Poststratification into many categories using hierarchical logistic regression

2019-06-25

Our paper for journal club today was the first to introduce the concept known as "MRP": multilevel regression with poststratification.[Gelman and Little, 1997] The purpose of the method is to provide better estimates of population quantities when the number of poststratification[1] cells is large, leading to cells with low numbers of observations or no observations. At a more general level, this paper serves as a great example of how to make model-based estimates for a population which was sampled using a complex survey design (e.g., multi-stage stratified random sample with unequal probabilities of selection across strata).

Let's assume a simple example for the discussion today. Suppose we had a population of 100 people, with 50 men and 50 women. We performed stratified random sampling, sampling the men with $p = 0.1$ and the women with $p = 0.2$. After performing the sampling, we would like to estimate the total number of participants in the combinations of sex and age (young vs. old).[2] Let's assume that 80 people in the population are young and 20 are old. Then are population would look like this

	Men	Women
Young	40	40
Old	10	10

and the sample dataset, due to chance, might look like this

	Men	Women
Young	5	6
Old	1	2

Gelman and Little say that poststratification is anything "that adjusts to population totals." That seems to be a very general definition, that covers raking and smoothing of weights. In the examples they give, it seems like weight construction is performed after survey sampling, with only knowledge of the population counts in cells of interest that matter.

Discussion question: If the population cells sizes are known, why would you ever use the sampling weights? In our example, the sampling weight for every man would be $1/0.1 = 10$ and $1/0.2 = 5$ for every woman. Assuming no non-response, in design-based inference we would simply use the associated sample weights for each participant, constructed on only sampling probabilities by sex. This approach would lead to estimated cell sizes of

[1] "Poststratification" refers to partitioning the target population into groups that you care about making individual predictions for.

[2] Therefore, we are performing poststratification on two variables: one that was used to construct the sample and one that was not.

	Men	Women
Young	50	30
Old	10	10

with marginal totals of 60 for men and 40 for women. The benefit of raking is that the marginal counts of men and women will always be equal to their known population totals, regardless of the sampling probabilities or actual number of observations sampled in each cell. Raking simply makes the weight for each cell N_j/n_j. For men the weight would be $50/6 \approx 8.3$ and for women the weight would be $50/8 = 6.25$. This weights would lead to estimated population sizes of (rounded)

	Men	Women
Young	42	37.5
Old	8	13.0

We could also rake on the joint distribution of age and sex, which would lead us to population estimates that are nearly identical to the known population. However, raking doesn't assume any interactions between variables (i.e., cell counts are a product of the margins).

Discussion question: Why are design-based standard errors larger than model-based standard errors? Where's the extra uncertainty coming from?

Poststratification

Assume that we partition the population into R categorical variables, where the rth variable has J_r levels. The total number of categories is $J = \Pi_{r=1}^{R} J_r, j = 1, \ldots, J$. Assume that N_j is known for all j. The overall population mean $\bar{Y} = \frac{\Sigma_j N_j \pi_j}{\Sigma N_j}$. n_j is the number of units sampled. Assume that non-response depends only on the R variables. R could include anything to construct the survey weights, as well as any information informative about Y.

The model

$$\text{logit}(\pi_j) = X_j \beta,$$

with a uniform prior on β, corresponds closely to classical weighting schemes.[3] Then Gelman claims the following relationships between model-based inference and design-based inference:

- $X_{JxJ} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \cdots & \cdots & \cdots & 0 \\ 0 & \cdots & \cdots & \cdots & 1 \end{bmatrix}$ is the same as weighting each unit in each cell by N_j/n_j

[3] It's not actually identical with a non-linear link function, but the differences are minor with large samples.

- $X_{Jx\sum_{j=1}^{J}R_j} = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}$ is the same as raking on all R 1-way tables
- Adding columns that specify interactions between sex and age amounts to raking on the 2-way table between sex and age
- $X_{Jx1} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$ is equal to the sample mean estimate.

Hierarchical model

Now assume that we want to set up a hierarchical model with parameter vector $\beta = (\alpha, \gamma_1, \ldots, \gamma_L)$, were α are fixed effects and each $\gamma_l, l = 1, \ldots, L$ is a subvector of coefficients γ_{lk} where we fit $\gamma_{lk} \sim N(0, \tau_l^2), k = 1, \ldots, K_l$.[4] If we make a matrix C_{nxJ} and a matrix $Z_{nx?} = C_{nxJ}X_{Jx?}$, then we can assume

$$y_i \sim \text{Bernoulli}(p_i)$$

$$\text{logit}(p_i) = C_{nxJ}X_{Jx?}\beta_{?x1},$$

where $\beta \sim MVN(0, \Sigma_\beta)$. So, we will end up with a p_i for each observation with $?$ distinct values. The population π is arrived at by the formula

$$\pi = \frac{\sum_{j=1}^{J} N_j \pi_j}{\sum_{j=1}^{J} N_j}.$$

Discussion question: How do you arrive at the weighted estimates for each π_j, or a combination of some j (e.g., all males)?

Using the Bayesian paradigm to fit these models just allows for additional shrinkage, and potentially much less variable estimates in cells with low sample size counts.

Discussion question: How do you do MRP if you don't know the cell counts in the population?

References

Andrew Gelman and Thomas C Little. Poststratification into many categories using hierarchical logistic regression. *Surv. Methodol.*, 23(2):127–135, 1997.

[4] In other words, there are L difference random effects and K_l levels to each type of random effect. Another way to think of L is a higher level of grouping among all of the LK cells.