

Supplementary Material: Uncertainty Quantification for Multi-level Models Using the Survey-Weighted Pseudo-Posterior

MATTHEW R. WILLIAMS^{1,*}, F. HUNTER MCGUIRE², AND TERRANCE D. SAVITSKY³

¹*RTI International, USA*

²*PicnicHealth, USA*

³*U.S. Bureau of Labor Statistics, USA*

1 Summary of Bernstein-von Mises Result

We summarize the main Bernstein-von Mises result and assumptions from [Williams and Savitsky \(2021\)](#). This result demonstrates that the asymptotic covariance of the pseudo-posterior is distinct from the asymptotic covariance of the maximum likelihood estimator (MLE). The former is just the Hessian of the log-likelihood while the latter has the Godambe ‘sandwich’ form. This motivates the need for a curvature adjustment for the pseudo-posterior to achieve proper asymptotic frequentist coverage.

Following [Ghosal et al. \(2000\)](#), we define expectation functionals with respect to empirical distributions by $\mathbb{P}_{N_\nu}^\pi f = \frac{1}{N_\nu} \sum_{i=1}^{N_\nu} \frac{\delta_{y_i}}{\pi_{\nu i}} f(\mathbf{X}_{\nu i})$. Similarly, $\mathbb{P}_{N_\nu} f = \frac{1}{N_\nu} \sum_{i=1}^{N_\nu} f(\mathbf{X}_{\nu i})$. We denote the associated centered empirical processes as $\mathbb{G}_{N_\nu}^\pi = \sqrt{N_\nu} (\mathbb{P}_{N_\nu}^\pi - P_0)$ and $\mathbb{G}_{N_\nu} = \sqrt{N_\nu} (\mathbb{P}_{N_\nu} - P_0)$. From the main text we have defined:

$$\begin{aligned} H_{\theta_0}^\pi &= H_{\theta_0} - \frac{1}{N} \sum_{i \in U} \mathbb{E}_{P_{\theta_0}} \ddot{\ell}_{\theta_0}(y_i), \\ J_{\theta_0} &= \frac{1}{N_\nu} \sum_{i \in U_\nu} \mathbb{E}_{P_{\theta_0}} \dot{\ell}_{\theta_0}(\mathbf{X}_{\nu i}) \dot{\ell}_{\theta_0}(\mathbf{X}_{\nu i})^T, \\ J_{\theta_0}^\pi &= \mathbb{E}_{P_{\theta_0}, P_\nu} \left[\left(\mathbb{P}_{N_\nu}^\pi \dot{\ell}_{\theta_0} \right) \left(\mathbb{P}_{N_\nu}^\pi \dot{\ell}_{\theta_0} \right)^T \right], \end{aligned}$$

with $\dot{\ell}_{\theta_0}$ and $\ddot{\ell}_{\theta_0}$ representing the first and second order derivatives of the log-likelihood, P_{θ_0} representing the probability distribution underlying the population generation of the data, P_ν representing the distribution of the sample design (sample inclusion indicators), and P_{θ_0}, P_ν representing their joint distribution.

The three main theorems in [Williams and Savitsky \(2021\)](#) are summarized below followed by a summary of the conditions needed. Theorem 1 establishes that the (pseudo-)MLE estimator (scaled by the population size) is a Gaussian process plus an error term that goes to zero. Theorem 2 establishes that the covariance of this Gaussian process has the sandwich form $H_{\theta_0}^{-1} J_{\theta_0}^\pi H_{\theta_0}^{-1}$. In contrast, Theorem 3 demonstrates that the pseudo-posterior distribution $\Pi_{N_\nu}^\pi(\theta)$ converges to a Gaussian distribution centered at the MLE but with a different covariance matrix $H_{\theta_0}^{-1}$. This motivates the approach for a post-hoc adjustment to the pseudo-posterior such that the resulting distribution asymptotically has the same covariance as the (pseudo-)MLE. Demonstrating a one-to-one connection between asymptotic distributions of the (pseudo-)posterior and the (pseudo-)MLE is referred to as a ‘Bernstein-von Mises’ result.

*Corresponding author. Email: matthew.dunn.williams@gmail.com.

Theorem 1 (Asymptotic Normality of the Pseudo-MLE). *Suppose conditions A1-A7 hold. Then*

$$\sqrt{N_\nu} \left(\hat{\theta}_{\pi, N_\nu} - \theta_0 \right) = -H_{\theta_0}^{-1} \mathbb{G}_{N_\nu}^\pi \dot{\ell}_{\theta_0} + \mathcal{O}_{P_{\theta_0}, P_\nu}(1).$$

Theorem 2 (Asymptotic Variance of the Pseudo-MLE). *Suppose conditions A1-A5 hold. Then*

$$\text{Var}_{P_{\theta_0}, P_\nu} \{ -H_{\theta_0}^{-1} \sqrt{N_\nu} \mathbb{P}_{N_\nu}^\pi \dot{\ell}_{\theta_0} \} = H_{\theta_0}^{-1} J_{\theta_0}^\pi H_{\theta_0}^{-1}.$$

Theorem 3 (Asymptotic Distribution of the Pseudo-Posterior). *Suppose conditions A1-A7 hold. Then*

$$\sup_{B \in \Theta} \left| \Pi_{N_\nu}^\pi(\theta \in B \mid \mathbf{X}_\nu \boldsymbol{\delta}_\nu) - \mathcal{N}_{\hat{\theta}_{\pi, N_\nu}, N_\nu^{-1} H_{\theta_0}^{-1}}(B) \right| \xrightarrow{P_{\theta_0}, P_\nu} 0.$$

A high-level description of the conditions needed for these theorems are below. More details are available in [Williams and Savitsky \(2021\)](#).

(A1) (Continuity): The log-likelihood is continuous in the data space X and differentiable and Lipschitz continuous in the parameter space θ .

(A2) (Local Quadratic Expansion): The Kullback-Liebler divergence $\left(\mathbb{E}_{P_{\theta_0}} \log \frac{p_\theta}{p_{\theta_0}} \right)$ has a second order Taylor expansion about θ_0 .

(A3) (Bartlett's First Identity): $\mathbb{E}_{P_{\theta_0}} \dot{\ell}_{\theta_0} = 0$.

(A4) (Consistency of the MLE for the population): $\hat{\theta}_{N_\nu} \rightarrow \theta_0$.

(A5) (Non-zero Inclusion Probabilities): The minium probability of selection for any individual is bounded away from zero (and thus its inverse probability weight is finite and bounded from above).

(A6) (Growth of Dependence is Restricted) Most pairs of individuals must be selected independently (dependence shrinks asymptotically). The number of pairs which can have arbitrary joint selection cannot grow too fast (order N_ν instead of N_ν^2).

(A7) (Constant Sampling Fraction). Asymptotically the sample size n_ν becomes a constant fraction of the population size N_ν .

2 Additional Analyses for the Simulation Studies

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.

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

In the main text, our simulation studies generated samples from different designs including a simple random sample (SRS) and an informative sample with unequal weights from probability proportional to size (PPS) design. As our focus is on better uncertainty quantification, we evaluated the frequentist coverage of model-based intervals over repeated sampling (100 samples). To complement these results, we include comparisons of the empirical sampling distribution of the posterior means of each parameter over the 100 samples. This provides an empirical approximation of the ‘true’ sampling distributions. We contrast these with the full (pseudo-) posterior distributions from each of a subsample of of the full set of 100 samples. Qualitatively we can compare the distributions to get a sense of similarity of shape and scale.

2.1 Simple Random Sample

Figure 1 compares these two sets of distributions under an SRS, with the unadjusted pseudo-posterior being the exact posterior in this case. As expected, we see variability across the posterior samples from different survey samples. However the general shape and scale is in agreement with the empirical distribution of the posterior mean. We note that for some samples, the agreement is much closer than others, reflecting the variability across samples. The overall alignment in shape is consistent with the coverage assessment in the main text in which coverage for the SRS sample is close to nominal (Table 1). We contrast this with our adjusted posterior. Figure 2 shows reasonable agreement for the fixed effects (b0 and b1) but increased variability for the random effects. The random effects variance shows less variability within each sample posterior. These results provide additional insights into the slight over-coverage for the random effects and the notable under-coverage of the random effects variance observed in the main text (Table 1). Ideally, an adjustment would perform no worse than no-adjustment when it was not needed (e.g. for an SRS). This suggests room for improvement for our proposed adjustment.

Adjustment	Interval Length			Interval Coverage %		
	Fixed	Random	σ_α	Fixed	Random	σ_α
Unadjusted	0.499	1.277	0.622	93.0	96.5	96.0
Yeo-Johnson	0.566	1.217	0.377	97.0	95.6	78.0

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 two alternative variance approaches.

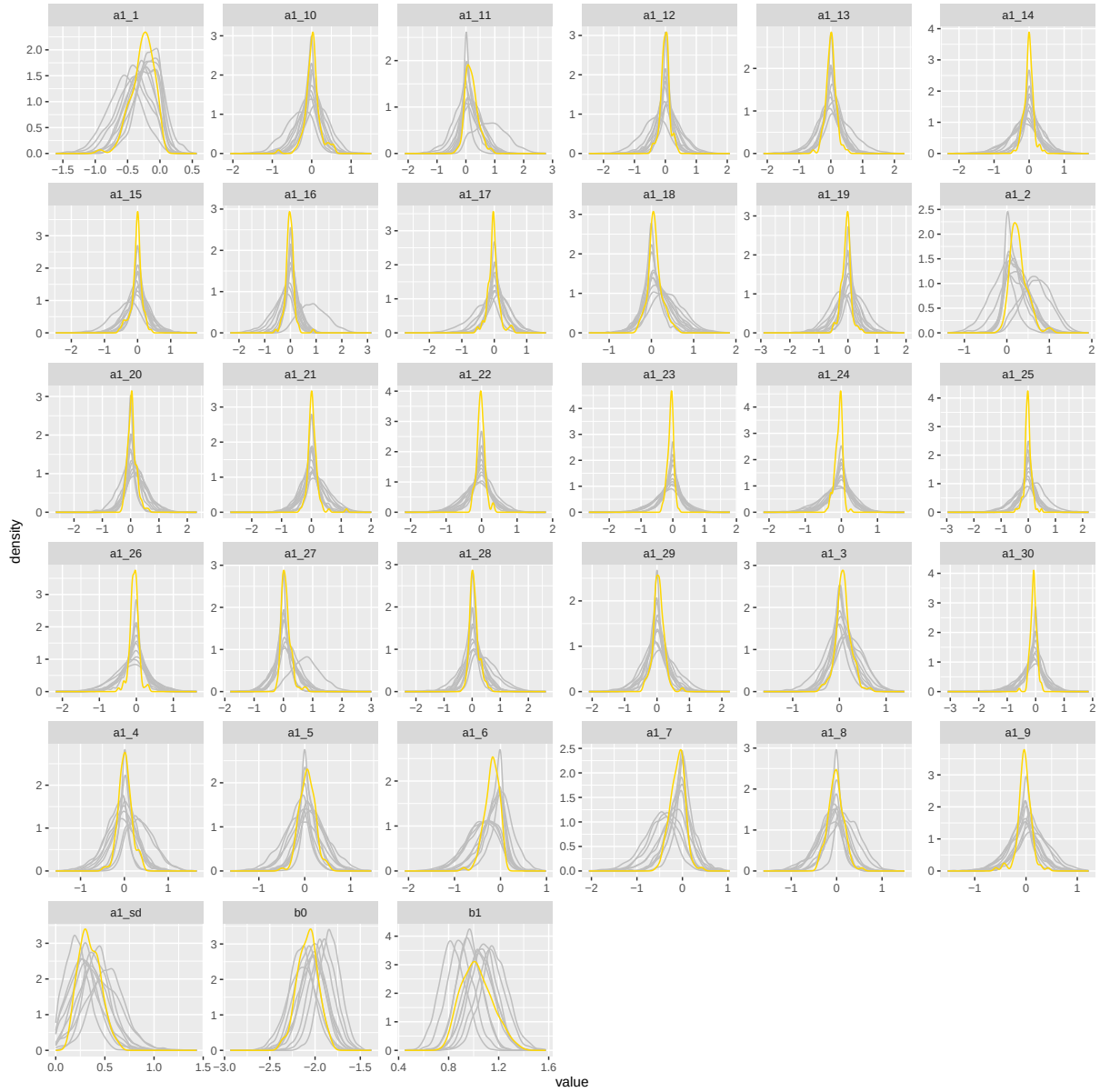


Figure 1: Comparison of the empirical sampling distribution (gold) of the posterior mean from 100 SRS samples with the unadjusted posterior distributions (grey) for each of 10 SRS samples (subset from 100 for ease of display). Group level random effects $a1_1$ through $a1_{30}$ with variance parameter $a1_{sd}$, intercept $b0$ and slope $b1$.

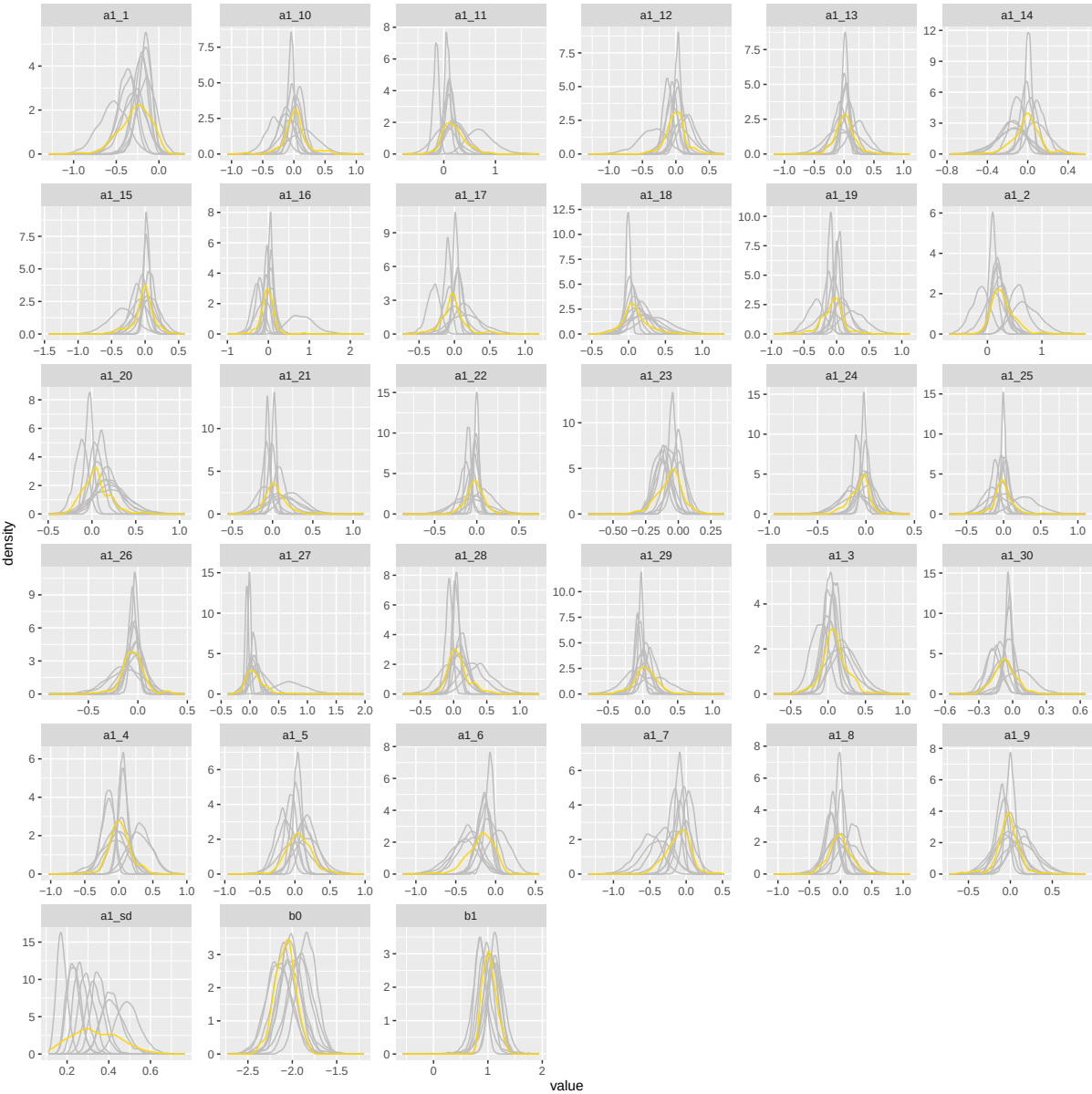


Figure 2: Comparison of the empirical sampling distribution (gold) of the posterior mean from 100 SRS samples with the adjusted posterior distributions (grey) for each of 10 SRS samples (subset from 100 for ease of display). Group level random effects a1_1 through a1_30 with variance parameter a1_sd, intercept b0 and slope b1.

2.2 Probability Proportional to Size Sample

For the PPS sample, Figure 3 shows variability for the random effect variance and the fixed effects (slope and intercept). The variance parameter distributions are too narrow and more variable across samples. The adjusted samples (Figure 4) mitigate some of this variability but also have similar differences between the posterior distribution of the samples and the sampled distribution of the posterior mean. Table 2 provides a summary of coverage that is consistent with the distribution plots.

Adjustment	Interval Length			Interval Coverage %		
	Fixed	Random	σ_α	Fixed	Random	σ_α
Unadjusted	0.531	1.239	0.557	64.5	86.8	89.0
Yeo-Johnson	0.658	1.677	0.766	72.0	88.9	90.0

Table 2: Comparison of the empirical sampling distribution (gold) of the posterior mean from 100 PPS samples with the adjusted pseudo-posterior distributions (grey) for each of 10 SRS samples (subset from 100 for ease of display). Group level random effects a1_1 through a1_30 with variance parameter a1_sd, intercept b0 and slope b1.

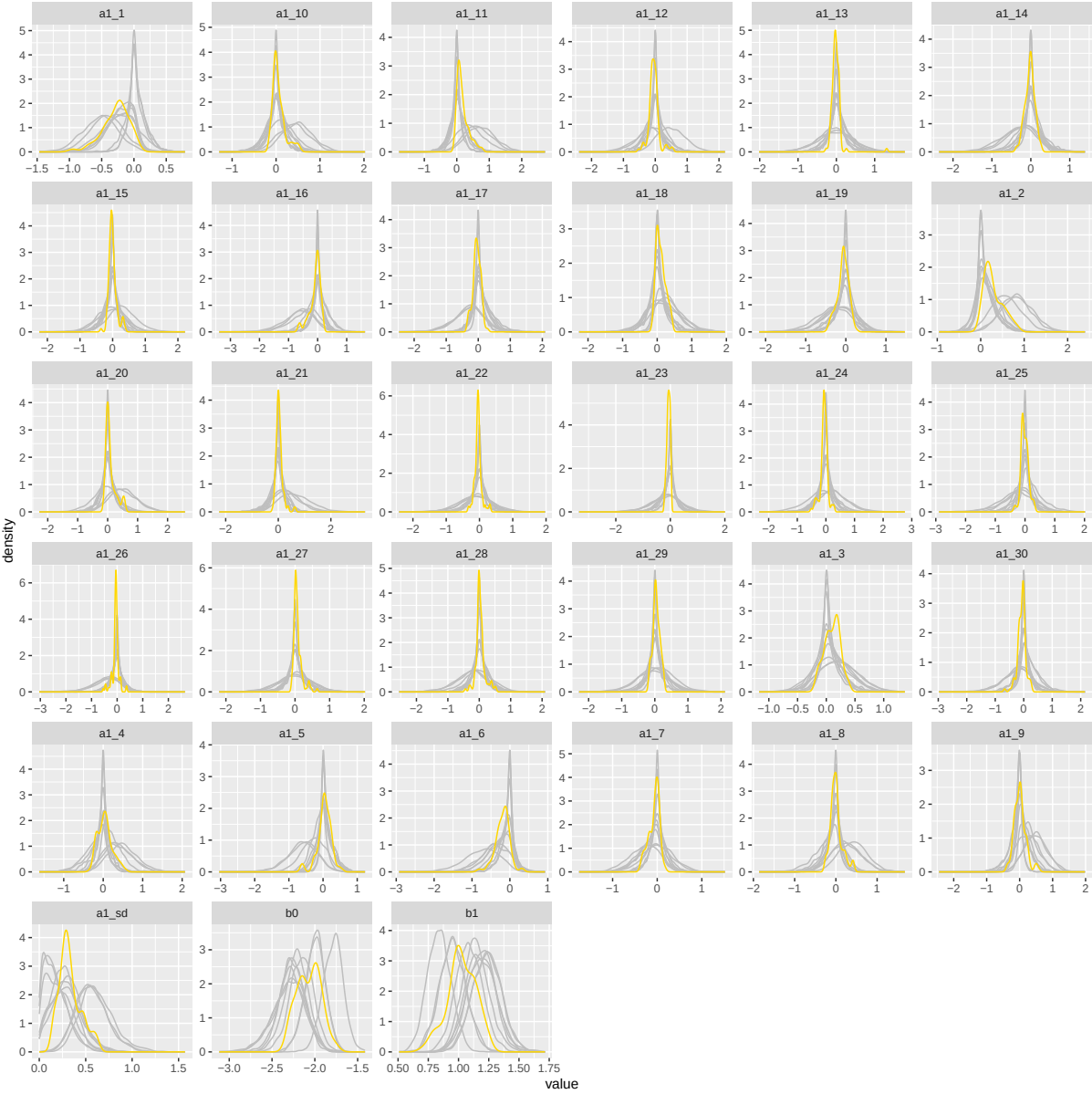


Figure 3: Enter Caption

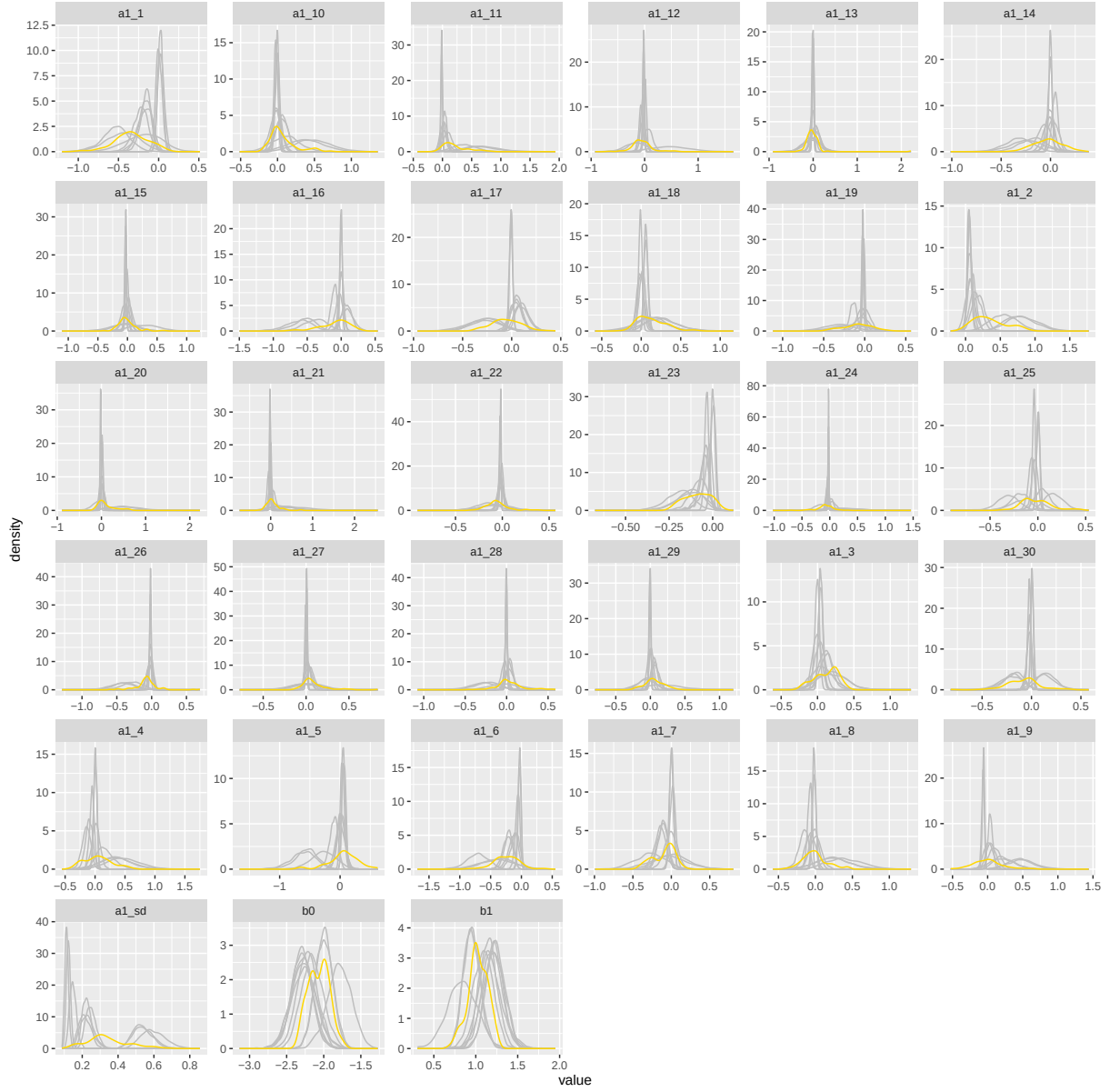


Figure 4: Comparison of the empirical sampling distribution (gold) of the posterior mean from 100 PPS samples with the unadjusted pseudo-posterior distributions (grey) for each of 10 SRS samples (subset from 100 for ease of display). Group level random effects $a1_1$ through $a1_{30}$ with variance parameter $a1_sd$, intercept $b0$ and slope $b1$.

3 NSDUH subsample comparisons

Using just a subsample of 14,000 from the combined NSDUH data, we compare the unadjusted estimates (direct from pseudo-posterior) with those from the Yeo-Johnson adjustment approach and with estimates obtained by fitting a new Stan model for each of the 100 replicates. The latter is roughly 100 times slower but provides a reasonable alternative comparison. Figure 5 compares the three alternatives in terms of their 95% intervals for global parameters (fixed effects and random effect variance). The unadjusted posterior consistently has narrower intervals than either the YJ adjusted or the naive replication approach. This is to be expected, because a ‘design effect’ reduces the effective sample size due to clustered and unequal sampling. Of note, the credible interval for the random effect variance (Level 2) is significantly narrower for the unadjusted method, while the two alternative approaches show similar results. We note that for the fixed effect for sex (women), the YJ adjusted interval is notably smaller than the replication interval but still slightly larger than the unadjusted interval.

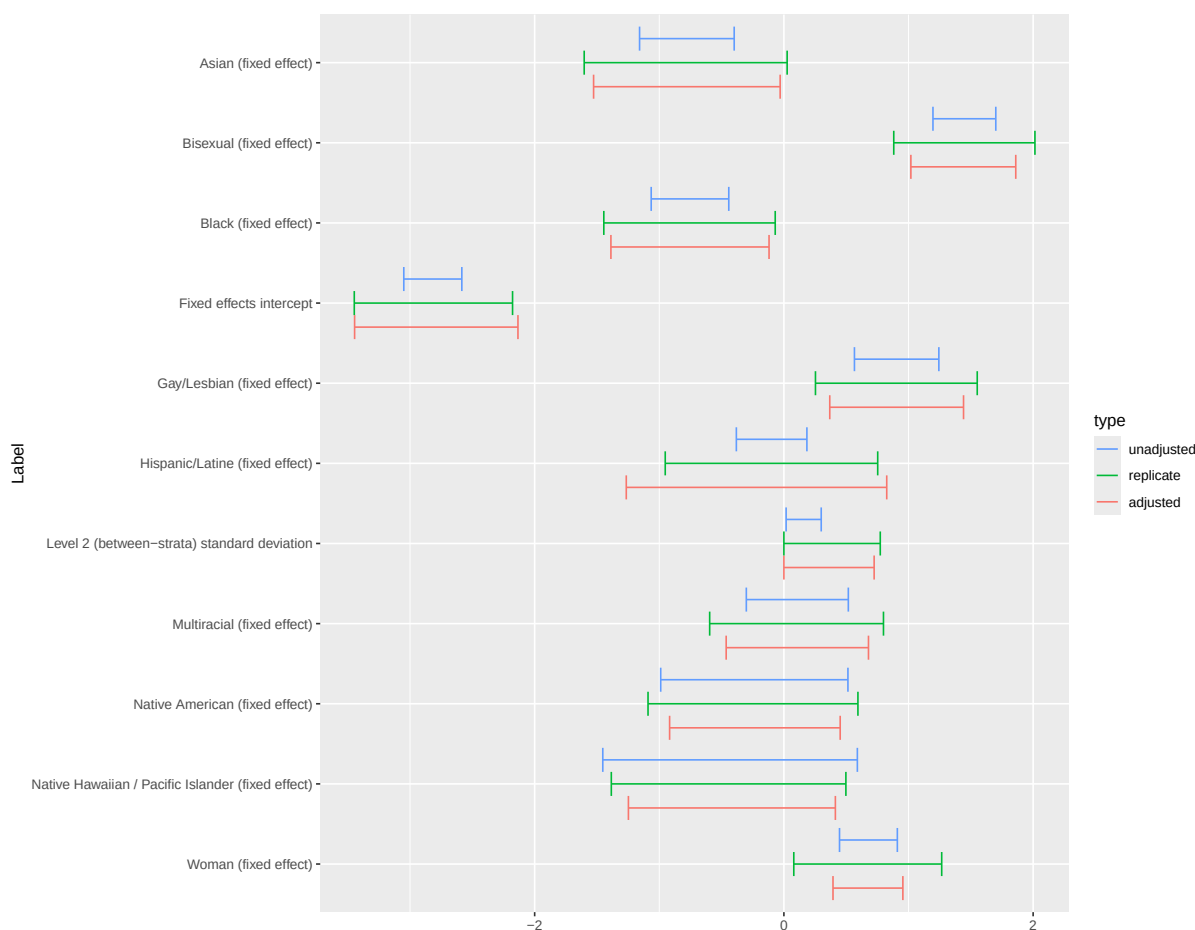


Figure 5: Comparison of 95% intervals for global regression parameters for the multi-level logistic regression for past year depression from NSDUH.

We next compare the three approaches for local parameters or random effects (Figure 6). We see that the unadjusted intervals are the narrowest, with the naive replication approach

leading to wider intervals. The YJ intervals are consistently wider than those for the replication approach. This is in contrast to the global parameters in which both intervals behaved similarly, including for the random effects (Level 2) variance. The wider intervals from the YJ adjustment translate into wider intervals for the domain estimates. Figure 7 compares the three model-based alternatives to the direct estimates.

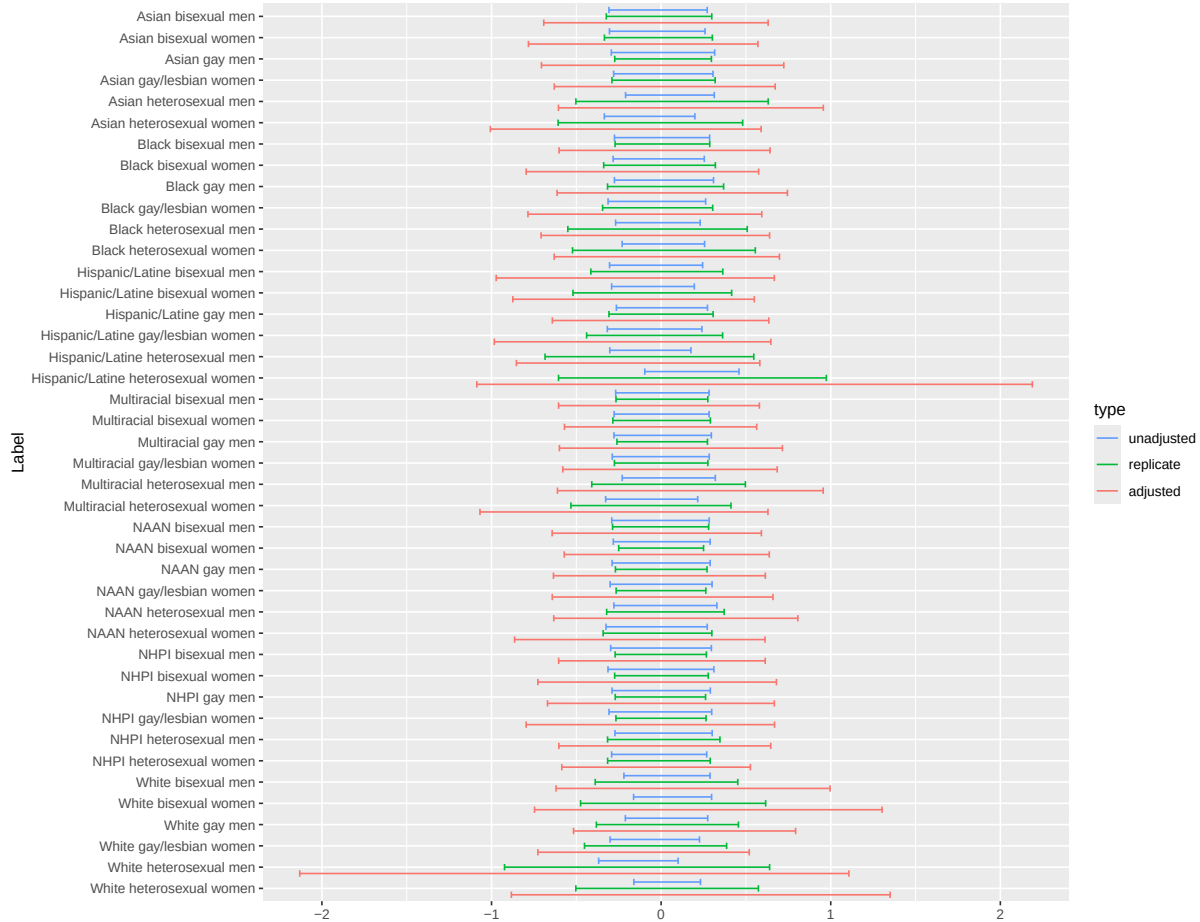


Figure 6: Comparison of 95% intervals for local random effects for the multi-level logistic regression for past year depression from NSDUH.

In contrast to the reasonable performance of the YJ intervals for the full sample, the instability of the YJ intervals for the subsample may be due to the extreme weights from the down-sampling of the larger groups. This indicates that the YJ adjustment for random effects may be sensitive to the scaling and stability of weights.

4 Global-Local Priors for Interactions

Our motivating hierarchical model (1) is relatively simple or ‘shallow’. Instead of placing a simple prior on the group level random effects α_j defined by cross-classification of three variables we could instead follow Si et al. (2020) and create a hierarchical prior structure that penalizes main effects as well as second-order and third-order interaction terms. These priors are linked, with

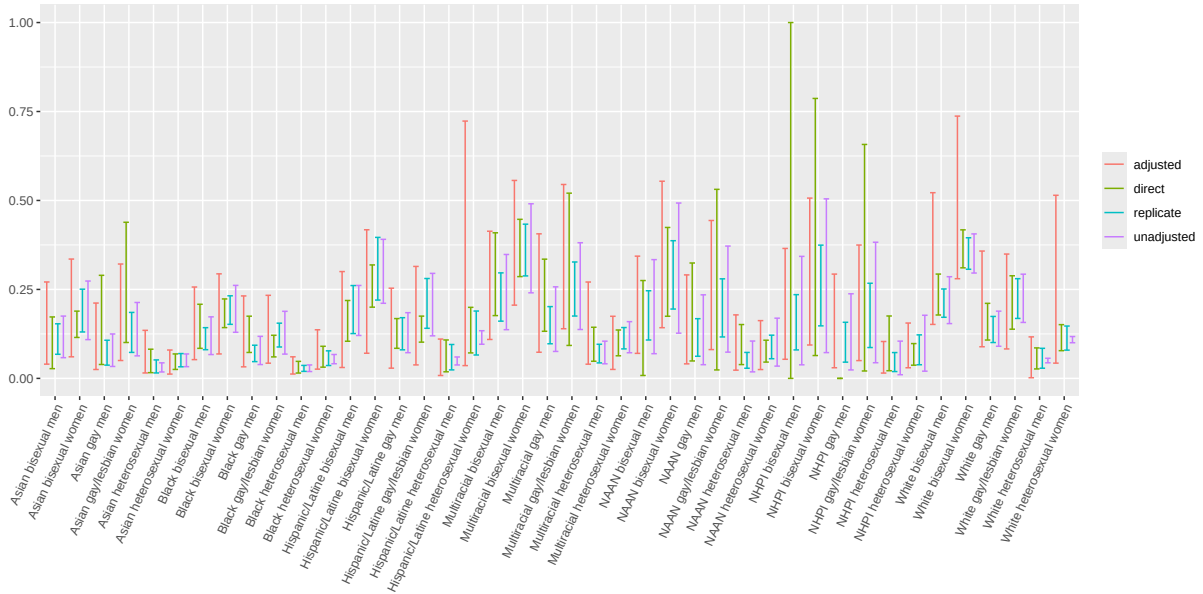


Figure 7: Comparison of 95% intervals for strata/group level estimates for past year depression from NSDUH.

second-order and third-order priors being amplified by priors from corresponding main effects. The resulting structure induces sparsity, particularly for higher-level interaction terms.

$$\begin{aligned}
 y_{ij} &\sim \text{Bern}(p_{ij}), \\
 \text{logit}(p_{ij}) &= \mu_{ij} = X_{ij}\beta + \alpha_j, \\
 \beta &\sim N(0, \Sigma), \\
 \alpha_j &= \sum_{k \in S^{(1)}} \alpha_{j,k}^{(1)} + \sum_{k \in S^{(2)}} \alpha_{j,k}^{(2)} + \sum_{k \in S^{(3)}} \alpha_{j,k}^{(3)},
 \end{aligned} \tag{2}$$

where $S^{(\ell)}$ is all ℓ -way interactions and $\alpha_{j,k}^{(\ell)}$ is the k^{th} interaction term of order ℓ that corresponds to group j . Following Si et al. (2020), the prior structure is

$$\begin{aligned}
 \alpha_{j,k}^{(\ell)} &\sim N\left(0, (\lambda_k^{(\ell)} \sigma)^2\right), \\
 \sigma &\sim \text{Cauchy}_+(0, 1), \\
 \lambda_k^{(\ell)} &= \delta^{(\ell)} \prod_{\nu \in M^{(k)}} \lambda_{\nu}^{(\ell)}, \\
 \lambda_k^{(1)} &\sim N_+(0, 1), \\
 \delta^{(\ell)} &\sim N_+(0, 1),
 \end{aligned} \tag{3}$$

where σ is a ‘global’ scale and each $\lambda_k^{(\ell)}$ is a local scale adjustment, such that higher order local scales depend on the scales from corresponding main effect terms from the set $M^{(k)}$. Lastly, $\delta^{(\ell)}$ is an additional scaling specific to the order of the interaction term.

For demonstration, we estimate this more complex hierarchical model on the NSDUH stratified subsample of 14,000. Figure 8 shows (pseudo-) posterior mean and 95% intervals for the random effects, color-coded by order (main, second, third). We see that the structured prior pulls most higher order random effects towards zero. However some second-level and third-level interactions are still noticeable. Comparing the adjusted (pseudo-) posterior estimates for domains

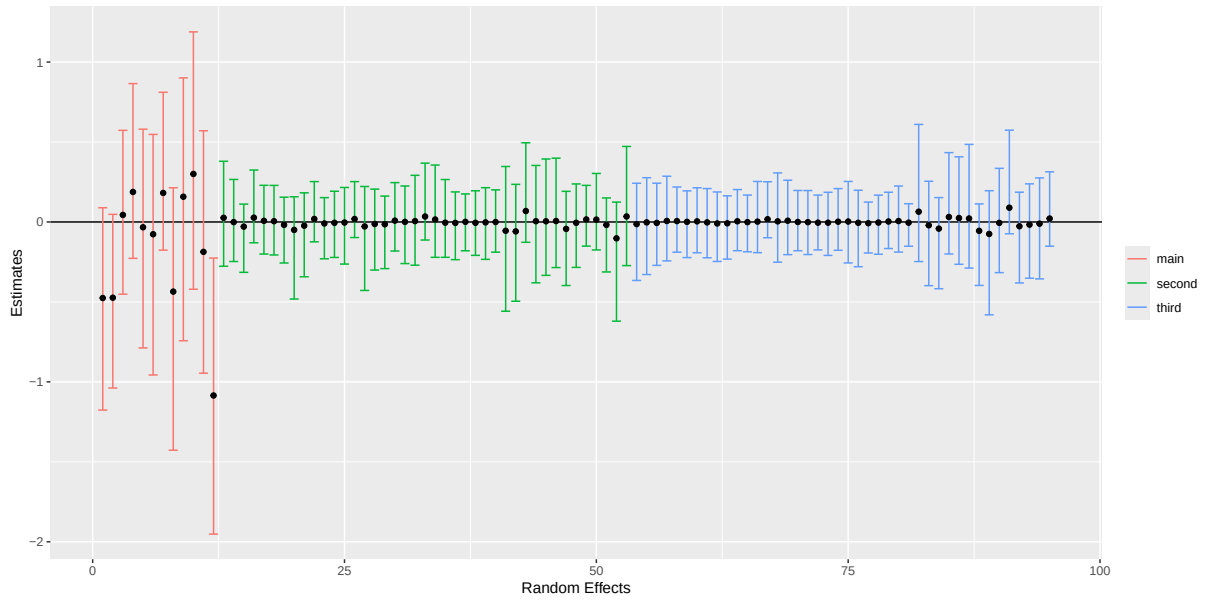


Figure 8: (Pseudo-)Posterior distribution (mean and 95% interval) of random effects from the hierarchical global-local model estimated on a stratified subsample from NSDUH.

between our simpler model (1) and the more complex global-local (2), we see qualitatively very similar estimates for domains (Figure 9). From Appendix 3, we know that the YJ adjustment is unstable for the extreme subsampling for this NSDUH subset. The adjusted version of the global-local model is extremely unstable, leading to unusable results (Figure 10). This suggests a real limit to the ‘automatic’ adjustment for deep hierarchical models. However, future work could investigate a tiered approach in which some hierarchical parameters are excluded from the adjustment.

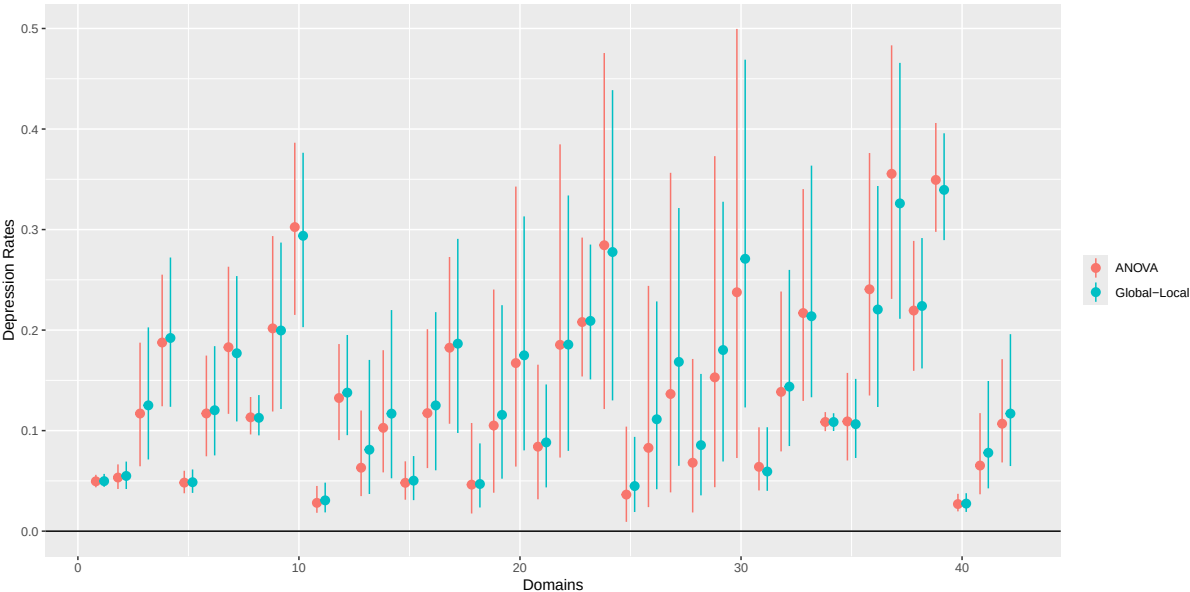


Figure 9: Comparison of domain level estimates of depression (mean and 95% interval) from the *unadjusted* survey-weighted ANOVA model and the hierarchical global-local model for a stratified subsample from NSDUH.

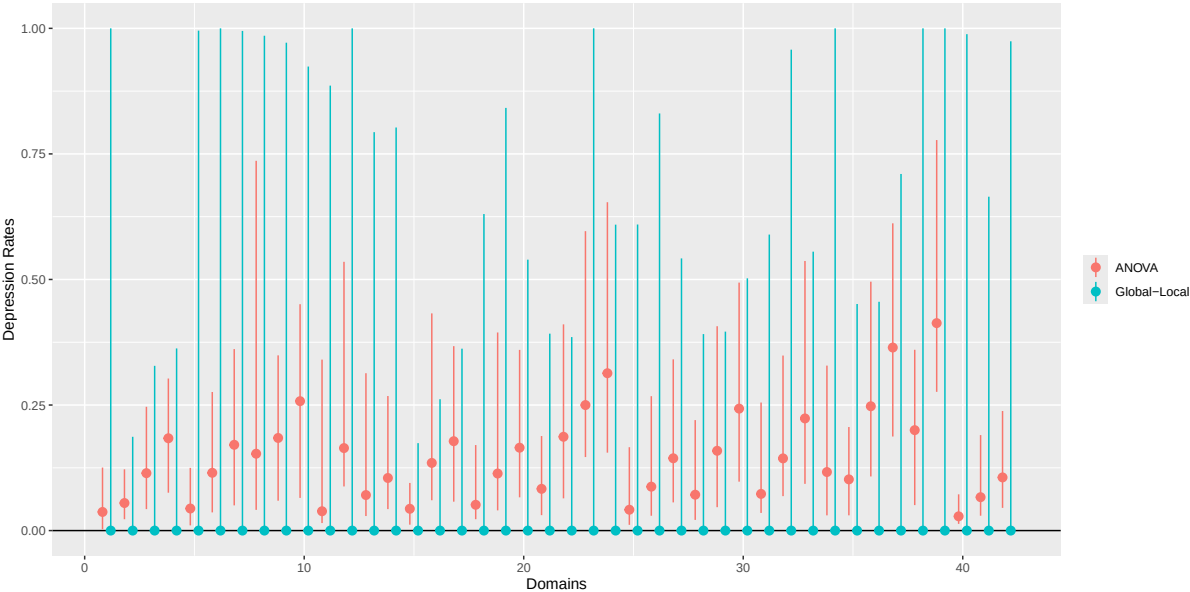


Figure 10: Comparison of domain level estimates of depression (mean and 95% interval) from the *adjusted* survey-weighted ANOVA model and the hierarchical global-local model for a stratified subsample from NSDUH.

References

- Ghosal S, Ghosh JK, Vaart AWVD (2000). Convergence rates of posterior distributions. *Ann. Statist.*, 500–531.
- Si Y, Trangucci R, Gabry JS, Gelman A (2020). Bayesian hierarchical weighting adjustment and survey inference. *Survey Methodology*, 46(2): 181–215.
- Williams MR, Savitsky TD (2021). Uncertainty estimation for pseudo-bayesian inference under complex sampling. *International Statistical Review*, 89(1): 72–107.