Cross validation to compare stage 2 models

Austin Schumacher

2022-03-10

Model

Let $r = 1, \dots, R = 47$ index admin 1 regions in Kenya. We have a model that estimates the mean HAZ ($c = 1$) and WAZ ($c = 2$) in each regions, denoted as $\hat{\mu}_{rc}$, using as input data the direct estimates in each region $\hat{\boldsymbol{y}}_r = [\hat{y}_{r1}, \hat{y}_{r2}]$ which have asymptotic normal design-based covariance matrix $\hat{\boldsymbol{V}}_r^{des}$.

Our data likelihood is

$$\hat{\boldsymbol{y}}_r | \boldsymbol{\mu}_r, \boldsymbol{V}_r \sim N_2(\boldsymbol{\mu}_r, \boldsymbol{V}_r)$$

with $\boldsymbol{V}_r$ assumed fixed and known.

The model to estimate $\hat{\boldsymbol{\mu}}_r$ is

$$\begin{aligned}
\mu_{r1} &= \beta_1 + v_{r1} + u_{r1} + \lambda(u_{2r}) \\
\mu_{r2} &= \beta_2 + v_{r2} + u_{r2} \\
v_{r1} | \sigma_1^2 &\overset{iid}{\sim} N(0, \sigma_1^2) \\
v_{r2} | \sigma_2^2 &\overset{iid}{\sim} N(0, \sigma_2^2) \\
\boldsymbol{u}_1 &\sim ICAR(1) \\
\boldsymbol{u}_2 &\sim ICAR(1)
\end{aligned}$$

using a BYM2 parameterization for the ICAR and IID random effects in each model.

Cross validation

We will devise a leave-one-out cross validation approach that accounts for the correlation between HAZ and WAZ. Briefly, for region r, we will delete $\hat{\boldsymbol{y}}_r$ from the data, fit a second stage model, and use this to predict $\hat{\boldsymbol{y}}_r$ as well as to estimate its predicted distribution. In order to compare models, we can calculate, for HAZ and WAZ separately, the prediction error, the posterior probability of observing the held out data given the model fit the the rest of the data (CPO).

Specifically, we loop through regions and do the following for each region:

1. replace the HAZ and WAZ direct estimates for region r with missing values
2. fit the second stage model in INLA

3. extract the predicted medians of the latent means, $\widehat{med}(\mu_{-rc})$, which is the value x such that $\int_{-\infty}^{x} \pi(\mu_{rc}|\hat{\boldsymbol{y}}_{-r}) = 0.5$ (INLA calculates this automatically), and use this to calculate prediction error for HAZ and WAZ separately as $\hat{y}_{rc} - \widehat{med}(\mu_{-rc})$
4. draw $S = 10000$ samples from the joint marginal distributions $\pi(\boldsymbol{\mu}_r|y_{-r})$ using `inla.rmarginal()` and call these $\boldsymbol{\mu}_{-r}^{(s)}$
5. calculate the CPO for region r as $\frac{1}{S}\sum_{s=1}^{S} p(\hat{\boldsymbol{y}}_r|\boldsymbol{\mu}_{-r}^{(s)}, \hat{\boldsymbol{V}}_r^{des})$ where $p(\cdot|\boldsymbol{\mu}_{-r}^{(s)}, \hat{\boldsymbol{V}}_r^{des}) = N_2(\boldsymbol{\mu}_{-r}^{(s)}, \hat{\boldsymbol{V}}_r^{des})$
6. calculate $-\sum_{r=1}^{R} \log(CPO_r)$.

QUESTION: Will `inla.posterior.sample()` work to sample from $p(\hat{\boldsymbol{y}}_r|\hat{\boldsymbol{y}}_{-r})$?

Results