Computational Statistics and Data Analysis 82 (2015) 207–222

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Variable selection in general multinomial logit models

Gerhard Tutz [a], Wolfgang Pößnecker [a,*], Lorenz Uhlmann [b]

[a] *Department of Statistics, Ludwig-Maximilians-University, 80539 München, Germany*

[b] *Department of Medical Biometry, Ruprecht-Karls-University, 69120 Heidelberg, Germany*

ARTICLE INFO

Article history:
Received 21 October 2013
Received in revised form 8 September 2014
Accepted 10 September 2014
Available online 21 September 2014

Keywords:
Logistic regression
Multinomial logit model
Variable selection
Lasso
Group Lasso
CATS Lasso

ABSTRACT

The use of the multinomial logit model is typically restricted to applications with few predictors, because in high-dimensional settings maximum likelihood estimates tend to deteriorate. A sparsity-inducing penalty is proposed that accounts for the special structure of multinomial models by penalizing the parameters that are linked to one variable in a grouped way. It is devised to handle general multinomial logit models with a combination of global predictors and those that are specific to the response categories. A proximal gradient algorithm is used that efficiently computes stable estimates. Adaptive weights and a refitting procedure are incorporated to improve variable selection and predictive performance. The effectiveness of the proposed method is demonstrated by simulation studies and an application to the modeling of party choice of voters in Germany.

1. Introduction

The multinomial logit model is the most frequently used model in regression analysis for un-ordered multi-category responses. The maximum likelihood (ML) method, which is typically used for estimation, has the drawback that it requires more observations than parameters to be estimated. The amount of parameters, however, increases rapidly when the number of predictors grows, as multinomial logit models employ several coefficients for each explanatory variable. Therefore, ML estimates tend to deteriorate quickly and interpretability suffers as well, so that the number of predictors in the model is severely limited. For these reasons, variable selection is necessary to obtain multinomial logit models that are both interpretable and reliable.

As a motivating example, our application uses data from the German Longitudinal Election Study (GLES) about party choice of voters during the 2009 parliamentary elections for the German Bundestag. Modeling the decision of voters for specific political parties and determining the major factors behind their preference are of great interest in political sciences. The available parties to choose from are the Christian Democratic Union (CDU), the Social Democratic Party (SPD), the Green Party (Bündnis 90/Die Grünen), the Liberal Party (FDP) and the Left Party (Die Linke). As explanatory variables, various individual characteristics of the voter are considered, like, for example, gender, age or education, see Section 5 for a complete list. The main goal of our analysis is to select those predictors that influence party choice and to remove the rest. Besides improving interpretability, this is beneficial for polling firms and in opinion research. If, for example, gender was found to be irrelevant for party preference, one could save time and money while performing opinion polls as one would not have to care about a representative gender ratio among the interviewed persons.

* Correspondence to: Seminar of Applied Stochastics, Department of Statistics, Ludwig-Maximilians-University Munich, Ludwigstraße 33, 80539 München, Germany.

E-mail address: wolfgang.poessnecker@stat.uni-muenchen.de (W. Pößnecker).

http://dx.doi.org/10.1016/j.csda.2014.09.009
0167-9473/

While variable selection for such individual-specific predictors requires specific methodology due to the particular structure of multinomial logit models, this dataset offers another challenge in the form of predictors that are party-specific. For various topics like immigration or nuclear energy, participants of the study were asked how they perceive the parties' stance on this issue. Additionally, they stated their personal position on the topic. From this information, the distance between the personal point of view and the perceived position of the parties can be computed. These distances are then included into the model. Since they take different values for different categories of the response, they are an example of so-called category-specific predictors. To be able to deal with the challenges of such a dataset, a method for variable selection in such general multinomial logit models is developed in this paper. It accounts for the categorical and multivariate structure of these models and works with both global and category-specific predictors.

The standard method for variable selection are forward/backward strategies which have been used for a long time, but are notoriously unstable and computationally costly, so that they cannot be recommended (see, for example, Hastie et al., 2009). An established alternative is penalty approaches for regularized variable selection. For linear and generalized linear models (GLMs), a variety of such methods has been proposed. The most prominent example is the Lasso (Tibshirani, 1996) and its extensions to Fused Lasso (Tibshirani et al., 2005) and Group Lasso (Yuan and Lin, 2006). Alternative regularized estimators that enforce variable selection are the Elastic Net (Zou and Hastie, 2005), SCAD (Fan and Li, 2001), the Dantzig selector (Candes and Tao, 2007) and boosting approaches (Bühlmann and Yu, 2003; Bühlmann and Hothorn, 2007; Tutz and Binder, 2006).

These methods, however, were developed for models with univariate response. Because the multinomial logit model is not a common univariate GLM, these methods cannot be applied directly. As mentioned previously, the effect of one predictor variable is represented by several parameters. Therefore, one has to distinguish between variable selection and parameter selection, where variable selection is only achieved if all effects/parameters that belong to one variable are simultaneously removed from the model. Early suggestions for regularization in multinomial logit models (Krishnapuram et al., 2005; Friedman et al., 2010) use L_1-type penalties that shrink all the parameters individually. Thus, they do not use the natural grouping of coefficients that is available, with each group containing the parameters that belong to the same explanatory variable. In particular, they pursue the goal of parameter selection and cannot directly promote variable selection.

Therefore, a more explicit approach to variable selection in multinomial logit models is to use a grouped penalization with groups that consist of all coefficients belonging to the same predictor variable. Such a grouped variable selection was used in general multivariate regression, among others, by Turlach et al. (2005) and Argyriou et al. (2007). The present paper extends and complements more recent work by Simon et al. (2013), Vincent and Hansen (2014) and Chen and Li (2013), who were among the first to explicitly embed the idea of the group lasso into the multinomial regression framework. However, all three of these papers only consider global predictors that are constant/independent from the observed response category. By contrast, we consider the general multinomial logit model in which a mix of global and category-specific predictors is used. Moreover, we extend our penalization approach to the case of categorical predictors with more than two categories. Since all dummy variables belonging to such a predictor should be selected jointly, this creates an additional level of grouping on top of the grouping across response categories. Additionally, the concept of adaptive weights (Zou, 2006; Wang and Leng, 2008) is adopted. We demonstrate that this distinctly improves variable selection and prediction accuracy. From an algorithmic point of view, the three aforementioned papers all use variations of a coordinate descent algorithm for the computation of numerical estimates. By contrast, we use the fast iterative shrinkage thresholding algorithm (FISTA) of Beck and Teboulle (2009), which is an accelerated version of proximal gradient methods.

This paper is organized as follows: In Section 2 we introduce the general multinomial logit model and suggest a penalty that yields proper variable selection. Extensions to incorporate category-specific variables and multi-categorical predictors both in the model and in the penalization are discussed separately. Regularized estimation is considered in Section 3, where a proximal gradient algorithm is derived that efficiently solves the corresponding estimation problem. The performance of our estimator is investigated in simulation studies in Section 4. Then, in Section 5, the real data example from the German Longitudinal Election Study is analyzed using the developed methodology.

2. Model and regularization

2.1. The multinomial logit model with category-specific covariates

For data $(y_i, \boldsymbol{x}_i),\ i = 1, \dots, n$, with y_i denoting an observation of the categorical response variable $Y \in \{1, \dots, k\}$ and $\boldsymbol{x}_i$ the p-dimensional vector of predictors, the multinomial logit model in its generic form specifies

$$\pi_{ir} = P(Y = r|\boldsymbol{x}_i) = \frac{\exp(\beta_{r0} + \boldsymbol{x}_i^T\boldsymbol{\beta}_r)}{\sum_{s=1}^{k}\exp(\beta_{s0} + \boldsymbol{x}_i^T\boldsymbol{\beta}_s)} = \frac{\exp(\eta_{ir})}{\sum_{s=1}^{k}\exp(\eta_{is})}, \quad (1)$$

where $\boldsymbol{\beta}_r^T = (\beta_{r1}, \dots, \beta_{rp})$. Since parameters $\beta_{10}, \dots, \beta_{k0}, \boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_k$ are not identifiable, additional constraints are needed. Typically, one of the response categories is chosen as reference category. We use category k as the reference category by