Measurement Error

Susanna Makela

August 5, 2015

Why Measurement Error?

This project was inspired by the working paper of Geruso and Spears (2015), in which the authors investigate the effects of household- and community-level open defecation (OD) on child mortality. Using data from the three rounds of the National Family Health Survey (NFHS) in India, the authors define community-level OD as the proportion of households in a primary sampling unit (PSU) practicing OD. The NFHS is a two-stage survey, sampling PSUs in the first stage and households in selected PSUs in the second. PSUs are defined as villages in rural areas and census enumeration blocks in urban areas. In both cases, PSUs consist of approximately 100-200 households, with 15-60 households sampled per PSU. Estimates of community-level OD rates, then, are based on the sampled households and may not be accurate estimates of the true OD rate. As Geruso and Spears (2015) note, this measurement error may affect their estimates of the community-level OD effect.

PSU-level estimates of risk factors and health determinants are particularly subject to measurement error when they are measured for a subset of the population that may not exist in each sampled household. For example, while toilet facility information can be collected for each sampled household, whether an infant is being exclusively breastfed can only be measured for households with women who had a recent live birth. The number of such households may be considerably less than the total number of sampled households, rendering the resulting estimates even less reliable.

Measurement error is well known to affect the coefficient estimates of imprecisely measured covariates. However, it can also affect coefficient estimates for perfectly measured covariates that are correlated with the imprecisely measured one (Gustafson, 2003). In addition, when a regression includes multiple mismeasured covariates, coefficient estimates are not always attenuated. Depending on the magnitude of the measurement errors and the correlation between the covariates, the estimated coefficients may be attenuated or inflated compared to the truth.

In this project, we first do a simulation study to understand the effects of measurement error in a situation analogous to that in Geruso and Spears (2015). In particular, we consider a multilevel model with a continuous outcome, a binary individual-level predictor, and two continuous PSU-level predictors, one of which is the PSU-level mean of the individual-level predictor. We are interested in the effect of measurement error in the PSU-level predictor on the coefficients for that predictor and for the individual-level predictor.

Simulation Study

The population and the sample

Consider a population consisting of J primary sampling units (PSUs). Each of the $j = 1, \ldots, J$ PSUs consists of N_j individuals in a particular target demographic group, with a total population size of $N = \sum_{j=1}^{J} N_j$. A survey is conducted that samples n_I PSUs and n_j individuals in each PSU. Let S_j denote the set of all individuals in the target demographic group in PSU j and let s_j denote the set of *sampled* individuals in PSU j so that $|S_j| = N_j$ and $|s_j| = n_j$.

The survey collects information on $n = \sum_{j=1}^{n_I} n_j$ individuals. Specifically, the survey data consist of an individual-level disease status outcome Y_i and a binary individual-level covariate Z_i. The binary covariate Z_i is the presence/absence of a risk factor for individual i and is assumed to be drawn from a Bernoulli distribution with parameter $\rho_{j[i]} \in [0, 1]$ (the notation $j[i]$ denotes the area j to which individual i belongs):

$$Z_i | \rho_{j[i]} \overset{ind}{\sim} \text{Bern}(\rho_{j[i]}).$$

Thus, ρ_j is the latent prevalence of or propensity for the risk factor Z in PSU j. We assume that it has a normal distribution (on the logit scale):

$$\text{logit}(\rho_j) \sim N(\mu, \tau^2).$$

In contrast, the true unobserved prevalence of Z in PSU j is

$$p_j = \frac{1}{N_j} \sum_{i \in S_j} Z_i,$$

from which it follows that the true underlying number of individuals in the PSU with the risk factor, $T_j = \sum_{i \in S_j} Z_i$, is distributed as

$$T_j | \rho_j \sim \text{Bin}(N_j, \rho_j).$$

Under this data generating mechanism, people moving in and out of area j will change the unobserved finite population prevalence p_j (since they will cause N_j to change), but they won't affect the superpopulation prevalence/propensity ρ_j. This distinction between the latent propensity ρ_j for the risk factor and the underlying prevalence p_j is useful because in reality, we know that it is the Z_i's that determine p_j, not the other way around.

Because we only sample $n_j < N_j$ individuals in each PSU, we only observe the imperfectly measured prevalence p_j^* instead of the true prevalence p_j, where

$$p_j^* = \frac{1}{n_j} \sum_{i \in s_j} Z_i.$$

The observed number of individuals with the risk factor is $T_j^* = \sum_{i \in s_j} Z_i$. The distribution of T_j^* is hypergeometric with parameters n_j, T_j, and N_j: we have a PSU with N_j individuals, $T_j = \sum_{i \in S_j} Z_i$ of whom have the risk factor of interest, and in a sample of n_j individuals *without replacement*, we want to know the number T_j^* of them with the risk factor.

For now, we assume that the outcome Y_i is continuous and normally distributed. Specifically, we

assume that

$$
\begin{aligned}
Y_i|\alpha_{j[i]}, Z_i &\sim N(\alpha_{j[i]} + \beta Z_i, \sigma_y^2) \\
\alpha_j|p_j, N_j &\sim N(\gamma_0 + \gamma_1 p_j + \gamma_2 \log(N_j), \sigma_\alpha^2),
\end{aligned}
$$

Note that the (log of the) PSU population sizes N_j are assumed to be predictive of the outcome Y_i. Here this is a slight simplification of the scenario where population size is correlated with a variable that is itself predictive of the outcome. In many health applications, this is a realistic assumption: village size may be correlated with health determinants such as access to major roads, quality of local health facilities, or geography (e.g. less malaria-prone highlands or more fertile agricultural areas).

This model assumes that it is the true finite population prevalence p_j rather than ρ_j that drives the variation in PSU-specific intercepts. In other words, the average value of Y_i in PSU j among individuals without the risk factor is $\mathbb{E}[Y_i|Z_i = 0] = \gamma_0 + \gamma_1 p_j + \gamma_2 \log(N_j) = \gamma_0 + \gamma_1 \frac{1}{N_j} \sum_{i \in S_j} Z_i + \gamma_2 \log(N_j)$.

However, we do not observe p_j and can only use the imperfect surrogate p_j^*. In epidemiology, measurement error models are often broken down into three submodels (Richardson and Gilks, 1993). The first is a disease model that describes the relationship between the outcome or disease status Y and the true risk factor p, and possibly other accurately measured risk factors. Next is a measurement model that relates the true risk factor p to the mismeasured surrogate p^*, and last is the exposure model that describes the distribution of the true risk factor p in the population.

In these submodels, the risk factor p and the disease status Y are both measured at the individual level. In particular, the measurement error applies to an individual-level risk factor. In our case, however, the individual-level risk factor Z_i is measured accurately, but the PSU-level prevalence p_j is not because we only observe Z_i for n_j out of N_j individuals in each PSU. In our scenario, the disease, measurement, and exposure models are as follows:

$$
\begin{aligned}
\text{disease model:} \quad & Y_i|Z_i, \alpha_{j[i]} \sim N(\alpha_{j[i]} + \beta Z_i, \sigma_y^2) && (1) \\
& \alpha_j|p_j, N_j \sim N(\gamma_0 + \gamma_1 p_j + \gamma_2 \log(N_j), \sigma_\alpha^2) \\
& Z_i|\rho_{j[i]} \sim \text{Bern}(\rho_{j[i]}) \\
\text{measurement model:} \quad & T_j^*|n_j, T_j, N_j \sim \text{Hypergeom}(n_j, T_j, N_j) && (2) \\
& T_j|\rho_j \sim \text{Bin}(N_j, \rho_j) \\
\text{exposure model:} \quad & \text{logit}(\rho_j) \sim N(\mu, \tau^2) && (3)
\end{aligned}
$$

Some simulations

Generating finite population data

Set parameters/hyperparameters:

$$
\begin{aligned}
J &= 5000 \quad & (N^{low}, N^{high}) &= (200, 500) \\
\mu &= 0 & \gamma_0 &= 1 \\
\beta &= 4 & \gamma_1 &= 2 \\
\tau &= 1 & \gamma_2 &= 1 \\
\sigma_y &= 1 & \sigma_a &= 0.5
\end{aligned}
$$

Generate population data:

1. Draw PSU population sizes N_j from $N_j \sim \text{Unif}(N^{low}, N^{high})$.
2. Draw the superpopulation prevalences ρ_j from $\text{logit}(\rho_j) \sim N(\mu, \tau^2)$.
3. Set individual-level risk factor to present/absent according to $Z_i|\rho_{j[i]} \sim \text{Bern}(\rho_j)$.
4. Calculate finite population prevalence $p_j = 1/N_j \sum_{i=1}^{N_j} Z_i$.
5. Draw PSU-specific intercepts from $\alpha_j|p_j \sim N(\gamma_0 + \gamma_1 p_j + \gamma_2 \log(N_j), \sigma_\alpha^2)$.
6. Draw individual-level outcome from $Y_i|\alpha_{j[i]}, Z_i \sim N(\alpha_{j[i]} + \beta Z_i, \sigma_y^2)$.

Sampling

We then sample from the population:

1. Sample n_I out of J PSUs with probability proportional to size (PPS).
2. From each sampled PSU, take a simple random sample (SRS) of n_j individuals.

We use $n_I \in \{10, 50, 100, 500\}$ and $n_j \in \{2, 20, 200\}$ for all j. In this way, the sample is self-weighting. (Of course, in reality we know that surveys like the DHS come with weights that aren't exactly 1 like you'd expect from a self-weighting survey because the weights also account for things like nonresponse. This is an interesting area of further research: given a dataset that comes with survey weights, how should we incorporate those weights into a measurement error context?)

Models

We fit two types of models to the data as shown in Table 1. Our simulations initially focus on the case where the N_j's are known.

	N_j unknown	N_j known
Naive model	$Y_i \sim N(\alpha_{j[i]} + \beta Z_i, \sigma_y^2)$	$Y_i \sim N(\alpha_{j[i]} + \beta Z_i, \sigma_y^2)$
	$\alpha_j \sim N(\gamma_0 + \gamma_1 p_j^*, \sigma_\alpha^2)$	$\alpha_j \sim N(\gamma_0 + \gamma_1 p_j^* + \gamma_2 \log(N_j), \sigma_\alpha^2)$
Full model	$Y_i \sim N(\alpha_{j[i]} + \beta Z_i, \sigma_y^2)$	$Y_i \sim N(\alpha_{j[i]} + \beta Z_i, \sigma_y^2)$
	$\alpha_j \sim N(\gamma_0 + \gamma_1 \rho_j, \sigma_\alpha^2)$	$\alpha_j \sim N(\gamma_0 + \gamma_1 \rho_j + \gamma_2 \log(N_j), \sigma_\alpha^2)$
	$Z_i \sim \text{Bern}(\rho_{j[i]})$	$Z_i \sim \text{Bern}(\rho_{j[i]})$

Table 1: Summary of models.

- **"Naive" model.** This model ignores measurement error and assumes that p_j^* is a good enough approximation for p_j.

  N_j's known.

$$
\begin{aligned}
Y_i|\alpha_{j[i]}, Z_i &\sim N(\alpha_{j[i]} + \beta Z_i, \sigma_y^2) \\
\alpha_j|p_j^* &\sim N(\gamma_0 + \gamma_1 p_j^* + \gamma_2 \log(N_j), \sigma_\alpha^2)
\end{aligned}
$$

N_j**'s unknown.**

$$Y_i|\alpha_{j[i]}, Z_i \sim N(\alpha_{j[i]} + \beta Z_i, \sigma_y^2)$$
$$\alpha_j|p_j^* \sim N(\gamma_0 + \gamma_1 p_j^* + \gamma_2 \log(N_j), \sigma_\alpha^2)$$

- **"Full" model.** This model accounts for measurement error. Here we make the simplification of using the superpopulation parameter ρ_j in place of the finite population parameter p_j in the model for α_j. This approximation is very good in the context of our simulation. By the Central Limit Theorem, $(p_j - \rho_j) \xrightarrow[n\to\infty]{d} N(0, \rho_j(1-\rho_j)/N_j)$. The values of N_j that we are using are 200 or greater, so the normal approximation is reasonable, and the variance of p_j around ρ_j is therefore at most $.5 * .5/200 = 0.00125$. This is equivalent to a standard deviation of $\sqrt{0.00125} = 0.035$. Assuming the normal approximation holds, we can then say that $\mathbb{P}(|p_j - \rho_j| < 0.07) = 0.95$ – that is, the observed p_j's are within 0.07 of ρ_j with 95% probability.

  N_j**'s known.**

$$Y_i|\alpha_{j[i]}, Z_i \sim N(\alpha_j + \beta Z_i, \sigma_y^2)$$
$$\alpha_j|\rho_j \sim N(\gamma_0 + \gamma_1 \rho_j + \gamma_2 \log(N_j), \sigma_\alpha^2)$$
$$Z_i|\rho_{j[i]} \sim \text{Bern}(\rho_{j[i]})$$

  N_j**'s unknown.**

$$Y_i|\alpha_{j[i]}, Z_i \sim N(\alpha_{j[i]} + \beta Z_i, \sigma_y^2)$$
$$\alpha_j|\rho_j \sim N(\gamma_0 + \gamma_1 \rho_j, \sigma_\alpha^2)$$
$$Z_i \sim Bern(\rho_{j[i]})$$

Replication

We generate a total of 20 populations using the hyperparameters and parameters specified above. For each population, we generate 100 samples and fit the models to all 100 samples. In this way, we obtain distributions of posterior means for the parameters of interest. These distributions are over 20 populations, so the peculiarlities of any particular population are averaged out.

Summary

Our simulation study thus has two goals. First, our interest is in the effects of both p_j and Z_i on individual disease status Y_i, so we want to understand the effect of ignoring measurement error on these coefficients (γ_1 and β, respectively). Second, we want to understand the performance of the naive and full models in the cases that the N_j's are known and unknown.

References

Michael Geruso and Dean Spears. Neighborhood sanitation and infant mortality. Technical report, research institute for compassionate economics, 2015. URL `http://ssrn.com/abstract=2605479`.

Paul Gustafson. *Measurement error and misclassification in statistics and epidemiology: impacts and Bayesian adjustments*. CRC Press, 2003.

Sylvia Richardson and Walter R. Gilks. Conditional independence models for epidemiological studies with covariate measurement error. *Statistics in Medicine*, 12(18):1703–1722, 1993. ISSN 1097-0258. doi: 10.1002/sim.4780121806. URL `http://dx.doi.org/10.1002/sim.4780121806`.