Small Area Estimation of Hearing Loss at the County Level Combining Estimates from NHANES and ACS

Carolina Franco, NORC at Univeristy of Chicago

February 9, 2022

1 Bridging Model Approach

1.1 The NHANES model

1.1.1 Sampling Model

In small area estimation, the sampling model captures the relationship between the noisy direct survey estimates and the "truth" they are trying to estimate. For the NHANES data, the sampling model could be:

$$n_g^N \left(\hat{y}_{1g}^N, \hat{y}_{2g}^N, \hat{y}_{3g}^N\right) \sim \text{Multinom}(n_g^N ;\ \theta_{1g}^N, \theta_{2g}^N, \theta_{3g}^N), \tag{1}$$

Here, $g = 1, \ldots, 48$, represents the demographic characteristics of interest, which includes 4 race categories, 2 sex categories, and 6 age group categories. The parameters θ_{1g}^N, θ_{2g}^N, and θ_{3g}^N represent the true proportions of individuals that have no hearing loss for demographic group g, partial but not total hearing loss, and complete hearing loss, respectively. The quantities $\hat{y}_{1g}^N$, $\hat{y}_{2g}^N$, and $\hat{y}_{3g}^N$ are corresponding NHANES direct estimates for counts of these. These can computed as

the corresponding direct estimate of the proportion, which we'll call $\hat{\theta}_{1g}^N$, $\hat{\theta}_{2g}^N$, and $\hat{\theta}_{3g}^N$, multiplied by the NHANES sample size for group g, which we denote n_g^N, or alternatively by the effective sample size. For complex surveys with informative sampling, the effective sample size approximately accounts for the effect of having a complex design rather than SRS (Simple Random Sampling). An estimate for the effective sample size when there is a single proportion p being estimated is:

$$\tilde{n} = \frac{\hat{p}(1-\hat{p})}{\widehat{var}(\hat{p})} \tag{2}$$

Here, $\widehat{var}(\hat{p})$ is an estimate of the sampling variance of the direct survey-weighted estimator $\hat{p}$. When dealing with a multinomial random variable, defining an effective sample size is more complex. There has been some literature on the subject, but the literature is limited. We may use the definition of effective sample size in McAllister and Ianelli (1998) or try to come up with a better one. Or we may use the sample size n_g^N without accounting for the informative sampling.

The sampling model attempts to approximately capture both the sampling variability of the direct estimates, as well as the dependence of the sampling errors due to the fact that the categorical, mutually exclusive estimates for a given area used in the sampling model are derived from the same sample and are hence not independent.

1.1.2 The linking model

The linking model attempts to capture the relationship between the true underlying quantities, θ_{dg}, $g = 1, \ldots G$, $d = 1, 2, 3$, and any auxiliary information that is available. In this case, what we attempt to capture with the linking model is the dependence among the underlying categories–e.g., proportions with no hearing loss, partial hearing loss, total hearing loss. We may also wish to use covariates. A reasonable strategy here might be to use a logit transofrmation on the true underlying proportions in this example. We will show two ways in which you can

parametrize such a linking model for this example. In a flavor similar to Molina et al. (2015) we can express the model as follows:

$$log(\theta_{dg}/\theta_{3g}) = \mathbf{x}_d\boldsymbol{\beta}_{dg} + u_{dg} \quad d = 1, 2, g = 1, \ldots, G \tag{3}$$

where $\sum_{d=1}^{3} \theta_{gd} = 1$. Here Note that here, we arbitrarily chose the last dimension of the multinomial as a "baseline" category.

An alternative parametrization requires new definitions. This parametrization is inspired by Slud et al. 2018. Let's call μ_g the probability of having any hearing loss. Let ρ_g the probability of deafness among those who have hearing loss. We note some relationships between the θ's above and the newly defined parameters. For instance, θ_{1g} represents the probability of not having any type of hearing loss. Hence,

$$\theta_{1g} = 1 - \mu_g \tag{4}$$

Since θ_{2g} represents the probability of having hearing loss but not being deaf, and can be expressed as

$$\theta_{2g} = \mu_g(1 - \rho_g) \tag{5}$$

Lastly, θ_{3g} represent the probability of complete hearing loss/deafness, and can be expressed as:

$$\theta_{3g} = \mu_g\rho_g \tag{6}$$

Note that, automatically from the definitions above $\sum_{d=1}^{3} \theta_{dg} = 1$. The rest of the model then becomes:

$$log(\mu_g) = \mathbf{x}_{g1}\boldsymbol{\beta}_1 + u_{g1}, \quad g = 1, \ldots, G \tag{7}$$

$$log(\rho_g) = \mathbf{x}_{g2}\boldsymbol{\beta}_2 + u_{g2} \tag{8}$$

This parametrization avoids having to choose a baseline category and may be numerically more stable. It may also more naturally handle covariates.

If we don't have any good covariates for this cross-classification, then we would use a common intercept instead of regression functions $\mathbf{x}_{g1}\boldsymbol{\beta}_1$ and $\mathbf{x}_{g2}\boldsymbol{\beta}_2$.

Under either parametrization, note that there are some assumptions being made. We are assuming that the true proportions can be explained by covariates (or a common mean, if there aren't any covariates) plus a random "model error." Above, note that we are assuming common regression coefficients for all demographic categories. These linking models allows us to approximately capture the model error, and the dependence among the true partial vs total hearing loss that is not explained by covariates.

1.2 The ACS Model–Sampling and Linking Models

For ACS, we could explore models like the Fay-Herriot model (FH, 1979), or a Binomial-Logit Normal Model (BLN, see Franco and Bell 2013, 2015, 2021, for instance). The FH model can behave better numerically and converge faster because it makes normality assumptions both in the linking and sampling model. However, these normality assumptions may not be the best option when modeling proportions, which often display skewness and many estimates of zero. The BLN model can naturally handle skewness and zeros in the data. Either of these models could be formulated as either univariate or bivariate, where the bivariate models would try to borrow strength from the previous (non-overlapping) vintage of ACS to improve the estimates of the vintage of interest. Borrowing strength from the previous vintages may have moderate gains because the two sets of estimates have similar sample sizes. Furthermore, if the auxiliary covariates we use are predictive, this also tends to diminish benefits from a bivariate model compared to a univariate model. For more discussion on this, see Franco and Bell (2021).

For an example of one of the small area models mentioned above, let i be the counties, $i = 1, \ldots, m$. Let y_{ig} be the ACS sample count of people with hearing loss for county i and demographic group g, computed by multiplying the sample size or effective sample size by the ACS survey weighted proportion. Then the univariate BLN model would be as follows:

$$y_{ig}|p_{ig}^{ACS}, n_{ig} \sim \text{Bin}(n_{ig}, p_{ig}^{ACS}) \qquad i = 1, \ldots, m, \qquad g = 1, \ldots, G, \tag{9}$$

$$\text{logit}(p_{ig}^{ACS}) = \mathbf{x}'_{ig}\beta + u_{ig} \tag{10}$$

where $\text{logit}(p_{ig}) = \log[p_{ig}/(1 - p_{ig})]$, $u_{ig} \sim i.i.d.\ N(0, \sigma_u^2)$, and n_{ig}^{ACS} is the sample size for county i and group g. See Franco and Bell (2013, 2015, 2021) for the bivariate version.

One could also allow β and/or σ_u to depend on g, since there are many counties to use in the estimation. We would analyze the data to see if this assumption is necessary. We can compare model diagnostics among alternative models (i.e. FH, BLN) like leave one out validation (loo), posterior predictive checks, etc.

1.3 The Bridging Model

Let's start with a simple bridging model, similar to Erciulescu and Opsomer (2021). Let θ_{dgi}^N be the underlying truths at the county level from the NHANES, where the subscripts are the same as above. Here once again the subscript i is county, g is the demographic group, and $d = 1, 2, 3$ corresponds to no hearing loss, partial but not total hearing loss, and total hearing loss. Note that we don't observe estimates of these at the county level, but we still write expressions for them.

First, we postulate that the true values of these theoretical county quantities that would be measured by NHANES add up to the corresponding underlying true values measured by

NHANES at the demographic level g. That is,

$$\sum_{i=1}^{m} N_{gi}\theta^N_{dgi} = N_g * \theta^N_{dg}, \quad d = 1, 2, 3; \quad g = 1, \ldots, G$$

Here N_{gi} are population totals for county i and demographic group g, which we would need to assume known (Erciulescu et al 2021 also assume this). In practice we would would need to use survey estimates or some other public source. Likewise for N_g, which is the population total for demographic group g.

Next, we need to assume a relationship between the ACS and NHANES truths at the county level. Recall that ACS only asks whether survey subjects have hearing loss, without getting at total or partial. It might be reasonable to postulate a relationship between that and the combined "partial" and "total" hearing loss from NHANES. Then, along the lines of Ericulescu et. al (2021) we assume:

$$\theta^{ACS}_{ig} = \alpha + \gamma * (\theta^N_{2ig} + \theta^N_{3ig}) \qquad g = 1, \ldots, 3, \qquad m = 1, \ldots m$$

The assumption of perfect correlation is strong (made also in Erciulescu et al. 2021), and it might be possible to relax it.

1.4 Computational Challenges of Bridging Model

I have previously fit models like the ACS sampling and linking models above, and models like the NHANES model above, using STAN. However I have not fit a bridging model, and the model above is a bit more complex than the bridging model of Erciulescu and Opsomer (2021, also fit via STAN). That, combined with the large number of estimates of zero than I expect to see in the data, might pose computational problems.

If running one model for all the small areas of interest (counties cross-classified with

demographics) is computationally prohibitive due to the large number of counties, we can't fit bridging models separately by region as we don't have totals for this in NHANES. So we would have to, say, fit separate models for each age group or race, to see if this can ease the computational burden. However, suppose, for instance, we fit different models for different age groups. Then inference related to differences among ages would be inaccurate as we would not be capturing sampling and modeling dependencies among different ages. Another alternative is to fit separate models for hearing loss and partial hearing loss, ignoring their dependence. Yet another alternative is to assume normality, which tends to be numerically more stable. In this case, we would be using the bridging model of Erciulescu and Opsomer (2021). However, if there is a large number of estimates of zero, a normal distribution might not be ideal.

2 Alternative Strategies–Calibration

2.1 Ignoring Error of survey estimates

Instead of doing bridging models, we can use calibration. In the simplest calibration technique, we can fit models for ACS such as the ones mentioned, above in Section 1.2 e.g.,

$$y_{ig}|\theta_{ig}^{ACS}, n_{ig} \sim \text{Bin}(n_{ig}, \theta_{ig}^{ACS}) \qquad i = 1, \ldots, m, \qquad g = 1, \ldots, G, \tag{11}$$

$$\text{logit}(p_{ig}^{ACS}) = \mathbf{x}'_{ig}\beta + u_{ig} \tag{12}$$

where $\text{logit}(\theta_{ig}) = \log[\theta_{ig}/(1-\theta_{ig})]$, $u_{ig} \sim i.i.d.\ N(0, \sigma_u^2)$, and n_{ig}^{ACS} is the sample size for county i and group g. Fay-Herriot type models can also be tried. In this case, it's possible to fit models separately by state, or separately by some kind of demographic group division. Later, we apply calibration to the complete set of estimates so that sums across geographies sum at to the

appropriate NHANES estimates by group g. However, recall that if we want to do inference on a difference of two quantities of interest, and properly capture their dependence, they need to be jointly estimated.

In the model above, we are postulating the truth as measured by the ACS estimates, but not the truth as measured by the NHANES estimates. Hence we cannot compute the posterior variance of a parameter that is not expressed. We need to make some kind of assumption. For instance, suppose the calibration weights for a cell is w_{ig}. Then we assume $\theta_{ig}^N = w_{ig}\theta_{ig}^{ACS}$. E.g, the same calibration weights that make our model predictions based on ACS add up to the direct NHANES survey estimate of the national total translate the true underlying truth of the ACS to the true underlying truth of NHANES. Under this assumption, we can compute the desired posterior variances simply as the posterior variance of $w_{ig}p_{ig}^{ACS}$ via MCMC.

A related alternative is to fit a small area model to the NHANES data, such as the one mentioned in Section 1.1. Then, instead of calibrating to the direct estimates, we will calibrate to the model estimates. The idea is that the model might provide better estimates, especially if covariates are available.

2.2 Taking into account error of calibration totals

Suppose you want to take into account that the NHANES estimates (whether direct or modeled) we are calibrating to have error. We will drop the subscrip d, which indicates whether we are talking about total hearing loss, partial hearing loss, or deafness. This idea applies to either. Let η_g be the underlying trouth of the total you are calibrating to in NHANES, and let $\hat{\eta}$ be the estimate of it. Then

$$\hat{\eta}_g = \eta_g + e_g \tag{13}$$

The you would add the constraint:

$$\sum_i p_{gi} N_{gi} = N \times p_g \tag{14}$$

and fit the model with such constraints, again using STAN.

With this, the predicted estimates do not really add up to the NHANES estimates because it posits that the NHANES estimates have error.