Simulation with first stage variability

Austin Schumacher

2022-01-31

Overview

We will perform a simulation study to explore the recovery of parameters. This simulation will include the variability from the first-stage model in order to validate the full model (rather than taking the first stage results as fixed and just validating the second stage model). Instead of simulating individual-level data and calculating first-stage direct estimates and fitting a second stage smoothing model to those, we will set known means and variances for each region and then draw area-level sample means from their sampling distribution, and additionally draw area-level sample variances their sampling distribution. The known mean parameters will be parameterized by smoothing models that we will fit in the second stage. Pseudo-direct estimates will be these area-level sample means with associated variance equal to the these sample variances. For each region, we will assess the bias of the sample means and coverage of uncertainty intervals created with the sample variances. We will then fit our second-stage smoothing models on these sample means and sample variances and assess the bias and coverage of the estimated latent means in each region.

1-dimensional data

We will start with the simplest case of one-dimensional data.

IID latent mean model

Before we try simulating data with both IID and ICAR components, we will first look at the IID-only case.

Data generating mechanism

Let $r = 1, \dots, R$ index region. The region-level means will be generated as

$$\begin{aligned} \mu_r &= \beta_1 + v_r \\ v_r &\overset{iid}{\sim} N(0, \sigma_v^2). \end{aligned}$$

with β_1 set to be the estimated intercept and σ_v^2 set to be the estimated variance of the random effects from a model fit to the 2014 KDHS HAZ data. The random effects will only be simulated once and the same ones will be used across all simulations.

Once we have these μ_r values, we will simulate the area-level sample means as

$$\hat{y}_r | \mu_r, V_r \sim N(\mu_r, V_r)$$

where the V_r are set to be equal to the asymptotic design-based variance estimate, $\hat{V}_r^{des}$, of the area-level mean HAZ from the 2014 KDHS data.

Now, for an SRS, the sampling distribution of the area-level sample variances is

$$\hat{V}_r^{srs}|V_r, n_r \sim \Gamma\left(\frac{n_r - 1}{2}, \frac{n_r - 1}{2V_r}\right)$$

where n_r is the sample size of an SRS in area r. The 2014 KDHS uses a stratified cluster design rather than an SRS, which has a different sampling variance than an SRS. The ratio of the variance for a statistic calculated using a specific survey design to the variance of that statistic calculated using an SRS of the same sample size is called the design effect,

$$d^2 = \frac{\hat{V}}{\hat{V}^{srs}}.$$

In the a typical DHS, the average design effect across all indicators is $d^2 = 1.5^2 = 2.25$ (taken from this answer on the DHS user forum: https://userforum.dhsprogram.com/index.php?t=msg&goto=3448&S=Google). Thus, for our simulation we will let n_r be the number of sampled children with HAZ scores in each region, and we will calculate the effective sample size that would be needed from an SRS in order to have the same sampling variance as the DHS sampling design.

Since the variance of a $\Gamma(\alpha, \beta)$ distribution is α/β^2, we have

$$\begin{aligned}\text{Var}(\hat{V}_r^{srs}) &= \frac{\left(\frac{n_r-1}{2}\right)}{\left(\frac{n_r-1}{2V_r}\right)^2} \\ &= \frac{2(V_r)^2}{(n_r - 1)}.\end{aligned}$$

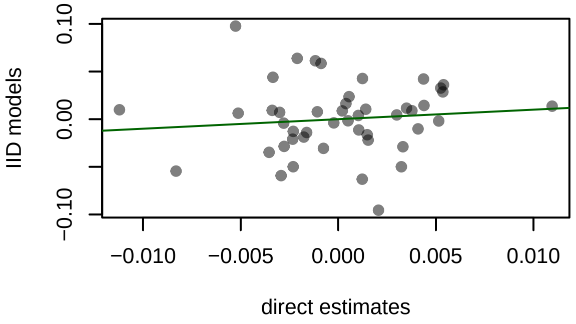
Now, we want $\hat{V}_r = 2.25 \times \hat{V}_r^{srs} = 2.25 \times \frac{2(V_r)^2}{(n_r-1)} = \frac{2(V_r)^2}{((n_r-1)/2.25)} \approx \frac{2(V_r)^2}{((n_r/2.25-1))}$, which yields an effective sample size of $n_r^* = n_r/2.25$ (as long as n_r is sufficiently large).

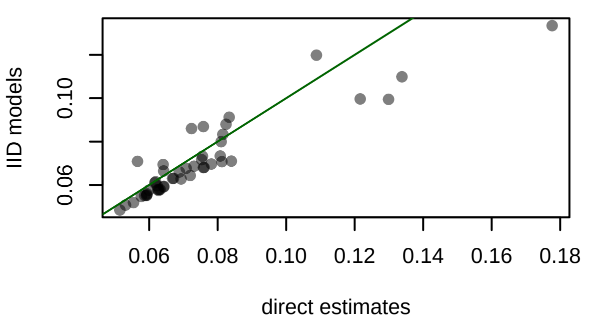
Thus, we will simulate the area-level sample variances as

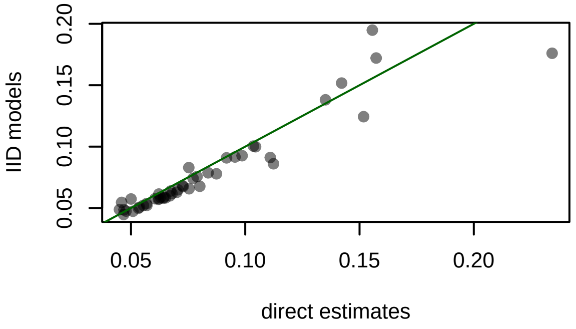
$$\hat{V}_r|V_r, n_r^* \sim \Gamma\left(\frac{(n_r^* - 1)}{2}, \frac{n_r^* - 1}{2V_r}\right).$$

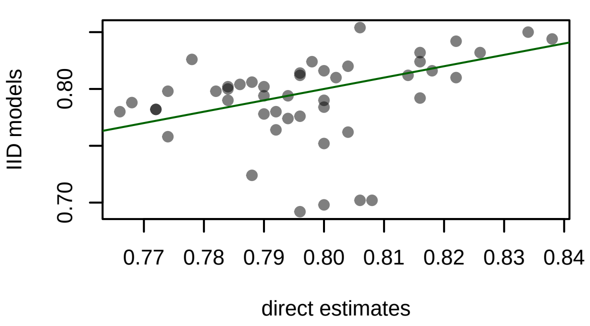
In each simulation, we will have the same values of μ_r and V_r, and we use these to simulate $\hat{y}_r$ and $\hat{V}_r$. We will treat the $\hat{y}_r$ as pseudo-direct estimates and calculate asymptotic 95% confidence intervals as $\hat{y}_r \pm t_{n_r^*-1,0.975}\sqrt{\hat{V}_r}$. We will also fit an IID smoothing model to the pseudo-direct estimates to estimate the latent means in each region with corresponding 95% credible intervals. This smoothing model will be correctly specified to match the data generating mechanism and be fit via a Bayesian model with relatively uninformative priors.

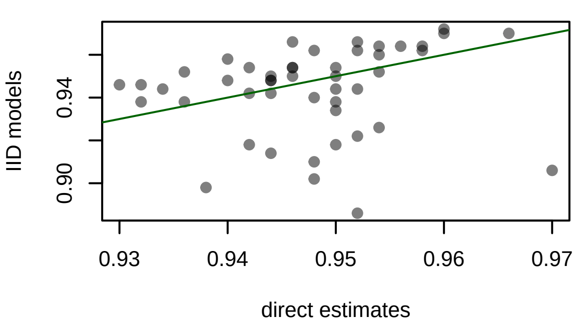
Results

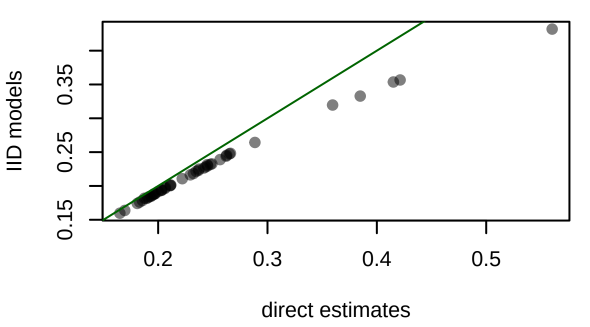

mean bias


mean absolute bias


mean relative bias


80% coverage


95% coverage


80% width

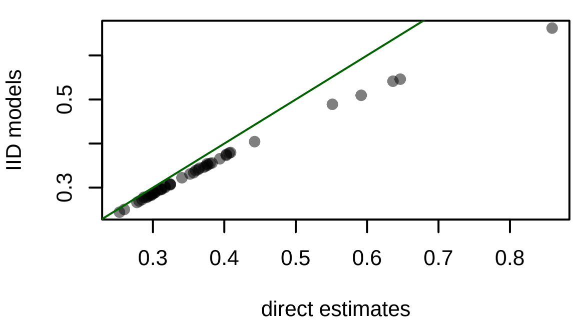

95% width

Table 1: Simulation results for 500 simulations, univariate DGM

measure	Direct	IID model
mean bias	0.0001	0.0007
mean absolute bias	0.0751	0.0707
mean relative bias	0.0820	0.0778
80% coverage	0.7974	0.7912
95% coverage	0.9477	0.9436
80% width	0.2400	0.2224
95% width	0.3683	0.3402