Mixed effects model for SAE

Tomáš Košlab

FNSPE CTU in Prague

June 22, 2018

Outline

Task

- N individuals in D domains, N_d individuals in the d-th domain
- every individual modelled by a Bernoulli-distributed r.v. Y_{dj} (above / below poverty line), realization y_{dj}
- sample of size n_d from every domain
- task: predict the population mean in every domain, $\overline{y}_d = \frac{1}{N_d} \sum_{j=1}^{N_d} y_{dj}$
- if n_d is too small for a direct estimate of sufficient quality - small area

Proposed model

For $Y_{dj} \sim Be(p_{dj})$ we propose the following logit regression model

$$\begin{aligned} &\text{(F)} \quad \text{logit}(p_{dj}) = \boldsymbol{x}_{dj}^T\boldsymbol{\beta} + \mu_d, \quad d = 1, \ldots, D_F, \quad j = 1, \ldots, n_d, \\ &\text{(R)} \quad \text{logit}(p_{dj}) = \boldsymbol{x}_{dj}^T\boldsymbol{\beta} + u_d, \quad d = D_F + 1, \ldots, D, \quad j = 1, \ldots, n_d \end{aligned} \tag{1}$$

where

- $\boldsymbol{x}_{dj}^T = (x_{dj,1}, \ldots, x_{dj,p})$ is the vector of covariates belonging to the j-th individual in the d-area,
- $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is the vector of unknown fixed parameters,
- $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_{D_F})^T$ is the vector of fixed area effects,
- $\boldsymbol{u} = (u_{D_F+1}, \ldots, u_D)^T$ is the vector of random area effects.

Model - assumptions & notes

- data in different areas are independent (each area has its own effect)
- data in areas modelled by a fixed effect are independent
- probability p_{dj} can be expressed as

$$\text{(F)} \quad p_{dj} = \frac{\exp(\boldsymbol{x}_{dj}^T\boldsymbol{\beta} + \mu_d)}{1 + \exp(\boldsymbol{x}_{dj}^T\boldsymbol{\beta} + \mu_d)}, \quad d = 1, \ldots, D_F, \quad j = 1, \ldots, n_d,$$

$$\text{(R)} \quad p_{dj} = \frac{\exp(\boldsymbol{x}_{dj}^T\boldsymbol{\beta} + u_d)}{1 + \exp(\boldsymbol{x}_{dj}^T\boldsymbol{\beta} + u_d)}, \quad d = D_F + 1, \ldots, D, \quad d = 1, \ldots, n_d \tag{2}$$

PQL method

- log-likelihood function

$$l(\boldsymbol{\beta}, \boldsymbol{\mu}, \sigma^2; \boldsymbol{y}) = \sum_{d=1}^{D_F} \sum_{j=1}^{n_d} [y_{dj} \log p_{dj} + (1 - y_{dj}) \log(1 - p_{dj})]$$

$$+ \sum_{d=D_F+1}^{D} \log \int_R \prod_{j=1}^{n_d} p_{dj}^{y_{dj}} (1 - p_{dj})^{1-y_{dj}} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{u_d^2}{2\sigma^2}} du_d \tag{3}$$

- PQL can be derived from the Laplace approximation of the log-likelihood function, $l_{Laplace}$
- omission of the last term in $l_{Laplace}$ (for computational reasons) leads to l_{PQL}
- estimates of $\boldsymbol{\beta}, \boldsymbol{\mu}$ are obtained by N.-R. algorithm, estimate of σ^2 by fixed-point

Empirical Best Predictor (EBP)

- closely related to the Best Predictor (BP)
- BP of $\hat{\bar{y}}_d$ can be expressed as

$$\hat{\bar{y}}_d = \frac{1}{N_d}\left(\sum_{j\in s_d} y_{dj} + \sum_{j\in r_d} \hat{y}_{dj}\right) = \frac{1}{N_d}\left(\sum_{j\in s_d} y_{dj} + \sum_{j\in r_d} \hat{p}_{dj}\right) \tag{4}$$

  where
  - predictions are denoted by hats
  - s_d and r_d denote the indices of observations from the d-th area that are inside and outside the sample respectively
- EBP is obtained by substituting parameter estimates into formulas for BP

Plug-in predictor

- the formula for plug-in predictor coincides with (4)
- the difference lies in the term $\hat{p}_{dj}$ - while for EBP (BP) it is calculated by a formula, plug-in is based on the formulas (2)

$$p_{dj} = \frac{\exp(\boldsymbol{x}_{dj}^T\boldsymbol{\beta} + \mu_d)}{1 + \exp(\boldsymbol{x}_{dj}^T\boldsymbol{\beta} + \mu_d)}, \quad d = 1, \ldots, D_F, \quad j = 1, \ldots, n_d,$$

$$p_{dj} = \frac{\exp(\boldsymbol{x}_{dj}^T\boldsymbol{\beta} + u_d)}{1 + \exp(\boldsymbol{x}_{dj}^T\boldsymbol{\beta} + u_d)}, \quad d = D_F + 1, \ldots, D, \quad d = 1, \ldots, n_d \tag{5}$$

where the estimates of $\boldsymbol{\beta}$, μ_d and the predictions of u_d are substituted into the equations

- requires predictions of u_d

Setup

- $D = 30$ domains
- $D_F = 5$ areas modelled by a fixed effect
- $N_d = 1000$, $d = 1, \ldots, D$
- design matrix $\rightarrow$ 3 parameters β_1, β_2, β_3
  - $X_{dj,1} \sim Be(0.48)$
  - $X_{dj,2} \sim Be(0.6)$
  - if $x_{dj,2} = 1$, then $X_{dj,3} \sim Be(0.5)$, else $x_{dj,3} = 1$
- $Y_{dj} \sim Be(p_{dj})$, values of y_{dj} are generated using equation (2)
- different sample sizes for areas with fixed (n_d^F) and random effects n_d^R
- task: prediction of $\overline{y}_d$ for every domain for the whole population

Results

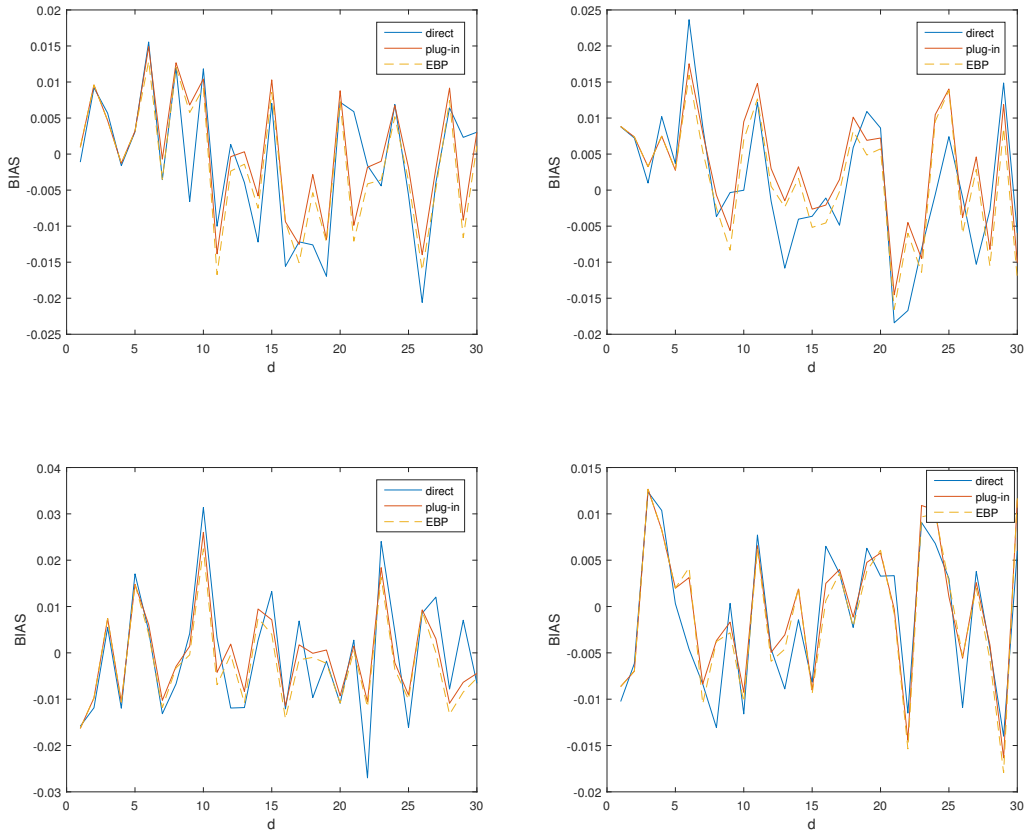

Figure: BIAS of predictions using the respective methods.

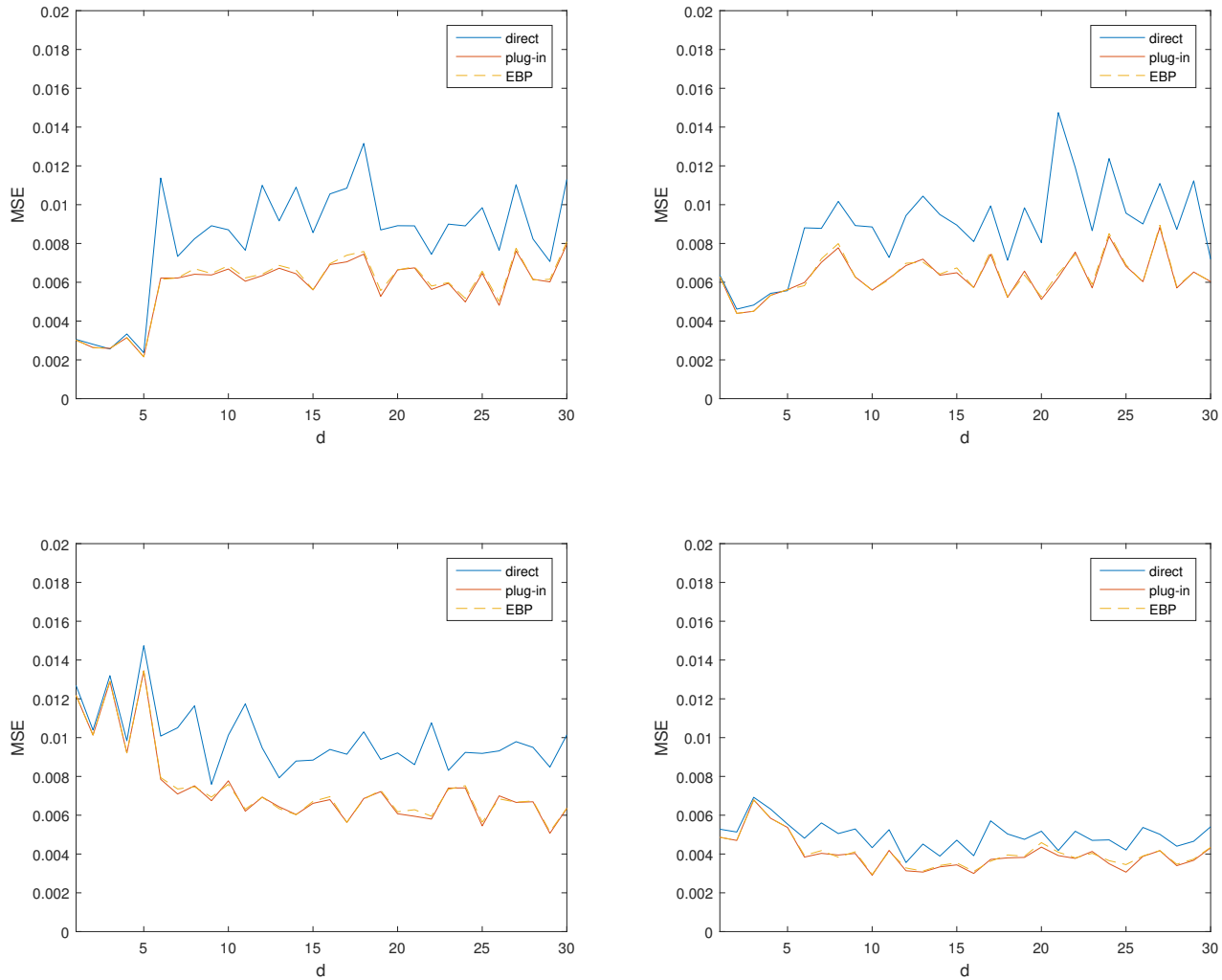

Figure: MSE of predictions using the respective methods.

Conclusion & future tasks

- predictions obtained by EBP and plug-in predictor are of comparable quality
- both are superior compared to the direct estimate, especially for small amounts of data
- estimate of prediction error - parametric bootstrap
- application on real data and comparison with other models

Thank you for your attention!