Journal of Data Science 24 (3), 564–583
July 2026
DOI: 10.6339/26-JDS1238
Statistical Data Science

Uncertainty Quantification for Multi-Level Models Using the Survey-Weighted Pseudo-Posterior

Matthew R. Williams[1,*], F. Hunter McGuire[1], and Terrance D. Savitsky[2]

[1]*RTI International, Durham, NC, 27709, USA*
[2]*U.S. Bureau of Labor Statistics, Suitland, MD 20746, USA*

Abstract

Parameter estimation and inference from complex survey samples typically focuses on global model parameters whose estimators have asymptotic properties, such as from fixed effects regression models. The central challenge is to both mitigate bias induced from potentially unbalanced samples and to incorporate adjustments for differences in effective sample size to get correct variance and interval estimates. We present a motivating example of Bayesian inference for a multi-level or mixed effects model in which estimates of both the local parameters (e.g. group level random effects) and the global parameters need to be adjusted for the complex sampling design. We evaluate the limitations of the survey-weighted pseudo-posterior and an existing automated post-processing method to improve the uncertainty quantification. We propose modifications to the automated process and demonstrate their improvements for multi-level models via a simulation study and a motivating example from the National Survey on Drug Use and Health. Reproduction examples are included in the supplementary material and the updated R package is available via github: https://github.com/RyanHornby/csSampling

Keywords *Bayesian inference; complex survey data;* **R***;* **Stan***; survey weights*

1 Introduction

Linear regression models are the backbone of statistical analyses for a variety of fields including the social, economic, behavioral, and health sciences. Furthermore, multi-level regression models which allow for hierarchical specification of sources of randomness (e.g. group vs. individual level variations) have become ubiquitous. Multipurpose statistical modeling packages such as *brms* (Bürkner, 2017) provide a powerful suite of models and features for performing Bayesian analysis for multilevel models, while using syntax closely related to base linear regression. However, key social and economic data sources are collected not as simple random samples, but through complex sampling designs that include sampling of clusters and unequal probabilities of selection. Current software packages can be coerced to produce estimates for these models that are asymptotically consistent by including survey weights. However, the resulting standard errors and credible intervals will not produce correct coverage (e.g. a 95% interval will not contain the true value 95% of the time).

Our motivation is to extend the use of variance component and other hierarchical models to the case where data are collected under a complex survey sampling design associated with a finite population. For simple regression models, well established theory and tools exist (e.g. the *survey*

*Corresponding author. Email: mrwilliams@rti.org.

Received October 7, 2025; Accepted June 19, 2026

package (Lumley, 2016)), allowing for the incorporation of features of the survey sampling design (unequal probabilities of selection, stratification, and clustering) to provide asymptotically correct point and interval estimation. Williams and Savitsky (2021) develop the basic framework Bayesian analysis via a post-estimation adjustment of the survey-weighted posterior via a sandwich adjustment to the covariance matrix. However, their framework only directly applies to so-called global parameters which asymptotically have increasing information about them from the data. For example, we may expect global parameter variances to decrease proportional to the sample size $\propto n^{-1}$. While Bayesian models may capture an arbitrarily complex dependence or covariance structure, this complexity is typically achieved with local parameters dependent on a small portion of the sample observations. These local features might have variances proportional to a slower rate such as $\propto n^{-1/2}$ if we believe that these small groups would grow at a slower rate than the total sample. Some hyper-parameters for deeper models might have very slow rates, close to constant. Achieving asymptotic normality of the joint distribution of both global and local parameters may require much larger sample sizes than is practical. Under this setting, the normality assumption may be violated resulting in under-coverage of intervals when using the sandwich adjustment of Williams and Savitsky (2021). Yet, Bayesian analysts seek correct uncertainty quantification under small samples.

We demonstrate this challenge of identifying which parameters can be reasonably approximated by a normal distribution and which cannot through a simple group-level random effects model. In particular, we are motivated by variance decomposition approaches to model intersectionality effects on health outcomes (McGuire et al., 2024). We consider a two-level generalized linear mixed model. In particular, we consider a logistic regression with fixed effects and random intercepts indexed by groups within the population. This will be our main analysis model, which we will evaluate through simulation and on a large scale complex survey:

$$\begin{aligned} y_{ij} &\sim Bern(p_{ij}), \\ logit(p_{ij}) &= \mu_{ij} = X_{ij}\beta + \alpha_j, \\ \beta &\sim N(0, \Sigma), \\ \alpha_j &\sim N(0, \sigma_\alpha), \end{aligned} \tag{1}$$

where y_{ij} is the binary response for individual i in group j defined by stratifying the population by crossing several demographic variables. We need reliable estimates of both the global parameters β and σ_α as well as the group level estimates α_j to produce marginal estimates of prevalence for each group j averaged over the covariate values X_{ij}. Reliable estimates of β will provide estimates for the main effects or 'single' group effects. Reliable estimates for the group level estimates α_j will provide insights into how different the groups is from the sum of its main effects. For example a person with a combination of race, gender, and sexual orientation have a different expected outcome than is predicted by simply combining separate (fixed) effects from race, gender, and sex. We also wish to approximate the variance contribution, comparing total variation in the outcomes to the contribution coming from between group variation (Goldstein et al., 2002).

In Section 2, we review the motivation behind the survey-weighted pseudo-posterior and a post-processing sandwich adjustment to the posterior variance. In Section 3, we compare different implementation options for the automating such an adjustment. We then compare these approaches in a series of simulation studies (Section 4). We then apply our methods to the National Survey on Drug Use and Health (Section 5). We conclude in Section 6 with a discussion. Additional results are included in the Appendices.

2 Approaches to Adjusting for Survey Design

Suppose that we wish to estimate a statistical model with parameters θ, where θ_0 are the true unknown parameter values and P_{θ_0} describes the data generating mechanism given those parameters. In our motivating example, the parameters $\theta = \{\beta, \Sigma, \sigma_\alpha\}$. We wish to make inference on data Y for a population $U = (1, \ldots N)$ of size N. However, we only have a selected sample $S = (1, \ldots, n \leqslant N)$ drawn from the underlying population. The sampling process is governed by a distribution P_ν, which is the joint distribution over the inclusion indicator variables $(\delta_i, \ldots \delta_N)$. The resulting joint distribution P_{θ_0}, P_ν governs the observed sample we have.

The design P_ν defines the marginal probabilities of selection (or inclusion) $\pi_i = Pr(\delta_i = 1)$ as well as all higher order joint inclusions, e.g. $\pi_{ii'} = Pr(\delta_i = 1, \delta_{i'} = 1)$. A design P_ν is called 'informative', if the distribution of Y is different between the sample and the population. Informative sample designs pose two challenges for inference: (i) Estimation on the sample without accounting for unequal selection π_i will lead to biased estimates. (ii) Failure to account for dependence induced by sample design will lead to uncertainty estimates (standard errors and confidence intervals) that are incorrect. It may be the case that conditioning on covariates X, the conditional distribution of Y (e.g. the relationship between Y and X) *is* the same in the sample and the population. Thus the 'informativeness' of the sample depends on the outcome, and the covariates available in the sample, and the analysis model.

The first challenge can be addressed by using the survey weights $w_i \propto 1/\pi_i$. In order to approximate the population posterior

$$\pi\left(\boldsymbol{\theta}|\mathbf{y}\right) \propto \prod_{i=1}^{N} \pi\left(y_i|\boldsymbol{\theta}\right) \pi\left(\boldsymbol{\theta}\right),$$

we use the survey-weighted pseudo-posterior for the observed sample resulting in a pseudo-posterior:

$$\pi^{\pi}\left(\boldsymbol{\theta}|\mathbf{y}, \mathbf{w}\right) \propto \left[\prod_{i=1}^{n} \pi\left(y_i|\boldsymbol{\theta}\right)^{w_i}\right] \pi\left(\boldsymbol{\theta}\right). \tag{2}$$

In practice we follow Savitsky and Toth (2016) by using weights scaled to sum to the sample size $\sum_i w_i = n$. This provides for a coarse adjustment to the posterior for uncertainty quantification. However, further post-hoc adjustments are often needed.

The incorporation of survey weights is a relatively minor change to the underlying model and easy to implement, especially in general modeling software which already allows for weights. This modification leads to consistent estimation (i.e. the posterior mean converges to the true value θ_0 and posterior distribution (or intervals) shrink, collapsing around the true value θ_0). This consistency is with respect to the joint process of the population generation P_{θ_0} and the taking of samples P_ν. The assumptions needed for consistency of the posterior fall into two classes: (i) conditions on the model and prior and (ii) conditions on the sample design. For the model, there are restrictions on the complexity (e.g. number of parameters) compared to the sample size. For the prior, there are restrictions on having enough prior mass on the true parameter (to exclude cases of very strong and very wrong priors). For the sample design, the assumptions require all individuals in the population to have a non-zero chance of being selected. Selection of individuals needs to be close to independent, with deviations from this being a small fraction, and the overall sample size needs to grow as a rate close to the population size. For more technical details, see Savitsky and Toth (2016) and Williams and Savitsky (2020). It is

worth noting that the consistency for the survey-weighted pseudo-maximum likelihood estimator (MLE) has been known for many years and has been demonstrated for multiple applications with similar restrictions on the sample design but different conditions on the models. See Isaki and Fuller (1982) for an example of an earlier reference and Han and Wellner (2021) for a recent comprehensive review.

The second challenge is asymptotically correct uncertainty quantification. While consistent estimation for P_{θ_0} is possible, the models are misspecified with respect to P_ν. These kinds of 'working' models lead to misleading posterior (credible) intervals, in the sense that the intervals are often too narrow compared to those obtained by a correctly specified joint model for P_{θ_0} and P_ν (León-Novelo and Savitsky, 2019). See Ribatet et al. (2012) for a closely related example for posterior inference based on composite likelihoods. For correctly specified models, we expect asymptotically that the sampling distribution of the maximum likelihood estimator and the posterior distribution will both be normal with the same mean and variance $N(\theta_0, n^{-1}H_{\theta_0}^{-1})$ (called a Bernstein-von Mises result). For a textbook level presentation of Bernstein-von Mises results for many classes of Bayesian models, see Ghosal and Van der Vaart (2017). For misspecified or 'working' models such as the survey-weighted pseudo-posterior, the two will have different covariance matrices. See Kleijn and van der Vaart (2012) for more technical details. In the next section 2.1 we define these covariance matrices for our pseudo-posterior example. In the supplementary materials, we provide a high-level summary of the Bernstein-von Mises result from Williams and Savitsky (2021) for the setting of the pseudo-posterior.

2.1 Asymptotic Covariances

We briefly review the different covariance matrices that arise from complex survey sampling. To facilitate comparisons between these covariance matrices, we first use the following empirical distribution approximation for the joint distribution over population generation and the drawing of an informative sample that produces the observed sample. This empirical distribution construction follows Breslow and Wellner (2007) and incorporates inverse inclusion probability weights, $\{1/\pi_i\}_{i=1,\ldots,N}$, to account for the informative sampling design (Savitsky and Toth, 2016).

$$\mathbb{P}^{\pi}_{N_\nu} = \frac{1}{N}\sum_{i\in U}\frac{\delta_i}{\pi_i}\delta\left(y_i\right), \tag{3}$$

where $\delta\left(y_i\right)$ denotes the Dirac delta function, with probability mass 1 on y_i. This construction contrasts with the usual empirical distribution, $\mathbb{P}_N = \frac{1}{N}\sum_{i\in U}\delta\left(y_i\right)$. Using the notation of Ghosal et al. (2000) we define expectation functionals with respect to these empirical distributions by $\mathbb{P}^{\pi}_N f = \frac{1}{N}\sum_{i=1}^{N}\frac{\delta_i}{\pi_i} f\left(y_i\right)$. Similarly, $\mathbb{P}_N f = \frac{1}{N}\sum_{i=1}^{N} f\left(y_i\right)$. Starting with the Fisher Information matrix for a simple random sample:

$$H_{\theta_0} = -\mathbb{E}_{P_{\theta_0}}\left[\mathbb{P}_N\ddot{\ell}_{\theta_0}\right] = -\frac{1}{N}\sum_{i\in U}\mathbb{E}_{P_{\theta_0}}\ddot{\ell}_{\theta_0}(y_i),$$

which is the negative of the expected value of the second derivative of the log-likelihood $\ddot{\ell}_{\theta_0}$. Under an SRS, $H_{\theta_0}^{-1}$ is the asymptotic variance for the MLE estimate $\hat{\theta}$. It turns out that the survey-weighted Fisher Information matrix is the same as the SRS version, $H^{\pi}_{\theta_0} = H_{\theta_0}$:

$$H^{\pi}_{\theta_0} = -\mathbb{E}_{P_{\theta_0},P_\nu}\left[\mathbb{P}^{\pi}_{N_\nu}\ddot{\ell}_{\theta_0}\right] = -\frac{1}{N}\sum_{i\in U}\mathbb{E}_{P_{\theta_0}}\left[\mathbb{E}_{P_\nu}\frac{\delta_i}{\pi_i}\ddot{\ell}_{\theta_0}(y_i)\right] = -\frac{1}{N}\sum_{i\in U}\mathbb{E}_{P_{\theta_0}}\ddot{\ell}_{\theta_0}(y_i).$$

The survey-weighted MLE has a sandwich form to its variance $H_{\theta_0}^{-1} J_{\theta_0}^{\pi} H_{\theta_0}^{-1}$ with

$$\begin{aligned} J_{\theta_0}^{\pi} &= \mathbb{V}_{P_{\theta_0}, P_\nu}\left[\mathbb{P}_{N_\nu}^{\pi} \dot{\ell}_{\theta_0}\right] = \mathbb{E}_{P_{\theta_0}, P_\nu}\left[\left(\mathbb{P}_{N_\nu}^{\pi} \dot{\ell}_{\theta_0}\right)\left(\mathbb{P}_{N_\nu}^{\pi} \dot{\ell}_{\theta_0}\right)^T\right], \\ J_{\theta_0} &= \mathbb{V}_{P_{\theta_0}}\left[\mathbb{P}_N \dot{\ell}_{\theta_0}\right] = \mathbb{E}_{P_{\theta_0}}\left[\left(\mathbb{P}_N \dot{\ell}_{\theta_0}\right)\left(\mathbb{P}_N \dot{\ell}_{\theta_0}\right)^T\right], \\ J_{\theta_0}^{\pi} &\neq J_{\theta_0}, \end{aligned}$$

where $J_{\theta_0}^{\pi}$ is the variance of the gradient of the log-likelihood with respect to the joint distribution of P_{θ_0}, P_ν. Under correctly specified models, $J_{\theta_0} = H_{\theta_0}$ (Bartlett's 2^{nd} identity) and the sandwich collapses. However for misspecified models, such as those using a pseudo-likelihood, these two matrices are distinct $J_{\theta_0}^{\pi} \neq H_{\theta_0}$, leading to the sandwich form (Godambe Information matrix). In contrast, the survey-weighted posterior behaves as if it were taken from a simple random sample, with variance $H_{\theta_0}^{-1}$. The key challenge then is to adjust the pseudo-posterior such that its variance will match that of the MLE ($H_{\theta_0}^{-1} J_{\theta_0}^{\pi} H_{\theta_0}^{-1}$).

2.2 Curvature Adjustments

Ribatet et al. (2012) propose a curvature adjustment to the pseudo-posterior sample and embed it into their MCMC procedure. However, Williams and Savitsky (2021) perform this adjustment post-hoc, instead of inserting significant complexities into the underlying MCMC sampling process. Let $\hat{\theta}_m$ represent the sample from the pseudo-posterior for $m = 1, \ldots, M$ draws with sample mean $\bar{\theta}$. Define the adjusted sample:

$$\hat{\theta}_m^a = \left(\hat{\theta}_m - \bar{\theta}\right) R_2^{-1} R_1 + \bar{\theta}, \tag{4}$$

where R_1 and R_2 are 'square root' matrices such that $R_1' R_1 = H_{\theta_0}^{-1} J_{\theta_0}^{\pi} H_{\theta_0}^{-1}$ and $R_2' R_2 = H_{\theta_0}^{-1}$. Asymptotically, $\hat{\theta}_m \overset{\text{a}}{\sim} N(\theta_0, n^{-1} H_{\theta_0}^{-1})$. Then we now have $\hat{\theta}_m^a \overset{\text{a}}{\sim} N(\theta_0, n^{-1} H_{\theta_0}^{-1} J_{\theta_0}^{\pi} H_{\theta_0}^{-1})$, which is the asymptotic distribution of the MLE under the pseudo-likelihood.

In survey statistics, we often estimate a 'design effect' (see, for example Kish, 1995) which is a ratio of the (asymptotic) variance of an estimate under the complex survey design vs. its (asymptotic) variance under a simple random sample. This helps us understand the penalty or efficiency of the design (in terms of variance per fixed sample size) compared to a simple random sample. In this way, the adjustment $R_2^{-1} R_1$ is a multivariate estimate of a design effect, providing a parameter-specific adjustment for effective sample size for variances and as well as for correlations between parameters. Adjusting the parameters simultaneously allows for us to adjust all quantities derived from adjusted parameters in downstream inference, such as predicted values. The key focus of this work is to (1) investigate whether or not both global and local parameters need an adjustment for design effect and (2) whether an adjustment can be effectively applied to both global and local parameters in an automated implementation.

In order to apply these adjustments, we must estimate the resulting variance matrices H_θ and J_θ^{π}. The weighted pseudo-posterior samples alone are not sufficient to estimate $J_{\theta_0}^{\pi}$. Instead we can only obtain estimates of $J_\theta = H_\theta$. To remedy this, we must find a way to use survey design-based variance estimation strategies. To achieve this, we use a hybrid approach that is a variation on Taylor linearization (For a review of Taylor linearization methods, see Binder (1996)). Instead of directly estimating the variance of $\hat{\theta}$, Williams and Savitsky (2021) target the variance of the transformed parameters:

$$(\hat{\phi} - \phi_0) = H_{\theta_0}(\hat{\theta} - \theta_0) \approx \sum_{i \in S} w_i \dot{\ell}_{\hat{\theta}}(y_i) = \sum_{i \in S} w_i z_i(\hat{\theta}), \tag{5}$$

where $Var_{P_{\theta_0},P_\nu}(\hat{\phi}-\phi_0) \approx J^\pi_{\theta_0}$. Once we estimate $J^\pi_{\theta_0}$, we back transform to solve for the variance of $\hat{\theta}$ as the sandwich $H^{-1}_{\theta_0} J^\pi_{\theta_0} H^{-1}_{\theta_0}$ In order to estimate $J^\pi_{\theta_0}$, we use a variance estimation method for the linearized term $\sum_{i\in S} w_i \dot{\ell}_{\hat{\theta}}(y_i)$ using a plug-in estimate $\hat{\theta}$. We recall that a replication based approach, such as a jackknife, creates structured subsamples of the data, estimates the target of interest, and calculates the between replicate variance as an estimate of the overall variance. There are many variations of this approach (Rao et al., 1992), with the key being the preservation of the sample design features (e.g. clustering, stratification, and unequal weights) across all replicates.

This hybrid variance estimation approach is implemented in *csSampling*. A replication design in used to estimate $J^\pi_{\theta_0}$ but the use of the resulting sandwich adjustment from $H^{-1}_{\theta_0} J^\pi_{\theta_0} H^{-1}_{\theta_0}$ is a linearization approach. We seek to perform the intense MCMC computation to make posterior draws once, followed by a post-hoc correction to get approximately correct uncertainty quantification.

3 Adjustment for Local Parameters

There are many different options for replicate designs (Rao et al., 1992) ranging from pure bootstrap resampling to structured balanced designs (eg. jackknife and balanced repeated replication). In practice, the results are often qualitatively similar, particularly for means and single regression coefficients. Williams and Savitsky (2021) chose the half-sample bootstrap approach which selects half the primary sampling units with equal probability and doubles their weights, which is implemented as the default in *csSampling*. For global parameters, the use of the half-sample bootstrap appears to provide adequate uncertainty estimates. However, for random effects model with a skewed population, some groups j with small sample size n_j may have small sample representation in several of the bootstrap replicates. That could lead to biased estimation of parameters. Instead, we investigate using a delete-a-group jack-knife approach in which one primary sampling unit (or some aggregated collection of these) is removed, while the remaining units are included. We find that this leads to larger and more stable sample sizes for small groups across all the replicates.

Using the same common delete-a-group jackknife replicates, we compare three alternatives of this basic sandwich approach which could impact the effectiveness of automatic adjustment across both global and local parameters.

3.1 Naïve Adjustment (Baseline)

We first describe the default approach implemented from Williams and Savitsky (2021) in the initial versions of *csSampling*. Stan allows for the extraction of the log of the posterior density and the gradient of the log posterior density. These must be available in order to implement Hamiltonian Monte Carlo (HMC) sampling. However, the gradient of the log-likelihood is not readily available. Let $\xi_\theta = \ell_\theta(y) + \log \pi_\theta$ be the log posterior density with gradient $\dot{\xi}_\theta = \dot{\ell}_\theta(y) + \dot{\log \pi_\theta}$. The simplest approach is to use these in place of the log-likelihood ℓ_θ and $\dot{\ell}_\theta$. The Hessian of log posterior $H_\xi = -\mathbb{E}_{P_{\theta_0},P_\nu}\left[\mathbb{P}^\pi_{N_\nu} \ddot{\xi}_{\theta_0}\right] = H_{\theta_0} + H^0_{\theta_0}$. We assume that the curvature of the prior $H^0_{\theta_0}$ is a negligible contribution to the precision matrix H_θ. Then for convenience we use $\dot{\xi}_\theta$ instead of $\dot{\ell}_\theta(y)$ for estimating $J^\pi_{\theta_0}$ and H_{θ_0}. This is often a reasonable assumption for global parameters such as regression coefficients. However, it is a dubious assumption for local parameters and hyper-parameters.

3.2 Using the Prior Curvature in the Adjustment

While the Hessian of the log posterior $H_\xi = H_{\theta_0} + H^0_{\theta_0}$, the between replicate variance does not contain an extra term from the prior: $J^\pi_\xi = \mathbb{V}_{P_{\theta_0}, P_\nu}\left[\mathbb{P}^\pi_{N_\nu}\dot{\xi}_{\theta_0}\right] = J^\pi_{\theta_0}$. The prior density does not contain any data y or sample indicators or weights. Then for a subset of local or hyper-parameters (e.g. α_j), entries in J^π_ξ will be too small compared to H_ξ, leading to an artificial reduction in scale. To compensate, we can first estimate H^0 directly and use $H_\xi = H_\theta + H^0$ which is computed directly via gradients in Stan, and we can use $J^\pi_* = J^\pi_\xi + H^0$. Thus we use estimates of H_ξ and J^π_* in any curvature adjustments. In practice many Stan models, such as those created by the *brms* package, have options to sample from the priors only, which allows us to construct an estimate for H^0 using the same approach as for H_θ with little additional effort.

Alternatively, we could use $H_\xi - H^0$ and J^π_ξ instead of H_ξ and J^π_* for the sandwich adjustment. As noted above for global parameters, the two are asymptotically equivalent. However, for local parameters and smaller sample sizes, keeping the prior curvature H^0 provides a stabilizing effect, as the variation in some local parameters will have a large contribution from the prior.

3.3 Using Transformations Toward Normality

For MCMC sampling, Stan transforms all model parameters to $(-\infty, \infty)$. For example, a variance parameter specified as $\sigma_\alpha \in [0, \infty)$ is given a log transformation to map to $(-\infty, \infty)$. By default, *csSampling* estimates the H_θ and J^π_θ and produces an adjustment on this 'unconstrained' parameter space. For variance components and other hierarchical model parameters, it is plausible that even unconstrained parameterizations in Stan (e.g. $\log(\sigma_\alpha)$) maybe still be far from normal. This has several implications:

- We can use univariate transformations to get each parameter marginally closer to a Gaussian distribution (ignoring multivariate normality).
- Estimates for the Hessian H_θ may be a poor estimate of the (inverse) of the posterior variance. We can consider using the posterior sample itself to get this variance since the two are equivalent asymptotically.
- We expect that estimates for J^π_θ should be more robust due to the resampling approach.

We propose the following approach which automates the first two points.

- Use the Yeo and Johnson (2000) testing and transformation approach (available in the *car* package Fox and Weisberg (2019)). This is a modification of the Box-Cox approach (Box and Cox, 1964) that is extended to negative valued variables. We call this transformation $\psi(\theta) = \eta$. We will also need this inverse $\psi^{-1}(\eta) = \theta$ (available from the *VGAM* package (Yee, 2025)) as well as its derivative $\partial\psi^{-1}(\eta)/\partial\eta$, which we can estimate numerically.

$$\psi(\lambda, x) = \begin{cases} \{(x+1)^\lambda - 1\}/\lambda, & (x \geqslant 0, \lambda \neq 0), \\ \log(x+1), & (x \geqslant 0, \lambda = 0), \\ -\{(-x+1)^{2-\lambda} - 1\}/(2-\lambda), & (x < 0, \lambda \neq 0), \\ -\log(-x+1), & (x < 0, \lambda = 0). \end{cases}$$

  In implementation, we estimate a unique hyperparameter $\lambda \in (-3, 3)$ and apply a transformation for each of the model parameters independently.
- Approximate the model-based information $H_\eta \approx V^{-1}(\eta)$ as the inverse of the (pseudo-) posterior variance from the weighted sampler.
- Estimate the empirical information matrix $J^\pi_\eta(\eta) = J^\pi(\psi^{-1}(\eta))G(\eta)$ where the multiplication is element-wise and the cells $\{G(\eta)\}_{j\ell} = (\partial\psi^{-1}(\eta_j)/\partial\eta_j)(\partial\psi^{-1}(\eta_\ell)/\partial\eta_\ell)$. We do this by first

estimating $J^{\pi}(\hat{\theta})$ with $\bar{\theta} = \psi^{-1}(\bar{\eta})$ where $\bar{\eta}$ is the posterior mean. We then apply $G(\bar{\eta})$ due to the chain rule.

- Adjust the transformed parameters

$$\eta_m^a = (\eta_m - \bar{\eta})\, R_2^{-1} R_1 + \bar{\eta}.$$

Then transform back to $\hat{\theta}_m^a = \psi^{-1}(\eta_m^a)$.

3.4 Summary of Alternatives

We now present the three approaches in the form of pseudo-code. Algorithm 1 is an adaptation of the original approach from Williams and Savitsky (2021) where we (1) use a different replication approach (jackknife instead of bootstrap) and (2) are more explicit about the estimation of H_θ and J_θ^{π} using posterior gradients ξ_θ vs. likelihood gradients ℓ_θ. Algorithm 2 modifies Algorithm 1 by using an offset from the prior curvature and adding that to J_θ^{π} (changes in blue). Algorithm 3 modifies Algorithm 2 by including a variable transformation $\psi(\theta) = \eta$, applying adjustments to the transformed parameters, then back-transforming. In addition, Algorithm 3 replaces estimates of H_θ with the inverse of the empirical variance of the posterior draws $V^{-1}(\hat{\theta})$ (changes in blue).

Algorithm 1: Naïve adjustment of pseudo-posterior.

input : $\hat{\theta}_m$ from the pseudo-posterior (2);
$\{j, k\}$ indicators for PSUs $j = 1, \ldots, J_k$ and Strata $k = 1, \ldots, K$;
$\{w_{ijk}, \mathbf{X}_{ijk}\}$ for all i in $1, \ldots, I_{jk}$ for every $\{j, k\}$;
R the total number of PSUs.

output: Adjusted sample $\hat{\theta}_m^a$.

Calculate the posterior mean $\bar{\theta} = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m$.
Calculate Monte Carlo average $\hat{H}_\xi = \frac{-1}{M} \sum_{m=1}^{M} \sum_{i \in S} w_i \ddot{\xi}_{\theta_m}(\mathbf{X}_i)$.
for PSUs $r \leftarrow 1$ **to** R **do**
  Create replicate weights $w_{ijk}^r = w_{ijk}$.
  Set weight to zero for each member in r^{th} PSU: $w_{irk}^r = 0$.
  Scale weights for remaining PSU's $\tilde{w}_{ir'k}^r = w_{ir'k} \left(\sum_{ij} w_{ijk} / \sum_{ij} w_{ijk}^r \right)$.
  Evaluate $j_r = \sum_{ijk} \tilde{w}_{ijk}^r \dot{\xi}_{\bar{\theta}}(\mathbf{X}_{ijk})$.
end
Calculate $\hat{J}_{\theta_0}^{\pi} = \frac{C}{R-1} \sum_{r=1}^{R} (j_r - \bar{j})(j_r - \bar{j})^t$ with $\bar{j} = \frac{1}{R} \sum_{r=1}^{R} j_r$.
Calculate $\hat{R}_1$ via Cholesky decomposition: $\hat{R}_1' \hat{R}_1 = \hat{H}_\xi^{-1} \hat{J}_{\theta_0}^{\pi} \hat{H}_\xi^{-1}$.
Calculate $\hat{R}_2$ via Cholesky decomposition: $\hat{R}_2' \hat{R}_2 = \hat{H}_\xi^{-1}$.
Calculate inverse $\hat{R}_2^{-1}$.
Evaluate Eq. 4: $\hat{\theta}_m^a = \left(\hat{\theta}_m - \bar{\theta} \right) \hat{R}_2^{-1} \hat{R}_1 + \bar{\theta}$.

4 Simulations

We present three simulation studies to compare the unadjusted survey-weighted posterior to the three adjustment methods (e.g. naïve, prior-curvature, YJ transformation + prior-curvature).

Algorithm 2: Adjustment of pseudo-posterior using prior curvature (changes from Algorithm 1 in blue).

input : $\hat{\theta}_m$ from the pseudo-posterior (2);
$\{j, k\}$ indicators for PSUs $j = 1, \ldots, J_k$ and Strata $k = 1, \ldots, K$;
$\{w_{ijk}, \mathbf{X}_{ijk}\}$ for all i in $1, \ldots, I_{jk}$ for every $\{j, k\}$;
R the total number of PSUs.

output: Adjusted sample $\hat{\theta}^a_m$.

Calculate the posterior mean $\bar{\theta} = \frac{1}{M}\sum_{m=1}^{M} \hat{\theta}_m$.

Calculate the plug-in estimate for prior curvature $\hat{H}^0 = -\ddot{\log}\,\pi(\bar{\theta})$.

Calculate Monte Carlo average $\hat{H}_\xi = \frac{-1}{M}\sum_{m=1}^{M}\sum_{i\in S} w_i \ddot{\xi}_{\theta_m}(\mathbf{X}_i)$.

for PSUs $r \leftarrow 1$ **to** R **do**

  Create replicate weights $w^r_{ijk} = w_{ijk}$.

  Set weight to zero for each member in r^{th} PSU: $w^r_{irk} = 0$.

  Scale weights for remaining PSU's $\tilde{w}^r_{ir'k} = w_{ir'k}\left(\sum_{ij} w_{ijk} / \sum_{ij} w^r_{ijk}\right)$.

  Evaluate $j_r = \sum_{ijk} \tilde{w}^r_{ijk} \dot{\xi}_{\bar{\theta}}(\mathbf{X}_{ijk})$.

end

Calculate $\hat{J}^\pi_{\theta_0} = \frac{C}{R-1}\sum_{r=1}^{R}(j_r - \bar{j})(j_r - \bar{j})^t$ with $\bar{j} = \frac{1}{R}\sum_{r=1}^{R} j_r$.

Adjust $\hat{J}^\pi_\xi = \hat{J}^\pi_{\theta_0} + \hat{H}^0$.

Calculate $\hat{R}_1$ via Cholesky decomposition: $\hat{R}_1'\hat{R}_1 = \hat{H}_\xi^{-1}\hat{J}^\pi_\xi\hat{H}_\xi^{-1}$.

Calculate $\hat{R}_2$ via Cholesky decomposition: $\hat{R}_2'\hat{R}_2 = \hat{H}_\xi^{-1}$.

Calculate inverse $\hat{R}_2^{-1}$.

Evaluate Eq. 4: $\hat{\theta}^a_m = \left(\hat{\theta}_m - \bar{\theta}\right)\hat{R}_2^{-1}\hat{R}_1 + \bar{\theta}$.

Our first study compares the methods for the case of a simple random sample, under which the unadjusted approach should be ideal and other methods should perform similarly in order to be seriously considered. Our second study induces a single-stage informative sampling design in which the data and group composition are skewed compared to the general population. Our third study utilizes a two-stage cluster sample with both stages inducing an informative design. In these last two scenarios, we seek methods that improve upon the unadjusted pseudo-posterior. In general, these designs are 'strongly' informative, in that both the distribution of responses y_{ij} and the proportion of group membership j in the sample are much different from that of the population.

4.1 SRS Simulation

We first generate a population of $N = 10{,}000$ individuals, each assigned to one of $G = 30$ groups ranging in size n_G from 2,000 to 100. We then generate binary outcomes from the following model:

$$\begin{aligned} y_{ij} &\sim Bern(p_{ij}), \\ logit(p_{ij}) &= -2 + x_{ij} + \alpha_j, \\ \alpha_j &\sim N(0, 0.25), \\ x_{ij} &\sim N(0, 1). \end{aligned}$$

Algorithm 3: Adjustment of pseudo-posterior using transformations toward normality and prior curvature (changes from Algorithm 2 in blue).

input : $\hat{\theta}_m = \{\hat{\theta}_1, \ldots, \hat{\theta}_D\}_m$ from the pseudo-posterior (2);
$\{j, k\}$ indicators for PSUs $j = 1, \ldots, J_k$ and Strata $k = 1, \ldots, K$;
$\{w_{ijk}, \mathbf{X}_{ijk}\}$ for all i in $1, \ldots, I_{jk}$ for every $\{j, k\}$;
R the total number of PSUs.

output: Adjusted sample $\hat{\theta}^a_m$.

for Parameters $d \leftarrow 1$ **to** D **do**
  Estimate $\hat{\lambda}_d \in (-3, 3)$.
  Apply transformation $\{\eta_d\}_m = \psi(\hat{\lambda}_d, \{\hat{\theta}_d\}_m)$.
end
Compile $\eta_m = [\{\eta_1\}_m, \ldots, \{\eta_D\}_m]$ and $\psi(\theta) = \left[\psi(\hat{\lambda}_1, \{\hat{\theta}_d\}_m), \ldots \psi(\hat{\lambda}_1, \{\hat{\theta}_d\}_m)\right]$.
Calculate the posterior mean $\bar{\eta} = \frac{1}{M}\sum_{m=1}^{M} \eta_m$.
Estimate the posterior variance $V(\eta) = \frac{1}{M-1}\sum_{m=1}^{M}(\eta_m - \bar{\eta})(\eta_m - \bar{\eta})^t$.
Estimate the posterior curvature $\hat{H}_\eta = V^{-1}(\eta)$.
Calculate the plug-in estimate for prior curvature $\hat{H}^0 = -\ddot{\log}\,\pi(\psi^{-1}(\bar{\eta}))$.
for PSUs $r \leftarrow 1$ **to** R **do**
  Create replicate weights $w^r_{ijk} = w_{ijk}$.
  Set weight to zero for each member in r^{th} PSU: $w^r_{irk} = 0$.
  Scale weights for remaining PSU's $\tilde{w}^r_{ir'k} = w_{ir'k}\left(\sum_{ij} w_{ijk} / \sum_{ij} w^r_{ijk}\right)$.
  Evaluate $j_r = \sum_{ijk} \tilde{w}^r_{ijk} \dot{\xi}_{\psi^{-1}(\bar{\eta})}(\mathbf{X}_{ijk})$.
end
Calculate $\hat{J}^\pi_\theta = \frac{C}{R-1}\sum_{r=1}^{R}(j_r - \bar{j})(j_r - \bar{j})^t$ with $\bar{j} = \frac{1}{R}\sum_{r=1}^{R} j_r$.
Adjust on θ space $\hat{J}^\pi_\xi = \hat{J}^\pi_\theta + \hat{H}^0$.
Calculate $G(\bar{\eta}) = \left(\frac{\partial\psi^{-1}(\eta)}{\partial\eta}|_{\bar{\eta}}\right)\left(\frac{\partial\psi^{-1}(\eta)}{\partial\eta}|_{\bar{\eta}}\right)^t$.
Convert to η space $\hat{J}^\pi_\eta = \hat{J}^\pi_\xi G(\eta)$.
Calculate $\hat{R}_1$ via Cholesky decomposition: $\hat{R}_1'\hat{R}_1 = \hat{H}^{-1}_\eta \hat{J}^\pi_\eta \hat{H}^{-1}_\eta$.
Calculate $\hat{R}_2$ via Cholesky decomposition: $\hat{R}_2'\hat{R}_2 = \hat{H}^{-1}_\eta$.
Calculate inverse $\hat{R}^{-1}_2$.
Evaluate Eq. 4: $\eta^a_m = (\eta_m - \bar{\eta})\,\hat{R}^{-1}_2\hat{R}_1 + \bar{\eta}$.
Transform back $\hat{\theta}^a_m = \psi^{-1}(\eta^a_m)$.

Figure 1 shows the distribution of the binary outcome across the 30 groups in the population. The proportion increases across groups, roughly doubling from the first to the last group.

We then take 200 simple random samples (SRS) sample of 1000 individuals. We estimate the survey-weighted pseudo-posterior on each and apply the different adjustments to the posterior variance. Table 1 compares the results over the 100 samples. As expected, the unadjusted pseudo-posterior distribution shows close to nominal coverage (95%) on average for the fixed effects, random effects, and the random effects variance parameter. The three adjustments achieve or exceed coverage for fixed effects, incurring a cost in the increased size of the intervals. The naïve approach severely under-covers both the random effects and their variance. Both the prior

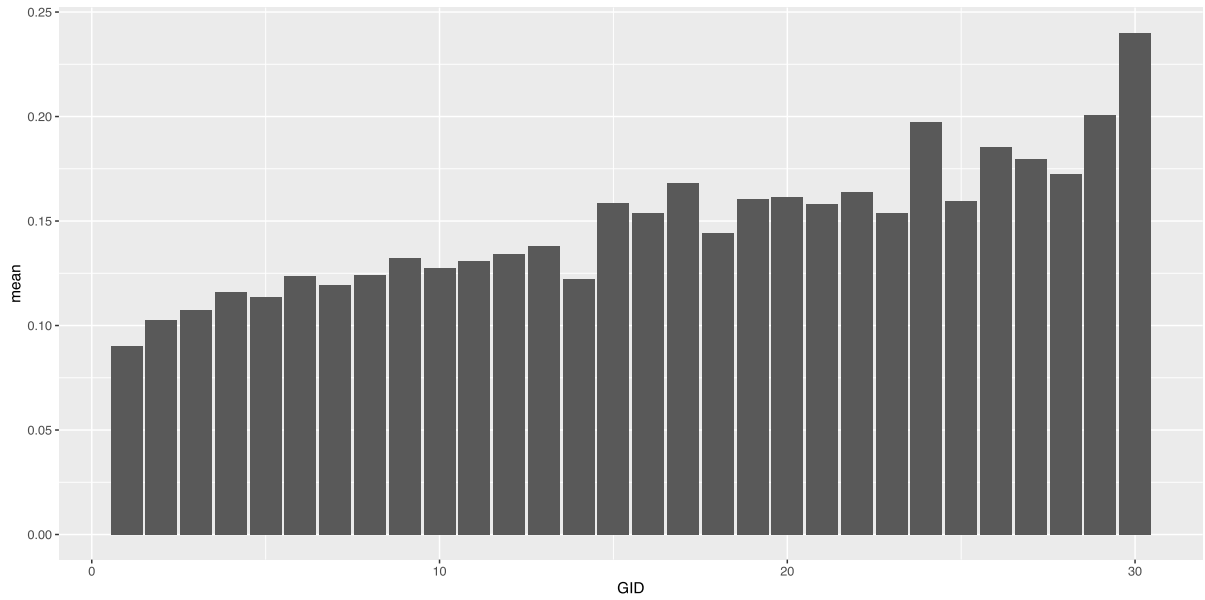

Figure 1: Proportion of binary outcomes in the simulated population across 30 groups.

Table 1: Results for 100 SRS samples, comparing mean interval length and coverage over fixed effects (slope and intercept), random effects (30 groups), and the random effect variance for the four alternative variance approaches.

	Interval Length			Interval Coverage %		
Adjustment	Fixed	Random	σ_α	Fixed	Random	σ_α
Unadjusted	0.499	1.277	0.622	93.0	96.5	96.0
Naïve	0.532	0.405	0.146	95.5	52.7	38.0
Prior Curvature	0.958	0.980	0.189	99.5	89.6	46.0
Yeo-Johnson	0.566	1.217	0.377	97.0	95.6	78.0

curvature and the Yeo-Johnson approach (with prior curvature) dominate the naïve method in terms of coverage. Yeo-Johnson achieves nominal coverage on average for the random effects, but still falls somewhat short for the random effect variance. However, it gets much higher coverage than the naïve and prior curvature adjustments.

4.2 PPS Simulation

Similar to the SRS simulation, we first generate a population of N =10,000 individuals, each assigned to one of $G = 30$ groups ranging in size n_G from 2,000 to 100. We use the same population generation model for y_{ij}. To create a Probability Proportional to Size (PPS) sample, we develop a size measure for each individual:

$$s_{ij} = \max\{0.1, (\mu_{ij} - \min_{i,j}(\mu_{ij})) + 5\alpha_j\},$$

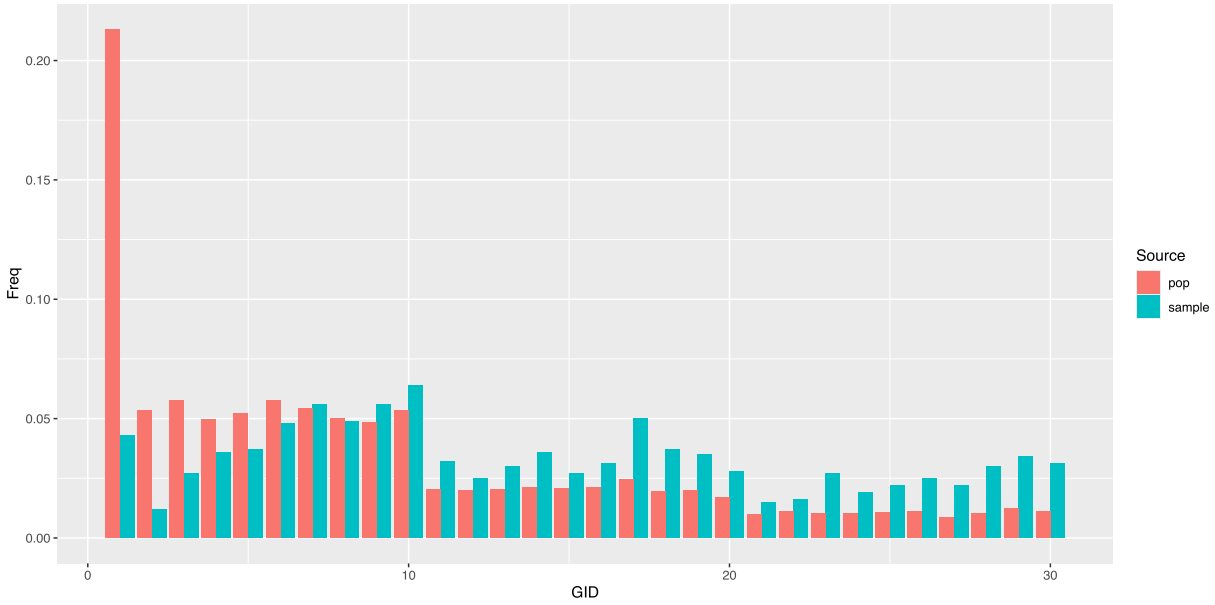

Figure 2: Comparison of proportion of group membership between the population and a PPS sample.

Table 2: Results for 100 PPS samples, comparing mean interval length and coverage over fixed effects (slope and intercept), random effects (30 groups), and the random effect variance for the four alternative variance approaches.

	Interval Length			Interval Coverage %		
Adjustment	Fixed	Random	σ_α	Fixed	Random	σ_α
Unadjusted	0.531	1.239	0.557	64.5	86.8	89.0
Naïve	0.686	0.377	0.180	71.0	39.0	39.0
Prior Curvature	1.111	0.981	0.232	88.5	79.1	51.0
Yeo-Johnson	0.658	1.677	0.766	72.0	88.9	90.0

where $\mu_{ij} = logit(p_{ij})$. The size measure is larger for individuals more likely to have a value of 1 and is larger for those belonging to a group with a large random effect. Figure 2 compares the distribution of group membership in the population with that from one PPS sample. Compared to the population, the PPS sample has a higher proportion of membership in groups with larger group IDs. These same groups also have a larger proportion of positive indicators (Figure 1). Together this creates samples that have both a different proportion of positive responses and a larger representation of some groups compared to the population.

Table 2 compares the results over the 100 repeated samples. As expected, the unadjusted pseudo-posterior distribution falls short of coverage due to the unequal weighting and the dependence induced by the sampling design. The naïve adjustment leads to some improvement in coverage for the fixed effects. It also has severe under-coverage for the random effects and the random effect variance. The Yeo-Johnson adjustment dominates both the unadjusted and naïve adjusted approaches, leading to some improvements for both the fixed effects and the random

effects. The prior curvature approach has the highest coverage for fixed effects, but reduced coverage for the random effects and random effects variance compared to the unadjusted method. All approaches fall short of the nominal 95% coverage level. This may be due to the relatively small sample size and strongly informative sample design.

4.3 Two-Stage Sample Simulation

We first generate a larger population of N =100,000 individuals, each assigned to one of $G = 30$ groups ranging in size n_G from 20,000 to 1,000. In order to create an informative two-stage design, we first sort the individuals by their true expected value p_{ij}. We then assign 1000 primary sampling units using the sort order: the largest 100 get assigned to the first PSU, the next largest 100 get assigned to the second PSU and so on. Thus individuals within each PSU are more similar to each other than to a randomly selected member of the population. We develop a size measure for each PSU:

$$s_k = \max\{0.1, (\mu_k - \min_k(\mu_k))^2\},$$

where $\mu_k = \frac{1}{100}\sum_{i,j}\mu_{ij}1_k(i, j)$ is the average over the 100 individuals in the k^{th} PSU. We then select 100 of the 1,000 clusters (PSUs) from the population.

For the second stage, within each PSU indexed by k, we select individuals with probability proportional to size:

$$s_{ij}^k = \max\{0.1, (\mu_{ij} - \min_{i,j}(\mu_{ij})) + 5(\alpha_j - \min_j(\alpha_j))\}.$$

We select 10 individuals from each cluster for a total sample size of 1,000.

At the first stage, the size measure is larger for clusters that have a higher proportion of 1's as well as a higher proportion of groups with large random effects. Together this should induce a design that has both a different proportion of positive responses and a larger representation of some groups compared to the population. The second stage of sampling amplifies this imbalance by further selecting individuals within each cluster which are more likely to have a response of 1 and more likely to be from a group with a larger random effect. This can be seen in Figure 3 in which the group membership in the multistage sample is further skewed towards the right (the rarer groups with larger proportions of the binary response).

Table 3 compares the results over the 100 repeated samples. The unadjusted pseudo-posterior distribution falls short of coverage for the fixed effects parameters and the random effect variance. However, it seems to cover the random effects at a close to nominal rate. The naïve adjustment leads to almost no improvement for the fixed effects and much worse coverage for the random effects and the random effects variance. The Yeo-Johnson adjustment dominates both the unadjusted and naïve adjusted approaches, leading to modest improvements for the fixed effects and the random effects. Coverage for the random effects variance is much improved, but still far from nominal. The prior curvature approach improves coverage for fixed effects and random effects but shows little improvement in coverage for the random effects variance compared to the naïve approach. All approaches fall short of the nominal 95% coverage level for the fixed effect and the random effects variance. This may be due to the relatively small sample size and strongly informative sample design.

Figure 3: Comparison of proportion of group membership between the population and a multi-stage PPS sample.

Table 3: Results for 100 two-stage samples, comparing mean interval length and coverage over fixed effects (slope and intercept), random effects (30 groups), and the random effect variance for the four alternative variance approaches.

	Interval Length			Interval Coverage %		
Adjustment	Fixed	Random	σ_α	Fixed	Random	σ_α
Unadjusted	0.643	1.896	0.678	62.5	92.2	21.0
Naïve	0.662	1.277	0.417	65.0	76.6	8.0
Prior Curvature	0.883	1.908	0.459	84.5	93.6	9.0
Yeo-Johnson	0.744	2.538	1.091	72.0	94.6	40.0

5 NSDUH Application

Our motivating example comes from the analysis of depression in adults from the National Survey on Drug Use and Health (NSDUH) by McGuire et al. (2024). The NSDUH is a nationally representative survey of the civilian, non-institutionalized population aged 12 years or older residing within the United States. The multi-stage design stratifies states and regions within states, then takes a series of nested samples of increasing geographic resolution leading to the selection of individual dwelling units and up to two individuals within each. The survey over-samples smaller population states and younger age groups (Center for Behavioral Health Statistics and Quality, 2020). National level analyses across large age-ranges (such as all adults) will need to use the survey weights to mitigate bias due to differences in substance use and mental health rates across age groups and across states. The clustering of the sample by geographic areas (census block groups) and the sampling of up to two persons per household reduces the effective sample size of the design relative to a simple random sample. Together, the unequal selection

probability and the clustering are non-ignorable features of the survey design. They should be included during analysis to mitigate bias and to obtain more accurate uncertainty quantification.

The full example uses the public use files from 2015 through 2020, resulting in a sample size of about 236,000 adults. There are 42 unique groups defined by the intersection of race/ethnicity, gender, and sexual orientation. For analysis, we use past year depression status and a multilevel logistic regression (1) with fixed effects as the main effects (race/ethnicity, gender, sexual orientation) and the random effects as the intersectional groups. Our focus of estimation and uncertainty adjustment will be on the model parameters. However, this will also impact derived quantities such as group-level estimates. In the full model of McGuire et al. (2024), they also include age as a fixed effect to use in age-adjusted estimates. For ease of discussion we exclude age from our example analysis.

We compare the unadjusted estimates (directly from pseudo-posterior) with the Yeo-Johnson (with prior curvature) adjustment approach. We compare results with fitting a posterior for each of the 100 replicate data sets and then taking the variance of the posterior means across replicates. Because this is computationally intensive (roughly 100 times slower), we first randomly subset the data. We subsample 500 adults from each of the 42 groups (taking all members of groups with fewer than 500 individuals). We then adjust the weights of each group to account for the sub-sampling. This leads to a sample of approximately 14,000 adults while still preserving the sample size of the smallest groups (For a comparison of the unadjusted and YJ adjusted estimates for this subsample, see the supplementary material. See Appendix A for approximate run times).

Figure 4 compares the three alternative approaches in terms of their 95% intervals for global parameters (fixed effects and random effect variance). The unadjusted sample has narrower intervals that the YJ adjusted sampled. This is to be expected, as typically we expect a 'design effect' to reduce the effective sample size due to clustered and unequal sampling. Of note, the interval for random effect variance (Level 2) is narrowest for the unadjusted sample. We further note that the replication based approach results in larger intervals across all these global parameters due to the reduced size of the subsample (14,000 vs. 236,000).

We next compare the performance of the three alternatives for random effects estimation (Figure 5). We see that the unadjusted intervals are the narrowest. The YJ intervals are consistently wider than those of the unadjusted approach. The naive replication method (using a subsample) has wider intervals for some random effects corresponding to larger groups, but similar intervals for smaller groups with original sample sizes less than 500, for example Native Hawaiian and Pacific Islander (NHPI) and Native American and Alaskan Native (NAAN) groups.

Figure 6 compares the resulting domain level estimates for each group to direct estimates from the full sample obtained by taking the weighted average depression rate within each group. In general, we see that for smaller sized groups (such as NHPI and NAAN) the direct estimates have extremely wide intervals compared to the model-based approaches. This is the main motivation for using models to gain precision for smaller domains. In contrast, for groups with larger sample sizes (such as White) the direct estimates have intervals of similar or even smaller widths compared to the model-based estimates. As expected, the YJ intervals are wider than the unadjusted intervals. Similar to the comparison of the random effects, the naive replication approach on the subsample leads to similar results to the YJ approach for smaller sized groups, but leads to wider intervals for larger and mid-sized groups (such as Hispanic and White).

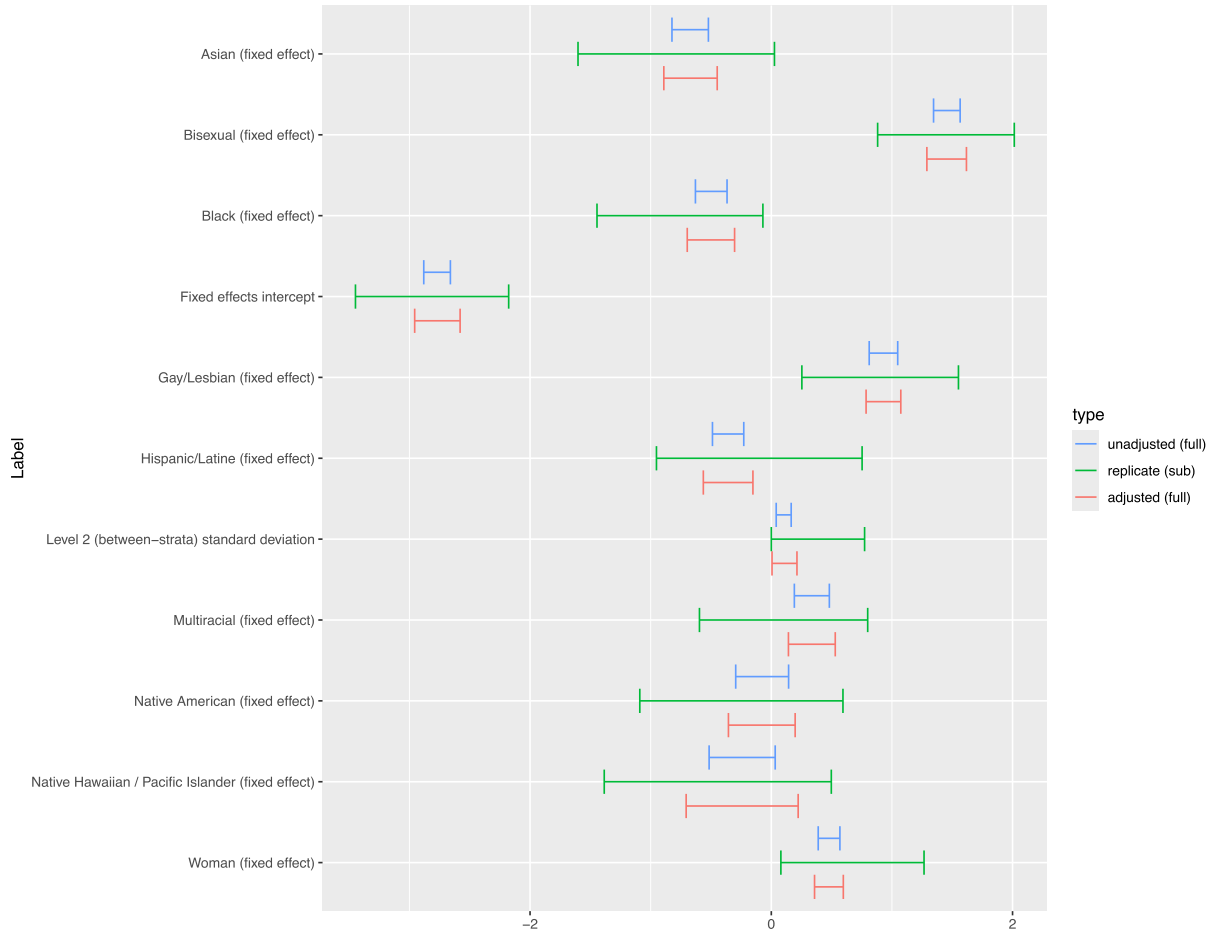

Figure 4: Comparison of 95% intervals for global regression parameters for the multi-level logistic regression for past year depression from NSDUH.

6 Discussion and Conclusions

In this work, we demonstrated how an approach for automated adjustment of Bayesian models for survey data - which works well for global or fixed effects models - has challenges for fitting multi-level models with local parameters. Our motivating example was a variance component model for a logistic regression with random intercepts for different subpopulations. We discussed how the automation approach could be modified to be less reliant on asymptotics (e.g. large sample sizes) through the incorporation of prior curvature and through automated transformations of the parameter space towards symmetry and normality. Through a series of simulations, we see that these modifications mitigate the underestimation of variance from the existing approach. Our application to past year depression from the NSDUH compares our best approach (YJ) to the brute force replication approach in which the Bayesian model is re-estimated for each replicate (e.g. 100 times) but on a stratified subsample of the data. Given a similar computational level of effort, our proposed approach on the full sample (YJ) provides more efficient intervals compared to a replication method applied to a subsample of the data. Both approaches provide improvements over direct estimates for small and mid-sized domains.

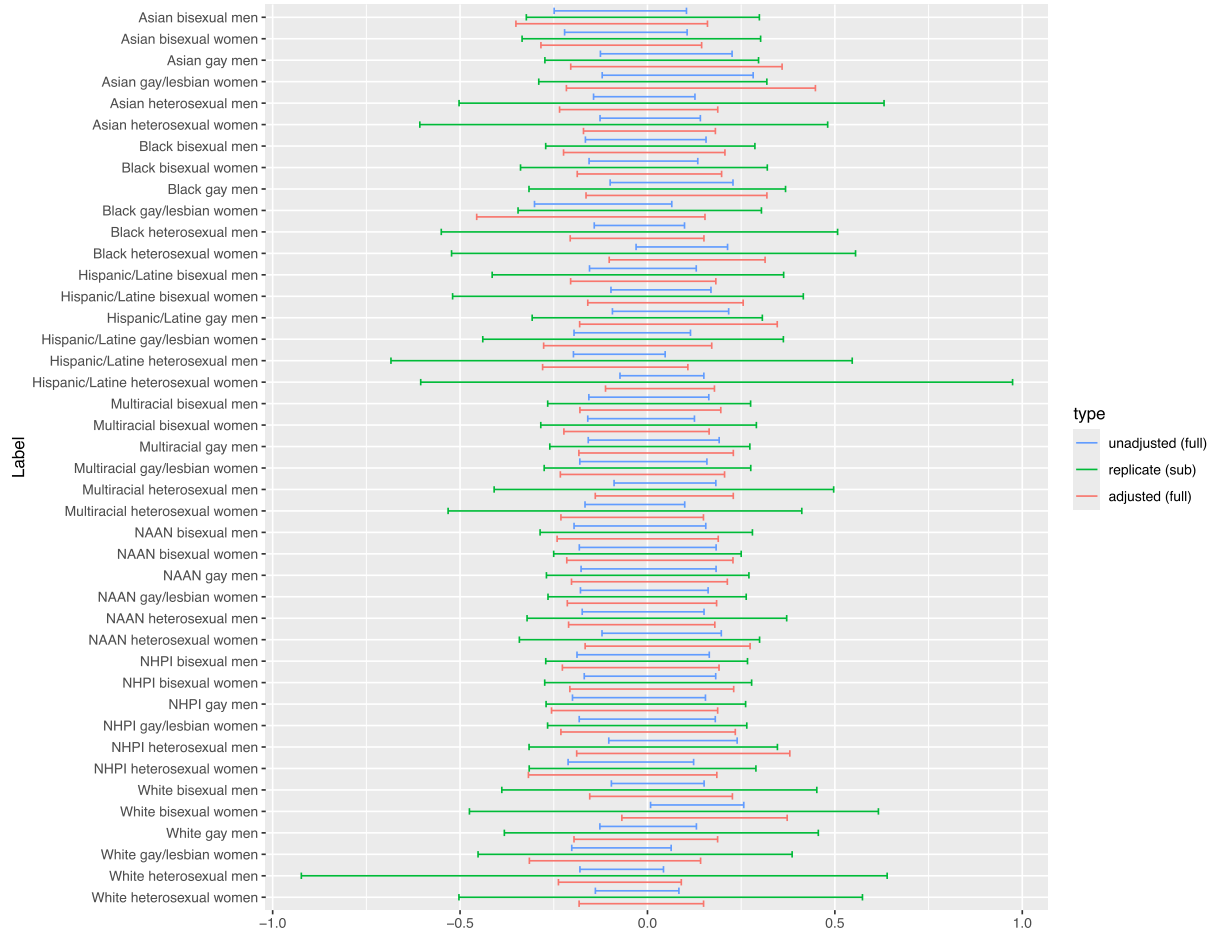

Figure 5: Comparison of 95% intervals for local random effects for the multi-level logistic regression for past year depression from NSDUH.

Our motivating ANOVA-like model is a 'shallow' hierarchical model. We also evaluated our approach on a 'deeper' hierarchical selection model (global-local). In the supplementary material, we investigate a more complex model with a deeper hierarchy using global-local priors for a subset of NSDUH data. While the pseudo-likelihood approach leads to similar group-level point estimates, the YJ adjustment is extremely unstable for the more complex model. While the unadjusted estimates from the pseudo-posterior lead to similar domain level estimates between the simpler and more complex models, the YJ adjustment is very unstable and leads to unusable intervals.

Future work could incorporate refinements to the YJ transformation. For example, we may treat the YJ transformed variables as skew-normal and provide one additional transformation towards normality before creating the sandwich adjustment (See Robbins et al. (2013) for an example used in imputation). Future work could also focus on diagnosing when a model is 'too deep' or when parameters are close to mixed type (point mass at zero) such that the resulting sandwich adjustment (even after transformation) is not appropriate. The 'automatic' adjustment could be applied differentially, adjusting only a subset of parameters. For example, Lee (2026) develop a per-parameter design effect ratio diagnostic which can inform which fixed and random

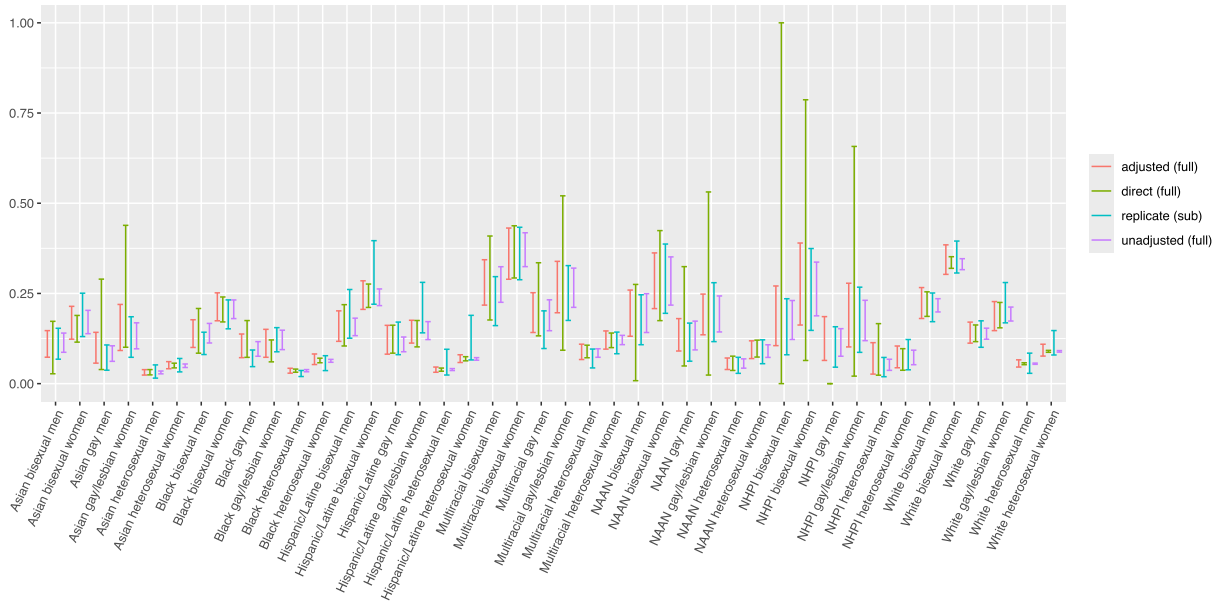

Figure 6: Comparison of 95% intervals for group level estimates for past year depression from NSDUH.

effects are survey sensitive. We can also consider data cloning (Lele et al., 2010), which attempts to estimate the rate of decrease of the variance to identify parameters that are not estimable. We could re-purpose this diagnostic to identify which parameters have variance reduction with increasing sample size in the global (n^{-1}), local ($n^{-1/2}$), or 'deep' hyper-parameter (e.g. $1/\log(n)$) regimes.

Supplementary Material

The supplementary material includes a summary of the Bernstein-von Mises result from Williams and Savitsky (2021). It also includes additional diagnostic summaries for the simulation study and alternative analyses of the NSDUH application. Also included is R code for replicating the simulation studies and R code and data for replicating the NSDUH example. The *csSampling* package has been updated with the new functionality described in this paper and can be downloaded from github: https://github.com/RyanHornby/csSampling

A Run Time Comparisons

In this section, we summarize approximate run times for the different examples and alternative approaches. All runs were performed on the same personal computer using a single core per run. Table 4 summarizes the approximate run times. For example, the simulation studies of Section 4 each had 100 samples of size 1,000 individuals selected. The run time for a single sample of 1,000 estimating the survey-weighted pseudo-posterior was approximately 17s. Running the naïve post-hoc adjustment resulted in a total run time of approximately 50s, which includes the initial estimation of the unadjusted pseudo-posterior. This was similar for the prior curvature adjustment. However, the YJ adjustment only added a small additional amount of run time

Table 4: Approximate run time comparisons for data sets and adjustment approaches for the variance component model (1). Times scale with sample size and the number of Hamiltonian Monte Carlo (HMC) warm-up and sample (draws).

Data Set	Size	Warm-up	Draws	Method	run time
Simulation Sample	1,000	3,000	2,000	Unadjusted	17s
		3,000	2,000	Naïve	50s
		3,000	2,000	Prior Curvature	50s
		3,000	2,000	Yeo-Johnson	20s
NSDUH subsample	14,000	3,000	2,000	Unadjusted	7m
		3,000	2,000	Yeo-Johnson	7m 10s
		3,000	2,000	Replication	12h
Full NSDUH	236,000	3,000	2,000	Unadjusted	3h
		3,000	2,000	Yeo-Johnson	3h 10m

resulting in a total duration of 20s. The YJ adjustment is faster because it avoids estimating the H_θ matrix on every MCMC draw.

For the NSDUH subsample of size 14,000, the additional time added for the post-hoc YJ adjustment is small. In contrast, running 100 models for the replication approach is two orders of magnitude slower. On the full NSDUH sample, only the YJ adjustment is feasible given time constraints. It still only requires a relatively smaller additional burden compared to the overall run time for estimating the pseudo-posterior model.

References

Binder DA (1996). Linearization methods for single phase and two-phase samples: A cookbook approach. *Survey Methodology*, 22: 17–22.

Box GE, Cox DR (1964). An analysis of transformations. *Journal of the Royal Statistical Society, Series B, Statistical Methodology*, 26(2): 211–243. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x

Breslow NE, Wellner JA (2007). Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scandinavian Journal of Statistics*, 34(1): 86–102. https://doi.org/10.1111/j.1467-9469.2006.00523.x

Bürkner PC (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, 80(1): 1–28. https://doi.org/10.18637/jss.v080.i01

Center for Behavioral Health Statistics and Quality (2020). 2019 national survey on drug use and health: Methodological summary and definitions. Retrieved from SAMHSA website.

Fox J, Weisberg S (2019). *An R Companion to Applied Regression*. Sage, Thousand Oaks CA, third edition.

Ghosal S, Ghosh JK, Vaart AWVD (2000). Convergence rates of posterior distributions. *The Annals of Statistics*, 28(2): 500–531.

Ghosal S, Van der Vaart A (2017). *Fundamentals of Nonparametric Bayesian Inference*, volume 44. Cambridge University Press.

Goldstein H, Browne W, Rasbash J (2002). Partitioning variation in multilevel models. *Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences*, 1(4): 223–231. https://doi.org/10.1207/S15328031US0104_02

Han Q, Wellner JA (2021). Complex sampling designs: Uniform limit theorems and applications. *The Annals of Statistics*, 49(1): 459–485. https://doi.org/10.1214/20-AOS1964

Isaki CT, Fuller WA (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77: 89–96. https://doi.org/10.1080/01621459.1982.10477770

Kish L (1995). Methods for design effects. *Journal of Official Statistics*, 11(1): 55.

Kleijn B, van der Vaart A (2012). The Bernstein-von-Mises theorem under misspecification. *Electronic Journal of Statistics*, 6: 354–381.

Lee J (2026). Design effect ratios for bayesian survey models: A diagnostic framework for identifying survey-sensitive parameters. arXiv preprint arXiv:2603.07791.

Lele SR, Nadeem K, Schmuland B (2010). Estimability and likelihood inference for generalized linear mixed models using data cloning. *Journal of the American Statistical Association*, 105(492): 1617–1625. https://doi.org/10.1198/jasa.2010.tm09757

León-Novelo LG, Savitsky TD (2019). Fully Bayesian estimation under informative sampling. *Electronic Journal of Statistics*, 13(1): 1608–1645.

Lumley T (2016). survey: analysis of complex survey samples. R package version 3.32.

McGuire FH, Beccia AL, Peoples JE, Williams MR, Schuler MS, Duncan AE (2024). Depression at the intersection of race/ethnicity, sex/gender, and sexual orientation in a nationally representative sample of us adults: A design-weighted intersectional maihda. *American Journal of Epidemiology*, 193(12): 1662–1674. https://doi.org/10.1093/aje/kwae121

Rao JNK, Wu CFJ, Yue K (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology*, 18: 209–217.

Ribatet M, Cooley D, Davison AC (2012). Bayesian inference from composite likelihoods, with an application to spatial extremes. *Statistica Sinica*, 22(2): 813–845.

Robbins MW, Ghosh SK, Habiger JD (2013). Imputation in high-dimensional economic data as applied to the agricultural resource management survey. *Journal of the American Statistical Association*, 108(501): 81–95. https://doi.org/10.1080/01621459.2012.734158

Savitsky TD, Toth D (2016). Bayesian estimation under informative sampling. *Electronic Journal of Statistics*, 10(1): 1677–1708. https://doi.org/10.1214/16-EJS1153

Williams MR, Savitsky TD (2020). Bayesian estimation under informative sampling with unattenuated dependence. *Bayesian Analysis*, 15(1): 57–77. https://doi.org/10.1214/18-BA1143

Williams MR, Savitsky TD (2021). Uncertainty estimation for pseudo-Bayesian inference under complex sampling. *International Statistical Review*, 89(1): 72–107. https://doi.org/10.1111/insr.12376

Yee TW (2025). VGAM: Vector Generalized Linear and Additive Models. R package version 1.1-13.

Yeo I, Johnson RA (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, 87(4): 954–959. https://doi.org/10.1093/biomet/87.4.954