Survey Methodology, June 2007
Vol. 33, No. 1, pp. 23-34
Statistics Canada, Catalogue No. 12-001



# Bayesian weight trimming for generalized linear regression models

**Michael R. Elliott** [1]

Abstract

In sample surveys where units have unequal probabilities of inclusion in the sample, associations between the probability of inclusion and the statistic of interest can induce bias. Weights equal to the inverse of the probability of inclusion are often used to counteract this bias. Highly disproportional sample designs have large weights, which can introduce undesirable variability in statistics such as the population mean estimator or population regression estimator. Weight trimming reduces large weights to a fixed cutpoint value and adjusts weights below this value to maintain the untrimmed weight sum, reducing variability at the cost of introducing some bias. Most standard approaches are ad-hoc in that they do not use the data to optimize bias-variance tradeoffs. Approaches described in the literature that are data-driven are a little more efficient than fully-weighted estimators. This paper develops Bayesian methods for weight trimming of linear and generalized linear regression estimators in unequal probability-of-inclusion designs. An application to estimate injury risk of children rear-seated in compact extended-cab pickup trucks using the Partners for Child Passenger Safety surveillance survey is considered.

Key Words: Sample survey; Sampling weights; Weight Winsorization; Bayesian population inference; Weight smoothing; Generalized linear mixed models.

## 1. Introduction

Analysis of data from samples with differential probabilities of inclusion typically use case weights equal to the inverse of the probability of inclusion to reduce or remove bias in the estimators of population quantities of interest. Replacing implicit means and totals in statistics with their case-weighted equivalents yields unbiased linear estimators and asymptotically unbiased non-linear estimators of population values (Binder 1983). Case weights may also incorporate non-response adjustments, which typically are equal to the inverse of the estimated probability of response (Gelman and Carlin 2002, Oh and Scheuren 1983), or calibration adjustments, which constrain case weights to equal known population totals, either jointly, as in poststratification or generalized regression estimation, or marginally, as in generalized raking estimation (Deville and Särndal 1992, Isaki and Fuller 1982).

There is little debate that sampling weights be utilized when considering descriptive statistics such as means and totals obtained from unequal probability-of-selection designs. However, when estimating "analytical" quantities (Cochran 1977, page 4) that focus on associations between, *e.g.*, risk factors and health outcomes estimated via linear and generalized linear models, the decision to use sampling weights is less definitive (*cf* Korn and Graubard 1999, pages 180-182). In a regression setting, discrepancies between weighted and unweighted regression slope estimators can occur either because the data model is misspecified or there is an association between the residual errors and/or the probability of inclusion (sampling is informative). When the data model is misspecified, one option is to improve the model specification. However, it may be difficult to determine the exact functional form; or it may be that the degree of misspecification is very modest but is magnified by the sample design; or it may be that an approximation to the true model is desired to simplify explanation (linearly approximating a quadratic trend). In the case of informative or non-ignorable sampling, design weights may be required to obtain consistent estimators of regression parameters (Korn and Graubard 1995). More formally, fully-weighted estimators of regression parameters are "pseudo-maximum likelihood" estimators (PMLEs) (Binder 1983, Pfeffermann 1993) in that they are "design consistent" for MLEs that would solve the score equations for the regression parameters under the assumed superpopulation regression model if we had observed data for the entire population. Design consistency implies that the difference between the population target quantity and the estimate derived from the sample tends to zero as the sample size and population size jointly increase, or that these differences will on average tend to 0 from repeated sampling of the population, where samples are selected in an identical fashion from $t \to \infty$ replicates of the population: see Särndal (1980) or Isaki and Fuller (1982). If observations are clustered, more care must be taken to develop design consistent estimators of PLMEs, although nested multi-stage designs allow for the census log-likelihood estimates to be approximated using weighted score equations if care is taken to account for the fact that the within-cluster sample sizes typically are small and remain so even if the number of clusters increases

1. Michael R. Elliott is Assistant Professor, Department of Biostatistics, University of Michigan School of Public Health, 1420 Washington Heights, Ann Arbor, MI. E-mail: mrelliot@umich.edu.

(Pfeffermann, Skinner, Holmes, Goldstein and Rabash 1998, Korn and Graubard 2003).

Although PMLEs are popular because of design consistency, this property is purchased at the cost of increased variance. This increase can overwhelm the reduction in bias, so that the MSE actually increases under a weighted analysis. This is particularly likely if a) the sample size is small, b) the differences in the inclusion probabilities are large, or c) the model is approximately correctly specified and the sampling is approximately noninformative. Perhaps the most common approach to dealing with this problem is *weight trimming* (Potter 1990, Kish 1992, Alexander, Dahl and Weidman 1997), in which weights larger than some value $w_0$ are fixed as $w_0$. Typically $w_0$ is chosen in an *ad hoc* manner - say 3 or 6 times the mean weight - without regard to whether the chosen cutpoint is optimal with respect to MSE. Thus bias is introduced to reduce variance, with the goal of an overall reduction in MSE.

Other design-based methods have been considered in the literature. Potter (1990) discusses systematic methods for choosing $w_0$, including 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 $1 - F(w_i) < 0.01$, are trimmed to $w_0$ such that $1 - F(w_0) = 0.01$. The MSE trimming procedure determines the empirical MSE at trimming level $w_t$, where the trimmed weight $w_i^* = w_t I(w_i \geq w_t) + w_i I(w_i < w_t)$, $i = 1, ..., n$ under the assumption that the fully weighted estimate is unbiased for the true mean. In practice, one considers a variety of trimming levels $t = 1, ..., T$, where $t = 1$ corresponds to the unweighted data $(w_1 = \min_i(w_i))$ and $t = T$ to the fully-weighted data $(w_T = \max_i(w_i))$, and $\hat{\theta}_t$ is the value of the statistic using the trimmed weights at level $t$. The trimming level chosen is then given by $w_0 = w_{t^*}$, where $t^* = \text{argmin}_t(\text{MSE}_t)$ for $\text{MSE}_t = (\hat{\theta}_t - \hat{\theta}_T)^2 + \hat{V}(\theta_t)$.

In the calibration literature, techniques have been developed that allow generalized poststratification or raking adjustments to be bounded to prevent the construction of extreme weights (Deville and Särndal 1992, Folsom and Singh 2000). Beaumont and Alavi (2004) extend this idea to develop estimators that focus on trimming large weights of highly influential or outlying observations. While these bounds trim extreme weights to a fixed cutpoint value, the choice of this cutpoint remains arbitrary.

An alternative approach to the direct weight trimming procedures has been developed in the Bayesian finite population inference literature (Elliott and Little 2000, Holt and Smith 1979, Ghosh and Meeden 1986, Little 1991, 1993, Lazzeroni and Little 1998, Rizzo 1992). These approaches account for unequal probabilities of inclusion by 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. Standard weighted estimates are then obtained when the weight stratum means of survey outcomes are treated as fixed effects, and trimming of the weights is achieved by treating the underlying weight stratum means as random effects. These methods allow for the possibility of "partially-weighted" data that uses the data itself to appropriately modulate the bias-variance tradeoff, and also allows estimation and inference from data collected under unequal probability-of-inclusion sample designs to be based on models common to other fields of statistical estimation and inference.

This paper extends these random-effects models, which we term "weight smoothing" models, to include estimation of population parameters in linear and generalized linear models. Section 2 briefly reviews Bayesian finite population inference, formalizes the concept of ignorable and non-ignorable sampling mechanisms, and develops the weight smoothing models for linear and generalized linear regression models in a fully Bayesian setting. Section 3 provides simulation results to consider the repeated sampling properties of the weight smoothing estimators of linear and logistic regression parameters in a disproportional-stratified sample design and compares them with standard design-based estimators. Section 4 illustrates the use of the weight smoothing estimators in an analysis of risk of injury to children in passenger vehicle crashes. Section 5 summarizes the results of the simulations and considers extensions to more complex sample designs.

## 2. Bayesian finite population inference

Let the population data for a population with $i = 1, ..., N$ units be given by $Y = (y_1, ..., y_N)$, with associated covariate vectors $X = (\boldsymbol{x}_1, ..., \boldsymbol{x}_N)$ and sampling indicator variable $I = (I_1, ..., I_N)$, where $I_i = 1$ if the $i^{\text{th}}$ element is sampled and 0 otherwise. As in design-based population inference, Bayesian population inference focuses on population quantities of interest $Q(Y)$, such as population means $Q(Y) = \bar{Y}$ or population least-squares regression parameters $Q(Y, X) = \min_{B_0, B_1} \sum_{i=1}^{N} (y_i - B_0 - B_1 x_i)^2$. In contrast to design-based inference, but consistent with most other areas of statistics, one posits a model for the population data $Y$ as a function of parameters $\theta$:

$Y \sim f(Y|\theta)$. Inference about $Q(Y)$ is made based on the posterior predictive distribution of $p(Y_{\text{nob}} | Y_{\text{obs}}, I)$, where $Y_{\text{nob}}$ consists of the elements of $Y_i$ for which $I_i = 0$:

$$p(Y_{\text{nob}} | Y_{\text{obs}}, I) = \frac{\iint p(Y_{\text{nob}} | Y_{\text{obs}}, \theta, \phi) p(I | Y, \theta, \phi) p(Y_{\text{obs}} | \theta) p(\theta, \phi) d\theta d\phi}{\iiint p(Y_{\text{nob}} | Y_{\text{obs}}, \theta, \phi) p(I | Y, \theta, \phi) p(Y_{\text{obs}} | \theta) p(\theta, \phi) d\theta d\phi dY_{\text{nob}}} \quad (1)$$

where $p(I | Y, \theta, \phi)$ models the inclusion indicator.

If we assume that $\phi$ and $\theta$ are *a priori* independent and if the distribution of sampling indicator $I$ is independent of $Y$, the sampling design is said to be "unconfounded" or "noninformative"; if the distribution of $I$ depends only on $Y_{\text{obs}}$, then the sampling mechanism is said to be "ignorable" (Rubin 1987), equivalent to the standard missing data terminology (the unobserved elements of the population can be thought of as missing by design). Under ignorable sampling designs, $p(\theta, \phi) = p(\theta)p(\phi)$ and $p(I | Y, \theta, \phi) = p(I | Y_{\text{obs}}, \phi)$, and thus (1) reduces to

$$\frac{\int p(Y_{\text{nob}} | Y_{\text{obs}}, \theta) p(Y_{\text{obs}} | \theta) p(\theta) d\theta}{\iint p(Y_{\text{nob}} | Y_{\text{obs}}, \theta) p(Y_{\text{obs}} | \theta) p(\theta) d\theta dY_{\text{nob}}} = p(Y_{\text{nob}} | Y_{\text{obs}}), \quad (2)$$

allowing inference about $Q(Y)$ to be made without explicitly modeling the sampling inclusion parameter $I$ (Ericson 1969, Holt and Smith 1979, Little 1993, Rubin 1987, Skinner, Holt and Smith 1989). Noninformative sample designs are a special case of ignorable sample designs, equivalent to missing completely at random mechanisms being a special case of missing at random mechanisms.

In the regression setting, where inference is desired about parameters that govern the distribution of $Y$ conditional on fixed and known covariates $X$, (1) becomes

$$p(Y_{\text{nob}} | Y_{\text{obs}}, X, I) = \frac{\iint p(Y_{\text{nob}} | Y_{\text{obs}}, X, \theta, \phi) \times p(I | Y, X, \theta, \phi) p(Y_{\text{obs}} | X, \theta) p(\theta, \phi) d\theta d\phi}{\iiint p(Y_{\text{nob}} | Y_{\text{obs}}, X, \theta, \phi) \times p(I | Y, X, \theta, \phi) p(Y_{\text{obs}} | X, \theta) p(\theta, \phi) d\theta d\phi dY_{\text{nob}}}$$

which reduces to

$$p(Y_{\text{nob}} | Y_{\text{obs}}, X) = \frac{\int p(Y_{\text{nob}} | Y_{\text{obs}}, X, \theta, \phi) p(Y_{\text{obs}} | X, \theta) p(\theta, \phi) d\theta}{\iint p(Y_{\text{nob}} | Y_{\text{obs}}, X, \theta, \phi) p(Y_{\text{obs}} | X, \theta) p(\theta, \phi) d\theta dY_{\text{nob}}}$$

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

### 2.1 Accommodating unequal probabilities of inclusion

Maintaining the ignorability assumption for the sampling mechanism often requires accounting for the sample design in both the likelihood and prior model structure. In the case of the unequal probability-of-inclusion sample designs, this can be accomplished by developing an index $h = 1, ..., H$ of the probability of inclusion (Little 1983, 1991); this could either be a one-to-one mapping of the case weight order statistics to their rankings, or a preliminary "pooling" of the case weights using, *e.g.*, the $100/H$ percentiles of the case weights. The data are then modeled by

$$y_{hi} | \theta_h \sim f(y_{hi}; \theta_h), \; i = 1, ..., N_h$$

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

To make this concrete, assume we are interested in estimating a population mean $Q(Y) = \bar{Y} = N^{-1}\sum_{i=1}^{N} y_i$ from a unequal probability-of-inclusion sample with a simple random sample within inclusion strata. Rewriting as $Q(Y) = \sum_h P_h \bar{Y}_h$ where $\bar{Y}_h = N_h^{-1} y_{hi}$ is the population inclusion stratum mean and $P_h = N_h/N$, we have

$$E(\bar{Y} | Y_{\text{obs}}) = \sum_h P_h E(\bar{Y}_h | Y_{\text{obs}}) = N^{-1} \sum_h \{ n_h \bar{y}_{h,\text{obs}} + (N_h - n_h) E(\bar{Y}_{h,\text{nob}} | Y_{\text{obs}}) \}$$

where $\bar{Y}_h$ is decomposed into the observed inclusion stratum mean $\bar{y}_{h,\text{obs}} = n_h^{-1} \sum_{i=1}^{N_h} I_{hi} y_{hi}$ and the unobserved inclusion stratum mean $\bar{Y}_{h,\text{nob}} = (N_h - n_h)^{-1} \sum_{i=1}^{N_h} (1 - I_{hi}) y_{hi}$. If we assume

$$y_{hi} | \mu_h, \sigma_h^2 \overset{\text{ind}}{\sim} N(\mu_h, \sigma_h^2)$$

$$p(\mu_h, \sigma_h^2) \propto 1$$

then

$$E(\bar{Y}_{h,\text{nob}} | Y_{\text{obs}}) = E(E(\bar{Y}_{h,\text{nob}} | Y_{\text{obs}}) | Y_{\text{obs}}, \mu_h, \sigma_h^2) = E(\mu_h | Y_{\text{obs}}) = \bar{y}_{h,\text{obs}}.$$

and the posterior predictive mean of the population mean is given by the weighted sample mean:

$$E(\bar{Y} | Y_{\text{obs}}) = \sum_h P_h E(\bar{Y}_h | Y_{\text{obs}}) = N^{-1} \sum_h N_h \bar{y}_{h,\text{obs}} = N^{-1} \sum_h \sum_{i=1}^{N_h} I_{hi} w_h y_{hi}$$

where $w_{hi} \equiv w_h = N_n/n_h$ for all the observed elements in inclusion stratum $h$. Further, the weighted mean will be the posterior predictive expectation of the population mean for any assumed distribution of $Y$ as long as $E(y_{hi} \mid \mu_h) = \mu_h$. In contrast, a simple exchangeable model for the data

$$y_i \mid \mu, \sigma^2 \overset{\text{ind}}{\sim} N(\mu, \sigma^2)$$

$$p(\mu, \sigma^2) \propto 1$$

yields $E(\bar{Y} \mid Y_{\text{obs}}) = n^{-1}\sum_{i=1}^{N} I_i y_i$, the unweighted estimator of the mean, which may be badly biased if exchangeability fails to hold, as would be the case if there is an association between the probability of inclusion and $Y$.

### 2.2 Weight smoothing models

In its general form, our proposed "weight smoothing method" stratifies the data by the probability of inclusion and then uses a hierarchical model to effect trimming via shrinkage. A general description of such a model is given by

$$y_{hi} \mid \theta_h \sim f(y_{hi}; \theta_h) \tag{3}$$

$$\theta_h \mid M_h, \mu, R \sim N(\hat{y}_h, R), \hat{y}_h = g(M_h, \mu)$$

$$\mu, R \mid M_h \sim \Pi.$$

where $h = 1, ..., H$ indexes the probability of inclusion from the highest to the lowest probabilities, $g(M_h, \mu)$ is a function linking information $M_h$ from the inclusion probability stratum and a smoothing parameter $\mu$ to the data distribution parameter $\theta_h$ indexed by the inclusion stratum, and $\Pi$ is a flat or weakly informative hyperparameter distribution (Little 2004).

The particulars of the likelihood and prior specifications will depend on the population parameter of interest, the sample design, distributional assumptions about $y$, and efficiency-robustness tradeoffs. Positing an exchangeable model on the inclusion stratum means from the previous example yields (Lazzeroni and Little 1998, Elliott and Little 2000)

$$y_{hi} \mid \theta_h \overset{\text{ind}}{\sim} N(\theta_h, \sigma^2)$$

$$\theta_h \overset{\text{ind}}{\sim} N(\mu, \tau^2).$$

Assuming for the moment $\sigma^2$ and $\tau^2$ known, we have

$$E(\bar{Y} \mid Y_{\text{obs}}) = N^{-1}\sum_h \{n_h \bar{y}_{h,\text{obs}} + (N_h - n_h)E(\mu_h \mid Y_{\text{obs}})\}$$

where $E(\mu_h \mid Y_{\text{obs}}) = w_h\bar{y}_h + (1 - w_h)\tilde{y}$ for $w_h = \tau^2 n_h/(\tau^2 n_h + \sigma^2)$ and $\tilde{y} = (\sum_h n_h/(n_h\tau^2 + \sigma^2))^{-1}\sum_h n_h/(n_h\tau^2 + \sigma^2)\bar{y}_h$. As $\tau^2 \to \infty$, $w_h \to 1$ so that $E(\bar{Y} \mid Y_{\text{obs}}) = \sum_h P_h \bar{y}_h$; thus a flat prior recovers the fully-weighted estimator, as we showed previously. On the other hand, as $\tau^2 \to 0$, $w_h \to 0$ so that $E(\mu_h \mid Y_{\text{obs}}) \to \tilde{y} \mid_{\tau^2=0} = \bar{y}$, the unweighted mean; thus the excluded units of the sample are estimated at the pooled mean since the model assumes that all $y_{hi}$ are drawn from a common mean. Hence this weight smoothing model allows compromise between the design-consistent estimator which may be highly inefficient, and the unweighted estimator that is fully efficient under the strong assumption that the inclusion probability and mean of $Y$ are independent. By assuming a weak hyperprior distribution on $\tau^2$, the degree of compromise between the weighted and unweighted mean will be "data-driven," albeit under the modeling assumptions.

### 2.3 Weight smoothing for linear and generalized linear regression models

Generalized linear regression models (McCullagh and Nelder 1989) postulate a likelihood for $y_i$ of the form

$$f(y_i; \theta_i, \phi) = \exp\left[\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi)} + c(y_i, \phi)\right] \tag{4}$$

where $a_i(\phi)$ involves a known constant and a (nuisance) scale parameter $\phi$, 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 $\text{Var}(y_i \mid \theta_i) = a_i(\phi)V(\mu_i)$, where $V(\mu_i) = b''(\theta_i)$. The link is canonical if $\theta_i = \eta_i$, in which case $g'(\mu_i) = V^{-1}(\mu_i)$. Well-known examples are the normal distribution, where $a_i(\phi) = \sigma^2$ and the canonical link is $g(\mu_i) = \mu_i$; the binomial distribution, where $a_i(\phi) = n_i^{-1}$ and the canonical link is $g(\mu_i) = \log(\mu_i/(1 - \mu_i))$; and the Poisson distribution, where $a_i(\phi) = 1$ and the canonical link is $g(\mu_i) = \log(\mu_i)$.

Indexing the inclusion stratum by $h$, we have $g(E[y_{hi} \mid \boldsymbol{\beta}_h]) = \boldsymbol{x}_{hi}^T\boldsymbol{\beta}_h$. We assume a hierarchical model of the form

$$(\boldsymbol{\beta}_1^T, ..., \boldsymbol{\beta}_H^T)^T \mid \boldsymbol{\beta}^*, G \sim N_{Hp}(\boldsymbol{\beta}^*, G). \tag{5}$$

where $\boldsymbol{\beta}^*$ is an unknown vector of mean values for the regression coefficients and $G$ is an unknown covariance matrix.

We consider the target population quantity of interest $\boldsymbol{B} = (B_1, ..., B_p)^T$ to be the slope that solves the population score equation $U_N(\boldsymbol{B}) = 0$ where

$$U_N(\boldsymbol{\beta}) = \sum_{i=1}^{N} \frac{\partial}{\partial \boldsymbol{\beta}} \log f(y_i; \boldsymbol{\beta}) = \sum_{h=1}^{H} \sum_{i=1}^{N_h} \frac{(y_{hi} - g^{-1}(\mu_i(\boldsymbol{\beta})))x_{hi}}{V(\mu_{hi}(\boldsymbol{\beta}))g'(\mu_{hi}(\boldsymbol{\beta}))}. \quad (6)$$

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 misspecificed (*i.e.*, $\eta_i$ is not exactly linear with respect to the covariates), since it is the linear approximation of $\boldsymbol{x}_i$ to $\eta_i = g(\mu_i)$. Under the model given by (4) and (5), a first-order approximation (assuming a negligible sampling fraction) to $E(\boldsymbol{B} \mid y, X)$ is given by $\hat{B}$ where

$$\sum_{h=1}^{H} W_h \sum_{i=1}^{n_h} \frac{(\hat{y}_{hi} - g^{-1}(\mu_i(\hat{\boldsymbol{B}})))\boldsymbol{x}_{hi}}{V(\mu_{hi}(\hat{\boldsymbol{B}}))g'(\mu_{hi}(\hat{\boldsymbol{B}}))} = 0 \quad (7)$$

where $W_h = N_h/n_h$, $\hat{y}_{hi} = g^{-1}(\boldsymbol{x}_{hi}^T\hat{\boldsymbol{\beta}}_h)$, and $\hat{\boldsymbol{\beta}}_h = E(\boldsymbol{\beta}_h \mid y, X)$, as determined by the form of (5). (If $N_h$ is unknown, it can be replaced with $\hat{N}_h = \sum_{i \in h} w_{hi}$, and the $\hat{N}_1, ..., \hat{N}_H$ treated as a multinomial distribution of size $N$ parameterized by unknown inclusion stratum probabilities $q_1, ..., q_H$ with, *e.g.*, a Dirichlet prior.) Thus, in the example of linear regression, where $V(\mu_i) = \sigma^2$ and $g'(\mu_i) = 1$, (7) resolves to

$$\hat{\boldsymbol{B}} = E(\boldsymbol{B} \mid y, X) = \left[\sum_h W_h \sum_{i=1}^{n_h} \boldsymbol{x}_{hi}\boldsymbol{x}_{hi}^T\right]^{-1}\left[\sum_h W_h \left(\sum_{i=1}^{n_h} \boldsymbol{x}_{hi}\boldsymbol{x}_{hi}^T\right)\hat{\boldsymbol{\beta}}_h\right]. \quad (8)$$

In the example of logistic regression, where $V(\mu_i) = \mu_i(1-\mu_i)$ and $g'(\mu_i) = \mu_i^{-1}(1-\mu_i)^{-1}$, $E(\boldsymbol{B} \mid y, X)$ is given by solving for the population regression parameters $B_j$, $j = 1, ..., p$

$$\sum_{h=1}^{H} W_h \sum_{i=1}^{n_h} x_{hij} \frac{\exp(x_{hij}B_j)}{1 + \exp(x_{hij}B_j)} = \sum_{h=1}^{H} W_h \sum_{i=1}^{n_h} x_{hij} \frac{\exp(x_{hij}\hat{\beta}_{hj})}{1 + \exp(x_{hij}\hat{\beta}_{hj})}. \quad (9)$$

This can be accomplished via simple root-finding numerical methods such as Newton's Method.

We consider four forms of $\boldsymbol{\beta}^*$ and $G$ in (5) in this paper:

1. Exchangeable Random Slope (XRS):
   $\boldsymbol{\beta}_h^* = (\beta_0^*, ..., \beta_p^*)$ for all $h$, $G = I_H \otimes \sum$. (10)
2. Autoregressive Random Slope (ARS):
   $\boldsymbol{\beta}_h^* = (\beta_0^*, ..., \beta_p^*)$ for all $h$,
   $G = A \otimes \sum$, $A_{jk} = \rho^{|j-k|}$, $j, k = 1, ..., H$.
3. Linear Random Slope (LRS):
   $\boldsymbol{\beta}_h^* = (\beta_{00}^* + \beta_{01}^* h, ..., \beta_{p0}^* + \beta_{p1}^* h)$,
   $G = I_H \otimes \Lambda$.
4. Nonparametric Random Slope (NPRS):
   $\boldsymbol{\beta}_h^* = (f_0(h), ..., f_p(h))$, $G = 0$.

$$\left\{ f_j : f_j^{v} \text{absolutely continuous, } v = 0,1, \int (f_j^{(2)}(u))^2 du < \infty, \min_{f_j} \sum_h (\beta_{hj}^* - f_j(h))^2 + \lambda_j \int (f_j^{(2)}(u))^2 du \right\}$$

where $h$ again indexes the probability of inclusion, $I_H$ is an $H \times H$ identity matrix, $\rho$ is an autocorrelation parameter that controls the degree of shrinkage across the weight strata, $\sum$ is an unconstrained $p \times p$ covariance matrix, $\Lambda$ is a $p \times p$ diagonal matrix, and $f_j(h)$ is a twice differentiable smooth function of $h$ that minimizes the residual sum of squares plus a roughness penalty parameterized by $\lambda_j$ (Wahba 1978, Hastie and Tibshirani 1990). Reformulating the NPRS model as in Wang (1998) we have

$$y_{hi} \mid \boldsymbol{\beta}_h \overset{\text{ind}}{\sim} N(\boldsymbol{x}_{hi}^T\boldsymbol{\beta}_h, \sigma^2)$$

$$\beta_{hj} = \beta_{j0}^* + \beta_{j1}^* h + \boldsymbol{\omega}_h \boldsymbol{u}_j$$

$$\boldsymbol{u}_j \overset{\text{ind}}{\sim} N_{H-1}(0, I\tau_j^2), \tau_j^2 = \sigma^2/(H\lambda_j)\ j = 0, ..., p$$

where $\boldsymbol{\omega}_h$ is the $h^{\text{th}}$ row of Choleski decomposition of the cubic spline basis matrix $\Omega$ where $\Omega_{hk} = \int_0^1 ((h-1)/(H-1) - t)_+((k-1)/(H-1) - t)_+ dt$, $(x)_+ = x$ if $x \geq 0$ and $(x)_+ = 0$ if $x < 0$, $h, k = 1, ..., H$. The NPRS model can be extended into the generalized linear model form as in Lin and Zhang (1999), where the first-stage normality assumption is replaced with a link function that is linear in the covariates: $g(E(y_{hi} \mid \boldsymbol{\beta}_h)) = \boldsymbol{x}_{hi}^T\boldsymbol{\beta}_h$, for $g(\cdot)$ as in (4).

Assuming for the moment that the second stage parameters are known, we see that, in the case of the XRS model with normal data, as $|G| \to \infty$, sharing of information across inclusion strata ceases, and $\hat{\boldsymbol{\beta}}_h \approx (\boldsymbol{x}_h^T\boldsymbol{x}_h)^{-1}\boldsymbol{x}_h^T\boldsymbol{y}_h$, the regression estimator within the inclusion stratum. Replacing this into (8) yields $\hat{\boldsymbol{B}} \approx \hat{\boldsymbol{B}}^w$, the fully weighted estimator of the population slope. Similarly, as $|G| \to 0$, the within-inclusion-stratum slopes $\hat{\boldsymbol{\beta}}_h \approx \boldsymbol{\beta}^*$ the common prior slope, yielding $\hat{\boldsymbol{B}} \approx \boldsymbol{\beta}^*$ when replaced in (8), or $\hat{\boldsymbol{B}}^u$ if a non-informative hyperprior distribution is placed on $\boldsymbol{\beta}^*$ and its posterior mean obtained as $(\boldsymbol{x}^T\boldsymbol{x})^{-1}\boldsymbol{x}^T\boldsymbol{y}$. Empirical or fully Bayesian methods that allow the data to estimate the second stage parameters thus allow for data-driven "weight smoothing," compromising between the unweighted and fully-weighted estimators.

In practice, of course, the second-stage mean and variance components are usually not known; hence we

complete the model specification by postulating a hyperprior distribution for the second-stage parameters:

$$p(\phi, \boldsymbol{\beta}^*, G) \propto p(\boldsymbol{\zeta}).$$

Typically the hyperprior distribution $p(\boldsymbol{\zeta})$ is either weakly informative or non-informative. Gibbs sampling (Gelfand and Smith 1990; Gelman and Rubin 1992) can then be utilized to obtained draws from the full joint posterior of $(\boldsymbol{\beta}, \boldsymbol{\beta}^*, \phi, G)^T \mid \boldsymbol{y}, X$. In the XRS model, we consider $p(\sigma, \boldsymbol{\beta}^*, \Sigma) \propto \sigma^{-2} \mid \Sigma \mid^{-(p+1/2)} \exp(-1/2 \text{ tr } \{r\Sigma^{-1}\})$, that is, non-informative prior distributions on the scale and prior mean parameters and an independent inverse-Wishart hyperprior distribution on the prior variance $G$ centered at the identity matrix scaled by $r$ with $p$ degree of freedom. The same prior distribution is used for the ARS model, with the additional assumption that $\rho \overset{\text{ind}}{\sim} U(0,1)$ (non-negative autocorrelation between inclusion strata). In the LRS and NPRS models, $p(\sigma, \boldsymbol{\beta}^*, \Lambda) \propto \sigma^{-2}$ and $p(\sigma, \boldsymbol{\beta}^*, \boldsymbol{\tau}) \propto \sigma^{-2}$ (standard non-informative scale prior distribution and hyperprior distribution). Description of the conditional draws of the Gibbs sampler are available at http://www.sph.umich.edu/mrelliot/trim/meth2.pdf.

The degree of compromise is a function of the mean and variance structure of the chosen model. The XRS and ARS models assume exchangeable slope means; the ARS model is more flexible in that its variance structure allows units with more nearly equal probabilities of inclusion to be smoothed more heavily than units with very unequal probabilities of inclusion. The LRS model assumes an underlying linear trend in slopes, whereas the NPRS model assumes only an underlying trend smooth up to its second derivative. Note that, in the LRS and NPRS models, we assume *a priori* independence for the regression parameters associated with a given covariate, *i.e.*, $(\beta_{1j}, \ldots \beta_{Hj}) \perp (\beta_{1j'}, \ldots \beta_{Hj'})$, $j \neq j'$. This is because we model trends in these parameters across the inclusion stratum, and do not wish to "link up" these trends across the covariates.

Shrinkage will be greatest, corresponding to the most severe weight trimming, when the weight stratum slopes have little variability, or when the lowest probability-of-inclusion stratum are poorly estimated. Little shrinkage should occur when weight stratum slopes are precisely estimated and when they are systematically associated with their probability of inclusion. Based on Elliott and Little (2000), we would expect the XRS model to be the most efficient when large amounts of weight trimming are required to minimize MSE, but to be the most vulnerable to "overshrinking" when bias correction is most important. Increasing structure, particularly in the mean portion of the model as in LRS and NPRS, will provide more robust estimation in the sense that overshrinkage will occur only in near-pathological situations (*e.g.*, when mean trends are non-monotonic and highly discontinuous), and even then may only lead to slightly less bias correction than the data warrant. The price to be paid for this robustness, however, will be a reduction in efficiency relative to the exchangeable models.

## 3. 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 linear and logistic regression, under a misspecified model with a non-informative sampling design.

### 3.1 Linear regression

For the linear regression model in the presence of model misspecification, we generated population data as follows:

$$Y_i \mid X_i, \sigma^2 \sim N(\alpha X_i + \beta X_i^2, \sigma^2), \qquad (11)$$

$$X_i \sim U(0, 10), i = 1, \ldots, N = 20,000.$$

A noninformative, disproportionally stratified sampling scheme sampled elements as a function of $X_i$ ($I_i$ equals 1 if sampled and 0 otherwise):

$$h_i = \lceil X_i \rceil$$

$$P(I_i = 1 \mid h_i) = \pi_i \propto (1 + h_i/2.5) h_i$$

This created 10 strata, defined by the integer portions of the $X_i$ values. Elements $(Y_i, X_i)$ had $\approx 1/36^{\text{th}}$ the selection probability when $0 < X_i \leq 1$ as when $9 < X_i < 10$. We sampled $n = 500$ elements without replacement for each simulation. The object of the analysis is to obtain the population slope $B_1 = \sum_{i=1}^N (Y_i - \bar{Y})(X_i - \bar{X}) / \sum_{i=1}^N (X_i - \bar{X})^2$. We fixed $\alpha = \beta = 1$, yielding a positive bias in the estimate of $B_1$, and varied $\sigma^2$. 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 $\sigma^2 \to \infty$. We considered values of $\sigma^2 = 10^l$, $l = 1, \ldots, 5$; 200 simulations were generated for each value of $\sigma^2$.

Here and below we utilized an inverse-Wishart hyperprior distribution on the prior variance $G$, centered at the identity matrix with 2 degree of freedom.

In addition to the exchangeable random slope (XRS), autoregressive random slope (ARS), linear random slope (LRS), and nonparametric random slope (NPRS) models

discussed in Section 2.3, we consider the standard designed-based (fully weighted) estimator, as well as trimmed weight and unweighted estimators. For the fully-weighted (FWT) estimator, we use the PMLE $\boldsymbol{B}_w = (X'WX)^{-1}X'W\boldsymbol{y}$ where, denoting by lower case the sampled elements ($I_i = 1$), $w_{hi} \equiv w_h$ for $h = 1, ..., H$, $i = 1, ..., n_h$, $W = \text{diag}(w_{hi})$, $\boldsymbol{x}_{hi} = (1\ x_{hi})'$, $X_h$ contains the stacked rows of $\boldsymbol{x}'_{hi}$ and $X$ contain the stacked matrices $X_h$. We obtained inference about $\hat{\boldsymbol{B}}_w$ via the standard Taylor Series approximation (Binder 1983):

$$\text{Var}(\hat{\boldsymbol{B}}_w) = \hat{S}_{XX}^{-1}\hat{\sum}(\hat{\boldsymbol{B}}_w)\hat{S}_{XX}^{-1}$$

where $\hat{S}$ is a design-consistent estimator of the population total $\sum_{i=1}^{N}\boldsymbol{x}'_i\boldsymbol{x}_i$ given by $X'WX$ and $\hat{\sum}(\hat{\boldsymbol{B}}_w)$ is a design-consistent estimate of the variance of the total $\sum_{i=1}^{N}\epsilon_i\boldsymbol{x}_i$ where $\epsilon_i = y_i - \boldsymbol{x}_i\boldsymbol{B}$ is the difference between the value of $y_i$ and its estimated value under the true population slope $\boldsymbol{B}$: $\hat{\sum}(\hat{\boldsymbol{B}}_w) = \sum_h n_h/(n_h-1)\sum_{i=1}^{n_h}(\tilde{\boldsymbol{x}}_{hi} - \bar{\tilde{\boldsymbol{x}}}_h)'(\tilde{\boldsymbol{x}}_{hi} - \bar{\tilde{\boldsymbol{x}}}_h)$, where $\tilde{\boldsymbol{x}}_{hi} = w_{hi}e_{hi}\boldsymbol{x}_{hi}$ for $e_{hi} = y_{hi} - \boldsymbol{x}_{hi}\hat{\boldsymbol{B}}_w$. We also consider the trimmed (TWT) estimator obtained by replacing the weights $w_{hi}$ with trimmed values $w^t_{hi}$ that set the maximum normalized value to 3: $w^t_{hi} = N\tilde{w}^t_{hi}/\sum_{h=1}^{H}n_h\tilde{w}^t_h$, where $\tilde{w}^t_{hi} = \min(w_{hi}, 3N/n)$, and the unweighted (UNWT) estimator obtained by fixing $w_{hi} = N/n$ for all $h, i$.

Table 1 shows the relative bias, root mean square error (RMSE), and nominal 95% coverage for the three design-based and four model-based estimators of the population slope (second component of $\hat{\boldsymbol{B}}$) under consideration, as a function of the variance $\sigma^2$.

The fully-weighted estimator of the population slope is essentially design-unbiased under model misspecification; the unweighted and trimmed estimators are biased. The biases of the exchangeable and autoregressive models increase as variance increases, as these models trade unbiasedness of the fully-weighted estimator for the reduced variance of the unweighted estimator. The linear and nonparametric model were approximately unbiased.

The unweighted and trimmed weight estimators perform poorly with respect to MSE for small values of $\sigma^2$, where the bias due to model misspecification is critical, and well for larger values of $\sigma^2$, where the instability of the fully-weighted estimator is more important than bias reduction. The exchangeable model-based estimator has good RMSE properties for small and large values of $\sigma^2$, with MSE reductions of over 30%, but oversmooths for intermediate degrees of model specification. The autoregressive model performance equals that of the exchangeable model for small and large values of $\sigma^2$, but is largely protected against the oversmoothing of the exchangeable models at intermediate levels. The linear and nonparametric models essentially dominated the fully weighted estimators with respect to MSE under all of the simulations considered, although MSE reductions were only on the order of 10%.

The unweighted and trimmed estimators have poor coverage except when model misspecification in nearly absent. The failure of the bias-variance tradeoff for the exchangeable estimator in the presence of model misspecification is evident in the poor coverage of the estimator for intermediate values of $\sigma^2$; this effect is ameliorated, but not completely removed, for the autoregressive estimator. The linear and non-parametric estimators have good coverage when model misspecification is less important but undercover to some degree when model misspecification is more important.

**Table 1**
**Relative bias (%), square root of mean square error (RMSE) relative to RMSE of fully-weighted estimator, and true coverage of the 95% confidence interval or posterior predictive interval of population linear regression slope estimator under model misspecification. Population slope and intercept are estimated via design-based unweighted (UNWT), fully-weighted (FWT), and weight-trimmed estimators (TWT), and as the posterior mean in (8) under an exchangeable (XRS), autoregressive (ARS), linear (LRS), and non-parametric (NPRS) prior for the regression parameters. MSE relative to the fully-weighted estimator less than 1 in boldface**

| | Relative bias (%) Variance $\log_{10}$ | | | | | RMSE relative to FWT Variance $\log_{10}$ | | | | | True Coverage Variance $\log_{10}$ | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Estimator | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 | 1 | 2 | 3 | 4 | 5 |
| UNWT | 21.5 | 21.8 | 22.2 | 20.8 | 22.3 | 12.1 | 4.57 | 1.76 | **0.75** | **0.67** | 0 | 0 | 6 | 78 | 92 |
| FWT | 0.0 | 0.1 | 1.4 | 1.6 | -0.2 | 1 | 1 | 1 | 1 | 1 | 94 | 95 | 96 | 94 | 96 |
| TWT | 8.3 | 8.4 | 9.6 | 8.8 | 7.8 | 4.74 | 1.88 | 1.02 | **0.71** | **0.75** | 0 | 13 | 78 | 95 | 96 |
| XRS | 0.2 | 2.2 | 11.4 | 15.1 | 18.3 | 1.00 | 1.17 | 1.18 | **0.73** | **0.68** | 87 | 86 | 64 | 91 | 96 |
| ARS | 0.1 | 1.4 | 9.6 | 14.5 | 17.4 | 1.00 | 1.03 | 1.11 | **0.74** | **0.69** | 87 | 89 | 78 | 90 | 96 |
| LRS | -0.2 | -0.4 | 1.1 | 1.6 | -0.3 | **0.99** | **0.91** | **0.91** | **0.91** | **0.93** | 85 | 91 | 96 | 95 | 94 |
| NPRS | -0.1 | -0.3 | 0.9 | 1.5 | -0.4 | **0.89** | **0.90** | **0.95** | **0.90** | **0.95** | 86 | 92 | 96 | 94 | 94 |

### 3.2 Logistic regression

For the logistic regression model, we generated population data as follows:

$$P(Y_i = 1 \mid X_i) \sim B(\text{expit}(3.25 - 0.75X_i + \gamma X_i^2)), \quad (12)$$

$$X_i \sim U(0, 10), i = 1, ..., N = 20,000.$$

where $B(p)$ is a Bernoulli distribution with probability of "success" $p$, $\text{expit}(\cdot) = \exp(\cdot)/(1 + \exp(\cdot))$. The object of the analysis is to obtain the logistic population regression slope, defined as the value $B_1$ in the equation $\sum_i^N (y_i - \text{expit}(B_0 + B_1 x_i))\begin{pmatrix} 1 \\ x_i \end{pmatrix} = 0$. An unequal probability of selection sampling scheme was implemented as described in the linear regression simulations. We consider values of $\gamma = 0, 0.0158, 0.0273, 0.0368, 0.0454,$ corresponding to curvature measures of $K = 0, 0.02, 0.04, 0.06, 0.08$ at the midpoint 5 of the support for $X$, where $K(X; \gamma) = |2\gamma/[1 + (2\gamma X - 0.75)^2]^{3/2}|$; 200 simulations were generated for each value of $\gamma$. As in the linear regression simulations, elements were sampled without replacement with probability proportional to $(1 + h_i/2.5)h_i$; a total of 1,000 elements were sampled for each simulation. We again considered the PMLE-based the fully weighted (FTW), unweighted (UNWT), and trimmed weight estimator (TWT), along with the exchangeable random slope (XRS), autoregressive random slope (ARS), linear random slope (LRS), and nonparametric random slope (NPRS) estimators. Inference about the PMLE estimators is obtained via Taylor Series approximations (Binder 1983), as discussed in the previous section.

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 seven estimators of the population slope $(B)$ for different values of curvature $K$, corresponding to increased degrees of misspecification.

The undersampling of small values of $X$ meant that the maximum likelihood estimator of $B$ in the model misspecification setting was unbiased for $K = 0$ and biased downward for $K = 0.02, 0.04, 0.06,$ and 0.08 unless the sample design was accounted for. The trimmed estimator's bias was intermediate between the unweighted and fully weighted estimator. The exchangeable estimator's bias was between the trimmed weight estimator and fully weighted estimator; the autoregressive estimator's bias between that of the exchangeable and fully weighted estimator; while the linear and nonparametric estimators were essentially unbiased.

The unweighted estimator had substantially improved MSE (40% reduction) when the linear slope model was approximately correctly specified, but failed with moderate to large degree of misspecification. The trimmed weight, autoregressive, and nonparametric estimators all dominated the standard fully-weighted estimator, and the exchangeable and linear estimators nearly so, over the range of simulations considered. The crude trimming estimator yielded up to 30% reduction in MSE, the nonparametric, exchangeable and autoregressive estimators reductions of up to 20-25%, and the linear estimator reductions of only 10% or less.

The unweighted estimator had poor coverage except when the linear slope model was correctly specified, or nearly so. The model-based estimators had generally good coverage properties when the linear model was correctly specified, with slight reductions in coverage when curvature was substantial.

**Table 2**
**Relative bias (%), square root of mean square error (RMSE) relative to RMSE of fully-weighted estimator, and true coverage of the 95% confidence interval or posterior predictive interval of population logistic regression slope estimator under model misspecification. Population slope and intercept are estimated via design-based unweighted (UNWT), fully-weighted (FWT), and weight-trimmed estimators (TWT), and as the posterior mean in (8) under an exchangeable (XRS), autoregressive (ARS), linear (LRS), and non-parametric (NPRS) prior for the regression parameters. MSE relative to the fully-weighted estimator less than 1 in boldface**

| | Relative bias (%) | | | | | RMSE relative to FWT | | | | | True Coverage | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Curvature $K$ | | | | | Curvature $K$ | | | | | Curvature $K$ | | | | |
| Estimator | 0 | 0.02 | 0.04 | 0.06 | 0.08 | 0 | 0.02 | 0.04 | 0.06 | 0.08 | 0 | 0.02 | 0.04 | 0.06 | 0.08 |
| UNWT | 1.0 | -4.9 | -11.9 | -21.6 | -34.6 | **0.57** | **0.73** | **0.88** | 1.19 | 1.61 | 96 | 89 | 66 | 32 | 17 |
| FWT | 1.1 | 2.2 | 1.3 | -0.3 | 1.6 | 1 | 1 | 1 | 1 | 1 | 95 | 94 | 90 | 94 | 94 |
| TWT | 0.5 | -1.0 | -3.5 | -7.2 | -12.1 | **0.70** | **0.77** | **0.77** | **0.78** | **0.95** | 98 | 97 | 94 | 84 | 92 |
| XRS | 1.3 | -0.8 | -1.9 | -5.6 | -8.7 | **0.75** | **0.82** | **0.85** | **0.88** | 1.02 | 97 | 94 | 92 | 91 | 90 |
| ARS | 1.3 | -0.5 | -2.2 | -4.8 | -7.5 | **0.78** | **0.85** | **0.84** | **0.84** | **0.95** | 94 | 92 | 90 | 92 | 90 |
| LRS | 0.8 | 1.7 | 1.5 | -0.4 | 1.1 | **0.89** | **0.97** | **0.94** | **0.91** | 1.02 | 95 | 91 | 88 | 92 | 89 |
| NPRS | 0.3 | 1.5 | 1.1 | 0.9 | 0.5 | **0.87** | **0.88** | **0.87** | **0.80** | **0.90** | 95 | 92 | 88 | 94 | 96 |

## 4. 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 claims since December 1998 involving at least one child occupant $\leq$ 15 years of age riding in a model year 1990 or newer State Farm-insured vehicle (Durbin, Bhatia, Holmes, Shaw, Werner, Sorenson and Winston 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 stratified cluster sample is used to select vehicles (the unit of sampling) for the conduct of a telephone survey with the driver. Vehicles containing children who received medical treatment following the crash were over-sampled so that the majority of injured children would be selected while maintaining the representativeness of the overall population. (Medical treatment is defined as treatment by paramedics, treatment at a physician's office or emergency room, or hospitalization.) If a vehicle was sampled, all child occupants in that vehicle were included in the survey. Drivers of sampled vehicles were contacted by phone and, if medical treatment had been received by a passenger, screened via an abbreviated survey to verify the presence of at least one child occupant with an injury. All vehicles with at least one child who screened positive for injury and a 10% random sample of vehicles in which all child occupants who were reported to receive medical treatment but screened negative for injury were selected for a full interview; a 2% (later 2.5%) sample of crashes where no medical treatment was received were also selected. 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 probability-of-selection category), the sampling design is informative, with unweighted odds ratios biased toward the null (Korn and Graubard 1995). In addition, the weights for this dataset are quite variable: $1 \leq w_i \leq 50$, where 9% of the weights have normalized values greater than 3.

Winston, Kallan, Elliott, Menon and Durbin (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 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, 48.03). 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 part because of the presence of one consequential-injured child in the compact extended cab pickups had a very low probability of selection (0.025). In Winston *et al.* (2002), this child was removed before conducting the analysis.

Table 3 shows the results for the unadjusted and adjusted odds ratios of consequential injury risk using the unweighted, fully-weighted, and trimmed-weight design-based estimators, along with the model-based exchangeable, autoregressive, and linear regression slope models. (Results for the model-based estimators from 250,000 draws of a single chain after a 50,000 draw burn-in; convergence was assessed via Geweke (1992).) For the XRS and ARS models, $p(\sum) \sim \text{INVERSE-WISHART}(p, 0.1I)$, where $p = 2$ for the unadjusted model and $p = 13$ for the adjusted model. In the unadjusted results, the XRS and ARS estimators are intermediate between the unweighted and fully-weighted estimator, while the linear and nonparametric estimators tends to track the fully-weighted estimator. In the adjusted analysis, all three model-based estimators are intermediate between the unweighted and fully-weighted estimators, with the XRS estimator closest to the unweighted estimator and the LRS estimator closest to the fully-weighted estimator. Based on the results of the simulation, it appears that the ARS estimator, which suggest relative risks of injury on the order of 7 for children in compact extended cab pickups relative to other vehicles, may be a better estimator of relative risk than either the unweighted or fully weighted estimator. (As a "sanity check" of sorts, we note that an additional two years of data, not available at the time of Winston *et al.* (2002), 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.)

**Table 3**
**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), weights trimmed to a normalized value of 3 (TWT), exchangeable random slope (XRS), autoregression random slope (ARS), linear random slope (LRS), and nonparametric random slope (NPRS) estimators; unadjusted and adjusted for child age, crash severity, direction of impact, and vehicle weight. Point estimates for XRS, ARS, and LRS models from posterior median. 95% confidence interval or posterior predictive interval in subscript. Data from Partners for Child Passenger Safety**

| | UNWT | FWT | TWT | |
|---|---|---|---|---|
| Unadj. | $3.54_{(2.01, 6.23)}$ | $11.32_{(2.67, 48.02)}$ | $9.15_{(2.65, 31.57)}$ | |
| Adj. | $3.50_{(1.88, 6.53)}$ | $14.56_{(3.45, 61.40)}$ | $10.99_{(2.97, 34.64)}$ | |
| | **XRS** | **ARS** | **LRS** | **NPRS** |
| Unadj. | $6.70_{(2.51, 20.92)}$ | $6.69_{(2.64, 21.05)}$ | $11.17_{(3.21, 24.94)}$ | $10.34_{(3.27, 24.62)}$ |
| Adj. | $4.45_{(2.39, 8.67)}$ | $6.67_{(3.56, 11.94)}$ | $11.87_{(3.33, 36.93)}$ | $10.23_{(3.02, 37.93)}$ |

## 5. Discussion

The models discussed in this paper generalize the work of Lazzeroni and Little (1998) and Elliott and Little (2000), where population inference was restricted to population means under Gaussian distributional assumptions. Viewing weighting as an interaction between inclusion probability and model parameters opens up an alternative paradigm for weight trimming as a random effects model that smoothes model parameters of interest across inclusion classes. Models with exchangeable mean structures offer the largest degree of shrinkage or trimming but the most sensitivity to model misspecification; models with highly structured means are potentially less efficient but are more robust to model misspecification. This robustness property may be particularly important in light of the fact that elements of the large inclusion strata provide the largest degree of potential variance reduction in the model-based setting but are also subject to the largest degree of model bias and variance due to extrapolation.

We consider simulations under varying degrees of model misspecification and informative sampling for both linear and logistic regression models. The linear and non-parametric smoothing models nearly dominated fully-weighted estimators with respect to squared error loss in the simulations considered. The exchangeable model showed some tendency to oversmooth, favoring variance reduction over bias correction, especially in the linear regression setting. All of the weight smoothing estimators tended to have less than nominal coverage when models were highly misspecified, although in no case was the nominal coverage catastrophically low. The autoregressive smoothing model, which allows for differential degrees of local smoothing across weight strata, appeared to provide non-trivial increases in efficiency with limited risk of severe over-smoothing or undercoverage.

Applying the methods 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 was indeed conservative. Indeed, the ARS estimator, favored by MSE in simulations, suggests an adjusted excess risk of 6.7 with a 95% PPI of (3.6, 11.9), versus the 14.6 with 95% CI of (3.4, 61.4) of the fully-weighted estimator.

Although this paper utilizes a fully Bayesian approach to inference about the posterior predictive distribution of the population regression slope, empirical Bayes (EB) estimates can also be obtained via ML or REML estimation using standard linear or generalized linear mixed model methods. In the Gaussian setting, the EB estimates of $G$ and $\sigma^2$ can be "plugged into" the closed-form expressions for $E(\boldsymbol{B} \mid y, X)$ and $\mathrm{Var}(\boldsymbol{B} \mid y, X)$. The general exponential setting is more problematic. The plug-in estimates can be used to determine $E(\boldsymbol{B} \mid y, X)$ via root-finding methods; the lack of a closed form for $E(\boldsymbol{B} \mid y, X)$ makes it difficult to obtain model-based Empirical Bayes estimators for $\mathrm{Var}(\boldsymbol{B} \mid y, X)$. Also, standard Empirical Bayes estimators do not account for the uncertainty in the estimation of $G$.

We also note that, while computation of the actual trimming values of the case weights is unnecessary in this approach, it is possible to determine the revised design weights implied by the shrinkage. In the linear model setting, these can be obtained via a iterative application of a calibration weighting scheme such as generalized regression estimators or GREG (Deville and Särndal 1992). The

general exponential setting required embedding the calibration weight algorithm within the iterative reweighted least squares (IRWLS) algorithm used to fit a generalized linear model.

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

The methods discussed in this paper show the promise of adapting model-based methods to attack problems in survey data analysis. Our goal is not to develop a single hierarchical Bayesian model finely-tuned to a specific or question dataset at hand, but to develop robust yet efficient methods that can be applied in a fast-paced "automated" setting that many applied survey research analysts must sometimes work. Although computationally intensive, the methods considered are applications or extensions of the existing random-effect model "toolbox," and can either be implemented in existing statistical packages or executed with relatively simple MCMC methods. Our approach retains a design-based flavor in that we attempt to develop "automated" Bayesian model-based estimation techniques that yield robust inference in a repeated-sampling setting when the model itself is misspecified. 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 unequal 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 inclusion strata remains an important area for future work.

## Acknowledgements

The author thanks Roderick J.A. Little, along with the Editor, Associate Editor, and two anonymous reviewers, for their review and comments. 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. This research was supported by National Institute of Heart, Lung, and Blood grant R01-HL-068987-01.

## References

Alexander, C.H., Dahl, S. and Weidman, L. (1997). Making estimates from the American Community Survey. *Proceedings of the Social Statistics Section*, American Statistical Association, 2000, 88-97.

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

Beaumont, J.-F., and Alavi, A. (2004). Robust generalized regression estimation. *Survey Methodology*, 30, 195-208.

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

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., 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 Survey Research Methods Section*, American Statistical Association, 2000, 598-603.

Gelfand, A.E., and Smith, A.M.F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, 85, 389-409.

Gelman, A., and Carlin, J.B. (2002). Poststratification and weighting adjustments. *Survey Nonresponse*, (Eds., R.M. Groves, D.A. Dillman, J.L. Eltinge and R.J.A. Little), 289-302.

Gelman, A., and Rubin, D.B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, 7, 457-472.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. *Bayesian Statistics 4*, *Proceedings of the Fourth Valencia International Meeting*, (Eds., J.M. Bernardo, J.O. Berger, A.P. Dawid and A.F.M. Smith), 89-193.

Ghosh, M., and Meeden, G. (1986). Empirical Bayes estimation of means from stratified samples. *Journal of the American Statistical Association*, 81, 1058-1062.

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

Holt, D., and Smith, T.M.F. (1979). Poststratification. *Journal of the Royal Statistical Society*, Series A, 142, 33-46.

Isaki, C.T., and Fuller, W.A. (1982). Survey design under a regression superpopulation model. *Journal of the American Statistical Association*, 77, 89-96.

Kish, L. (1992). Weighting for unequal $P_i$. *Journal of Official Statistics*, 8, 183-200.

Korn, E.L., and Graubard, B.I. (1995). Examples of differing weighted and unweighted estimates from a sample survey. *The American Statistician*, 49, 291-295.

Korn, E.L., and Graubard, B.I. (1999). *Analysis of Health Surveys*. New York: John Wiley & Sons, Inc.

Korn, E.L., and Graubard, B.I. (2003). Estimating variance components using survey data. *Journal of the Royal Statistical Society*, Series B, 65, 175-190.

Lazzeroni, L.C., and Little, R.J.A. (1998). Random-effects models for smoothing post-stratification weights. *Journal of Official Statistics*, 14, 61-78.

Lin, X., and Zhang, D. (1999). Inference in generalized additive mixed models by using smoothing splines. *Journal of the Royal Statistical Society*, Series B, 61, 381-400.

Little, R.J.A. (1983). Estimating a finite population mean from unequal probability samples. *Journal of the American Statistical Association*, 78, 596-604.

Little, R.J.A. (1991). Inference with survey weights. *Journal of Official Statistics*, 7, 405-424.

Little, R.J.A. (1993). Poststratification: A modeler's perspective. *Journal of the American Statistical Association*, 88, 1001-1012.

Little, R.J.A. (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. CRC Press: Boca Raton, Florida.

Oh, H.L., and Scheuren, F.J. (1983). Weighting Adjustment for Unit Nonresponse, *Incomplete Data in Sample Surveys*, (Eds., W.G. Madow, I. Olkin and D.B. Rubin), 2, 143-184.

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

Pfeffermann, D., Skinner, C.J., Holmes, D.J., Goldstein, H. and Rabash, J. (1998). Weighting for unequal selection probabilities in multilevel models. *Journal of the Royal Statistical Society*, Series B, 60, 23-40.

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.

Rizzo, L. (1992). Conditionally consistent estimators using only probabilities of selection in complex sample surveys. *Journal of the American Statistical Association*, 87, 1166-1173.

Rubin, D.B. (1987). *Multiple Imputation for Non-Response in Surveys*. New York: John Wiley & Sons, Inc.

Särndal, C.-E. (1980). On $\pi$-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*. New York: John Wiley & Sons, Inc.

Wahba, G. (1978). Improper priors, spline smoothing, and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society*, Series B, 40, 364-372.

Wang, Y. (1998). Smoothing spline models with correlated random errors. *Journal of the American Statistical Association*, 93, 341-348.

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.