Bayesian inference for population prediction of individuals without health insurance in Florida

Neung Soo Ha

NISS
Outline

- Motivation
- Description of the Behavioral Risk Factor Surveillance System, BRFSS
- Description of posterior predictive distribution under informative/non-informative sampling
- Selection bias
- Inference for proportions of individuals without the health insurance
- Variation of maps
Motivation:

Interest:

1. Inference for the county-level proportions of persons without health insurance in FL, using the 2010 Behavioral Risk Factor Surveillance System, BRFSS.

2. Display the variation of maps
2010 BRFSS: Survey description

- The largest on-going telephone based survey; administered by the Centers for Disease Control and Prevention.
- Collect data on individual risk behaviors and preventive health practices for the adult population (18 years of age and older); collects state-specific data
  - e.g.: alcohol/cigarette consumption, general health status, health insurance,
- Uses a disproportionate stratified sample (DSS) design.
- In FL, 63 out of 67 counties were sampled.
- No cell phones for 2010.
Examples of observed data

Figure: Observed: All races

white=having insurance, blue=not having insurance

Figure: Observed: Hispanics
Complication

1. Small or non-existent number of observations for some sub-population groups in certain geographical areas.
2. Possible presence of selection bias in the sample
Model specification for the BRFSS data

\[ P(Y_{ik\ell} = 1|\theta_{ik}) = \theta_{ik} \]
\[ \text{logit}(\theta_{ik}) = X'_k\beta + \nu_i \]
\[ \nu_i \sim N(0, \sigma^2_\nu) \]

- \( Y_{ik\ell} = 1 \): not having insurance, \( Y_{ik\ell} = 0 \): having insurance
- \( i = 1, \ldots, M \): county
- \( k = 1, \ldots, K \): population class (9 age groups \( \times \) 3 races \( \times \) 2 genders )
- \( \ell = 1, \ldots, N_{ik} \): units in county \( i \) and group \( k \)
- \( N_{ik} \): population size in \( i \)th county and \( k \)th group
- \( X'_k \): vector of indicators for group \( k \).
- \( p(\beta) = \text{const on } (-\infty, \infty) \) Gelman et al.(2004)
- \( p(\sigma^2_\nu) = \text{const on } (0, \infty) \)
Model Evaluation

Model fit

▶ Bayes residual plots
▶ QQ plots
▶ Bayesian tests: Partial posterior predictive p-values, Bayarri and Berger (2002)

Predictions

▶ Cross-validation to simulate finite population inference
  ▶ Use $100(1-p)\%$ of observed data to make predictions for the $100*p\%$ that were “held-out”
Posterior predictive distribution with non-informative sampling

Use posterior predictive distribution to make inference for the remaining non-sampled units

Under the non-informative sampling, i.e. independence between the probability of selecting a person and the response:

- $Y_s =$ sampled units, $Y_{ns} =$ non-sampled units
- $Y_{ns} \perp \perp Y_s | \theta, X$
- $f(Y_{ns}|Y_s, X) = \int g(Y_{ns}|Y_s, X, \theta)p(\theta|Y_s, X)d\theta$
  
  $= \int g(Y_{ns}|\theta, X)p(\theta|Y_s, X)d\theta$

- $p(\theta|Y_s, X) \propto L(\theta|Y_s, X)p(\theta|\beta, \sigma_\nu^2)p(\beta, \sigma_\nu^2):$ posterior distribution

  - $\theta =$ a vector of model parameters
  - $(\beta, \sigma_\nu^2) =$ a vector of hyperparameters
Under informative sampling

- Selection probabilities are related to the outcome variables.
- The observed outcomes may not be representative of the population outcomes.
- May cause bias for inferences about the finite population parameters.
Selection bias

Let $I = (I_1, \ldots, I_N)$, $I_i = 1$ if $i \in s$, $I_i = 0$ if $i \not\in s$, $s = \text{sample}$

The posterior predictive distribution with $I$

$$f(Y_{ns}|Y_s, X, \theta, I) \propto g(Y_{ns}|\theta, X)p(I|Y, X, \theta) \quad (1)$$

Note:

- $p(I|X, Y, \theta) \neq p(I|X, \theta)$ i.e. selection bias.
- Selection information only comes from the sampled units.
Procedure: Ha and Sedransk (2015)

1. Pfeffermann and Sverchkov (2009):
\[ p(1|Y, X, \theta) = p(I|Y, X) = \prod_k f_1(I_k|Y, X_k) \]

2. Using the Poisson sampling approximation, Chambers et al. (1998)
\[ p(I|Y, X) = M \prod_{i=1}^{M} K \prod_{k=1}^{K} \left\{ \prod_{\ell \in s_{ik}} P(I_{ik\ell} = 1|Y, X_k) \right\} \left\{ \prod_{\ell \notin s_{ik}} (1 - P(I_{ik\ell} = 1|Y, X_k)) \right\} \]

3. \[ P(I_{ik\ell} = 1|Y, X_k) = E_U(\pi_{ik\ell}|Y, X_k) = \left\{ E_s(w_{ik\ell}|Y, X_k) \right\}^{-1}, \]
Pfeffermann and Sverchkov (2009),
- \( \pi_{ik\ell} = \text{probability of unit selected} \)
- \( w_{ik\ell} = 1/\pi_{ik\ell} \)
Posterior predictive distribution with inclusion probability

\[ g(Y_{ns} | Y_s, X, \theta, I) \approx g(Y_{ns} | \theta, X) \times \prod_{i=1}^{M} \prod_{k=1}^{K} (1 - a_k)^{M_{ik1}} (1 - b_k)^{M_{ik0}}, \]

\[ a_k = \frac{1}{\bar{w}_{k,y0}}, \quad b_k = \frac{1}{\bar{w}_{k,y1}} \]

\[ \bar{w}_{k,y0} = \text{average of inverse probability of selected units for class } k \text{ with response } Y_{ik\ell} = 0 \]

\[ \bar{w}_{k,y1} = \text{average of inverse probability of selected units for class } k \text{ with response } Y_{ik\ell} = 1 \]

\[ M_{ik1} : \# \text{ non-sampled units with } Y_{ik\ell} = 1 \]

\[ M_{ik0} : \# \text{ non-sampled units with } Y_{ik\ell} = 0 \]

\[ M_{ik1} + M_{ik0} = N_{ik} - n_{ik} \]

Use rejection sampling algorithm to obtain \( g(Y_{ns} | \cdot) \), Robert, C. and Casella, G. (2004)

In our data set, \( 0.96 \leq (1 - a_k)/(1 - b_k) \leq 1.01 \) with 43 of the 54 \( K \) groups.

\( \text{No substantial selection bias.} \)
Comparison between prediction to observation

County comparison

Figure: Observed: All races

Figure: Predicted: All races
Comparison between prediction to observation

Figure: Observed: Hispanics

Figure: Predicted: Hispanics
Comparison between races

Figure: Predicted: Whites

Figure: Predicted: Hispanics
Variation of maps

Objective: Assess the variation of the entire map as a unit rather than expressing the variation separately for each geographic unit.

- Produce a map, based on each MCMC replicate.
- Analyze variations from a set of maps

Illustrate variation between the counties from $\nu_i$, the county-level random effect.
Mean choropleth map of random effect

Figure: Mean map of $\hat{\nu}_i$

- Posterior means of county-level random effects, $\nu_i$, are partitioned into quintiles
- Each quintile is color coded
Provide 100(1 − a)% regions for each of \( \nu_i \):
\[
\nu_i(L) < \nu_i < \nu_i(U)
\]
Construct lower choropleth map of \( \nu_i(L) \) and \( \nu_i(U) \)

1. Denote the posterior MCMC sample by
\[
\{\nu_i^{(t)} : i = 1, \ldots, M, t = 1, \ldots, T\}
\]
2. Order \( \{\nu_i^{(t)} : t = 1, \ldots, T\} \) separately for each \( i \) to obtain order statistics \( x_i^{[t]} \) and ranks \( r_i^{(t)} \)
3. For fixed \( k \in \{1, \ldots, T\} \), let \( t^* \) be the smallest integer such that
\[
x_i^{[T+1-t^*]} \leq x_i^{(t)} \leq x_i^{[t^*]}, \text{ for all } i, \text{ for at least } k \text{ values of } t
\]
4. \( t^* = k \text{th order statistic from the set } a^{(t)} = \max\{\max_i r_i^{(t)}, T + 1 - \min_i r_i^{(t)}\}, t = 1, \ldots, T, \text{ i.e. } t^* = a^{[k]}
\]
5. Then, \( \{[x_i^{[T+1-t^*]}, x_i^{[t^*]}] : i = 1, \ldots, M\} \) are a set of simultaneous credible regions containing at least 100\( k / T \)% of the distribution
Lower and Upper maps

Figure: 90% Lower, Mean, 90% Upper
Goal: Illustrate the uncertainty of the choropleth maps.

- Use $T = 1000$ MCMC replications of $\nu_i$
- Each replication can be used to produce a choropleth map.
- Each quintile is color coded.
- For each replication, the value of $i$th county is color-coded according to its quintile.
Heat map

- Rows: replicates, columns: counties
- Variation in maps is evaluated by color changes
Summary

- Present a method for posterior predictive inference that includes selection information for sampled units
- Illustrate variation of maps
Reference

Bayesian benchmarking with application to small area estimation *Test* 20 574-588

