

Model Averaging Methods for Weight Trimming in Generalized Linear Regression Models

Michael R. Elliott¹

In sample surveys where units have unequal probabilities of inclusion, associations between the inclusion probability and the statistic of interest can induce bias in unweighted estimates. 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 nonignorable. 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 trade-offs. This article uses Bayesian model averaging to create “data driven” weight trimming estimators. We extend previous results for linear regression models (Elliott 2008) to generalized linear regression models, developing 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 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 nonlinear 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 squared 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. Consider a population

¹ Department of Biostatistics, School of Public Health, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109, U.S.A. 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. 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. The author thanks the Associate Editor and three anonymous reviewers for their assistance in improving the quality of the article.

generated from

$$Y_i|X_i \sim \text{BERNOULLI}(p_i) \text{ for } p_i = \frac{e^{A+BX_i+CX_i^2}}{1 + e^{A+BX_i+CX_i^2}}$$

while the superpopulation model of interest is the conditional distribution of Y_i given X_i modeled by

$$Y_i|X_i \sim \text{BERNOULLI}(p_i) \text{ for } p_i = \frac{e^{\alpha+\beta X_i}}{1 + e^{\alpha+\beta X_i}}$$

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 Z_i that are correlated with Y_i . (In the regression setting, an ignorable sample scheme is one in which the inclusion indicator I_i is independent of $Y_i|X_i$). We assume that the goal of the modeler is to describe the association between Y and X using the regression slope β 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 A, B such that

$$U \begin{pmatrix} A \\ B \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

for

$$U \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^N \partial/\partial \alpha \log f(y_i|\alpha, \beta) \\ \sum_{i=1}^N \partial/\partial \beta \log f(y_i|\alpha, \beta) \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^N \left(y_i - \frac{e^{\alpha+\beta X_i}}{1 + e^{\alpha+\beta X_i}} \right) \\ \sum_{i=1}^N X_i \left(y_i - \frac{e^{\alpha+\beta X_i}}{1 + e^{\alpha+\beta X_i}} \right) \end{pmatrix}$$

(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 OLS estimator and case-weighted WLS estimator of (A, B) are given by solving

$$U \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^N \partial/\partial \alpha S_i \log f(y_i|\alpha, \beta) \\ \sum_{i=1}^N \partial/\partial \beta S_i \log f(y_i|\alpha, \beta) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$U \begin{pmatrix} \hat{\alpha}^w \\ \hat{\beta}^w \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^N \partial/\partial \alpha S_i w_i \log f(y_i|\alpha, \beta) \\ \sum_{i=1}^N \partial/\partial \beta S_i w_i \log f(y_i|\alpha, \beta) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

respectively, where S_i is an indicator for inclusion in the sample, and $w_i = 1/\pi_i$.

Table 1 shows the results of an evaluation from 500 simulations for equivalent populations of $N = 10,000$, under correctly specified and misspecified models and ignorable and nonignorable sample designs, for sample sizes of both $n = 50$ and $n = 500$.

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

	Sampling ignorable?			
	Yes		No	
	Superpopulation model correctly specified?			
	Yes	No	Yes	No
$n = 50$				
$\hat{\beta}$	8.4(2.43)	-12.3(1.82)	-56.8(7.56)	-118.8(4.00)
$\hat{\beta}^w$	10.1(3.64)	3.3(2.83)	43.8(30.03)	-11.8(37.09)
$n = 500$				
$\hat{\beta}$	1.2(.19)	-13.3(.16)	-50.3(1.21)	-86.7(.93)
$\hat{\beta}^w$.7(.44)	2.4(.32)	6.9(.58)	5.8(.39)

The target quantity of interest is a logistic regression slope linearly relating a covariate X to the log-odds that $Y = 1$ for a dichotomous outcome Y in a population of $N = 10,000$; bias and mean squared error (MSE) are computed relative to this target quantity of interest. The correctly specified model has a linear association between $\log P(Y = 1)/(1 - P(Y = 1))$ and X , while the misspecified model has a quadratic association. The probability of selection is a function of X only for the ignorable sampling design, and a function of both X and Z , where $Z \sim N(Y, 1)$, for the nonignorable sample design.

$$\left(\text{Corr}(Y, Z) = \sqrt{\frac{\text{Var}(Y)}{\text{Var}(Y) + 1}} \right)$$

which equals .41 for the correctly specified population model and .43 for the misspecified population model.) We assume that both X and Z are available for the entire population, but Y is only observable for the sampled elements. 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 (though the bias remains in small sample settings, as Table 1 shows). Nonetheless, when the sample is small, the mean squared error of the weighted estimator is larger than that of the unweighted estimator for both the correctly specified and misspecified superpopulation models. When the sample design is ignorable for the population slope (probability of selection depends only on X), use of the weighted estimator provides protection against model misspecification, but can introduce so much variability into the estimator that the MSE is larger than for the unweighted estimator under all of the conditions considered.

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 that 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 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 trade-off is data-driven. By allowing for all contiguous inclusion strata to be considered for pooling, we induce a high degree of robustness into our model, protecting against the “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 a proper framework for population inference. In particular, we consider an alternative Bayesian modeling approach that treats Y as a random variable and focuses on population quantities of interest $Q(Y)$, such as population means $Q(Y) = \bar{Y}$ 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):

$$p(Q(Y)|y) = \int f(Q(Y)|\theta)p(\theta|y)d\theta = \frac{\int f(Q(Y)|\theta)f(y|\theta)p(\theta)d\theta}{\int f(y|\theta)p(\theta)d\theta} \quad (1.1)$$

If the sampling indicator I is independent of Y , as is the case in the probability sampling design, then the sampling mechanism is said to be unconfounded or noninformative (Little 2004), and inference about $Q(Y)$ can be made using $p(Q(Y)|y)$ alone. However, in order to make the assumption of unconfoundedness in (1.1) reasonable, the sample design needs to be accounted for in both the likelihood and prior model structures. For more detail about Bayesian survey inference in the context of regression models, see Elliott (2008).

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 illustrates the use of the weight pooling estimators in an analysis of risk of injury to children in passenger truck crashes. Section 6 summarizes the results of the simulations and considers extensions to more complex sample designs.

2. Standard Weight Trimming Procedures

Standard weight trimming approaches pick a single cutpoint w_0 at which all weights larger than this value are to be fixed, with the remaining weights usually 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 squared 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 $t = 1, \dots, T$ under the assumption that the true population mean is given by the fully weighted estimate: $M\hat{S}E_t = (\hat{\theta}_t - \hat{\theta}_T)^2 + \hat{V}(\theta_T)$, 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 $M\hat{S}E_t$ 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 Särndal 1992). Techniques have been developed that allow these adjustments to be bounded to prevent the construction of extreme weights (Deville and Särndal 1992; Folsom and Singh 2000), but these bounds involve 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 large 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 $h = 1, \dots, H$. We assume here that N_h is known, as when the weight strata come from a stratified or post-stratified random sample. The untrimmed (design-based) weighted mean estimator is then

$$\bar{y}_w = \frac{\sum_h \sum_i w_h y_{hi}}{\sum_h \sum_i w_h} = \sum_h \frac{N_h}{N} \bar{y}_h$$

Weight trimming typically proceeds by establishing an *a priori* cutpoint, say 3 for the normalized weights, and multiplying the remaining weights by a normalizing constant

$$\gamma = \frac{\left(N - \sum \kappa_i w_o\right)}{\sum (1 - \kappa_i) w_i}$$

where κ_i is an indicator variable for whether or not $w_i \geq w_0$. The trimmed mean estimator is thus given by

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

where

$$\gamma = \frac{N - w_0 \sum_{h=l}^H n_h}{\sum_{h=1}^{l-1} N_h}$$

and

$$\bar{y}^{(l)} = \left(\frac{1}{\sum_{h=l}^H n_h} \right) \sum_{h=l}^H n_h \bar{y}_h$$

Choosing

$$w_0 = \frac{\sum_{h=1}^H N_h}{\sum_{h=1}^H n_h}$$

yields $\gamma = 1$ and

$$\bar{y}_{wtr} = \sum_{h=1}^{l-1} \frac{N_h}{N} \bar{y}_h + \frac{\sum_{h=l}^H N_h}{N} \bar{y}^{(l)}$$

which corresponds to the estimate for a model that assumes distinct stratum means for the smaller weight strata and a common mean for the larger weight strata, that is:

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

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

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

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 $\mu_1 = \beta_0 + \beta_1, \dots, \mu_l = \beta_0 + \beta_{l-1}$. 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.

Elliott (2008) extended this model to consider the conditional distribution of y given covariates x (linear regression), and to allow for the pooling of all conterminous inclusion strata. The latter extension greatly increased the robustness of the model, preventing “overpooling” and increased MSE relative to the fully-weighted estimator due to bias. Elliott (2008) also found that use of fractional Bayes factor priors (O’Hagan 1995) could substantially increase the efficiency of the weight pooling models with little effect on robustness.

3.1. Weight Pooling Models for Generalized Linear Regression

Generalized linear regression models 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 \mathbf{x}_i through a link function $g(\cdot)$: $E(y_i|\theta_i) = \mu_i$, where $g(\mu_i) = g(b'(\theta_i)) = \eta_i = \mathbf{x}_i^T \boldsymbol{\beta}$ (McCullagh and Nelder 1989, p. 30). We also have $\text{Var}(y_i|\theta_i) = a_i(\phi)V(\mu_i)$, 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 allowing for the pooling of all conterminous inclusion strata, we have

$$g(E[y_{hi}|\boldsymbol{\beta}_l, \sigma^2, L = l]) = \mathbf{Z}_{li}^T \boldsymbol{\beta}_l$$

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

Figure 1 shows the set of pooling patterns when $H = 4$. (Note that patterns 2 and 5 correspond to a crude weight trimming procedure that simply cuts the maximum weight to

$$\left(\frac{\sum_{h=l}^H N_h}{\sum_{h=l}^H n_h} \right)$$

We assume priors of the form

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

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

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

$$U_N(\boldsymbol{\beta}) = \sum_{i=1}^N \frac{\partial}{\partial \boldsymbol{\beta}} \log f(y_i; \boldsymbol{\beta}) = \sum_{h=1}^H \sum_{i=1}^N \frac{(y_i - g^{-1}(\mu_i(\boldsymbol{\beta}))) \mathbf{x}_i}{V(\mu_i(\boldsymbol{\beta})) g'(\mu_i(\boldsymbol{\beta}))}$$

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

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

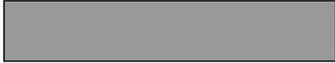
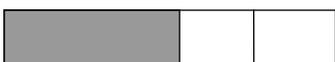
$l = 1$		$D_{hl} = (1)$ for all h
$l = 2$		$D_{hl} = (1\ 0)$ for $h = 1$ $D_{hl} = (0\ 1)$ for $h \geq 2$
$l = 3$		$D_{hl} = (1\ 0)$ for $h \leq 3$ $D_{hl} = (0\ 1)$ for $h = 4$
$l = 4$		$D_{hl} = (1\ 0)$ for $h \leq 2$ $D_{hl} = (0\ 1)$ for $h \geq 3$
$l = 5$		$D_{hl} = (1\ 0\ 0)$ for $h = 1$ $D_{hl} = (0\ 1\ 0)$ for $h = 2$ $D_{hl} = (0\ 1\ 0)$ for $h \geq 3$
$l = 6$		$D_{hl} = (1\ 0\ 0)$ for $h = 1$ $D_{hl} = (0\ 1\ 0)$ for $h = 2, 3$ $D_{hl} = (0\ 1\ 0)$ for $h = 4$
$l = 7$		$D_{hl} = (1\ 0\ 0)$ for $h \leq 2$ $D_{hl} = (0\ 1\ 0)$ for $h = 3$ $D_{hl} = (0\ 1\ 0)$ for $h = 4$
$l = 8$		$D_{hl} = (1\ 0\ 0\ 0)$ for $h = 1$ $D_{hl} = (0\ 1\ 0\ 0)$ for $h = 2$ $D_{hl} = (0\ 0\ 1\ 0)$ for $h = 3$ $D_{hl} = (0\ 0\ 1\ 0)$ for $h = 4$
	$h = 1$ $h = 2$ $h = 3$ $h = 4$	

Fig. 1. The set of $\{D_{hl}\}$ when four weight strata are present: all patterns of pooling coterminous strata

for $\theta_l = (\beta_l, \phi, L = l)$. Simulations from $p(\mathbf{B}|y, X)$ can be obtained by first obtaining a draw from $p(\theta_l|y, X)$ and then computing

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

where $W_h = N_h/n_h$ and $\hat{y}_{hi} = g^{-1}(\mathbf{z}_{li}^T \beta_l)$. 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 \mathbf{B} can be computed by solving for $B_j, j = 1, \dots, p$

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

where β_{hj} corresponds to the j th value of the β_l parameter for the h th inclusion stratum as a function of the l th pooling pattern. Thus $\beta_{hj} = \beta_j$ for all h when $l = 1$ (i.e., $\beta_j = B_j$ for the unweighted estimator); $\beta_{hj} = \beta_{1j}$ for $h = 1$, $\beta_{hj} = \beta_{2j}$ for $h > 1$ when $l = 2$; and so forth. This can be accomplished via simple root-finding numerical methods such as Newton's Method.

A direct draw of $p(\theta_l|y, X)$ is not generally possible outside of the Gaussian setting. We approximate a direct draw by using a Laplace approximation to obtain a draw from $p(L = l|y, X)$ and a Metropolis step to obtain a draw from $p(\beta_l|L = l, y, X)$; alternatively a Metropolis step (Gelman et al. 2004, pp. 289–290) may be used to obtain draws from

$p(L = l|\beta y, X)$ and a Markov Chain Monte Carlo algorithm implemented instead. See the Appendix for details.

3.2. Fractional Bayes Factors

In the absence of strong prior information to define $P(\theta_l)$, the Bayes Factors comparing weight pooling model l with weight pooling model l'

$$\begin{aligned} BF(y, X) &= \frac{p(L = l|y, X)}{p(L = l'|y, X)} = \frac{p(y|L = l, X)p(L = l)}{p(y|L = l', X)p(L = l')} \\ &= \frac{\iint p(y|\beta_l, \sigma^2 L = l, X) d\beta_l d\sigma^2 p(L = l)}{\iint p(y|\beta_{l'}, \sigma^2 L = l', X) d\beta_{l'} d\sigma^2 p(L = l')} \end{aligned}$$

can be quite sensitive to the choice of $p(\theta_l)$ (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 so as to provide a data-based proper prior for θ_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

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

Small values of b should be most efficient when 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 $h(\theta_l)$ in contrast to our proper prior, but very weakly informative priors, such 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

Because we desire models that are simultaneously more efficient than design-based estimators yet reasonably robust to model misspecification – and in general we feel that even Bayesian models should have good frequentist properties – we evaluate our proposed models in a repeated sampling context. We consider two settings where weights can be utilized to reduce bias in generalized linear regression models: model misspecification and informative sampling.

4.1. Model Misspecification

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

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

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

where $\text{expit}(\cdot) = \exp(\cdot)/(1 + \exp(\cdot))$. The object of the analysis is to obtain an estimate of the logistic population regression slope, defined as the value B_1 in the equation

$$\sum_i^N (y_i - \text{expit}(B_0 + B_1x_i)) \begin{pmatrix} 1 \\ x_i \end{pmatrix} = 0$$

A disproportional sampling scheme is 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

$$K(X; C) = \left| \frac{2C}{[1 + (2CX - .75)^2]^{3/2}} \right|$$

200 simulations are 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):

$$H_i = [X_i]$$

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

A total of $n = 1,000$ elements were sampled for each simulation (maximum normalized weight ≈ 7.5).

For priors, we consider a nearly noninformative prior of the form $\beta_l|L = l \sim N_2(0, 225I)$, which assumes that the logistic regression parameters lie between -30 and $+30$ with approximately 95% probability. We term the estimator of B_1 obtained under this model PWT. We also consider the Fractional 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. O'Hagan suggests that PWTF1 will be more efficient when 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 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 is 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 2 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 noninformative 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 trimmed weight estimator, while the weight pooling FBF estimator with the larger training fraction (PWTF2) had bias similar to the fully weighted estimator.

The unweighted estimator had substantially improved MSE (40% reduction) when the linear slope model was approximately correctly specified, but was highly biased with a moderate to large degree of misspecification. The weight pooling estimator with a noninformative prior had MSE very similar to the fully weighted estimator. The trimmed weight estimator dominated the standard fully-weighted estimator over the range of simulations considered, with MSE savings of 10–35%. The weight pooling estimators with the fractional Bayes factors had MSE reductions of more than 40% when the linear slope model was approximately correctly specified, and both were robust against model misspecification.

The unweighted estimator had poor coverage except when the linear slope model was correctly specified, or nearly so. Because of the modest sample size, the fully weighted estimator had somewhat below nominal coverage both when the model was correctly specified and when the model was badly misspecified; the trimmed weight estimator had below nominal coverage only when the linear model was misspecified. The weight pooling estimator with noninformative prior had below nominal coverage regardless of model misspecification. The fractional Bayes factor estimators generally had correct to somewhat conservative coverage for all ranges of model specification.

4.2. Informative Sampling

For the informative sampling setting, we consider a correctly specified mean model, but allow the probability of selection to be related to a known covariate that is correlated with the outcome Y . The stronger the association, the more informative the sample design will be. We generated population data as follows:

$$P(Y_i = 1|X_i) \sim \text{BERNOULLI}(\text{expit}(2 - .4 * X_i))$$

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

Table 2. Relative bias (%), squared root of mean squared error (RMSE) relative to RMSE of fully-weighted estimator, and true coverage of the 95% CI or PPI of the population logistic regression slope estimator under model misspecification

Estimator	Relative bias (%)				RMSE relative to FWT				True coverage			
	Curvature K				Curvature K				Curvature K			
	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	1.5	2.5	0.3	7.0	.99	1.00	.99	1.04	84	84	84	82
PWTF1	0.6	-5.4	-19.3	-127.5	.57	.72	.79	.92	96	94	92	94
PWTF2	-0.4	-1.2	5.6	9.0	.75	.82	.98	.80	95	94	96	98

As in the previous simulations, the object of the analysis is to obtain the logistic population regression slope, defined as the value B_1 in the equation

$$\sum_i^N (y_i - \text{expit}(B_0 + B_1 x_i)) \begin{pmatrix} 1 \\ x_i \end{pmatrix} = 0$$

An unequal probability of selection scheme is utilized:

$$(P(I_i|Y_i, X_i) \propto \exp(.5Z_i/(X_i/10 + .25) - 1)), \text{ where } Z_i \sim N(Y_i, 10^l)$$

where $l = -0.5, -0.25, 0, 0.25, 0.5$. The resulting asymptotic correlations between Y and Z are .664, .555, .447, .351, and .271. The maximum normalized weight ranges from approximately 7 for $l = -.5$ to 50 for $l = 0.5$. A total of $n = 1,000$ elements were sampled for each simulation. The priors, including the fractional Bayes factors, are identical to those used in the model misspecification simulations. The unweighted, fully weighted, and trimmed weighted estimators are also considered along with the weight pooling estimators. The FBF estimators again used training fractions of $n^{-1/2}$ for PWTF1 and .1 for PWTF2.

Table 3 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. The results were overall similar to what was observed in the model misspecification setting. The unweighted estimator had substantial bias that was only relieved when sampling was more nearly ignorable (modest correlation between Y and Z). The trimmed weight and weight pooling estimators both had very modest degrees of bias, and had equal or improved MSEs as compared with the fully weighted estimator with little reduction in coverage. Both the trimmed weighted and the fraction Bayes factor weight pooling estimators had substantial reductions in MSE as compared with the fully-weighted estimator when the sampling was more nearly ignorable.

5. Application: Estimation of Injuries to Children in Compact Extended-Cab Pickup Trucks

The Partners for Child Passenger Safety dataset consists of the disproportionate, known-probability sample from all State Farm Insurance claims since December 1998 involving at least one child occupant ≤ 15 years of age riding in a model year 1990 or newer State Farm-insured vehicle (Durbin et al. 2001). Because injuries, and especially “consequential” injuries defined as facial lacerations or other injuries rated 2 or more on the Abbreviated Injury Scale (AIS) (Association for the Advancement of Automotive Medicine 1990), are relatively rare even among children in the population of crash-related vehicle damage claims, a disproportional probability-of-inclusion sample design is utilized. The weights for this dataset are quite variable: $1 \leq w_i \leq 50$, where 9% of the weights have normalized values larger than 3. Because the treatment stratification is imperfectly associated with risk of injury (more than 15% of the population with consequential injuries are estimated to be in the lowest probability-of-selection category and nearly 20% of those without consequential injuries are in the highest

Table 3. Relative bias (%), squared root of mean squared 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 informative sampling. Correlation between dichotomous outcome Y and probability of selection covariate Z

Estimator	Relative bias (%)					RMSE relative to FWT					True coverage				
	Correlation					Correlation					Correlation				
	.66	.56	.45	.35	.27	.66	.56	.45	.35	.27	.66	.56	.45	.35	.27
UNWT	29.0	27.8	23.0	25.3	13.9	2.97	2.22	1.79	1.38	.72	4	7	22	18	69
FWT	-.7	.2	.2	3.6	2.8	1	1	1	1	1	96	94	95	89	93
TWT	3.2	5.0	3.5	2.7	0.2	.96	.90	.81	.65	.52	95	91	95	92	97
PWT	3.3	5.4	3.8	3.4	-9.0	.99	1.02	.92	.84	.96	93	91	96	96	93
PWTF1	3.8	4.0	4.1	3.5	-.7	.98	1.01	.93	.82	.71	94	88	90	92	96
PWTF2	3.2	4.0	5.8	3.6	2.1	.99	.98	.94	.89	.89	92	92	92	94	98

probability-of-selection category), the sampling design is informative, with unweighted odds ratios biased toward the null (Korn and Graubard 1995).

Winston et al. (2002) determined that children rear-seated in compacted extended-cab pickups are at greater risk of consequential injuries than children rear-seated in other vehicles. However, quantifying degree of excess risk, and thus the size of the public health problem, was somewhat problematic. The unweighted odds ratio (OR) of consequential injury for children riding in compacted extended-cab pickups versus other vehicles was 3.54 (95% CI 2.01,6.23), versus the fully weighted estimator of 11.32 (95% CI 2.67,48.02). Because both injury risk and compacted extended-cab pickup use were associated with child age, crash severity (passenger compartment intrusion and drivability), direction of impact, and vehicle weight, a multivariate logistic regression model that adjusted for these factors was also considered. The unweighted and fully weighted adjusted ORs for injury risk in rear seated children in compacted extended-cab pickups versus other vehicles are 3.50 (95% CI 1.88,6.53) and 14.56 (95% CI 3.45,61.40), respectively. Utilizing the unweighted estimator was problematic because of bias toward the null induced by the informative sample design; however, the fully weighted estimator appeared to be highly unstable. In Winston et al. (2002), injured children with high weights were deleted, yielding an estimated OR for injury risk of 3.87 (95% CI 1.42,10.57), close to the unweighted estimator.

Table 4 shows the unadjusted and the adjusted odds ratios of consequential injury risk using the unweighted, fully weighted and pooled weighted estimators of OR of injury for children in compact pickup trucks versus other vehicles. Pooled weighted estimators are obtained using both the noninformative and fractional Bayes factor priors. The noninformative pooled estimator generally falls between the unweighted and the fully weighted estimator, and reduces the unrealistically larger 95% upper limit of the OR from 50–60 down to 19–20. The FBF pooled estimator with fraction $n^{-1/2}$ reduced unadjusted OR to 6.75 (95% CI 2.51,17.99), intermediate between the unweighted and fully weighted estimator, but provided an adjusted OR estimator similar to that of the unweighted estimator. The FBF pooled estimator with fraction 0.1 reduced unadjusted OR to 7.6 (95% CI 2.6,21.8) and the adjusted OR to 7.0 (95% CI 2.8,18.2).

An additional two years of data, which included an additional 4,091 rear-seated children in passenger vehicles (44 in compact extended-cab pickup trucks), provided a fully weighted unadjusted odds ratio for injury for children in compact extended-cab pickups of 6.3, and an adjusted OR of 7.0 – very close to the FBF pooled estimator results with fraction 0.1.

Table 4. Estimated odds ratio of injury for children rear-seated in compacted extended-cab pickups ($n = 60$) versus rear-seated in other vehicles ($n = 8,060$), using unweighted (UNWT), fully weighted (FWT), and pooled weight (PWTF2) estimators; unadjusted and adjusted for child age, crash severity, direction of impact, and vehicle weight. Point estimates for PWT and PWTF2 models from posterior median. 95% CI or PPI in subscript. Data from Partners for Child Passenger Safety

	UNWT	FWT	PWT	PWTF1	PWTF2
Unadj.	3.54 _(2.01,6.23)	11.32 _(2.67,48.02)	9.03 _(2.89,20.09)	6.75 _(2.51,17.99)	7.61 _(2.64,21.76)
Adj.	3.50 _(1.88,6.53)	14.56 _(3.45,61.40)	10.59 _(4.17,50.40)	3.46 _(1.80,6.36)	6.96 _(2.77,18.19)

6. Discussion

The model discussed in this article generalizes the work of Elliott and Little (2000) and Elliott (2008), where under the former population inference was restricted to population means using a weight pooling model that mimicked weight trimming, and under the latter where population inference was restricted to the estimation of population slopes under the Gaussian error model. As in Elliott (2008), we consider a model that permits the pooling of all conterminous inclusion strata, as well as utilizing the data-based “fractional Bayes Factors” of O’Hagan (1995). Here we extend the weight pooling method to consider population regression slopes under the generalized linear model, allowing for regression models for binary, count, or other outcomes with nonnormal error terms. We obtained robust estimators that can still gain considerable efficiencies as compared with standard fully weighted estimators. The crude trimming estimators also performed well among the simulations considered, although previous work (Elliott and Little 2000; Elliott 2007) shows that these estimators are generally not robust.

We also applied the methods developed in this article to the Partners for Child Passenger Safety data to determine the excess risk of injury in a crash to rear-seated children in compacted extended-cab pickups relative to rear-seated children in other passenger vehicles. It appears that the decision in Winston et al. (2002) to eliminate a low probability-of-selection child from the analysis to stabilize the estimates led to an underestimation of the effect of exposure. The pooled estimates suggested adjusted risks of 7.2, versus the 14.6 obtained from the fully weighted sample – very close to the estimate of 7.0 that was obtained using an additional two years of data.

More generally, the methods discussed in this article 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.

7. Appendix: Obtaining the Posterior Distribution of Parameters from Weight Pooling Models

To simplify notation, we assume that the generalized linear model of interest does not contain an unknown scaling parameter ϕ (i.e., logistic or Poisson regression models where $\phi = 1$ or is fixed at an overdispersion value treated as known). Straightforward extensions of all algorithmic steps can accommodate ϕ .

7.1. Simulations from the Generalized Weight Pooling Model Using Direct Draws

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

Note that

$$p(L = l | \mathbf{y}, X) = \frac{f(\mathbf{y} | L = l, X)}{\sum_l f(\mathbf{y} | L = l, X)} \quad (7.1)$$

where

$$\begin{aligned} f(\mathbf{y} | X, L = l) &= \int f(\mathbf{y} | X, \boldsymbol{\beta}_l, L = l) p(\boldsymbol{\beta}_l | L = l) d\boldsymbol{\beta}_l \approx \int f(\mathbf{y} | X, \boldsymbol{\beta}_l, L = l) d\boldsymbol{\beta}_l \\ &\approx (2\pi)^{(pH^*)/2} |\hat{\Sigma}_{\hat{\boldsymbol{\beta}}_l}|^{1/2} f(\mathbf{y} | X, \hat{\boldsymbol{\beta}}_l, L = l) \end{aligned}$$

where $\hat{\boldsymbol{\beta}}_l$ is the MLE of a GLM regressing \mathbf{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 noninformative or nearly noninformative prior on $\boldsymbol{\beta}_l | L = l$, and the second from the Laplace approximation to the true marginal distribution of \mathbf{y} .

Draws from $p(\boldsymbol{\beta}_l | L = l, \mathbf{y}, X)$ are made by running a Metropolis algorithm (Metropolis et al. 1953; Gelman et al. 2004, pp. 289–290). Briefly, a Metropolis step is a single draw from the Metropolis algorithm, which generates a draw from a general distribution $p(\theta | \mathbf{y})$ by sampling a proposed θ^* from a “jumping” distribution $J_t(\theta^* | \theta^{(t-1)})$, where J_t is symmetric ($J_t(\theta_b | \theta_a) = J_t(\theta_a | \theta_b)$). The proposal draw is then accepted as a draw of θ_t with probability $\min(1, p(\theta^* | \mathbf{y}) / p(\theta^{(t-1)} | \mathbf{y}))$; otherwise $\theta^{(t)} = \theta^{(t-1)}$. In this setting $N(\boldsymbol{\beta}^{(t-1)}, k\hat{\Sigma}_{\hat{\boldsymbol{\beta}}_l})$ 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 $\boldsymbol{\beta}_l^{prop} = \hat{\boldsymbol{\beta}}_l + \mathbf{e}$, $\mathbf{e} \sim N(0, k\hat{\Sigma}_{\hat{\boldsymbol{\beta}}_l})$ 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

$$\min\left(1, \frac{f(\mathbf{y} | X, \{\boldsymbol{\beta}_l^{prop}, L = l\})}{f(\mathbf{y} | X, \{\boldsymbol{\beta}_l^{(0)}, L = l\})}\right)$$

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.

7.1.1. Fractional Bayes Factors

When using the FBF prior, we replace $f(\mathbf{y} | L = l, X)$ in (7.1) with

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

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

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

so that

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

7.2. 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 \mathbf{y} is independent of $\{\boldsymbol{\beta}_{k \neq l}\}$ given that $L = l$. Assuming also that $\{\boldsymbol{\beta}_l\}$ are independent for $l = 1, \dots, L$, we have that

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

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

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

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

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

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

using the Metropolis step described in 7.1.

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

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

Because computing $\prod_{j=1}^{2^{H-1}} \{p(\boldsymbol{\beta}_j|L = j)\}$ is prohibitive except when H is small, we instead used a Metropolis step suggested by Dellaportas, Forster, and Ntzoufras (2002) to obtain a draw from $L|\boldsymbol{\beta}, \mathbf{y}, X$.

1. Propose new model l' with probability $h(l, l')$.
2. Generate $\boldsymbol{\beta}_{l'}$ from the pseudoprior $p(\boldsymbol{\beta}_{l'}|L \neq l')$.
3. Accept the new model l' with probability

$$\min \left\{ 1, \frac{f(\mathbf{y}|X, \boldsymbol{\beta}_{l'}, L = l') p(\boldsymbol{\beta}_{l'}|L = l') p(\boldsymbol{\beta}_l|L = l') P(L = l') h(l', l)}{f(\mathbf{y}|X, \boldsymbol{\beta}_l, L = l) p(\boldsymbol{\beta}_l|L = l) p(\boldsymbol{\beta}_{l'}|L = l) P(L = l) h(l, l')} \right\}$$

Carlin and Chib (1995) note that poor choices for pseudo priors $p(\boldsymbol{\beta}_{k \neq l}|L = l)$ 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 \mathbf{y} on Z_l , and covariance $\hat{\Sigma}_{\hat{\boldsymbol{\beta}}_k}$ given by the inverse of the expected information

matrix. Jumping probabilities to the l' models that exclude L were always given by the uniform discrete distribution with probability $(2^{H-1} - 1)^{-1}$.

8. References

- Association for the Advancement of Automotive Medicine (1990). The Abbreviated Injury Scale, 1990 Revision. Association for the Advancement of Automotive Medicine, Des Plaines, IL, U.S.A.
- Binder, D.A. (1983). On the Variances of Asymptotically Normal Estimators from Complex Surveys. *International Statistical Review*, 51, 279–292.
- Carlin, B.P. and Chib, S. (1995). Bayesian Model Choice via Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society, Series B*, 57, 473–484.
- 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, U.S.A.
- Dellaportas, P., Forster, J., and Ntzoufras, I. (2002). On Bayesian Model and Variable Selection using MCMC. *Statistics and Computing*, 12, 27–36.
- Deville, J.C. and Särndal, 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., and 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. (2007). Bayesian Weight Trimming for Generalized Linear Regression Models. *Survey Methodology*, 33, 23–34.
- Elliott, M.R. (2008). Model Averaging Methods for Weight Trimming. *Journal of Official Statistics*, 24, 517–540.
- 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.
- Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2004). *Bayesian Data Analysis*, (2nd Edition). Boca Raton, FL: Chapman and Hall/CRC.
- 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.
- Kass, R.E. and Raftery, A.E. (1995). Bayes Factors. *Journal of the American Statistical Association*, 90, 773–795.
- Korn, E.L. and Graubard, B.I. (1995). Analysis of Large Health Surveys: Accounting for the Sampling Design. *Journal of the Royal Statistical Society, Series A*, 158, 263–295.

- Little, R.J.A. (1993). Poststratification: A Modeler's Perspective. *Journal of the American Statistical Association*, 88, 1001–1012.
- 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.
- McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*, (2nd Edition). London: Chapman and Hall.
- Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E. (1953). Equations of State Calculations by Fast Computing Machines. *Journal of Chemical Physics*, 21, 1087–1092.
- O'Hagan, A. (1995). Fraction Bayes Factors for Model Comparison. *Journal of the Royal Statistical Society, Series B*, 57, 99–138.
- Potter, F. (1990). A 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.
- Skinner, C.J., Holt, D., and Smith, T.M.F. (1989). *Analysis of Complex Surveys*. New York: Wiley.
- Tierney, L. and 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., and 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.

Received August 2006

Revised December 2007