A Bayesian hierarchical model for categorical data with nonignorable nonresponse

Paul E. Green, Taesung Park
Seoul National University, Department of Statistics,
San 56-1 Shin Lim-Dong Kwanak-gu,
Seoul 151-742, Korea
pgreen@stats.snu.ac.kr

1. Introduction

Log-linear models can be used to adjust for nonresponse when categorical data are subject to nonignorable nonresponse. The data are augmented by a latent indicator variable that designates whether subjects are respondents or nonrespondents. Maximum likelihood estimates calculated from the augmented data table have been shown to suffer from instability due to boundary problems. Specification of a prior distribution over cell parameters induces smoothing away from the boundary, leading to estimates that tend to compromise between the data and estimates based on the log-linear model. Park and Brown (1994) and Park (1998) developed empirical Bayes models by assuming a multinomial likelihood for the observed data and a conjugate Dirichlet prior for the cell parameters. Estimates for the nonrespondents, based on a log-linear model, were calculated by maximizing the resulting posterior distribution using an EM algorithm.

In this work we consider a Bayesian hierarchical model. Since there is uncertainty in the variable selection process associated with the log-linear model, a finite number of models can be considered simultaneously using Bayesian model averaging and Markov chain Monte Carlo (MCMC) simulation. A stochastic search variable selection (SSVS) procedure due to George and McCulloch (1996) is applied such that the most promising models are visited most frequently.

2. Model

The model is comprised of a likelihood and a three-stage prior. Let Y be the response variable, indexed by j, that may be missing, and let R be a latent indicator variable, indexed by k, that designates whether a subject is a respondent ($k = 1$ corresponds to a response and $k = 2$ corresponds to no response). Denote by X an s-dimensional explanatory variable that is always observed and indexed by $\boldsymbol{i} = (i_1, i_2, \ldots, i_s)$. The levels of Y and X are denoted by J and $I = I_1 \times \ldots \times I_s$, respectively. Let $N = (I \times J \times 2)$ be the total number of cells in the augmented contingency table.

The observed cell frequencies for the respondents for which $k = 1$, denoted by $y_{\boldsymbol{i}j1}$, are Poisson, such that the log of the Poisson mean is $\eta_{\boldsymbol{i}j1}$,

$$y_{\boldsymbol{i}j1} \,|\, \eta_{\boldsymbol{i}j1} \sim \text{Poisson}(\exp(\eta_{\boldsymbol{i}j1})).$$

Conditional on marginal totals $y_{\boldsymbol{i}+2}$ summed over the nonresponses, the unobserved cell frequencies for the nonrespondents for which $k = 2$, denoted by $y_{\boldsymbol{i}j2}$ are multinomial

$$y_{\boldsymbol{i}12}, \ldots, y_{\boldsymbol{i}J2} \,|\, \sum_j y_{\boldsymbol{i}j2} = y_{\boldsymbol{i}+2}, \pi_{\boldsymbol{i}12}, \ldots, \pi_{\boldsymbol{i}J2} \;\sim\; \text{Multinomial}(y_{\boldsymbol{i}+2}; \pi_{\boldsymbol{i}12}, \ldots, \pi_{\boldsymbol{i}J2})$$

under the usual constraints for multinomial sampling

$$\sum_j \pi_{\boldsymbol{i}_{j2}} = 1 \qquad \text{and} \qquad \pi_{\boldsymbol{i}_{j2}} = \frac{\exp(\eta_{\boldsymbol{i}_{j2}})}{\sum_j \exp(\eta_{\boldsymbol{i}_{j2}})} = \frac{\exp(\eta_{\boldsymbol{i}_{j2}})}{y_{\boldsymbol{i}_{+2}}} .$$

At the first prior a log-linear model is induced by allowing the log of expected cell frequencies $\eta_{\boldsymbol{i}_{jk}}$ over the entire table to be normal

$$\boldsymbol{\eta} \,|\, \boldsymbol{\beta}, \Sigma \sim N(Z\boldsymbol{\beta}, \Sigma) ,$$

where $\boldsymbol{\beta}$ is $p \times 1$, Σ is an $N \times N$ diagonal covariance matrix, and Z is the $N \times p$ design matrix. The covariance matrix Σ has only two values, σ_1^2 and σ_2^2, corresponding to $\eta_{\boldsymbol{i}_{j1}}$ and $\eta_{\boldsymbol{i}_{j2}}$, respectively, since we expect $\sigma_1^2 < \sigma_2^2$.

At the second prior $\boldsymbol{\beta}$ is assigned a p-variate normal in accordance with the procedures of stochastic search variable selection (SSVS) outlined in George and McCulloch (1996), and σ_1^2 and σ_2^2 follow conjugate inverse gamma distributions. In particular,

$$\boldsymbol{\beta} \,|\, \boldsymbol{\gamma} \sim N_p(0, D_\gamma^2) , \quad \sigma_1^2 \sim \text{Inv-Gamma}(\nu_1/2, \nu_1\lambda_1/2) , \quad \sigma_2^2 \sim \text{Inv-Gamma}(\nu_2/2, \nu_2\lambda_2/2)$$

where D_γ is a diagonal matrix with elements, say 10^3 when $\beta_l = 0$ is not being tested, and elements $[(1-\gamma_l)\tau_l + \gamma_l\, c_l\, \tau_l] , l = 1, \ldots m < p$, when $\beta_l = 0$ is being tested. The parameters $(\tau_l, c_l, \nu_1, \lambda_1, \nu_2, \lambda_2)$ are assumed known and fixed in advance. The binary hyperparamer $\gamma_l \in \{0, 1\}$. The idea is that τ_l is chosen small and c_l is chosen large so that β_l is different from zero and is included in the model if $\gamma_l = 1$.

At the final stage $\gamma_l \sim$ Bernoulli(0.5) and independent, giving $p(\boldsymbol{\gamma}) = 0.5^m$.

3. Methods

The posterior distribution can be calculated as $p(\boldsymbol{\eta}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma_1^2, \sigma_2^2, y_{\boldsymbol{i}_{j2}} \,|\, y_{\boldsymbol{i}_{j1}}, y_{\boldsymbol{i}_{+2}})$. Then it can be shown that the conditional for $\boldsymbol{\beta}$ is normal, the conditional for σ_1^2 is Inv-Gamma, the conditional for σ_2^2 is Inv-Gamma, the conditionals for each γ_l is Bernoulli, and the conditionals for $(y_{\boldsymbol{i}_{12}}, \ldots, y_{\boldsymbol{i}_{J2}})$ are multinomial. The conditionals for $\boldsymbol{\eta}$ can be sampled using a Metropolis sampler. Interest focuses on $p(\boldsymbol{\eta} \,|\, y_{\boldsymbol{i}_{j1}}, y_{\boldsymbol{i}_{+2}})$ for allocating nonrespondents to the missing cells.

REFERENCES

George, E.I. and McCulloch,R.E. (1996). Approaches for Bayesian variable selection. Technical Report, University of Texas, Austin, http:
bevo2. bus.utexas.edu/GeorgeE/.

Park, T. (1998). An approach to categorical data with nonignorable nonresponse. *Biometrics* **54**, 1579-1590.

Park, T. and Brown, M.B. (1994). Models for categorical data with nonignorable nonresponse. *Journal of the American Statistical Association* **89**, 44-52.