Variable Selection in Logistic Regression Models by the Gibbs Sampling

Runqiu Fei

The University of Melbourne

The Variable Selection Problem

Variable selection is a critical step in statistical inference. It directly relates to the goodness and utility of statistical models, which is the base of effective inference. However, when the number of candidate variables is large, it is not computationally feasible to use exhaustive search. (e.g. When number of candidate variables = 50, exhaustive search would have to consider 2^{50} candidate models.) To solve this variable selection problem involving a large number of candidates, we referred to the method proposed by Qian and Field (2002). The key idea of the method is using random sampling to obtain a sample of the candidate models (which is expected to be much smaller than the population), then choose the sample best model as the estimate of the true best model. In this poster, we mainly focus on testing the effectiveness of one version of this method on logistic regression model selection. We would use BIC as the model selection criteria and Gibbs Sampling to obtain the sample of candidate models. We would also consider the possible over- or under-dispersion of the data.

Logistic Regression Model

Suppose a response variable Y follows a binomial distribution $B(m, \pi)$, and the probability π may be influenced by p explanatory variables $x_1, \ldots, x_p$. Then we can model the relation as a logistic regression model:

$$\log \frac{\pi}{1-\pi} = \beta_0 + \boldsymbol{x}^t \boldsymbol{\beta} \tag{1}$$

where $\boldsymbol{x}^t = (x_1, \ldots, x_p)$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^t$. Variable selection in this case is to select a best subset of variables $\boldsymbol{x}^t_{\boldsymbol{\alpha}} \subseteq \boldsymbol{x}^t$, where $\boldsymbol{\alpha} \in \{0,1\}^p$ is an indicator vector denoting which variables to include in $\boldsymbol{x}^t_{\boldsymbol{\alpha}}$. Then, the model based on $\boldsymbol{x}^t_{\boldsymbol{\alpha}}$ should be:

$$\log \frac{\pi}{1-\pi} = \beta_0 + \boldsymbol{x}^t_{\boldsymbol{\alpha}} \boldsymbol{\beta}_{\boldsymbol{\alpha}} \tag{2}$$

where $\boldsymbol{\beta}_{\boldsymbol{\alpha}} \subseteq \boldsymbol{\beta}$. It is easy to see that each candidate model can be represented by a unique indicator vector $\boldsymbol{\alpha}$. Then, we can represent the set of all candidate models as $\mathcal{A} = \{0,1\}^p$.

The Dispersion Parameter

In real-world, the response variable Y may not exactly follow the assumed binomial distribution $B(m, \pi)$. Specifically, Y may be over- or under-dispersed. In this case, we say that Y follows a quasi-binomial distribution, with

$$\mathbb{E}(Y) = m\pi, \qquad \text{Var}(Y) = \phi m \pi (1-\pi)$$

where ϕ is the dispersion parameter (McCullagh & Nelder, 1989). Note that ϕ actually characterises the distance between the true distribution of Y and the assumed distribution $B(m, \pi)$. The closer ϕ is to 1, the closer Y will be to $B(m, \pi)$ in distribution, which actually indicates a better model fit.

Model Selection Criteria - BIC

The Bayesian Information Criteria (BIC) proposed by Schwarz (1978) is commonly used to access model utility. In the case of our logistic regression models, we would like to adjust the original BIC with the estimated dispersion parameter $\hat{\phi}$, and use the adjusted BIC in the following form:

$$BIC(\boldsymbol{\alpha}|Y, \boldsymbol{x}_{\boldsymbol{\alpha}}) = -\ell(Y, \boldsymbol{x}_{\boldsymbol{\alpha}}|\boldsymbol{\alpha}) + \frac{1}{2}(1 + |\hat{\phi} - 1|) p_{\boldsymbol{\alpha}} \log N \tag{3}$$

where $\ell(Y, \boldsymbol{x}_{\boldsymbol{\alpha}}|\boldsymbol{\alpha})$ is the maximized log-likelihood function; $(1 + |\hat{\phi} - 1|)$ is the adjust scale; $p_{\boldsymbol{\alpha}}$ is the number of parameters in model $\boldsymbol{\alpha}$; and N is the total number of observations from $(Y, \boldsymbol{x}_{\boldsymbol{\alpha}})$. We would consider the model $\boldsymbol{\alpha}^* \in \mathcal{A}$ with the smallest $BIC(\boldsymbol{\alpha}^*|Y, \boldsymbol{x}_{\boldsymbol{\alpha}^*})$ as the best candidate model.

Random Search by Gibbs Sampling

Referring to Qian and Field (2002), we define the following probability distribution for model $\boldsymbol{\alpha}$ on $\mathcal{A}$ with

$$P(\boldsymbol{\alpha}) = B \exp\{-BIC(\boldsymbol{\alpha}|Y, \boldsymbol{x}_{\boldsymbol{\alpha}})\} \tag{4}$$

where $B = (\sum_{\boldsymbol{\alpha} \in \mathcal{A}} \exp\{-BIC(\boldsymbol{\alpha}|Y, \boldsymbol{x}_{\boldsymbol{\alpha}}\})^{-1}$. By this definition, the best model $\boldsymbol{\alpha}^*$ would have the highest probability $P(\boldsymbol{\alpha}^*)$. Therefore, if we generate a random sample based on this probability distribution, it is highly likely that $\boldsymbol{\alpha}^*$ will appear in the sample and it will appear early. In this case, to find the best model, we only need to search the moderate size sample, rather than doing the computationally infeasible exhaustive search.

At this stage, the key is to find a way to generate random samples. Though we cannot directly compute the constant B, we can still find the following conditional distribution:

$$P(\alpha_i|\alpha_1, \alpha_2, ..., \alpha_{i-1}, \alpha_{i+1}, ..., \alpha_p) = \frac{P(\boldsymbol{\alpha})}{P(\boldsymbol{\alpha}, \alpha_i = 0) + P(\boldsymbol{\alpha}, \alpha_i = 1)} \tag{5}$$

where α_i denotes the i-th term of the indicator vector $\boldsymbol{\alpha}$. We can see that this is a Bernoulli distribution.

Therefore, we can use Gibbs sampling (Casella & George, 1992) to generate a random sample of candidate models, with sample size = K, as the following steps:

1. Choose a random starting model, represented by indicator vector $\boldsymbol{\alpha}^{(0)}$. (e.g. $\boldsymbol{\alpha}^{(0)} = (1, 1, ..., 1)$, a $1 \times p$ indicator vector.)
2. Repeat for $j = 1, 2, ..., K$: generate model $\boldsymbol{\alpha}^{(j)}$, where $\alpha_i^{(j)}$ (i.e. the i-th term of $\boldsymbol{\alpha}^{(j)}$) is generated from the Bernoulli distribution:
$$P(\alpha_i|\alpha_1^{(j)}, \alpha_2^{(j)}, ..., \alpha_{i-1}^{(j)}, \alpha_{i+1}^{(j-1)}, ..., \alpha_p^{(j-1)}), \qquad i = 1, 2, ..., p$$
3. Return the generated sequence of models $\{\boldsymbol{\alpha}^{(1)}, \boldsymbol{\alpha}^{(2)}, ..., \boldsymbol{\alpha}^{(K)}\}$.

This generated sequence is actually a Markov chain, with chain-size = K. Normally, if a Markov chain is ergodic, it will become stationary and can be used as a random sample, after its burn-in sequence (Robert & Casella, 2004). However, we would not worry too much about the burn-in sequence in our study. In fact, from our real data application in the next section, we found out that the best candidate models appear fairly early in the Markov chain. Also, there is evidence showing that the chain may become stationary at a fairly early stage.

An Application with Real-world Data

To test the effectiveness of the random search, we would use a real-world data set originally published by Schoener (1970). It recorded the habitat (perch) preferences of two species of lizard, $grahami$ (G) and $opalinus$ (O). Specifically, it recorded the number of G's and the number of O's on a particular perch. The height (H) and diameter (D) of the perch were also recorded, together with whether the perch was sunny or shaded (S) and whether it was at early, middle ($T_1 = 1$) or late ($T_2 = 1$) of a day. Let π be the probability of G occupying a random observed perch, then we consider the main effects and the first-order interactions to obtain the following full logistic model:

$$\begin{aligned}\log \frac{\pi}{1-\pi} = \beta_0 + \beta_1 H + \beta_2 D + \beta_3 S + \beta_4 T_1 + \beta_5 T_2 + \beta_6 H \cdot D + \beta_7 H \cdot S + \beta_8 H \cdot T_1 + \beta_9 H \cdot T_2 \\ + \beta_{10} D \cdot S + \beta_{11} D \cdot T_1 + \beta_{12} D \cdot T_2 + \beta_{13} S \cdot T_1 + \beta_{14} S \cdot T_2\end{aligned}$$

As there are 14 candidate variables, there will be $2^{14} = 16384$ candidate models in total. For testing purposes, we use exhaustive search to rank the candidate models by their BIC values, and also derive the theoretical (true) probability mass function (pmf) of the models. The possible over- or under-dispersion of the data is also being considered. Then, we use the Gibbs sampling to generate a Markov chain of size 1000. The results and discussions are as follows.

Table 1 lists the 5 best models with the smallest BIC values. With $\hat{\phi}$ very close to 1 (i.e. $|\hat{\phi} - 1| < 0.1$), the models also fit the data well in dispersion. We have recorded these best models' first appearance positions in the generated Markov chain. It turns out that they all appear in the first 200 runs of the chain, which is fairly early. Typically, the top-1 model appears at the 34-th position of the chain. Moreover, these best models also appear frequently in the chain. We can see that the total frequency of the best 2 models = $0.373 + 0.210 = 0.583$ > 50%.

Table 1: The 5 best models and their behavior in random search

Rank	Model	BIC	$\hat{\phi}$	First-appear	Frequency
1	$H + D + T_2 + H \cdot S$	23.988	0.982	34-th	0.373
2	$H + D + T_2$	24.510	0.932	35-th	0.210
3	$H + D + T_2 + H \cdot S + H \cdot T_1$	26.706	0.995	69-th	0.029
4	$H + D + T_2 + S \cdot T_2$	26.873	0.976	133-th	0.018
5	$H + D + T_2 + H \cdot D + H \cdot S$	26.920	0.993	37-th	0.023

Figure 1 shows the closeness between the true pmf of the models and the estimated pmf by the model frequencies (e.g. see Table 1) in the Markov chain. This actually supports that the Markov chain becomes stationary at a fairly early stage, within 1000 runs.

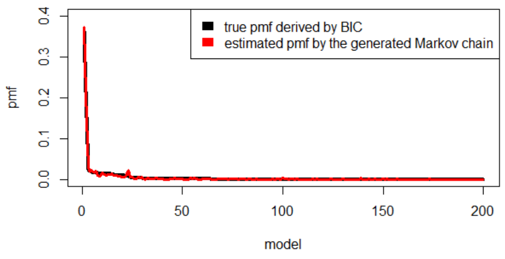

Fig. 1: pmf of the best 200 models

Based on the generated Markov chain of size 1000, we obtained the relative frequencies of the 14 variables. Table 2 shows the 5 most important variables with the highest relative frequencies. We can understand this frequency of variable in the following way: relative-frequency$(D) = 0.999$ means that 99.9% of the models in the Markov chain include variable D. We can see from Table 1 and Figure 1 that the best models appear more frequently in the chain, so their corresponding variables (i.e. the most important variables) should also appear more frequently. Therefore, the relative frequency of a variable actually indicates its importance.

Table 2: The 5 most important variables and their relative frequencies

Variable	D	H	T_2	$H \cdot S$	S
Frequency	0.999	0.997	0.905	0.536	0.105

If we consider these variable relative-frequencies as proportions, we can use normal approximation to construct confidence intervals (CI). We further define that a variable is *significant* when the CI of its relative-frequency is above 0.5. Then we find only the following set of variables are *significant*: $\{D,\ H,\ T_2,\ H \cdot S\}$. Unsurprisingly, these *significant* variables all appear in the best 5 models in Table 1. Interestingly, the top-1 model only consists of these *significant* variables.

Conclusion

In this poster, we have studied and examined the variable selection or model selection method proposed by Qian and Field (2002). We have employed BIC and the Gibbs sampling as tools for selection, and we have also considered the effect of over- or under-dispersion of the data. Applying the method with real-world data, we found that the method is efficient and effective, not only in selecting out the best candidate models, but also in filtering out the most important variables. Also, it might be useful to employ the dispersion parameter when applying this variable selection method.

Acknowledgements

I would like to thank my supervisor, Prof. Guoqi Qian, for all the thoughtful inspirations and strong supports. I would also like to thank the Vacation Scholarship program for offering me this valuable chance to conduct statistical research and learn from experience.

References

Casella, G., & George, E. I. (1992). Explaining the Gibbs sampler. *The American Statistician*, *46*(3), 167-174.

McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models*. London: Chapman and Hall.

Qian, G., & Field, C. (2002). Using MCMC for logistic regression model selection involving large number of candidate models. *Monte Carlo and Quasi-Monte Carlo Methods 2000*, 460-474.

Robert, C. P., & Casella, G. (2004). *Monte Carlo statistical methods* (Vol. 2). New York: Springer.

Schoener, T. W. (1970). Nonsynchronous spatial overlap of lizards in patchy habitats. *Ecology*, *51*(3), 408-418.

Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics*, *6*(2), 461-464.