Weight Smoothing for Generalized Linear Models Using a Laplace Prior

Xi Xia
PhD Candidate, Department of Biostatistics
School of Public Health, University of Michigan
1415 Washington Heights, Ann Arbor, MI, 48109

Michael R. Elliot
Professor, Department of Biostatistics
School of Public Health, University of Michigan
1415 Washington Heights, Ann Arbor, MI, 48109
Research Professor, Institute for Social Research
426 Thompson St, Ann Arbor, MI, 48106

Abstract

When analyzing data sampled with unequal inclusion probabilities, correlations between the probability of selection and the sampled data can induce bias. Weights equal to the inverse of the probability of selection are commonly used to correct this possible bias. When weights are uncorrelated with the sampled data, or more specifically the descriptive or model estimators of interest, highly disproportional sample design resulting in large weights can introduce unnecessary variability, leading to an overall larger root mean square error (RMSE) compared to the unweighted or Winsorized methods.

We describe an approach we term weight smoothing that models the interactions between the weights and the estimators of interest as random effects, reducing the overall RMSE by shrinking interactions toward zero when such shrinkage is supported by data. This manuscript adapts a more flexible Laplace prior distribution for the hierarchical Bayesian model in order to gain more robustness against model misspecification and considers this approach in the context of a generalized linear model. Both simulation and application suggest that under linear model setting, weight smoothing models with Laplace prior yield robust results when weighting is not necessary, and could reduce the RMSE by more than 25% if strong patterns exist in the data. Under logistic regression of same sample size, the estimates are still robust, but with less gain in efficiency.

Key Words: Weight Smoothing, Laplace Prior, Generalized Linear Model, Hierarchical Bayesian Model

1 Introduction

Studies based on data sampled with unequal inclusion probability typically apply case weights equal to the inverse of probability of inclusion to reduce or remove the bias in estimators of population quantities of descriptive interest, such as means or totals (Horvitz and Thompson 1952). This "fully weighted" approach can be extended to estimate analytical quantities that focus on associa-

tion between risk factors and outcomes, such as population slopes in linear and generalized linear models, by applying sampling weights to score equations, and solving for the resulting "pseudo-maximum likelihood" estimators (PMLEs) (Binder 1983, Pfeffermann 1993). Unweighted and weighted estimators generally correspond when the underlying model (either implicit or explicit) is correctly specified and the sampling scheme is noninformative. When the model is misspecified or the sampling scheme is informative, weighted estimators typically have reduced bias, often (although not always) at the cost of increased variance. As model assumptions improve and/or sampling better approximates noninformativeness, the increase in variance from weighted analysis could overwhelm the reduction in bias, and lead to an overall larger mean square error (MSE) than would be the case if the weights were ignored or at least controlled in some fashion.

Weight trimming, or "winsorization," is used to control the variation in weights, or more precisely, cap the weights at some value w_0, and redistribute the values above w_0 among the rest (Alexander et. al. 1997; Kish 1992; Potter 1990). Various approaches have been developed in creating different criteria to determine the cap value based on data. Some example includes NAEP method by Potter (1988), which set the cutoff point equal $\sqrt{c\sum_{i\in s} w_i^2/n}$, where c was chosen in an ad-hoc manner. Cox and McGrath (1981) approached it by estimating the cutoff point value which optimizes the empirical MSE estimated by $\hat{MSE}(\hat{\theta}_t) = (\hat{\theta}_t - \hat{\theta_w})^2 - \hat{Var}(\hat{\theta}_t) + 2\sqrt{\hat{Var}(\hat{\theta}_t)\hat{Var}(\hat{\theta_w})}$, where $\hat{\theta_w}$ is the fully weighted estimator, and $\hat{\theta}_t,\ t = 1, ..., T$, is the weight trimmed estimator, with t denoting various trimming levels, from 1 as the unweighted estimator to T as the fully-weighted estimator. Chowdbury et al. (2007) suggested treating the weights as coming from a skewed cumulative distribution (e.g., an exponential distribution), and using the upper 1% of the fitted distribution as a cut point for weight trimming. Beaumont (2008) proposed a generalized design-based method, replacing the actual weights with weights predicted on some form of response and design variables. Details of these design-based approaches are summarized in Henry's (2012) review.

An alternative to standard design-based weighted estimation is a model-based approach that accommodates disproportional probability-of-selection design in a finite population Bayesian inference setting. By creating dummy variables stratified by equal or approximately equal case weights, a fully weighted data analysis can be obtained by building a model that contains indicators for the weight strata together with interaction terms between the weight stratum indicators and model parameters of interest, then obtaining inference about the population quantity of interest from its posterior predictive distribution. Elliott and Little (2000) established two model-based approaches for weight-trimming: model averaging, or "weight pooling", and hierarchical modeling, or "weight smoothing". A weight pooling model collapses strata with similar weights together with their associated interaction terms, mimicking a data-driven weight trimming process. Weight smoothing treats the underlying weight strata as random effects, and achieves a balance between fully weighted and unweighted estimates using a shrinkage estimator: thus the strata are smoothed if data provide little evidence of difference between strata, and are separated if data suggest that interactions with strata are present. Under a Bayesian framework, a two-level model is implemented, assigning a multivariate normal prior for the random effects, with inference obtained from the posterior predictive distribution of the population parameter of interest. Elliott (2008, 2009) extended the application to linear and generalized linear models, and discussed different settings for the random effect priors, namely exchangeable, autoregressive, linear and nonparametric random slopes,

and evaluated their performances.

In this manuscript we consider extending the weight smoothing approach by use of Laplace priors for the random effect weight strata and interaction terms instead of multivariate normal priors, in order to achieve more robustness against "oversmoothing" in settings where weights are required to accommodate model misspecification or non-ignorable sampling. In addition, considering the prevailing performance of Laplace prior in sparse model selection, we expect the hierarchical model to properly smooth the strata when data provides no evidence in difference among strata, even under simplistic mean and covariance matrix settings such as exchangeable random priors, while maintaining its bias-reduction feature when it is needed. We evaluate the performance of our proposed model in a simulation study, under both model misspecification and informative sampling, for both numerical and dichotomous outcomes, and compare it with competing methods. The paper is organized as follows. In Section 2 we review the theory of model smoothing together with recently proposes model-assisted methods, and develop our model with Laplace priors. Section 3 provides a simulation study, and compares bias, coverage and MSE of the proposed method with competing methods. Section 4 demonstrates the method's performance for both linear and logistic scenarios by applications on Dioxin Dataset from NHANES and Partner of Child Passenger Safety Dataset. Section 5 provides a summary discussion.

2 Weight Smoothing Methodology

2.1 Finite Bayesian Population Inference

For finite Bayesian population inference, we model the the population data Y: $Y \sim f(Y|\theta, Z)$, where Z are the variables associated with the sample design (probabilities of selection, cluster indicators, stratum variables). Note that the parametric model f can either be highly parametric with a low dimension θ (e.g., a normal model with common mean and variance), or have a more semi-parametric or non-parametric flavor with a high-dimension θ (such as a spline or Dirichlet process model). Inference about some population quantity of interest $Q(Y)$ is based on the posterior predictive distribution of

$$p(Y_{nob} \mid Y_{obs}, I, Z) = \frac{\int\int p(Y_{nob} \mid Y_{obs}, Z, \theta, \phi)p(I \mid Y, Z, \theta, \phi)p(Y_{obs} \mid Z, \theta)p(\theta, \phi)d\theta d\phi}{\int\int\int p(Y_{nob} \mid Y_{obs}, Z, \theta, \phi)p(I \mid Y, Z, \theta, \phi)p(Y_{obs} \mid Z, \theta)p(\theta, \phi)d\theta d\phi Y_{nobs}} \tag{1}$$

where Y_{nob} consists of the $N - n$ unobserved cases in the population, and ϕ models the inclusion indicator I. Assuming that ϕ and θ have independent priors, the sampling mechanism is said to be "noninformative" if the distribution of I is independent of $Y|Z$, or "ignorable" if the distribution of I only depends on $Y_{obs}|Z$. When the sampling design is ignorable, $p(I \mid Y, Z, \theta, \phi) = p(I \mid Y_{obs}, Z, \phi)$, and thus (1) reduces to

$$\frac{\int p(Y_{nob} \mid Y_{obs}, Z, \theta)p(Y_{obs} \mid Z, \theta)p(\theta)d\theta}{\int\int p(Y_{nob} \mid Y_{obs}, Z, \theta)p(Y_{obs} \mid Z, \theta)p(\theta)d\theta dY_{nobs}} = p(Y_{nob} \mid Y_{obs}, Z),$$

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

Notice that if inference about quantities $Q(Y|X)$ involving covariates X is desired (e.g., regression slope), noninformative or ignorable sample designs can be relaxed to have distribution of I depend on X.

2.2 Weight Prediction

Beaumont (2008) proposed a model-assisted method, tamping down the extreme values in weights by replacing weights with their predicted values from a prediction model of weights regressed on response and design variables. Denote $I = (I_1, ...I_N)^T$ as the vector of sample inclusion indicators, i.e. $I_i = 1$ as ith unit sampled and $I_i = 0$ otherwise, $Y = (Y_1, ...Y_N)^T$ the vector of survey response variable, and $Z = (Z_1, ...Z_N)^T$ the vector of design variables. Assuming a noninformative sampling design, thus $P(I|Z, Y) = P(I|Z)$, the predicted weights are obtained by $\tilde{w}_i = E_M(w_i|I_i = 1, z_i, y_i)$, or sometimes reduced to $\tilde{w}_i = E_M(w_i|I_i = 1, y_i)$. Beaumont (2008) discussed two estimators, the linear form $E_M(w_i|I, Y) = H_i^T\beta + v_i^{1/2}\epsilon_i$, and the exponential form, $E_M(w_i|I, Y) = 1 + exp(H_i^T\beta + v_i^{1/2}\epsilon_i)$, where H_i and $v_i > 0$ are known functions of y_i. (The exponential form prevents the predicted weights from being negative.) He presented two examples of $H_i^T\beta$, one-degree polynomial and five-degree polynomial of y_i. The predicted weights are obtained by fitting the (unweighted) model on the sampled data, then the re-weighted estimator of the survey response variable of interest is obtained using the predicted weights.

2.3 Weight Smoothing

In general, weight smoothing stratifies the data by inclusion probability, and applies a hierarchical model treating strata means as random effects, thus achieves trimming via shrinkage. Considering the population mean as the quantity of interest, an example weight smoothing model is as following:

$$\begin{aligned} Y_{hi} &\overset{iid}{\sim} N(\mu_h, \sigma^2) \\ \mu &\sim N_H(\phi, G) \end{aligned}$$

where $\mu = (\mu_1, ...\mu_H)$, $\phi = (\phi_1, ...\phi_H)$, and $h = 1, ..., H$ indexes different "weight strata" defined, e.g., by same or similar inclusion probabilities. We assume ϕ, D, and σ^2 all have weak or non-informative priors. Notice that the weight strata are not necessarily ordered by inclusion probability, but could be in a more natural ordering, for example, if the weight strata represent a disproportionately stratified sample by age. Based on this model, the posterior mean of the population mean is derived as:

$$E(\bar{Y}|y) = \sum_{h=1}^{H}[n_h\bar{y}_h + (N_h - n_h)\hat{\mu}_h]/N$$

where $\hat{\mu}_h = E(\mu_h|y)$. Various assumptions can be made for the prior distribution of μ, such as

Exchangeable random effect (XRE): $\phi_h = \phi_0$ for all h, $G = \tau^2 I_H$
Autoregressive (AR1): $\phi_h = \phi_0$ for all h, $G = \tau^2 A$, $A_{jk} = \rho^{|j-k|}$, $j, k = 1, ..., H$

Linear (LIN): $\phi_h = \phi_0 + \phi' * h$, $G = \tau^2 I_H$
Nonparametric (NPAR): $\phi_h = g(h)$, $G = 0$ where g is an unspecified, twice-differentiable function.
See Elliott and Little (2000) for a detailed review.

The weight smoothing mechanism can be easily intuited in the simplest case of the exchangeable random effect (XRE) model (Holt and Smith 1979; Ghosh and Meeden 1986, Little 1991, Lazzaroni and Little 1998),where $\phi_h = \mu$ for all h, and $G = \tau^2 I_H$. The estimation of $\hat{\mu}_h$ is now a shrinkage estimator as $\hat{\mu}_h = w_h \bar{y}_h + (1 - w_h)\tilde{y}$, 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 $\tau^2 \to \infty$, $w_h \to 1$, and $E(\bar{Y}|y) = \sum_{h=1}^{H}[n_h\bar{y}_h + (N_h - n_h)\bar{y}_h]/N = \sum_{h=1}^{H}(N_h/N)\bar{y}_h$, the fully-weighted estimator. On the other hand, as $\tau^2 \to 0$, $w_h \to 0$, and the estimation shrinks toward the unweighted mean: since $\tilde{y} = \frac{\sum_h n_h\bar{y}_h/\sigma^2}{\sum_h n_h/\sigma^2} = \bar{y}$ if $\tau^2 = 0$, $E(\bar{Y}|y) = \sum_{h=1}^{H}[n_h\bar{y}_h + (N_h - n_h)\bar{y}]/N = (n/N)\bar{y} + \bar{y}(1 - n/N) = \bar{y}$ if $\tau^2 = 0$. Since τ^2 is itself estimated from the data, and is a measure of the information available to distinguish how the population means within a weight strata differ, the weight smoothing model achieves a "data-driven" compromise between the weighted estimator, which is design consistent but may be highly inefficient, and unweighted estimator, which is fully efficient when assumption of independent between inclusion probability and mean of Y holds, but is likely biased otherwise.

2.4 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

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

where $a_i(\phi)$ is a known function of (nuisance) scale parameter ϕ, and the mean of y_i given by $\mu_i = b'(\theta_i)$ is based on a linear combination of fixed covariates x_i through some link function $g()$ such that $E(y_i|\theta_i) = \mu_i$, and $g(\mu_i) = g(b'(\theta_i)) = \eta_i = x_i^T\beta$. In the meantime, $Var(y_i|\theta_i) = a_i(\phi)V(\mu_i)$, where $V(\mu_i) = b''(\theta_i)$; thus the variance is usually a function of the mean, with the exception of normal distribution, for which $b''(\theta_i) = 1$. The link is considered canonical if $\theta_i = \eta_i$, with the simplifying results that $V(\mu_i) = 1/g'(\mu_i)$. Some examples include Gaussian (linear) regression, where $a_i(\phi) = \sigma^2$ and the canonical link $g(\mu_i) = \mu_i$; logistic regression, where $a_i(\phi) = n_i^{-1}$ and the canonical link $g(\mu_i) = log(\mu_i/(1 - \mu_i))$, and Poisson regression, where $a_i(\phi) = 1$ and the canonical link $g(\mu_i) = log(\mu_i)$.

When considering weighted estimators, we index by the inclusion stratum h, thus $g(E[y_{hi}|\beta_h]) = x_{hi}^T\beta_h$. For weight smoothing models, the hierarchical structure is considered as

$$(\beta_1^T, ...\beta_H^T)^T|\beta^*, G \sim N_{HP}(\beta^*, G)$$

where β^* is an unknown vector of mean values for the regression coefficients and G is an unknown covariance matrix. Our interest is to estimate the target population quantity $B = (B_1, ...B_p)^T$,

which is the slope that solves the population score equation $U_N(B) = 0$ where

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

Notice that the quantity B that satisfies $U(B) = 0$ is always a meaningful population quantity even if the model is misspecified, since it is a linear approximation of x_i to η_i. A first-order approximation of $E(B|y, X)$ is given based on $\hat{B}$ where

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

where $W_h = N_h/n_h$, $\hat{y}_{hi} = g^{-1}(x_{hi}^T \hat{\beta}_h)$, and $\hat{\beta}_h = E(\beta_h|y, X)$. For linear regression, where $V(\mu_i) = \sigma^2$ and $g'(\mu_i) = 1$,

$$\begin{aligned}\hat{B} =& E(B|y, X) \\ =& [\sum_h W_h \sum_{i=1}^{n_h} x_{hi} x'_{hi}]^{-1} [\sum_h W_h (\sum_{i=1}^{n_h} x_{hi} x'_{hi}) \hat{\beta}_h]\end{aligned}$$

In case of logistic regression, $V(\mu_i) = \mu_i(1 - \mu_i)$ and $g'(\mu_i) = \mu_i^{-1}(1 - \mu_i)^{-1}$, $E(B|y, X)$ is obtained by solving the weighted score equation for population regression parameter β

$$\sum_{h=1}^{H} W_h \sum_{i=1}^{n_h} x_{hi} \Big(expit(x'_{hi}\beta) - expit(x'_{hi}\hat{\beta}_h) \Big) = 0$$

where $expit(.) = exp(.)/(1 + exp(.))$. In practice, approximate posterior distributions of B can be obtained by replacing the observed y_{hi} with the predicted values $g(x'_{hi}\hat{\beta}_h)$ for each draw of $\hat{\beta}_h$ and obtaining the pseudo-MLE for the chosen regression model.

2.5 Laplace Prior for Weight Smoothing

Instead of using a multivariate normal distribution as the prior of βs, we propose using a multivariate Laplace distribution. Unlike normal distribution prior which restricts the variation between random effect term and prior mean in an $L2$ manner, Laplace measures by the $L1$ distance. According to Eltoft(2006), the general form of Multivariate Laplace distribution is given by:

$$p_Y(y) = \frac{1}{(2\pi)^{d/2}} \frac{2}{\lambda} \frac{K_{(d/2)-1}(\sqrt{\frac{2}{\lambda} q(y)})}{(\sqrt{\frac{2}{\lambda} q(y)})^{(d/2)-1}}$$

where y is a d-dimensional random variables $y = (y_1, ..., y_d)$; $K_m(x)$ denotes the modified Bessel function of the second kind and order m, evaluated at x; $q(y) = (y-\mu)^t \Gamma^{-1}(y-\mu)$; $\Gamma = \{\gamma_{jk}\}$, $j, k = 1, ..., d$ is a $d \times d$ matrix defining the internal covariance structure of the variable Y, $\mu = (\mu_1, ..., \mu_d)$ is the vector of means, and λ an overall scale parameter. However, such a format is inconvenient for application. The alternative approach is to represent Laplace distribution as a scale mixture of normals with an exponential mixing density. By creating a set of latent mixing variables $D_\tau = diag(\tau_1^2, ..., \tau_{Hp}^2)$, and applying exchangeable random slope(XRS) setting, we reach the two level hierarchical form of Laplace Prior for β:

$$\begin{aligned}
(\beta_1^T, ..., \beta_H^T)^T | \beta_h^*, D_\tau, \sigma^2 &\sim MVN(\beta_h^*, \sigma^2 D_{\tau h}) \\
\beta_h^* | \sigma_0^2 &\sim MVN(0, \sigma_0^2 I_p) \\
D_{\tau h} &= diag(\tau_{h1}^2, ..., \tau_{hp}^2) \\
\sigma^2, \tau_1^2, ...\tau_{Hp}^2 &\sim 1/\sigma^2 \prod_{j=1}^{Hp} \frac{\lambda^2}{2} e^{-\lambda^2 \tau_j^2/2} \\
\lambda^2 &\sim Gamma(r, \delta)
\end{aligned}$$

The first level of the model depends on the distribution assumption of the generalized linear model used. In this paper, we take linear regression and logistic regression as examples, and provide the full hierarchical Bayesian model and related Gibbs Sampler algorithm.

For linear regression, Y conditional on all other parameters follows a normal distribution. Assuming that the residual variance σ^2 is independent from the latent mixing variables τ_i, the hierarchical model is as follows:

$$\begin{aligned}
y_{hi} | x_{hi}, \beta_h, \sigma^2 &\sim N(x_{hi}^T \beta_h, \sigma^2) \\
(\beta_1^T, ..., \beta_H^T)^T | \beta_h^*, D_\tau, \sigma^2 &\sim MVN(\beta_h^*, \sigma^2 D_{\tau h}) \\
\beta_h^* | \sigma_0^2 &\sim MVN(0, \sigma_0^2 I_p) \\
D_{\tau h} &= diag(\tau_{h1}^2, ..., \tau_{hp}^2) \\
\sigma^2, \tau_1^2, ...\tau_{Hp}^2 &\sim 1/\sigma^2 \prod_{j=1}^{Hp} \frac{\lambda^2}{2} e^{-\lambda^2 \tau_j^2/2} \\
\lambda^2 &\sim Gamma(\gamma = 1, \delta = 1.78)
\end{aligned}$$

Following the deduction in Park & Casella(2008), the analytical forms of all fully conditional distributions of β, σ^2 etc are achievable, and the posterior predictive distribution could be obtained

through a Gibbs Sampler as below. A detailed derivation is attached in the Appendix 1.

$$\beta_h|rest \sim MVN(A^{-1}(X_h^T Y_h + D_{\tau h}^{-1}\beta_h^*), \sigma^2 A^{-1}), A = X_h^T X_h + D_{\tau h}^{-1}$$
$$\beta_h^*|rest \sim MVN((\sigma^2 D_{\tau h})^{-1}((\sigma^2 D_{\tau h})^{-1} + (\sigma_0^2 I)^{-1})^{-1}\beta_h, ((\sigma^2 D_{\tau h})^{-1} + (\sigma_0^2 I)^{-1})^{-1})$$
$$\sigma^2|rest \sim InvGamma((n+Hp)/2, \frac{1}{2}[\sum_{h=1}^{H}(Y_h - X_h\beta_h)^T(Y_h - X_h\beta_h) +$$
$$\sum_{h=1}^{H}(\beta_h - \beta_h^*)^T(D_{\tau h})^{-1}(\beta_h - \beta_h^*)])$$
$$1/\tau_{hi}^2|rest \sim InvGaussian(\sqrt{\frac{\lambda^2\sigma^2}{(\beta_h - \beta_h^*)^2}}, \lambda^2)$$
$$\lambda^2 \sim Gamma(Hp + \gamma, \frac{1}{2}\sum_{h=1}^{H}\sum_{i=1}^{p}\tau_{hi}^2 + \delta)$$

For logistic regression, the model is similar to that for linear regression, except that Y follows a binomial distribution, and estimation of σ^2 is no longer necessary:

$$y_{hi}|x_{hi}, \beta_h, \sim \prod_{h=1}^{H}\prod_{i=1}^{n_h}\left(\frac{exp(x_{hi}\beta_h)}{1+exp(x_{hi}\beta_h)}\right)^{y_{hi}}\left(\frac{1}{1+exp(x_{hi}\beta_h)}\right)^{1-y_{hi}}$$
$$(\beta_1^T, ..., \beta_H^T)^T|\beta_h^*, D_\tau, \sim MVN(\beta_h^*, D_{\tau h})$$
$$\beta_h^*|\sigma_0^2 \sim MVN(0, \sigma_0^2 I_p)$$
$$D_{\tau h} = diag(\tau_{h1}^2, ..., \tau_{hp}^2)$$
$$\tau_1^2, ...\tau_{Hp}^2 \sim \prod_{j=1}^{Hp}\frac{\lambda^2}{2}e^{-\lambda^2\tau_j^2/2}$$
$$\lambda^2 \sim Gamma(r=1, \delta=1.78)$$

When the first level is not normally distributed, the fully conditional distribution of β does not belong to any known distribution, and thus direct sampling is impossible. Instead we apply Metropolis method, and the proposed β_h is drawn from $N_p(\beta_h', c_\beta D_\beta)$, for $D_\beta = (V_{\beta h}^{-1} + D_{\tau h}^{-1})^{-1}$, where β_h' is the ML estimate of the logistic regression of y on Z from strata h, and $V_\beta h$ the associated covariance matrix obtained from the expected information matrix evaluated at β_h'. The proposed β_h is accepted with probability $r = max[1, \{f_\beta(\beta_{prop})\}/\{f_\beta(\beta)\}]$,where f_β is the posterior distribution of β proportional to $p(\beta_h)\prod_{i=1}^{n_h} f(y_{hi}|\beta_h)$. All other parameters follow the Gibbs Sampler algorithm, and are directly drawn from their fully conditional distributions as below: (full deriva-

tion in Appendix 2)

$$\beta_h^*|rest \sim MVN((D_{\tau h})^{-1}((D_{\tau h})^{-1} + (\sigma_0^2 I)^{-1})^{-1}\beta_h, ((D_{\tau h})^{-1} + (\sigma_0^2 I)^{-1})^{-1})$$

$$1/\tau_{hi}^2|rest \sim InvGaussian(\sqrt{\frac{\lambda^2}{(\beta_h - \beta_h^*)^2}}, \lambda^2)$$

$$\lambda^2 \sim Gamma(Hp + \gamma, \frac{1}{2}\sum_{h=1}^{H}\sum_{i=1}^{p}\tau_{hi}^2 + \delta)$$

3 Simulation Study

To evaluate the performance of weight smoothing models using Laplace priors, we created two scenarios for ordinary linear regression and logistic regression, generating separate populations with normally distributed outcome and dichotomise outcome accordingly. The target of interest is the population slope. In addition to our Laplace prior estimator, we consider an unweighted estimator, a fully-weighted estimator, a normal-prior (exchangable) estimator (Elliott and Little 2000; Elliott 2007), and several variations of the model-assisted estimator proposed by Beaumont (2008). For each scenario and estimator, we compute bias, square root of mean square error (RMSE) and coverage of 95% confidence or credible intervals.

3.1 Hierarchical weight smoothing model for ordinary linear regression

We generate a population of $N = 20,000$ for ordinary linear regression. The predictor X is uniformly distributed on the interval from 0 to 10, and is equally divided into 20 strata with a range of 0.5 each. The response variable Y is then generated as a spline function of X with cutpoints between strata as knots. Three sets of coefficients are applied separately, so the pattern of $Y \mid X$ varies from straight slope to increasing curve and decreasing curve.

$$Y_i|X_i, \beta, \sigma^2 \sim N(\beta_0 + \sum_{h=1}^{20}\beta_h(x_i - h)_+, \sigma^2)$$

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

$$\beta_a = c(0, 2, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$$

$$\beta_b = c(0, 0, 0, 0, 0, 0, 0, 0, .5, .5, .5, .5, 1, 1, 1, 1, 2, 2, 2, 2, 4, 4)$$

$$\beta_c = c(0, 11, -4, -4, -2, -2, -2, -2, -1, -1, -1, -1, -0.5, -0.5, -0.5, -0.5, 0, 0, 0, 0, 0)$$

From the population, a sample of $n = 1000$ is selected without replacement, according to inclusion probabilities equal to $\pi_i = (1 + i/30) * i/2$ for the ith stratum. Thus the ratio between the maximum and minimum of weights is about 35, and the sample size of each stratum is always greater than 3. Z is created as $Z = I \otimes X$,where $I = c(I_1, ..., I_h)$ is an indicator vector stating if the current observation belongs to ith stratum. Z is centered within each column with respect to each stratum(for computation convenience), and used as predictor in the simulations.

	$\sigma^2 = 10$			$\sigma^2 = 10^3$			$\sigma^2 = 10^5$		
	Bias	RMSE	cover	Bias	RMSE	cover	Bias	RMSE	cover
UNWT	-0.013	0.730	0.95	0.009	0.693	0.95	-0.766	0.713	0.96
FWT	-0.006	1	0.94	-0.012	1	0.93	-1.253	1	0.96
HWS	-0.006	1.047	0.96	-0.019	0.978	0.95	-1.107	0.768	0.99
XRS	-0.001	1.489	1	0.016	0.720	0.96	-0.765	0.716	0.96
PREDY	-0.014	0.805	0.96	0.025	0.698	0.95	-0.781	0.711	0.97
PREDY5	0.108	1.511	0.52	0.104	0.756	0.95	-0.785	0.717	0.97
PREDYX	-0.005	0.803	0.97	-0.016	0.783	0.94	-1.145	0.796	0.95
PREDYX5	-0.005	0.978	0.94	-0.019	0.953	0.94	-1.540	0.983	0.95

Table 1: Table 1: Comparison of various estimators of slope B_1 under β_a linear spline setting. Bias and RMSE under populations with residual variance 10, 10^3 and 10^5 from following model: unweighted, fully weighted, hierarchical weight smoothing, exchangeable random effect and weight prediction by y, degree 5 polynomial of y, linear combination of x and y, and degree 5 polynomial of x,y .

Our inferential target is $B = (\sum_{i=1}^{N} \tilde{X}_i \tilde{X}_i')^{-1} \sum_{i=1}^{N} \tilde{X}_i Y_i$ for $\tilde{X}_i = (1\ X_i)'$, the least-squares linear approximation of Y to X. Under β_b and β_c, weights correct bias from model misspecification. Under β_a, the model is correctly specified, suggesting that the unweighted estimator may be most efficient. Population variance σ^2 varies among 10, 10^3 and 10^5, creating varying level of variance influence compared to possible bias; note that under β_b, the curvature is largest where the data is most densely sampled, while the reverse is true under β_c, suggesting that varying degrees of trimming will be required to optimize the bias-variance tradeoff.

For the hyperprior parameters, σ_0^2 is arbitrarily defined as 1000 to approximate a non-informative prior; the prior for λ follows a gamma hyperprior with parameter $r = 1$ and $\delta = 1.78$, as suggested by Park and Casella (2008). All other parameters in simulation are initialized at zero, expect for variance estimator σ^2, which is initialized at one. A Gibbs Sampler method is applied, that is, for each iteration, all parameters are sequentially drawn from the full conditional distribution. Then to obtain the estimate from posterior predictive distribution, the unobserved Y are generated based on sampled parameters from each iteration, and the target population slope B is obtained by fully weighted regression on observed and predicted Y. The process iterates 10000 times, with a burn-in of 2000. Diagnostic plots are generated to assure the algorithm's convergence. Bias, RMSE and 95% coverage are recorded for comparison. Overall 200 samples are generated from each population to provide the empirical distribution for the repeated measures properties.

We compare the properties of our Laplace model (HWT) with major competitors, including the unweighted model (UNWT), fully weighted model (FWT), weight smoothing model with normal prior and exchangeable random slope assumption (XRS), and four variations of the model-assisted estimators of Beaumont (2008): predicted weights on y only (PREDY); predicted weights on degree 5 polynomial of y (PREDY5); predicted weights on y and x (PREDYX) and predicted weights on degree 5 polynomial of y, together with x (PREDYX5). Bias and nominal 95% coverage are recorded directly, while RMSE is rescaled according to fully weighted estimator. Results are provided in Table 1,2, and 3.

	$\sigma^2 = 10$			$\sigma^2 = 10^3$			$\sigma^2 = 10^5$		
	Bias	RMSE	cover	Bias	RMSE	cover	Bias	RMSE	cover
UNWT	1.980	10.201	0	1.993	2.441	0.02	1.204	0.726	0.95
FWT	-0.006	1	1	-0.005	1	0.92	-1.252	1	0.96
HWS	-0.006	0.453	0.97	-0.042	0.947	0.96	-1.354	0.774	0.99
XRS	-0.103	1.008	0.95	1.769	2.213	0.04	1.203	0.729	0.94
PREDY	0.963	4.977	0	1.794	2.228	0.03	1.174	0.722	0.94
PREDY5	1.059	5.466	0	1.791	2.212	0.03	1.184	0.730	0.95
PREDYX	0.368	2.050	0.32	0.385	0.919	0.90	-0.746	0.792	0.95
PREDYX5	0.018	1.000	1	0.012	0.955	0.96	-1.515	0.983	0.96

Table 2: Table 2: Comparison of various estimators of slope B_1 under β_b linear spline setting. Bias and RMSE under populations with residual variance 10, 10^3 and 10^5 from following model: unweighted, fully weighted, hierarchical weight smoothing, exchangeable random effect and weight prediction by y, degree 5 polynomial of y, linear combination of x and y, and degree 5 polynomial of x,y .

	$\sigma^2 = 10$			$\sigma^2 = 10^3$			$\sigma^2 = 10^5$		
	Bias	RMSE	cover	Bias	RMSE	cover	Bias	RMSE	cover
UNWT	-1.874	6.227	0	-1.836	2.177	0.01	-2.611	0.758	0.9
FWT	-0.006	1	1	-0.020	1	0.97	-1.257	1	0.95
HWS	-0.005	0.337	0.85	0.069	0.937	0.96	-0.772	0.772	0.99
XRS	-0.549	1.872	0.06	-1.721	2.052	0.01	-2.609	0.761	0.91
PREDY	-0.009	1.258	0.97	-1.738	2.075	0.01	-2.627	0.756	0.90
PREDY5	-0.167	1.131	0.99	-1.256	1.757	0.37	-2.593	0.761	0.88
PREDYX	-0.327	1.416	0.75	-0.729	1.091	0.75	-1.861	0.809	0.93
PREDYX5	-0.020	1.019	1	-0.055	0.965	0.96	-1.557	0.983	0.95

Table 3: Table 3: Comparison of various estimators of slope B_1 under β_c linear spline setting. Bias and RMSE under populations with residual variance 10, 10^3 and 10^5 from following model: unweighted, fully weighted, hierarchical weight smoothing, exchangeable random effect and weight prediction by y, degree 5 polynomial of y, linear combination of x and y, and degree 5 polynomial of x,y .

Under β_a, where the model is correctly specified, all methods yield unbiased results, and the unweighted estimator maintains the best efficiency, with an approximate 30% decrease in RMSE comparing to fully weighted estimator. The original weight smoothing method under XRS tends to provide unstable results, inflating the variance when population signal is strong, but achieving similar RMSE as the unweighted estimator when the population signal is weak relative to the noise. Our model, under the same XRS assumption but with a Laplace prior, gives more stable results that resemble the fully weighted estimator when variance is low, but increases in efficiency as population variance increases. Both the XRS and HWT estimators have correct to somewhat conservative coverage when the linear model is correctly specified. Most model-assisted estimators have improved RMSE comparing to the fully weighted estimator, with the exception of PREDY5, which has unstable results and poor nominal coverage when $\sigma^2 = 10$.

For scenarios under β_b and β_c, the unweighted estimator of B is biased, and the fully weighted estimator strongly prevails over unweighted estimator with respect to both RMSE and coverage for small to moderate levels of residual variances. The weight smoothing method under XRS remains biased at moderate levels of variance for β_b and β_c, and also at small levels of variance for β_c, raising RMSE relative to FWT and destroying nominal coverage, suggesting that the exchangable random slope structure is not sophisticated enough to capture the relation in mean and variance among strata. The weight smoothing estimator with Laplace prior has limited bias similar to that of the fully weighted estimator, but very substantially reduced RMSE, though it suffers a moderate drop in coverage under β_c and $\sigma^2 = 10$. Most of the model-assisted estimators are insufficiently structured to reduce bias in the small-to-medium residual variance settings, except for PREDYX5, which mimics the fully weighted estimator and thus has little savings in relative RMSE under any of the scenarios.

3.2 Hierarchical weight smoothing model for logistic regression

Following Elliott (2007), we set up population in two approaches: model misspecification and informative sampling. For model misspecification, the population is equally divided into 20 strata, and the predictor X is uniformly distributed within each stratum on an interval ranging from $0.5(h-1)$ to $0.5h$. The binary response variable is generated as following:

$$P(Y_i = 1|X_i) \sim BER(expit(1.5 - .75X_i + C * X_i^2)),$$
$$X_{hi} \sim UNI(0.5 * (h-1), 0.5 * h), h = 1, .., 20, i = 1, ..., 1000$$

Our inferential target is $B = (B_0\ B_1)'$, the value of $\beta = (\beta_0\ \beta_1)'$ that solves the score equation $U(\beta) = \sum_{i=1}^{N} \tilde{X}_i(Y_i - expit(\tilde{X}_i'\beta))$, corresponding to the best linear approximation to X_i and $log\ \frac{E(Y_i|X_i)}{1-E(Y_i|X_i)}$. For C, we consider values of 0, .027, .045, .061, .080, corresponding to increasing levels of model misspecification. The selection probability for each observation remains the same within each stratum, and increases linearly along strata, with a ratio between maximum and minimum probabilities equals to 20.

	$c = 0$			$c = .45$			$c = .80$		
	Bias	RMSE	cover	Bias	RMSE	cover	Bias	RMSE	cover
UNWT	0.014	0.697	0.96	0.063	1.151	0.47	0.132	2.819	0
FWT	0.006	1	0.84	-0.015	1	0.82	-0.014	1	0.94
HWS	0.011	0.915	0.84	-0.014	0.909	0.82	-0.013	0.860	0.88
XRS	0.038	1.125	1	0.078	1.696	0.94	0.042	1.644	0.98
PREDY	0.014	0.765	0.95	0.082	1.614	0.27	0.136	3.115	0
PREDYX	0.003	0.791	0.96	0.021	0.903	0.92	0.038	1.123	0.79
PREDYX5	-0.004	0.962	0.93	-0.005	0.965	0.94	-0.001	0.967	0.98

Table 4: Table 4: Comparison under model misspecification. Bias and RMSE under populations with underlying model quadratic coefficient 0, $.45$ and $.80$ from following model: unweighted, fully weighted, hierarchical weight smoothing, exchangeable random effect and weight prediction by y, degree 5 polynomial of y, linear combination of x and y, and degree 5 polynomial of x,y.

For the informative sampling setting, we follow the same formula of

$$P(Y_i = 1|X_i) \sim BER(expit(1.5 - .75X_i + C * X_i^2)),$$
$$X_{hi} \sim UNI(0.5 * (h-1), 0.5 * h), h = 1, .., 20, i = 1, ..., 1000$$

but fix $C = 0$, so the model is correctly specified. We also create a vector of binary value Z_i^* such that $Cor(Y_i, Z_i^*) = r$, and r range from 0.05 to 0.95 to represent different level of correlation with Y. Then we let $Z_i = Z_i^* U_i + (1 - Z_i^*) X_i$, where $U_i \sim U(0, 10)$ independent of X_i, and the selection probability is proportional to Z_i. Thus whether the selection probability is related to X or not is determined by the value of Z^*, which is correlated with Y to some level. The process results in a ratio of roughly 30 between maximum weight and minimum weight, and the correlation between selection probability and Y varies from 0 to 30 % as the correlation between Z^* and Y increases from .05 to .95. 20 strata of equal size are created by pooling observations with similar selection probabilities together.

From this population, samples with $n = 1000$ are selected without replacement, with the selection probability stated above. We create weight strata using the values of h. A total of 200 samples are generated to create the empirical distribution for inference. A single MCMC chain is built for each data set, and for each iteration in the algorithm, all parameters are sequentially drawn from the full conditional distribution, except for β, which is proposed from a normal distribution centered at MLE with inverse expected information as covariance matrix, and accepted according to likelihood ratio times prior distribution. Then the predicted Y is calculated based on drawn parameters, and the target population slope is obtained by fully weighted logistic regression. The initial values of parameters are assigned the same as linear regression setting, and the process iterates 10000 times, with a burn-in of 2000.

We compare the properties of our Laplace model (HWT) with same major competitors as in the linear regression setting, with the exception of (PREDY5): since Y is a binary variable, higher-order polynomials are not relevant. Bias and nominal 95% coverage are recorded directly, while RMSE is rescaled according to fully weighted estimator. Results are provided in Table 4 and 5.

While comparing different models under model misspecification setting, the unweighted model

	$r = .05$			$r = .50$			$r = .95$		
	Bias	RMSE	cover	Bias	RMSE	cover	Bias	RMSE	cover
UNWT	0.057	0.990	0.76	0.069	1.155	0.52	0.053	0.914	0.64
FWT	0.023	1	0.94	0.009	1	0.88	0.001	1	1
HWS	0.022	0.906	0.82	0.009	0.914	0.84	0.001	0.875	0.96
XRS	0.067	1.417	1	0.071	1.463	0.98	0.059	1.272	1
PREDY	0.055	1.034	0.70	0.071	1.233	0.54	0.062	1.079	0.64
PREDYX	0.021	0.832	0.94	0.023	0.859	0.91	0.031	0.880	0.93
PREDYX5	0.004	0.977	0.94	0.002	0.969	0.95	0.007	0.997	0.97

Table 5: Table 5: Comparison under informative sampling. Bias and RMSE under populations with correlation between Z and Y equal to $.05$, $.50$ and $.95$ from following model: unweighted, fully weighted, hierarchical weight smoothing, exchangeable random effect and weight prediction by y, degree 5 polynomial of y, linear combination of x and y, and degree 5 polynomial of x,y.

has increased bias as the population model is less correctly specified, resulting in a change from efficient estimate to a poor estimate(RMSE ratio from 69.7% to 281.9% of FWT's as C increases) and poor coverage as misspecification increases. The exchangeable random slope model estimator is not robust, with bias similar to unweighted model, and larger RMSE than the fully-weighted estimator, although coverage is conservative. The hierarchical weight smoothing model with Laplace prior provides a more robust estimator, with minimal bias, and RMSE reduced by up to 14% compared to the FWT estimator, although coverage suffers to a moderate degree. The weight prediction models PREDY and PREDYX perform similar to unweighted estimate, gaining efficiency when model's correctly specified, and suffering as misspecification increases. PREDYX5, which predicts weights with a degree five polynomial of both x and y, essentially mimics the fully-weighted estimator.

Under informative sampling, the unweighted estimator has only slightly larger RMSE than the fully weighted estimator, but is substantially biased with poor coverage. The exchangeable random effect model has a similar degree of bias compared to the unweighted estimator, but has increased variability that, while providing conservative coverage, yields substantially increased RMSE over the fully-weighted estimator. The hierarchical weight smoothing model with Laplace prior again provides a more robust estimator, with minimal bias, and RMSE reduced by up to 12% compared to the FWT estimator, although coverage suffers to a moderate degree except when the sampling is highly informative. PREDY is modestly biased but has poor coverage (perhaps not surprising given that Y is binary), while PREDYX improves RMSE by upp to 17% while having only slight undercoverage. PREDYX5 again mimics the fully-weighted model.

4 Application

4.1 Application on Dioxin data from NHANES

To demonstrate the performance of our method in linear regression setting, we consider its application on the dioxin dataset from the National Health and Nutrition Examination Survey(NHANES). During the 2003-2004 survey, 1250 representative adult subjects were selected under a probability sample of the US, and had their blood biomarkers measured, including 2,3,7,8-tetrachlorodibenzo-p-dioxin(TCDD), a compound usually formed through incomplete combustion such as incineration, paper and plastics manufacturing, and smoking. Other demographic variables like age and gender are also available from the survey. The sampled data is stratified into 25 strata, within each consist of 2 Masked Variance Units(MVU's) for proper variance estimation procedure, with survey weights provided as well. Due to technical limit, 674 readings are below limit-of detection, and are imputed through multiple imputation using the model described in Chen et al. (2010), resulting in 5 replicate data sets. Both survey structure and imputation are incorporated in analysis using a jackknife method and Rubin's formula(Rubin 1987).

To determine the connection between log of TCDD level and individual demographic information, four linear regression models are fitted as log TCDD on age, log TCDD on gender, log TCDD on age and gender, and log TCDD on age, gender and interaction. The hierarchical model is built as described before, with same initial value of parameters as those in the simulation. For each model setting, the unweighted (UNWT), fully-weighted (FWT), and the hierarchical weight smoothing (HWS) estimators are obtained(exchangeable random slope model fails to converge and is removed from the result). To estimate mean square error, the fully weighted version is treated as unbiased. Note that the fully weighted estimator is unbiased only in expectation, leading to the true estimated square bias of regression coefficient $\hat{\beta}$ given by $max((\hat{\beta} - \hat{\beta}_w)^2) - \hat{V}_{01}$, where $\hat{V}_{01} = \hat{Var}(\hat{\beta}) + \hat{Var}(\hat{\beta}_w) - 2\hat{Cov}(\hat{\beta}, \hat{\beta}_w)$. To fully account for the design feature, all variance/covariance estimates are calculated via jackknife as $\hat{Var}(\hat{\beta}_w) = \sum_h \frac{k_h - 1}{k_h} \sum_{i=1}^{k_h} (\hat{\beta}_{w_{(hi)}} - \hat{\beta}_w)^2$, $\hat{\beta}_{w_{(hi)}} = (X' W_{(hi)} X)^{-1} X W_{(hi)} y$, where $\hat{\beta}_{w_{(hi)}}$ denotes the weighted β estimater from sample excluding i_{th} MVU in h_{th} stratum, and $W_{(hi)}$ is a diagonal matrix consisting of case weight w_j for all elements $j \notin h, j \notin i$, $\frac{k_h}{k_h - 1} w_j$ for all elements $j \in h, j \notin i$, and 0 for elements $j \in h, j \in i$. $\hat{Var}(\hat{\beta})$ and $\hat{Cov}(\hat{\beta}_w, \hat{\beta})$ are calculated accordingly, and estimates from five imputed replicate datasets are combined with Rubin's formula. The result was based on 10000 iterations after discarding 2000 draws as burn-in. And the resulting Biasness and RMSE are summarized in Tables 6-9.

For the first two models of log TCDD on age and gender separately, the estimation of the single predictor from unweighted model appears to be biased comparing to fully weighted model, resulting in estimated bias about 40% and 70% of RMSE. However, the weighted model also fails to provide a efficient estimate for effect on age, supported by a RMSE of 3.888, larger than 3.265 from the unweighted model. Meanwhile, the hierarchical weight smoothing model shows its ability to improve efficiency, both reducing the biasness comparing to unweighted model, and maintaining a RMSE similar to or smaller than fully weighted model depend on the severity of

Model	Bias(10^{-3})	RMSE(10^{-3})
UNWT	-1.262	3.265
WT	0	3.888
HWT	-0.086	1.214

Table 6: Table 6: Regression of log TCDD on Age. Bias and RMSE for linear slope estimated for age: unweighted, fully weighted and hierarchical weight smoothing.

Model	Bias(10^{-2})	RMSE(10^{-1})
UNWT	-8.219	1.248
WT	0	0.637
HWT	0.589	0.607

Table 7: Table 7: Regression of log TCDD on Gender. Bias and RMSE for linear slope estimated for gender: unweighted, fully weighted and hierarchical weight smoothing.

	Age		Gender	
Model	Bias(10^{-4})	RMSE(10^{-3})	Bias(10^{-2})	RMSE(10^{-2})
UNWT	-9.067	3.296	-0.159	9.017
WT	0	3.895	0	6.161
HWS	-0.841	1.227	1.058	5.659

Table 8: Table 8: Regression of log TCDD on age and gender. Bias and RMSE for linear slope estimated for age and gender: unweighted, fully weighted and hierarchical weight smoothing.

	Age		Gender		Interaction	
Model	Bias(10^{-4})	RMSE(10^{-3})	Bias(10^{-2})	RMSE(10^{-1})	Bias(10^{-3})	RMSE(10^{-3})
UNWT	-5.063	3.758	2.882	1.591	-0.880	3.285
WT	0	2.661	0	3.259	0	7.335
HWS	-9.142	2.048	-6.530	1.282	1.646	2.667

Table 9: Table 9: Regression of log TCDD on age and gender, and interaction between age and gender. Bias and RMSE for linear slope estimated for age, gender and interaction: unweighted, fully weighted and hierarchical weight smoothing.

variance inflation.

As more predictors enter the model, the estimated bias rapidly decreases in scale, leading to a scenario that both bias and inflation in variance could dominate the overall RMSE, and neither unweighted model nor fully weighted model prevails in estimating all predictors. Hence the hierarchical weight smoothing model cannot reduce bias further, yet it succeeds in reducing variance, resulting in overall smaller RMSE comparing to either unweighted estimator or fully weighted estimator.

4.2 Application on Partner for Child Passenger Safety data

In this section, we use Partners for Child Passenger Safety dataset to demonstrate our method's performance under logistic regression setting. Unit observations in the dataset are damaged vehicles disproportionally sampled from State Farm claims records between December 1998 and December 2005, when at least one child occupant less than 15 years of age gets involved in a model year 1990 or newer State Farm-insured vehicle. The focus of the study is children's consequential injuries, defined by either facial lacerations or other injuries rated 2 or more on the Abbreviated Injury Scale (AIS) (Association for the Advancement of Automotive Medicine 1990). Due to the rare occurrence of the injury among all claims, to improve accuracy of the corresponding estimation on this rare outcome, the overall population is divided into three strata based on injury status – vehicles with at least one child occupant screened positive for injury, vehicles with all child occupants reported receiving medical treatment but screened negative for injury, and vehicles with no occupants receiving medical treatment – and crossed with two strata defined by whether the vehicle was driveable or not. Since the stratification was associated with risk of injury, and cannot be fully explained by other auxiliary variable, the sampling design is informative, with weights varies from 1 to 50, and 9% of weights lying outside 3 times their standard deviation.

As determined by Winston, Kallan, Elliott, Menon and Durbin(2002), children rear-seated in compacted extended cab pickups are at greater risk of consequential injuries than children rear-seated in other vehicles. To strengthen the conclusion, two models are applied, the unadjusted logistic model of injury status on car type(compacted extended cab pickups or others), and adjusted logistic model adapting control variables including child age (years), use of restraint (Y/N), intrusion into the passenger cabin in accident (Y/N), tow-away after accident (Y/N), direction of impact (front/side/rear/other), and weight of the vehicle (pounds). The logistic hierarchical weight smoothing model is set up as stated in previous section, then the Gibbs sampler is executed for 10,000 iterations with 2,000 burn-in, and odds ratios are compared with unweighted and fully weighted model.

The estimated odds ratios for compacted extended cab pickups indicator didn't vary much from unadjusted model to fully adjusted model, while unweighted regression and fully weighted regression lead to quite different result, from a OR of 3.534 to 11.317 for unadjusted model, and from 3.448 to 13.890 when all other control variables are included. Both hierarchical weight smoothing model and exchangeable random effect model provide estimates lies in between the unweighted and weighted estimates, although estimation from HWS model tends to match estimation from

	OR	
	Unadjusted	Adjusted
UNWT	$3.534_{(2.003,6.234)}$	$3.448_{(1.850,6.430)}$
FWT	$11.317_{(2.737,46.784)}$	$13.890_{(3.176,60.760)}$
HWS	$10.559_{(3.731,29.876)}$	$13.268_{(7.919,22.232)}$
XRS	$8.681_{(2.790,27.007)}$	$6.725_{(2.162,20.922)}$

Table 10: Table 10: Odds ratio and relevant 95% confidence interval for estimated effect on injury from compacted extended cab pickups: unweighted, fully weighted, hierarchical weight smoothing and exchangeable random effect.

fully weighted model. It is also worth noting that with similar point estimates, HWS model provides a considerable reduction in estimated standard deviation, leads to a smaller 95% confidence interval comparing to fully weighted model, an characteristic also presented in simulation study before.

5 Discussion

Generally, most methods for weight trimming, both design based and model based, handle sampling weights by achieving a balance between bias and variance, resulting in an estimate usually lying between those from the unweighted model and fully weighted model. However, the weight smoothing model with Laplace prior shows the potential to provide a more efficient estimate than either unweighted model and fully weighted model at same time. This occurs especially when the model is misspecified, and population variance is small so the weight smoothing model is able to model the underlying data structure precisely, and yielding an estimate greatly reduced in RMSE. However, this aggressive estimation comes at the cost of robustness, that is, the overly reduced variance could lead to poor coverage rate. As presented in the simulation, the HWS model suffers a moderate drop in the coverage rate when population variance is small. It is worth exploring in future the model's mechanism in reducing the overall RMSE, and the limit of the scenarios under which it still maintains reasonable coverage.

Comparing the results of the Laplace prior weight smoothing models with the model-assisted estimators of Beaumont (2008), we find that the Laplace estimators offer the promise of relatively simple estimators that can approximately fully-weighted estimators when weights are required for bias correction, but improve over weighted estimators in terms of variability while maintaining approximately correct nominal coverage of credible intervals. In contrast, the model-assisted estimators can in some settings "oversmooth" weights when bias correction is needed and yield unstable estimators when the weight prediction is weak. The predicted weights in the model-assisted approach incorporate information from design variables, thus yielding better predictions for weighted mean and population total estimates than unweighted estimators. However, in some settings even a degree five polynomial may fail to correctly approximate the relationship between the inverse of the probability of selection and the sample statistic of interest. Perhaps even more importantly, highly structured models for weight prediction such as high degree polynomials may

results in unstable estimates of weights, adding unnecessary variance rather than dampening it. Ultimately we find attempts to model weights rather than data misguided, as it focuses on design factors on which we should be conditioning, rather than assessing uncertainties in the data that may be fertile ground for mean square error reduction while preserving approximate nominal coverage: i.e., calibrated Bayes estimators (Little 2011).

References

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

Beaumont, J.P. (2008). A new approach to weighting and inference in sample surveys. *Biometrika*, **95**, 539-553.

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

Chen, Q., Garabrant, D.H., Hedgeman, E., Little R.J.A., Elliott, M.R., Gillespie, B., Hong, B., Lee, S-Y, Lepkowski, J.M., Franzblau, A., Adriaens, P., Demond, A.H., Patterson, D.G. (2010). Estimation of Background Serum 2,3,7,8-TCDD Concentrations Using Quantile Regression in the UMDES and NHANES Populations. *Epidemiology*, **21**, S51-S57.

Chen, Q., Elliott, M.R., Little, R.J.A. (2010). Bayesian Penalized Spline Model-Based Inference for Finite Population Proportions in Unequal Probability Sampling. *Survey Methodology*, **36**, 22-34.

Chen, Q., Elliott, M.R., Little, R.J.A. (2010). Bayesian Inference for Finite Population Quantiles from Unequal Probability Samples. *Survey Methodology*, **38**, 203-215.

Chowdhury, S., Khare, M., and Wolter, K. (2007). Weight Trimming in the National Immunization Survey. *Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods, American Statistical Association* pp. 2651-8.

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.

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. (2009). Model averaging methods for weight trimming in generalized linear regression models. *Journal of Official Statistics*, **25**, 1-20.

Elliott, M.R. and Little, R.J.A. (2000). Model-based approaches to weight trimming. *Journal of Official Statistics*, **16**, 191-210.

Eltoft, T., Kim, T., and Lee, T. W. (2006). On the multivariate Laplace distribution. *Signal Processing Letters, IEEE*, **13(5)**, 300-303.

Ericson, W.A. (1969). Subjective Bayesian modeling in sampling finite populations. *Journal of the Royal Statistical Society,* **B31**, 195-234.

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

Henry, K., and Valliant, R. V. (2012). Methods for Adjusting Survey Weights when Estimating a Total. Proceedings of the 2012 Federal Committee on Statistical Methodology?s Research Conference

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

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

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

Little, R.J.A. (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.

McCullagh, P., and Nelder, J.A. (1989). *Generalized Linear Models*, 2^{nd} Edition. CRC Press: Boca Raton, Florida.

Park, T., and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, **103(482)**, 681-686.

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

Potter, F.A. (1988). Survey of Procedures to Control Extreme Sampling Weights *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 453-458.

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

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

Skinner, C.J., Holt, D., and Smith, T.M.F (1989). *Analysis of Complex Surveys*. Wiley: New York.

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 pickup trucks. *Journal of the American Medical Association*, **287**, 1147-1152.

Appendix 1: Full Conditional Distribution for Linear Model

To derive the fully conditional distribution of the linear model for Gibbs sampler, first we start with the hierarchical model:

$$\begin{aligned}
Y_h &\sim MVN(X_h\beta_h, \sigma^2 I_{n_h}) \\
\beta_h &= (\beta_{h1}, ..., \beta_{hp})^T, h = 1, ...H \\
\beta_h &\sim MVN(\beta_h^*, \sigma^2 D_{\tau h}) \\
\beta_h^* &\sim MVN(0_p, \sigma_0^2 I_p) \\
D_{\tau h} &= diag(\tau_{h1}^2, ..., \tau_{hp}^2) \\
\sigma^2 &\sim 1/\sigma^2 \\
\tau_{hi}^2 &\sim \frac{\lambda^2}{2} e^{-\lambda^2 \tau_{hi}^2/2} \\
\lambda^2 &\sim Gamma(\gamma, \delta)
\end{aligned}$$

Ignoring all constants, we reduce the formula to the kernel of likelihood of y, and all other conditional probabilities:

$$\begin{aligned}
p(Y|\beta, \sigma^2) &\propto (\sigma^2)^{-n/2} \prod_{h=1}^{H} \exp\{-\frac{1}{2}(Y_h - X_h\beta_h)^T(\sigma^2 I_{n_h})^{-1}(Y_h - X_h\beta_h)\} \\
p(\beta|\beta^*, \sigma^2, D_\tau) &\propto (\sigma^2)^{-Hp/2} \prod_{h=1}^{H} |D_{\tau h}|^{-1/2} \exp\{-\frac{1}{2}(\beta_h - \beta_h^*)^T(\sigma^2 D_{\tau h})^{-1}(\beta_h - \beta_h^*)\} \\
f(\beta^*) &\propto \prod_{h=1}^{H} \exp\{-\frac{1}{2}\beta_h^{*T}(\sigma_0^2 I_p)^{-1}\beta_h^*\} \\
f(\sigma^2) &\propto 1/\sigma^2 \\
f(\tau^2|\lambda^2) &\propto (\lambda^2)^{Hp} \prod_{h=1}^{H}\prod_{i=1}^{p} \exp(-\lambda^2\tau_{hi}^2/2) \\
f(\lambda^2) &\propto (\lambda^2)^{\gamma-1} \exp(-\delta\lambda^2)
\end{aligned}$$

Since β_hs from different strata are independent, we write separately the kernel of posterior distribution of β_h, which is proportional to the product of likelihood of y and β_h prior.

$$\begin{aligned}
p(\beta_h|rest) \propto & \exp\{-\frac{1}{2\sigma^2}[(Y_h - X_h\beta_h)^T(Y_h - X_h\beta_h) + (\beta_h - \beta_h^*)^T D_{\tau h}^{-1}(\beta_h - \beta_h^*)]\} \\
\propto & \exp\{-\frac{1}{2\sigma^2}[\beta_h^T X_h^T X_h\beta_h - 2Y_h^T X_h\beta_h + \beta_h^T D_{\tau h}^{-1}\beta_h - 2\beta_h^{*T} D_{\tau h}^{-1}\beta_h]\} \\
= & \exp\{-\frac{1}{2\sigma^2}[\beta_h^T(X_h^T X_h + D_{\tau h}^{-1})\beta_h - 2(Y_h^T X_h + \beta_h^{*T} D_{\tau h}^{-1})\beta_h]\} \\
\propto & \exp\{-\frac{1}{2\sigma^2}[(\beta_h - (X_h^T X_h + D_{\tau h}^{-1})^{-1}(Y_h^T X_h + D_{\tau h}^{-1}\beta_h^*))^T(X_h^T X_h + D_{\tau h})^{-1} \\
& (\beta_h - (X_h^T X_h + D_{\tau h}^{-1})^{-1}(Y_h^T X_h + D_{\tau h}^{-1}\beta_h^*))]\}
\end{aligned}$$

Which suggests that $\beta_h|rest \sim MVN(A^{-1}(X_h^T Y_h + D_{\tau h}^{-1}\beta_h^*), \sigma^2 A^{-1}), A = X_h^T X_h + D_{\tau h}^{-1}$. Similarly, we derive the kernel of fully conditional distribution of other parameters as follows:

$$\begin{aligned}
p(\beta_h^*|rest) \propto & \exp\{-\frac{1}{2}[(\beta_h^* - \beta_h)^T(\sigma^2 D_{\tau h})^{-1}(\beta_h^* - \beta_h) + \beta_h^{*T}(\sigma_0^2 I)^{-1}\beta_h^*]\} \\
\propto & \exp\{-\frac{1}{2}[\beta_h^{*T}((\sigma^2 D_{\tau h})^{-1} + (\sigma_0^2 I)^{-1})\beta_h^* - 2\beta_h^T(\sigma^2 D_{\tau h})^{-1}\beta_h^*]\} \\
\propto & \exp\{-\frac{1}{2}(\beta_h^* - ((\sigma^2 D_{\tau h})^{-1} + (\sigma_0^2 I)^{-1})^{-1}(\sigma^2 D_{\tau h})^{-1}\beta_h)^T((\sigma^2 D_{\tau h})^{-1} + (\sigma_0^2 I)^{-1}) \\
& (\beta_h^* - ((\sigma^2 D_{\tau h})^{-1} + (\sigma_0^2 I)^{-1})^{-1}(\sigma^2 D_{\tau h})^{-1}\beta_h)\} \\
\beta_h^*|rest \sim & MVN((\sigma^2 D_{\tau h})^{-1}((\sigma^2 D_{\tau h})^{-1} + (\sigma_0^2 I)^{-1})^{-1}\beta_h, ((\sigma^2 D_{\tau h})^{-1} + (\sigma_0^2 I)^{-1})^{-1})
\end{aligned}$$

$$\begin{aligned}
p(\sigma^2|rest) \propto & (\sigma^2)^{-n/2}\exp\{-\frac{1}{2}(\sigma^2)^{-1}\sum_{h=1}^{H}(Y_h - X_h\beta_h)^T(Y_h - X_h\beta_h)\} * \\
& (\sigma^2)^{-Hp/2}\exp\{-\frac{1}{2}(\sigma^2)^{-1}\sum_{h=1}^{H}(\beta_h - \beta_h^*)^T(D_{\tau h})^{-1}(\beta_h - \beta_h^*)\} * (\sigma^2)^{-1} \\
= & (\sigma^2)^{-(n/2+Hp/2)-1}\exp\{-\frac{1}{2}(\sigma^2)^{-1}[\sum_{h=1}^{H}(Y_h - X_h\beta_h)^T(Y_h - X_h\beta_h) + \\
& \sum_{h=1}^{H}(\beta_h - \beta_h^*)^T(D_{\tau h})^{-1}(\beta_h - \beta_h^*)]\} \\
\sigma^2|rest \sim & InvGamma((n+Hp)/2, \frac{1}{2}[\sum_{h=1}^{H}(Y_h - X_h\beta_h)^T(Y_h - X_h\beta_h) + \\
& \sum_{h=1}^{H}(\beta_h - \beta_h^*)^T(D_{\tau h})^{-1}(\beta_h - \beta_h^*)])
\end{aligned}$$

$$
\begin{aligned}
p(1/\tau_{hi}^2|rest) \propto & (\tau_{hi}^2)^{-\frac{1}{2}} \exp(-\frac{1}{2}\frac{(\beta_{hi}-\beta_{hi}^*)^2}{\sigma^2\tau_{hi}^2}) * \exp(-\frac{\lambda^2\tau_{hi}^2}{2}) * d(\tau_{hi}^2) \\
\propto & (1/\tau_{hi}^2)^{\frac{1}{2}} \exp[-\frac{1}{2}(\frac{(\beta_{hi}-\beta_{hi}^*)^2(1/\tau_{hi}^2)}{\sigma^2} + \frac{\lambda^2}{1/\tau_{hi}^2})] * (1/\tau_{hi}^2)^{-2} \\
= & (1/\tau_{hi}^2)^{-\frac{3}{2}} \exp[-\frac{1}{2}(\frac{(\beta_{hi}-\beta_{hi}^*)^2(1/\tau_{hi}^2)^2 + \lambda^2\sigma^2}{\sigma^2(1/\tau_{hi}^2)})] \\
\propto & (1/\tau_{hi}^2)^{-\frac{3}{2}} \exp[-\frac{1}{2}\frac{((1/\tau_{hi}^2) - \sqrt{\lambda^2\sigma^2/(\beta_{hi}-\beta_{hi}^*)^2})^2}{(\beta_{hi}-\beta_{hi}^*)^{-2}\sigma^2(1/\tau_{hi}^2)}] \\
1/\tau_{hi}^2|rest \sim & InvGaussian(\sqrt{\frac{\lambda^2\sigma^2}{(\beta_h-\beta_h^*)^2}}, \lambda^2)
\end{aligned}
$$

$$
\begin{aligned}
p(\lambda^2|rest) \propto & (\lambda^2)^{Hp} \exp(-\frac{1}{2}\lambda^2 \sum_{h=1}^{H}\sum_{i=1}^{p} \tau_{hi}^2) * (\lambda^2)^{\gamma-1} \exp(-\delta\lambda^2) \\
= & (\lambda^2)^{Hp+\gamma-1} \exp[-\lambda^2(\frac{1}{2}\sum_{h=1}^{H}\sum_{i=1}^{p} \tau_{hi}^2 + \delta)] \\
\lambda^2 \sim & Gamma(Hp+\gamma, \frac{1}{2}\sum_{h=1}^{H}\sum_{i=1}^{p} \tau_{hi}^2 + \delta)
\end{aligned}
$$

Appendix 2: Full Conditional Distribution for Logistic Model

$$
\begin{aligned}
y_{hi}|X_{hi}, \beta_h, &\sim Binomial(p = logit(x_{hi}\beta_h)) \\
\beta_h &= (\beta_{h1}, ..., \beta_{hp})^T, h = 1, ...H \\
\beta_h &\sim MVN(\beta_h^*, \sigma^2 D_{\tau h}) \\
\beta_h^* &\sim MVN(0_p, \sigma_0^2 I_p) \\
D_{\tau h} &= diag(\tau_{h1}^2, ..., \tau_{hp}^2) \\
\tau_{hi}^2 &\sim \frac{\lambda^2}{2} e^{-\lambda^2\tau_{hi}^2/2} \\
\lambda^2 &\sim Gamma(\gamma, \delta)
\end{aligned}
$$

Similarly, we start with the hierarchical model, derive the kernel of the posterior distribution of all parameters, and reveal that they belongs to some known distribution families. The full conditional distribution of β_h doesn't below to any known distribution family, and the rest of parameters are

presented below:

$$p(Y|\beta) = \prod_{h=1}^{H}\prod_{i=1}^{n_h} \quad \frac{exp(x_{hi}\beta_h)}{1+exp(x_{hi}\beta_h)} \quad ^{y_{hi}} \quad \frac{1}{1+exp(x_{hi}\beta_h)} \quad ^{1-y_{hi}}$$

$$p(\beta|\beta^*, D_\tau) \propto \prod_{h=1}^{H} |D_{\tau h}|^{-1/2} \exp\{-\frac{1}{2}(\beta_h - \beta_h^*)^T (D_{\tau h})^{-1}(\beta_h - \beta_h^*)\}$$

$$f(\beta^*) \propto \prod_{h=1}^{H} \exp\{-\frac{1}{2}\beta_h^{*T}(\sigma_0^2 I_p)^{-1}\beta_h^*\}$$

$$f(\tau^2|\lambda^2) \propto (\lambda^2)^{Hp} \prod_{h=1}^{H}\prod_{i=1}^{p} \exp(-\lambda^2\tau_{hi}^2/2)$$

$$f(\lambda^2) \propto (\lambda^2)^{\gamma-1}\exp(-\delta\lambda^2)$$

$$\begin{aligned} p(\beta_h^*|rest) \propto & \exp\{-\frac{1}{2}[(\beta_h^* - \beta_h)^T D_{\tau h}^{-1}(\beta_h^* - \beta_h) + \beta_h^{*T}(\sigma_0^2 I)^{-1}\beta_h^*]\} \\ \propto & \exp\{-\frac{1}{2}[\beta_h^{*T}(D_{\tau h}^{-1} + (\sigma_0^2 I)^{-1})\beta_h^* - 2\beta_h^T D_{\tau h}^{-1}\beta_h^*]\} \\ \propto & \exp\{-\frac{1}{2}(\beta_h^* - (D_{\tau h}^{-1} + (\sigma_0^2 I)^{-1})^{-1} D_{\tau h}^{-1}\beta_h)^T (D_{\tau h}^{-1} + (\sigma_0^2 I)^{-1}) \\ & (\beta_h^* - (D_{\tau h}^{-1} + (\sigma_0^2 I)^{-1})^{-1} D_{\tau h}^{-1}\beta_h)\} \end{aligned}$$

$$\beta_h^*|rest \sim MVN((D_{\tau h})^{-1}((D_{\tau h})^{-1} + (\sigma_0^2 I)^{-1})^{-1}\beta_h, ((D_{\tau h})^{-1} + (\sigma_0^2 I)^{-1})^{-1})$$

$$\begin{aligned} p(1/\tau_{hi}^2|rest) \propto & (\tau_{hi}^2)^{-\frac{1}{2}} \exp(-\frac{1}{2}\frac{(\beta_{hi} - \beta_{hi}^*)^2}{\tau_{hi}^2}) * \exp(-\frac{\lambda^2\tau_{hi}^2}{2}) * d(\tau_{hi}^2) \\ \propto & (1/\tau_{hi}^2)^{\frac{1}{2}} \exp[-\frac{1}{2}((\beta_{hi} - \beta_{hi}^*)^2(1/\tau_{hi}^2) + \frac{\lambda^2}{1/\tau_{hi}^2})] * (1/\tau_{hi}^2)^{-2} \\ = & (1/\tau_{hi}^2)^{-\frac{3}{2}} \exp[-\frac{1}{2}(\frac{(\beta_{hi} - \beta_{hi}^*)^2(1/\tau_{hi}^2)^2 + \lambda^2}{(1/\tau_{hi}^2)})] \\ \propto & (1/\tau_{hi}^2)^{-\frac{3}{2}} \exp[-\frac{1}{2}\frac{((1/\tau_{hi}^2) - \sqrt{\lambda^2/(\beta_{hi} - \beta_{hi}^*)^2})^2}{(\beta_{hi} - \beta_{hi}^*)^{-2}(1/\tau_{hi}^2)}] \end{aligned}$$

$$1/\tau_{hi}^2|rest \sim InvGaussian(\sqrt{\frac{\lambda^2}{(\beta_h - \beta_h^*)^2}}, \lambda^2)$$

$$\begin{aligned} p(\lambda^2|rest) \propto & (\lambda^2)^{Hp}\exp(-\frac{1}{2}\lambda^2\sum_{h=1}^{H}\sum_{i=1}^{p}\tau_{hi}^2) * (\lambda^2)^{\gamma-1}\exp(-\delta\lambda^2) \\ = & (\lambda^2)^{Hp+\gamma-1}\exp[-\lambda^2(\frac{1}{2}\sum_{h=1}^{H}\sum_{i=1}^{p}\tau_{hi}^2 + \delta)] \end{aligned}$$

$$\lambda^2 \sim Gamma(Hp + \gamma, \frac{1}{2}\sum_{h=1}^{H}\sum_{i=1}^{p}\tau_{hi}^2 + \delta)$$