). A study of procedures to identify and trim extreme sample weights. Proceedings of the Survey Research Methods Section, American Statistical Association, 1990, 225-230.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. International Statistical Review, 61, 317-337.

Pfeffermann, D. (1996). The use of sampling weights for survey data analysis. Statistical Methods in Medical Research, 5, 239-261.

Sarndal, C.E. (1980). On π-inverse weighting verses 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, Wiley: New York.

Tierney, L., Kadane, J. (1986). Accurate approximations for posterior moments and marginal densities. Journal of the American Statistical Association, 81, 82-86.

Winston, F.K., Kallan, M.K., Elliott, M.R., Menon, R.A. Durbin, D.R. (2002). Risk of injury to child passengers in compact extended pick-up trucks. Journal of the American Medical Association, 287, 1147-1152.