Measurement Error

Susanna Makela

August 4, 2015

Quick and Dirty

Generating finite population data

Set the values of the hyperparameters and parameters to be:

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

The population data is then generated from:

$$\begin{aligned} N_j &\sim \text{Unif}(N^{low}, N^{high}) \\ \text{logit}(\rho_j) &\sim N(\mu, \tau^2) \\ Z_i|\rho_{j[i]} &\sim \text{Bern}(\rho_{j[i]}) \\ \alpha_j|p_j &\sim N(\gamma_0 + \gamma_1 p_j + \gamma_2 u_j, \sigma_\alpha^2) \\ Y_i|\alpha_{j[i]}, Z_i &\sim N(\alpha_{j[i]} + \beta Z_i, \sigma_y^2), \end{aligned}$$

where

S_j is the set of all individuals in PSU j, with $|S_j| = N_j$

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

u_j is an unobserved covariate with $\text{Corr}(u_j, \log(N_j)) = 0.75$ and $\text{mean}(u_j) = \text{mean}(\log(N_j))$.

Sampling

Sample n_I PSUs and n_j individuals within each PSU. Analogously to S_j, let s_j be the set of individuals sampled from PSU j, and define $p_j^* = \frac{1}{n_j} \sum_{i \in s_j} Z_i$.

Models we fit

We have four different models we can fit depending on whether we assume the N_j's are known to the analyst and whether we are accounting for measurement error.

	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.

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

- **"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.

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

N_j's unknown.

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

More Detail

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

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

where u_j is an unobserved covariate such that $\mathbb{E}(u_j) = \log(N_j)$ and $\text{Corr}(u_j, log(N_j)) = 0.75$. That is, the (log of the) PSU population sizes N_j is correlated with a covariate u_j that is predictive of the outcome Y_i. 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) citations for this???.

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. (Other risk factors C, assumed to be accurately measured, can also be included in this model, but we ignore them for now.) 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:

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

Some simulations

Generating finite population data

Set parameters/hyperparameters:

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

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. Generate $\mathbf{u} = (u_1, \ldots, u_J)$ such that $\text{Corr}(\mathbf{u}, \log(\mathbf{N})) = 0.75$ $(\log(\mathbf{N}) = (\log(N_1), \ldots, \log(N_J)))$ and $\overline{\mathbf{u}} = \overline{\log(\mathbf{N})}$.
6. Draw PSU-specific intercepts from $\alpha_j|p_j \sim N(\gamma_0 + \gamma_1 p_j + \gamma_2 \log(u_j), \sigma_\alpha^2)$.
7. 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 reference?? ask shira

Models

We fit two types of models to the data.

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

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

- **"Full" model.** This model accounts for measurement error. Here we have two options, depending on whether the PSU population sizes N_j are known or not.

  N_j's known.

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

In this case, we can estimate the true prevalence p_j with one of two methods:

a) Impute the unobserved Z_i's by drawing from the posterior distribution $p(\rho_j|Z_i)$ and calculate

$$p_j = \frac{1}{n_j}\sum_{i\in s_j} Z_i + \frac{1}{N_j - n_j}\sum_{i\in(S_j\setminus s_j)} Z_i$$

b) Sample p_j by using the hypergeometric distribution for $T_j^* = n_j p_j^*$: $T_j^* \sim \text{Hypergeom}(n_j, T_j, N_j)$, where $T_j = N_j\rho_j$.

Either way, we should also incorporate N_j into the model for α_j because of the PPS sampling of PSUs. The PPS sampling renders the N_j's part of the design information, so we need to include it to ensure that our model renders the survey design ignorable cite bda. BUT, now γ_1 does not have the same interpretation as in the other models! What to do???

N_j's unknown. (to the analyst, not to the survey designer!)

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

If the N_j's are unknown, we estimate α_j using ρ_j instead of p_j.

Replication

We generate a total of 20 populations using the hyperparameters and parameters specified in **??**. 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 that we have averaged out sampling variability due to having used a particular sample.

Summary

Our simulation study thus has three 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. Second, we want to compare the performance of the full model in the cases that the N_j's are known and unknown. Finally, we want to confirm that we can estimate p_j in the full model when the N_j's are known using either of the methods described above.

References

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`.