Weight smoothing models to estimate survey estimates from binary data

Weight smoothing models to estimate survey estimates from binary data

Vandendijck Yannick[1], Faes Christel[1], Hens Niel[1,2]

[1] Interuniversity Institute for Biostatistics and statistical Bioinformatics, Hasselt University, Diepenbeek, Belgium
[2] Centre for Health Economic Research and Modeling Infectious Diseases, Vaccine and Infectious Disease Institute, University of Antwerp, Wilrijk, Belgium

E-mail for correspondence: yannick.vandendijck@uhasselt.be

Abstract: In surveys, when the number of respondents in a post-stratum is small relative to the population size in that post-stratum, post-stratification weights are inflated and modifications are required to obtain less variable estimates. Weight smoothing models, random-effects models that induce shrinkage across post-stratum means, are such modifying methods. We describe the empirical Bayes weight smoothing model approach to estimate the overall mean of a binary survey outcome. The generalized linear mixed model formulation of this model allows easy fitting. Two extensions of the model are presented. The estimation of the prevalence and incidence trend of influenza-like illness using the Great Influenza Survey in Flanders, Belgium, is considered as an application.

Keywords: Binary data; Empirical Bayes; Post-stratification; Random-effects.

1 Introduction

Stratification is the process of dividing the population into homogeneous mutually exclusive strata before sampling to improve the representativeness of the sample. In observational studies post-stratification can be used to correct for known differences between the obtained sample and the population. This is done by equating the distribution of a secondary variable (*e.g.*, age) measured in the sample with its distribution in the population, and adjusting estimates using weighting techniques. This can improve both the accuracy and precision of estimates (Little, 1991).
Let Y denote a binary survey outcome variable and X a discrete post-stratifying variable with H levels and known population distribution. Let N_h and n_h denote the population and sample size in post-stratum h, respectively. We assume that N_h is known. Define $N = \sum_{h=1}^{H} N_h$ and $n = \sum_{h=1}^{H} n_h$. We consider inference for the finite population mean $\bar{Y} = \sum_{h=1}^{H} P_h \bar{Y}_h$, where $\bar{Y}_h$ is the population mean in post-stratum h and $P_h = N_h/N$ is the population proportion in post-stratum h.

An estimate for the population mean is of the form $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} w_{i(h)} y_i$, where $w_{i(h)}$ is the weight of observation i belonging to post-stratum h. The unweighted sample mean, $\bar{y}_{unw}$, is obtained when $w_{i(h)} = 1$ $(\forall i)$, and can be written as $\bar{y}_{unw} = \sum_{h=1}^{H} p_h \bar{y}_h$, where $p_h = n_h/n$ is the sample proportion and $\bar{y}_h$ is the sample mean in post-stratum h. Whenever p_h deviates from its population proportion P_h, the unweighted mean is a biased estimate. The post-stratified mean estimate, $\bar{y}_{ps}$, is obtained when $w_{i(h)} = P_h/p_h$ $(\forall i)$. While $\bar{y}_{ps}$ is an unbiased estimate of $\bar{Y}$, it has greater variance than $\bar{y}_{unw}$. This increase in variance can overwhelm the reduction in bias, so that the post-stratified mean estimate actually increases the mean squared error. This happens especially when some weights are large.
A common approach to deal with this problem is weight trimming. This procedure uses the bias-variance trade-off by introducing some bias in the estimate, but effectively reducing the variance. An alternative model-based strategy is to model the stratum means directly by random-effects. These so-called weight smoothing models make a distributional assumption for the Y_i and use the model to predict the non-sampled values of Y. For a Gaussian survey outcome these models are well explained in literature (see *e.g.*, Elliott and Little, 2000). For a binary survey outcome only the full Bayesian approach has been discussed (Elliott, 2007) in the context of generalized linear regression estimators. We describe the empirical Bayes estimation approach of weight smoothing models for binary data and present two extensions of these models.

2 Weight Smoothing Models for Binary Data

The general form of the weight smoothing models for a binary survey outcome is

$$Y_{i(h)}|p_h \sim \text{Binom}(1, p_h) \quad \text{and} \quad \boldsymbol{\delta}^* \sim \mathcal{N}_H(\boldsymbol{\delta}, \mathbf{D}), \tag{1}$$

where $g(E[Y_{i(h)}|p_h]) = \delta_h^*$, $\boldsymbol{\delta}^* = (\delta_1^*, ..., \delta_H^*)^{\mathrm{T}}$ and $\boldsymbol{\delta} = (\delta_1, ..., \delta_H)^{\mathrm{T}}$ are unknown vectors, $\mathbf{D}$ is an unknown $H \times H$ covariance matrix and $g(\cdot)$ is the logit-link function. Under model (1) the weight smoothed estimate of $\bar{Y}$ is

$$\bar{y}_{ws} = E[\bar{Y}|\mathbf{y}] = \frac{1}{N}\sum_{h=1}^{H}\{n_h\bar{y}_h + (N_h - n_h)\hat{\mu}_h\}, \tag{2}$$

where $\hat{\mu}_h = E[\bar{Y}_h|\mathbf{y}] = E[\mu_h|\mathbf{y}] = g^{-1}(\delta_h^*)$. The unweighted and post-stratified mean are obtained as estimates of (2) if $\mathbf{D} \to 0$ and $\mathbf{D} \to \infty$, respectively. We consider four other cases of model (1) (Little, 1991; Lazzeroni and Little, 1998; Elliott and Little, 2000):
(a) Exchangeable random effects (XRE): $\delta_h = \beta$ for all $h, \mathbf{D} = \sigma_D^2 \mathbf{I}_H$.
(b) First order autoregressive (AR1): $\delta_h = \beta$ for all $h, (\mathbf{D})_{ij} = \sigma_D^2 \rho^{|i-j|}$ for $i, j \in \{1, \ldots, H\}$.

(c) Linear (LIN): $\delta_h = \beta_0 + \beta_1 X_h$ for all $h, \mathbf{D} = \sigma_D^2 \mathbf{I}_H$.
(d) Nonparametric (NPAR): $\delta_h = f(X_h)$ for all $h, \mathbf{D} = \sigma_D^2 \mathbf{I}_H$, where f is a nonparametric spline function. We use the approximating thin plate spline family.
All these models can be cast in the generalized linear mixed model (GLMM) framework. The model is fit by pseudo-restricted maximum likelihood estimation based on linearisation. At convergence, estimates of $\hat{\mu}_h$ are obtained and $\bar{y}_{ws}$ can be calculated. Calculation of the variance for $\bar{y}_{ws}$ can be either done analytically or by a bootstrap method.

Extension 1: Assume a binary survey outcome is measured at different time points, and interest is in the estimation of the time trend of the overall mean, namely $\bar{Y}_t = \frac{1}{N_t} \sum_{i=1}^{N_t} Y_{it}$, for $t = 1, ..., T$. At each time point, the unweighted or post-stratified mean can be calculated. However, one can use a smoothed weight approach which exploits the time-trend. The general form of this model is

$$Y_{i(h),t} | p_{ht} \sim \text{Binom}(1, p_{ht}), \ \forall \ t, \quad \text{and} \quad \boldsymbol{\delta}_t^* \sim \mathcal{N}_H(\boldsymbol{\delta}_t, \mathbf{D}). \tag{3}$$

The unknown parameters $\boldsymbol{\delta}_t = (\delta_{h1}, ..., \delta_{hT})^{\mathrm{T}}$ are additively decomposed into $\delta_{ht} = \delta_h + \delta_t$. For δ_h and $\mathbf{D}$ we assume models (a)-(d). For the time trend a non-parametric trend, namely $\delta_t = f_t(t)$, is assumed.

Extension 2: Misspecification in (1) leads to biased estimates for $\hat{\mu}_h$ in (2) and consequently a biased estimate of $\bar{y}_{ws}$. We propose the use of a doubly robust weight smoothed estimate of the form

$$\bar{y}_{ws,dr} = \frac{1}{N} \sum_{h=1}^{H} \left\{ \frac{n_h}{\hat{\pi}_h} \bar{y}_h + (N_h - \frac{n_h}{\hat{\pi}_h}) \hat{\mu}_h \right\}, \tag{4}$$

in analogy with doubly robust estimates in the missing data context. The $\hat{\pi}_h$ represent inclusion probabilities and are calculated using a method that resembles a trimming weights approach.

3 Application

The Great Influenza Survey (GIS) is an observational survey based on the voluntary participation of individuals via the internet aiming at the monitoring of influenza-like illness (ILI). We use data from the Flemish GIS from the 2010-2011 influenza season ($n = 4551$). Interest is in the estimation of the overall prevalence and the incidence trend of ILI. The age distribution of the GIS population is very dissimilar to the overall Flemish population age distribution (Figure 1(a)). Post-stratification weights range from 0.46 to 35.70 (18 age groups of length 5 years as post-strata). The unweighted mean estimate of the prevalence is 5.12% (95% CI: 4.52-5.80%), whereas the post-stratified mean estimate is 7.10% (95% CI: 5.31-9.45%).

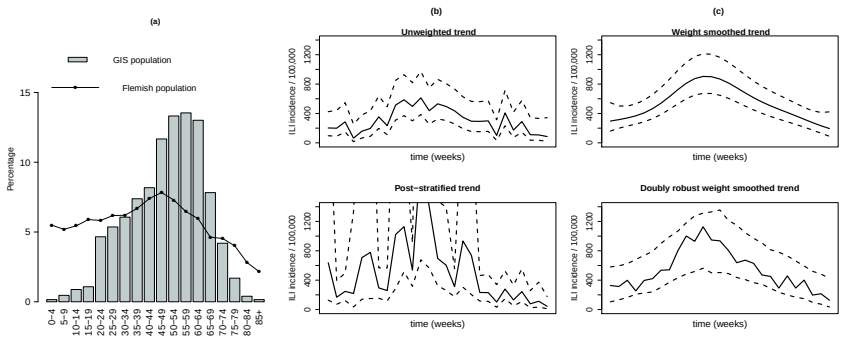

FIGURE 1. (a) Age distribution of the GIS and Flemish population. (b) Estimated unweighted and post-stratified trend with confidence intervals (CIs). (c) Estimated weight smoothed and doubly robust weight smoothed trend with CIs.

The weight smoothed estimate using the NPAR model yields 6.88% (95% CI: 5.69-8.30%). The doubly robust approach for the prevalence yields similar results, namely 6.82% (95% CI: 5.61-8.28%). The results of the incidence trend estimation (using the NPAR model) is shown in Figure 1(b) and Figure 1(c). It is seen that the weight smoothing results are a compromise between the unweighted and post-stratified trends.

4 Discussion

Weight smoothing models offer a good solution for inference of a binary survey outcome when some post-stratification weights are large. These models can be cast into the GLMM framework which allows for implementation in standard statistical software. In the real-life data application it was shown that the different approaches yield substantially different results. It is therefore important to use weight smoothing models in this specific data context.

References

Elliott, M.R. (2007). Bayesian weight trimming for generalized linear regression models. *Survey Methodology*, **33**, 23 – 34.

Elliott, M.R. and Little, R.J.A. (2000). Model-based alternatives to trimming survey weights. *Journal of Official Statistics*, **16**, 191 – 209.

Lazzeroni, L.C. and Little, R.J.A. (1998). Random-effects models for smoothing poststratification weights. *Journal of Official Statistics*, **14**, 61 – 78.

Little, R.J.A. (1991). Inference with survey weights. *Journal of Official Statistics*, **7**, 405 – 424.