A unifying framework for specifying generalized linear models for categorical data

Jean Peyhardi, Catherine Trottier, Yann Guédon

A unifying framework for specifying generalized linear models for categorical data

Jean Peyhardi[1] [2], Catherine Trottier[1], Yann Guédon[2]

[1] Université Montpellier 2, I3M, Montpellier, France
[2] CIRAD, UMR AGAP and Inria, Virtual Plants, Montpellier, France

E-mail for correspondence: jean.peyhardi@math.univ-montp2.fr

Abstract: In the context of categorical data analysis, the case of nominal and ordinal data has been investigated in depth while the case of partially ordered data has been comparatively neglected. We first propose a new specification of generalized linear models (GLMs) for categorical response variables which encompasses all the classical models such as multinomial logit, odds proportional or continuation ratio models but also led us to identify new GLMs. This unifying framework makes the different GLMs easier to compare and combine. We then define the more general class of partitioned conditional GLMs for categorical response variables. This new class enables to take into account the case of partially ordered data by combining nominal and ordinal GLMs.

Keywords: categorical data analysis; generalized linear model; partitioned conditional model; recursively partitioned categories.

1 Specification of generalized linear models for categorical response variables

Let Y denote the response variable with J categories ($J > 1$) and $X = (X_1, \ldots, X_p)$ be a vector of explanatory variables in a general form (a categorical variable being represented by an indicator vector). The definition of a GLM includes the specification of a link function g which is a C^1-diffeomorphism from $M = \{(\pi_1, ..., \pi_{J-1}) \in]0,1[^{J-1} | \sum_{j=1}^{J-1} \pi_j < 1\}$ to an open subset of $\mathbb{R}^{J-1}$, between the expectation $\pi = E[Y|X{=}x] = (\pi_1, ..., \pi_{J-1})^T$ and the linear predictor $\eta = (\eta_1, ..., \eta_{J-1})^T$. All the classical link functions $g = (g_1, \ldots, g_{J-1})$, described in the literature -see Agresti (2002) and Fahrmeir and Tutz (2001)- share the same structure which we propose to write as

$$g_j = F^{-1} \circ r_j \ , \quad j = 1, \ldots, J-1,$$

where F is a continuous and strictly increasing cumulative density function (cdf) and $r = (r_1, \ldots, r_{J-1})^T$ is a C^1-diffeomorphism from M to an open

subset of $]0,1[^{J-1}$. Thus we have

$$r_j(\pi) = F(\eta_j)\ , \qquad j = 1, \ldots, J-1.$$

In the following we describe in more details the components r, F and η.

Ratio r: The linear predictor η is not directly related to the expectation π but to a particular transformation r of the vector π which we call the ratio. In the following we will consider four particular C^1-diffeomorphism. The *adjacent*, *sequential* and *cumulative* ratios are respectively defined by $\pi_j/(\pi_j+\pi_{j+1})$, $\pi_j/(\pi_j+\ldots+\pi_J)$ and $\pi_1+\ldots+\pi_j$ for $j = 1, \ldots, J-1$, assume order among categories but with different interpretations. The *reference* ratio, defined by $\pi_j/(\pi_j + \pi_J)$ for $j = 1, \ldots, J-1$, is mainly useful for nominal response variables.

Latent variable cdf F: The most commonly used symmetric distributions are the *logistic* and *Gaussian* distributions but the *Laplace* and *Student* distributions may also be useful. The most commonly used asymmetric distributions are the *Gumbel max* and *Gumbel min* distributions. Playing on the symmetrical or asymmetrical character and the more or less heavy tails may markedly improve the model fit. In applications the Student(d) distribution will be approximated by a Gaussian distribution when $d > 30$.

Linear predictor η: It can be written as the product of the design matrix Z and the vector of parameters β (Fahrmeir and Tutz, 2001). Each explanatory variable can have its own design effect. For example, if X_1 has a *global* effect, X_2 a *local* effect, ... and X_p a *global* effect, the corresponding design matrix, with $J-1$ rows, is

$$Z = \begin{pmatrix} 1 & & & x_1^T & x_2^T & & & & x_p^T \\ & 1 & & x_1^T & & x_2^T & & & x_p^T \\ & & \ddots & \vdots & & & \ddots & \cdots & \vdots \\ & & & 1 \quad x_1^T & & & & x_2^T & x_p^T \end{pmatrix}.$$

This design will be denoted by the tuple (global, local,..., global) and a single word global or local will denote the same design for all the explanatory variables $X_1, \ldots, X_p$.

Finally, we propose to specify a particular GLM for categorical response variables by the (r, F, Z) triplet with

$$r(\pi) = \mathbf{F}(Z\beta),$$

where $\mathbf{F}(\eta) = (F(\eta_1), \ldots, F(\eta_{J-1}))^T$.

This specification eases the comparison of GLMs for categorical response variables; see examples in Table 1. Moreover, it enables to define an enlarged

TABLE 1. (r, F, Z) specification of some classical GLMs for categorical response variables.

Multinomial logit model $P(Y=j) = \frac{\exp(\alpha_j + x^T\delta_j)}{1+\sum_{k=1}^{J-1}\exp(\alpha_k + x^T\delta_k)}$	(reference, logistic, local)
Odds proportional logit model $\log\left\{\frac{P(Y\leq j)}{1-P(Y\leq j)}\right\} = \alpha_j + x^T\delta$	(cumulative, logistic, global)
Proportional hazard model (Grouped Cox Model) $\log\{-\log P(Y>j)\} = \alpha_j + x^T\delta$	(cumulative, Gumbel min, global)
Adjacent logit model $\log\left\{\frac{P(Y=j)}{P(Y=j+1)}\right\} = \alpha_j + x^T\delta_j$	(adjacent, logistic, local)
Continuation ratio logit model $\log\left\{\frac{P(Y=j)}{P(Y>j)}\right\} = \alpha_j + x^T\delta_j$	(sequential, logistic, local)

set of GLMs for nominal response variables by $\{(\text{reference}, F, Z)\}$ triplets, which includes the multinomial logit model. GLMs for nominal and ordinal response variables are usually defined with different design matrices Z; see the first two rows in Table 1. Fixing the desing matrix Z may ease the comparison of GLMs for nominal and ordinal response variables.

Finally, a single estimation procedure based on Fisher scoring algorithm can be applied to all the GLMs specified by (r, F, Z) triplets. Using the chain rule, the score function can be separated into two parts where the first depends on the triplet (r, F, Z), whereas the second does not.

$$\frac{\partial l}{\partial \beta} = \underbrace{Z^T \frac{\partial \mathbf{F}}{\partial \eta}\frac{\partial \pi}{\partial r}}_{(r,F,Z)\text{ dependant part}} \underbrace{\text{Cov}(Y|X=x)^{-1}\,[y-\pi]}_{(r,F,Z)\text{ independent part}}.$$

2 Partitioned conditional GLMs for categorical response variables

The main idea is to recursively partition the J categories and then to specify a GLM for each partition. Such combinations of GLMs have already been proposed such as the two-step model of Morawitz and Tutz (1990), that combines sequential and cumulative models, or the partitioned conditional model for partially ordered set (POS-PCM) of Zhang and Ip (2012) that combines multinomial logit and odds proportional logit models. Our proposal can be seen as a generalization of POS-PCMs that benefits from the genericity of the (r, F, Z) specification. In particular, our objective was not only to propose GLMs for partially-ordered response variables but also to differentiate the role of explanatory variables for each partition of categories using for instance different design matrices.

Definition: Let $J \geq 2$ and $1 \leq k \leq J-1$. A ***k*-partitioned conditional GLM** for categories $1, \ldots, J$ is defined by:

- A **partition tree $\boldsymbol{\mathcal{T}}$** of $\{1, \ldots, J\}$ with $\mathcal{V}^*$, the set of non terminal nodes of cardinal $\boldsymbol{k}$. Let Ω_j^V be the children of node $V \in \mathcal{V}^*$.
- A **collection $\{(\boldsymbol{r^V}, \boldsymbol{F^V}, \boldsymbol{Z^V}(\boldsymbol{x^V})) \mid \boldsymbol{V} \in \boldsymbol{\mathcal{V}^*}\}$** of GLM(s) for each conditional probability vector $\pi^V = (\pi_1^V, \ldots, \pi_{J_V-1}^V)$, where $\pi_j^V = P(Y \in \Omega_j^V | Y \in V, X^V = x^V)$ for $j = 1, \ldots, J_V$.

Model estimation: It can be shown that the log-likelihood of partitioned conditional GLMs can be decomposed into components such that each component can be maximised individually because GLMs attached to each partition of categories do not share common regression coefficients (Zhang and Ip, 2012). Each component corresponds to the partition of a parent node $V \in \mathcal{V}^*$, and therefore, each GLM $(r^V, F^V, Z^V(x^V))$ can be estimated separately using the procedure described in Section 1.

3 Application to back pain prognosis

Doran and Newell (1975) describe a back pain study with 101 patients. The response variable y was the assessment of back pain after three weeks of treatment using the six ordered categories: *worse, same, slight improvement, moderate improvement, marked improvement, complete relief.* The three selected explanatory variables observed at the beginning of the treatment period were x_1 = *length of previous attack* (1=short, 2=long), x_2 = *pain change* (1=getting better, 2=same, 3=worse) and x_3 = *lordosis* (1=absent/decreasing, 2=present/increasing).
The best model we obtained for this data set was a 2-partitioned conditional GLM (log-likelihood of -151.36 with 9 parameters); see figure 1.

Anderson (1984) obtained a log-likelihood of −154.39 with 9 parameters for the stereotype model. This gain is mainly due to the modularity of partitioned conditional GLMs (change of ratio r and design matrix Z between the two partitions).

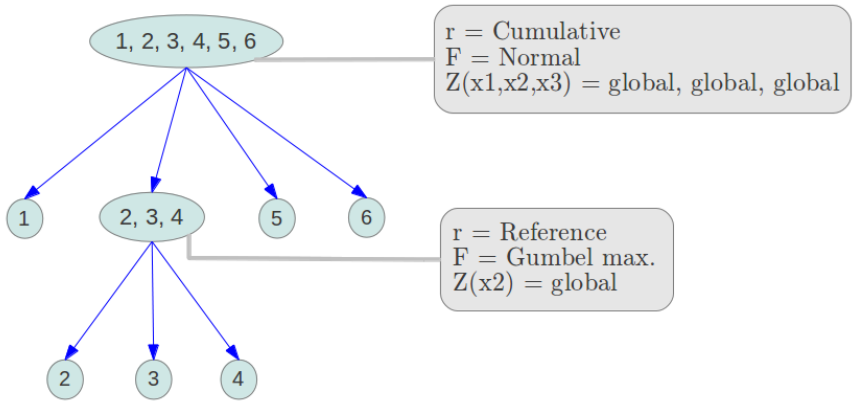

FIGURE 1. Representation of a 2-partitioned conditional GLM (partition tree $\mathcal{T}$ of six response categories and two associated GLMs for categorical response variables)

References

Agresti, A. (2002). Categorical Data Analysis. *John Wiley and Sons.*

Anderson, J.A. (1984). Regression and ordered categorical variables. *Journal of the Royal Statistical Society, Series B*, **46**, 1 – 30.

Doran, D.M.L. and Nowell, D.J. (1975). Manipulation in treatment of low back pain: a multicentre study. *British medical journal*, **2**, 161 – 164.

Fahrmeir, L. and Tutz, G. (2001). Multivariate Statistical Modelling Based on Generalized Linear Models. *Springer Verlag.*

Morawitz, B. and Tutz, G. (1990). Alternative parameterizations in business tendency surveys. *Mathematical Methods of Operations Research, Springer*, **34**, 143 – 156.

Zhang, Q. and Ip, E.H. (2012). Generalized linear model for partially ordered data. *Statistics in Medicine.*