BST 228 Project

Fuyu Guo
Qiuyue Kong
Zhuoran Wei

November 7, 2022

1 Statement of Problem and Specific Aims

Hearing loss in older populations is raising great research interest in public health for its high prevalence and serious consequences. National Health and Nutritional Examination Surveys report that among adults aged 60 to 69, about 45% report experiencing unilateral or bilateral hearing loss, with 68.1% and 89.1% among those who are 70-79 and over 80, respectively [1]. Hearing loss can cause disruptions to daily life, lower life quality [2], and negatively influence social well-being [3]. It was ranked as the third most common cause of years lived with disabilities according to the latest global burden disease 2019 study [4]. Formal audiometric testing is required for hearing loss diagnosis. However, it is expensive and time-consuming, requiring highly trained medical staff and specialized medical equipment [5]. Thus, it is more cost-effective to focus on the prevention and prediction of hearing loss. In particular, models based on commonly assessed and clinically modifiable factors such as body mass index and blood pressure can be useful in predicting individuals' risk and motivating screening and diagnosis in the early stages.

Previous studies found that hearing loss is a multifactorial disorder with underlying biological and environmental factors [6]. In addition to age, sex, and ethnicity [1, 7], modifiable factors including noise exposure and lifestyles (e.g., smoking, drinking, and diet) showed correlations with age-related hearing loss [8]. Comorbidities including cardiovascular diseases and type-2 diabetes are also predictive of hearing loss among older adults [9, 10]. However, previous work focused more on exploring potential risk factors, while less on hearing loss prediction. In this study, we plan to use household survey data to predict hearing loss risks. Identifying the population at high risk can help implement prevention and intervention strategies at an early stage, improving the prognostics. Using commonly measured socioeconomic variables frees practitioners from collecting detailed clinical information and allows them to apply the prediction model on a larger group of people.

In this study, we will employ a nationwide dataset, the American Community Survey data (ACS), with 8,798 subjects aged 65 or older [11]. We will build prediction models from over 20 potential covariates. Variable selection is needed since current domain expertise is insufficient for making accurate prediction. Moreover, most existing studies try to find a single best predictive model. For example, Least Absolute Shrinkage and Selection Operator (LASSO) regression is the most widely-used frequentist approach. However, it ignores the uncertainty about which variables to include in the model among all p potential variables [12]. Compared to frequentist LASSO, Bayesian modeling averaging (BMA) may have better prediction performance because it takes into account of model uncertainty. Additionally, Bayesian approach produces more easily interpretable results. The aims of our project include:

1. Identify significant predictors of hearing loss in American Community Survey data (2014 to 2016), using Bayesian model averaging along with Frequentist LASSO. Candidate predictors are age, sex, ethnicity, marital status, educational experience, insurance coverage, work status, income level, and veteran status.

2. Explore the influence of different priors on variable selection. We will try Zellner's g-prior, hyper-g prior, and EB-local in BMA.

3. Assess and compare the prediction performance among BMA with different priors and frequentist LASSO. We will use a separate testing dataset to compare these models' mean squared error (MSE) and area under the curve (AUC).

2 Methods

2.1 Data

We use a sample dataset from the 2014 to 2016 American Community Survey (ACS), which is a demographics survey conducted by the U.S. Census Bureau each year. We adopt the dataset provided by Dr. Lori Chibnik for education training use. Individuals in the dataset are older than the general population, with a mean age of around 75 years. Although the dataset includes individuals from 3 ACS rounds, it doesn't provide time information. Besides, the study is not designed to be a longitudinal study, so survival analysis is not applicable. It is likely to include repeated measurements for the same individuals in the data by chance. However, such a likelihood is relatively small, and thus, the dataset can be perceived as a cross-sectional study.

Hearing loss was defined as self-reported deaf or serious hearing difficulties. Socioeconomic information was collected, including sex, age, marital status (married, divorced, widowed, divorced, and never married), race (White, Black, Asian, and others), Hispanic ethnicity, health insurance coverage, educational experience (less than high school, high school, and college or greater), work hours per week, personal income, family income, wage and salary income, veteran status, having grandchildren in the house or not, and working status in the last year. For now, we exclude observations with missing values but may consider Bayesian methods to deal with missingness in the future.

After excluding missing values, we split data into a training dataset (N = 7038, 80%) and a testing dataset (N = 1760, 20%) randomly. We will use the training dataset to fit models and use the testing dataset to compare models.

2.2 Approach

We aim to model the risks of hearing loss with the collected variables in the data. The functional form of the model is unclear, and we rely on two variable selection approaches, BMA and LASSO. In order to achieve our goal, we will follow the steps below:

1. Choose priors for BMA.
2. Fit models in the training dataset, using BMA with different priors and LASSO.
3. Predict risks in the testing dataset with models from the training dataset.
4. Calculate and compare the MSE and AUC of each model.

2.3 Data types, likelihood, model, and prior selection

We will use logistic regressions to model the risk of hearing loss, with the binary outcome of hearing loss. The 23 predictors mentioned in the data section are categorical/continuous variables. We use Y_i to denote the outcome for subject i ($i = 1, 2, \ldots, n$), which can be characterized by the Bernoulli distribution. The expected value of Y_i is the probability of having hearing loss, denoted by π_i.

$$Y_i|\pi \sim Bern(\pi_i), \quad E(Y_i) = \pi_i. \tag{1}$$

The link function $g(\cdot)$ in the logistic regression is the logit link:

$$g(\boldsymbol{\pi}) = log(\frac{\boldsymbol{\pi}}{1-\boldsymbol{\pi}}) = \alpha \mathbf{1_n} + \mathbf{X}\boldsymbol{\beta}, \tag{2}$$

where α is the intercept, $\mathbf{X}$ is a $n \times p$ covariate matrix with p possible predictors in total, and $\boldsymbol{\beta}$ is the $p \times 1$ column vector of regression coefficients. Accordingly, there are 2^p possible sampling models.

Bayesian model averaging (BMA) is based upon the idea that the posterior predictive distribution of the parameter of interest is an average of the posterior distribution of all candidate models weighted by the models' posterior probabilities given the data [13]. Each logistic regression model (2) in the model space $\boldsymbol{\mathcal{M}}$ can be denoted as M_j ($j = 1, \ldots, 2^p$),

$$M_j : g(\boldsymbol{\pi}) = \alpha \mathbf{1_n} + \mathbf{X}_j\boldsymbol{\beta}_j, \tag{3}$$

where $\mathbf{X}_j$ is a $n \times p_j$ matrix of p_j predictors, and $\boldsymbol{\beta}_j$ is the $p_j \times 1$ column vector of corresponding regression coefficients. We can also express $\boldsymbol{\beta}_j$ as $\boldsymbol{\gamma\beta}$, where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_p)^T$ is a p-dimensional indicator vector. If $\gamma_j = 1$, then the predictor β_j is included in the model and $\gamma_j = 0$ otherwise.

Under a Bayesian framework, the Zellner g-prior [14] for $\boldsymbol{\beta}_j$ is

$$\boldsymbol{\beta}_j|\sigma^2, M_j \sim N(\mathbf{0}, gn(\mathbf{X}_j^\top \mathbf{X}_j)^{-1}). \tag{4}$$

The user-specified value g determines the prior variance of the regression parameter $\boldsymbol{\beta}_j$. There are a number of options for setting the value of g. In this study, we particularly focus attention to $g = \sqrt{n}$, where n is the sample size [12]. In this case, the prior sample size would be $\sqrt{n}$, which is an intermediate choice between the unit information prior (UIP) with $g = n$ and sample size of 1, and the default prior with $g = 1$ and sample size of n [15]. This Zellner g-prior has been found to have similar good performance as LASSO.

Alternatively, rather than assigning a value to g, we can estimate a different g for each model M_j based on the data, which is the Empirical Bayes (EB)–local method [16]. The corresponding estimate of g is the non-negative maximum marginal likelihood estimate:

$$\widehat{g}^{EBL} = max(F_j - 1, 0), \tag{5}$$

where

$$F_j = \frac{R_j^2/p_j}{(1 - R_j^2)/(n - 1 - p_j)} \tag{6}$$

is the F statistics for testing $\boldsymbol{\beta}_j = 0$ and R_j^2 is the coefficient of determination from model M_j [17].

Furthermore, we can adopt a fully Bayesian approach, such as the hyper-g method [17], which specifies a prior on g, or the corresponding shrinkage factor $g/(1 + g)$. For example,

$$\frac{g}{1+g} \sim Beta(1, \frac{a}{2} - 1), \tag{7}$$

in which $a \in (2, 4]$.

We decided to employ the three methods mentioned above because they are found to be the best performing among all other popular methods of BMA [13]. After specifying priors on $\boldsymbol{\theta}_j = (\alpha, \boldsymbol{\beta}_j, g) \in \boldsymbol{\Theta}_j$, we can obtain the marginal likelihood of the data under model M_j:

$$P(\boldsymbol{Y}|M_j) = \int_{\boldsymbol{\Theta}_j} P(\boldsymbol{Y}|\boldsymbol{\theta}_j, M_j) P(\boldsymbol{\theta}_j|M_j) d\boldsymbol{\theta}_j \tag{8}$$

The posterior probability of each candidate model is

$$P(M_j|\boldsymbol{Y}) = \frac{P(\boldsymbol{Y}|M_j)P(M_j)}{\sum_{j=1}^{2^p} P(\boldsymbol{Y}|M_j)P(M_j)}, \tag{9}$$

in which $P(M_j)$ is the prior probability for the model M_j [17]. The posterior predictive distribution is

$$P(\tilde{y}|\boldsymbol{Y}) = \int_{\boldsymbol{\Theta}} P(\tilde{y}|\boldsymbol{\theta}, \boldsymbol{Y}) P(\boldsymbol{\theta}|\boldsymbol{Y}) d\boldsymbol{\theta} \tag{10}$$

On the other hand, in the frequentist paradigm, model selection is usually achieved via the least absolute shrinkage and selection operator (LASSO) [18], which finds $\boldsymbol{\beta} = \{\beta_j\}$ by minimizing

$$\sum_{i=1}^{N} (y_i - \sum_j x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|. \tag{11}$$

LASSO has several weakness. For example, it is known to over-shrink the true significant coefficients to zero. In addition, if there are highly correlated covariates in the model, LASSO selection tends to be unstable since it will select one from the group of correlated variables [13].

In sum, we will compare 4 models:

1. BMA with $g = \sqrt{n}$
2. BMA with a estimated g through EB-local method
3. BMA with a hyper-g, $\frac{g}{1+g} \sim Beta(1, \frac{3}{2} - 1)$
4. LASSO

2.4 Metrics of interest

We will compare the models' performance in the test dataset. The models will be ranked based on their calibration and discrimination properties. We will use Brier Score, which is equivalent to mean squared error, to measure the model calibration [19],

$$Brier\ Score = \frac{\sum_{i=1}^{N}(E_i - Y_i)^2}{N} \tag{12}$$

where E_i is the predicted hearing loss risk for individual i.

We use the area under the receiver-operating characteristic curve (AUC) to measure models' discrimination. AUC is equivalent to the concordant statistic (C-statistic). A guess of 0.5 probability of getting hearing loss will yield an AUC = 0.5. The ideal prediction model will achieve both a high brier score and a high AUC. Secondary parameters of interest are $\boldsymbol{\gamma}$, i.e. which variables are included in the optimal model before averaging.

2.5 Prior setting, posterior calculation, and sample computation

As mentioned beforehand, we will use three g settings in the prior distributions of $\boldsymbol{\beta}$. The prior distribution for model weights is $Beta\text{-}Binomial(1, 10)$. We will use the R package *BMA* to calculate the posterior model probabilities (weights) and the estimation of $\boldsymbol{\beta}|\boldsymbol{Y}, M_j$. We choose the Markov chain Monte Carlo (MCMC) for the computation since a total of 23 parameters need to be estimated and Bayesian adaptive sampling is depreciated. We use a sample of 1000 individuals to fit BMA, with $g = \sqrt{n}$, EB-local g, and hyper-g and attach the results as supplementary files.

2.6 Estimation and inference of parameters

We would like to identify significant predictors and estimate effects of each of the predictors. The first part can be achieved by measuring the marginal posterior inclusion probabilities (PIP) of the candidate predictors: variables with high inclusion probabilities (e.g., greater than 0.7) are generally important and should be included in the variable selection. In the *BAS* package, we are able to get a list of the top 5 posterior models with the zero-one indicators for variable inclusion.

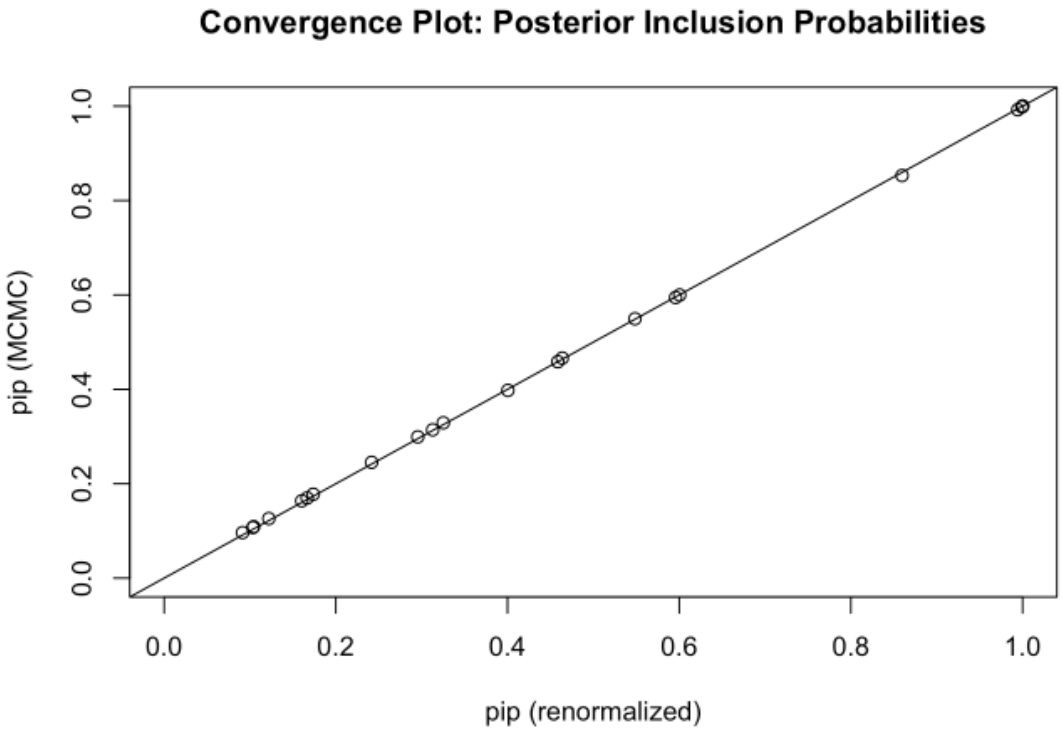
Besides, we also aim to estimate the effects of beta coefficients. Since we use the logistic regression with binary outcomes, the estimate of coefficients can be interpreted as the log odds ratio of hearing loss in exposed versus unexposed for a given binary/categorical variable, on average. As for a continuous covariate, it can be interpreted as the log odds ratio of hearing loss in those 1-unit increase versus not, on average. Large fluctuation in a certain beta parameter indicates the model uncertainty. Credible intervals at a certain significance level for coefficients can also be obtained in order to identify significant beta estimations.

2.7 Model diagnostics and convergence

A plot of residuals and fitted values under Bayesian Model Averaging can help us with the standard diagnostic checking. If model assumptions hold, we will not see outliers or non-constant variance in the residual plot.

We may create both visual and numerical diagnostics of the MCMC simulation. If all the coefficients converge to some ranges with parallel chains in the trace plots, the mixing will be good. The R-hat values that are very close to 1 indicates that the chains are stable, mixing quickly, and behaving much like an independent sample.

Noted that from the *BAS* package, we can only easily extract posterior inclusion probabilities (PIP) for each predictor and posterior model probabilities. The package developer suggests comparing the PIP/posterior model probabilities estimated by MCMC and by using the re-normalized posterior odds from sampled models. A line of $MCMC = renormalize$ indicated good convergence. Below we include a diagnostic plot showing PIP with a sample of 1000 individuals from the training dataset using $g = \sqrt{n}$ as the prior.


Convergence Plot: Posterior Inclusion Probabilities
pip (MCMC)
0.0
0.2
0.4
0.6
0.8
1.0
pip (renormalized)
0.0
0.2
0.4
0.6
0.8
1.0

References

[1] F. R. Lin, J. K. Niparko, and L. Ferrucci, "Hearing loss prevalence in the united states," *Archives of internal medicine*, vol. 171, no. 20, pp. 1851–1853, 2011.

[2] D. S. Dalton, K. J. Cruickshanks, B. E. Klein, R. Klein, T. L. Wiley, and D. M. Nondahl, "The impact of hearing loss on quality of life in older adults," *The gerontologist*, vol. 43, no. 5, pp. 661–668, 2003.

[3] D. Jung and N. Bhattacharyya, "Association of hearing loss with decreased employment and income among adults in the united states," *Annals of Otology, Rhinology & Laryngology*, vol. 121, no. 12, pp. 771–775, 2012.

[4] L. M. Haile, T. Bärnighausen, and J. B. Jonas, "Hearing loss prevalence and years lived with disability, 1990-2019," 2021.

[5] A. Bagai, P. Thavendiranathan, and A. S. Detsky, "Does this patient have hearing impairment?," *Jama*, vol. 295, no. 4, pp. 416–428, 2006.

[6] M. R. Bowl and S. J. Dawson, "Age-related hearing loss," *Cold Spring Harbor perspectives in medicine*, vol. 9, no. 8, p. a033217, 2019.

[7] P. M. Lantos, G. Maradiaga-Panayotti, X. Barber, E. Raynor, D. Tucci, K. Hoffman, S. R. Permar, P. Jackson, B. L. Hughes, A. Kind, and G. K. Swamy, "Geographic and racial disparities in infant hearing loss," *Otolaryngology– Head and Neck Surgery*, vol. 159, no. 6, pp. 1051–1057, 2018.

[8] P. Dawes, K. J. Cruickshanks, D. R. Moore, M. Edmondson-Jones, A. McCormack, H. Fortnum, and K. J. Munro, "Cigarette smoking, passive smoking, alcohol consumption, and hearing loss," *Journal of the Association for Research in Otolaryngology*, vol. 15, no. 4, pp. 663–674, 2014.

[9] P. Mitchell, B. Gopinath, C. M. McMahon, E. Rochtchina, J. Wang, S. Boyages, and S. Leeder, "Relationship of type 2 diabetes to the prevalence, incidence and progression of age-related hearing loss," *Diabetic Medicine*, vol. 26, no. 5, pp. 483–488, 2009.

[10] J. W. Hong, J. H. Jeon, C. R. Ku, J. H. Noh, H. J. Yoo, and D.-J. Kim, "The prevalence and factors associated with hearing impairment in the korean adults: the 2010–2012 korea national health and nutrition examination survey (observational study)," *Medicine*, vol. 94, no. 10, 2015.

[11] K. Tehranchi and A. Jeyakumar, "Hearing loss's incidence and impact on employment in the united states," *Otology & Neurotology*, vol. 41, no. 7, pp. 916–921, 2020.

[12] C. Fernandez, E. Ley, and M. F. Steel, "Benchmark priors for bayesian model averaging," *Journal of Econometrics*, vol. 100, pp. 381–427, 2001.

[13] A. Porwala and A. E. Raftery, "Comparing methods for statistical inference with model uncertainty," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 119, no. 16, 2022.

[14] A. Zellner, "On assessing prior distributions and bayesian regression analysis with g-prior distributions," *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, vol. 6, pp. 233–243, 1986.

[15] E. v. Zwet, "A default prior for regression coefficients," *Statistical Methods in Medical Research*, vol. 28, no. 12, pp. 3799–3807, 2019.

[16] E. I. George and D. P. Foster, "Calibration and empirical bayes variable selection," *Biometrika*, vol. 87, no. 4, pp. 731–747.

[17] F. Liang, R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger, "Mixtures of g priors for bayesian variable selection," *Journal of the American Statistical Association*, vol. 103, no. 481, pp. 410–423, 2008.

[18] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society*, vol. 73, pp. 273–282, 2011.

[19] Y. Huang, W. Li, F. Macheret, R. A. Gabriel, and L. Ohno-Machado, "A tutorial on calibration measurements and calibration models for clinical prediction models," *Journal of the American Medical Informatics Association*, vol. 27, no. 4, pp. 621–633, 2020.