VGAM Family Functions for Categorical Data

T. W. Yee

January 5, 2010

Version 0.7-10

© Thomas W. Yee

Department of Statistics,
University of Auckland,
New Zealand.
t.yee@auckland.ac.nz
http://www.stat.auckland.ac.nz/∼yee

Contents

[Important note: This document and code is not yet finished, but should be completed one day . . .]

1 Introduction

This document describes in detail VGAM family functions for a categorical response variable taking values $Y = 1, 2, \ldots, M+1$. Table 1 summarizes those current available. It is convenient to consider the two cases: when Y is *nominal* (no order) and when Y is *ordinal* (ordered). An example of the latter is Table 2 where the stages of a disease are $Y = 1$ for none, $Y = 2$ for mild, and $Y = 3$ for severe symptoms.

Regarding further documentation, Yee (2010) is a journal article version complementing this document and Yee (2008) gives an overview of VGLMs/VGAMs. General references for categorical data include Simonoff (2003), Agresti (2002), Fahrmeir and Tutz (2001), Leonard

(2000), Lloyd (1999), Long (1997) and McCullagh and Nelder (1989). An overview of models for ordinal responses is Liu and Agresti (2005). A manual for fitting common categorical models with various software found in Agresti (2002) is Thompson (2009). The website `http://www.ats.ucla.edu/stat/R` contains some useful material too.

Many of VGAM's features come from `glm()` and `gam()` so that readers unfamiliar with these functions are referred to Chambers and Hastie (1993). Additionally, the VGAM *User Manual* should be consulted for general instructions about the software. VGAM allows a categorical response to be inputted as a vector of factors, or a $n \times (M+1)$ matrix of counts. If the former, it will be converted to the latter form. For more information, see Section 5.1.

Table 1: Quantities defined in VGAM for a categorical response Y taking values $1, \ldots, M+1$. Covariates $\boldsymbol{x}$ have been omitted for clarity. The LHS quantities are η_j or η_{j-1} for $j = 1, \ldots, M$ and $j = 2, \ldots, M+1$ respectively. All except for `multinomial()` are suited to ordinal Y. Note that `propodds()` is a shortcut to `cumulative(parallel = TRUE, reverse = TRUE, link = "logit")`.

t

Quantity	Notation	Range of j	VGAM family function
$P(Y = j+1)/P(Y = j)$	ζ_j	$1, \ldots, M$	`acat()`
$P(Y = j)/P(Y = j+1)$	ζ_j^R	$2, \ldots, M+1$	`acat(reverse = TRUE)`
$P(Y > j \mid Y \geq j)$	δ_j^*	$1, \ldots, M$	`cratio()`
$P(Y < j \mid Y \leq j)$	δ_j^{*R}	$2, \ldots, M+1$	`cratio(reverse = TRUE)`
$P(Y \leq j)$	γ_j	$1, \ldots, M$	`cumulative()`
$P(Y \geq j)$	γ_j^R	$2, \ldots, M+1$	`cumulative(reverse = TRUE)`
$\log\{P(Y = j)/P(Y = M+1)\}$		$1, \ldots, M$	`multinomial()`
See caption			`propodds()`
$P(Y = j \mid Y \geq j)$	δ_j	$1, \ldots, M$	`sratio()`
$P(Y = j \mid Y \leq j)$	δ_j^R	$2, \ldots, M+1$	`sratio(reverse = TRUE)`

2 Nominal Responses

The *multinomial logit model* is the most common model in this case. We describe this below, as well as a variant called the *stereotype model*.

2.1 Multinomial logit model

The multinomial logit model (MLM), which is also known as the *multiple logistic regression model* or *polytomous logistic regression model*, is given by

$$p_j \;=\; P(Y = j|\boldsymbol{x}) \;=\; \frac{\exp\{\eta_j(\boldsymbol{x})\}}{\sum_{\ell=1}^{M+1} \exp\{\eta_\ell(\boldsymbol{x})\}}, \quad j = 1, \ldots, M+1, \tag{1}$$

where $\eta_j(\boldsymbol{x}) = \boldsymbol{\beta}_j^T \boldsymbol{x}$. The model is particularly useful for exploring how the relative chances of falling into the response categories depend upon the covariates as $p_j(\boldsymbol{x})/p_k(\boldsymbol{x}) = \exp\{\eta_j(\boldsymbol{x}) - \eta_k(\boldsymbol{x})\}$. Identifiability constraints, e.g., $\eta_{M+1}(\boldsymbol{x}) \equiv 0$, are required by the model. This implies

$$\log\left(\frac{p_j}{p_{M+1}}\right) \;=\; \eta_j, \quad j = 1, \ldots, M. \tag{2}$$

VGAM fits the multinomial logit model using the family function `multinomial()`. It uses the last column of the response matrix as baseline (by default), or if the response is a factor, the last level. The special case of $M = 1$ corresponds to logistic regression. The multinomial logit model is also related to neural networks and to classification—see, e.g., Ripley (1996).

2.1.1 Marginal effects

The *marginal effects* in the MLM are the derivatives of the probabilities with respect to the explanatory variables. In some applications, where even knowing the sign is important, this is particularly useful to know.

Consider a MLM without constraints. The marginal effects are given as

$$\frac{\partial\, p_j(\boldsymbol{x}_i)}{\partial\, \boldsymbol{x}_i} \;=\; p_j(\boldsymbol{x}_i) \left\{ \boldsymbol{\beta}_j - \sum_{s=1}^{M+1} p_s(\boldsymbol{x}_i)\, \boldsymbol{\beta}_s \right\}. \tag{3}$$

(Show this—it is an exercise). Marginal effects are implemented in `margeff()`, which will accept a `multinomial()` VGLM.

By the way, a related quantity know as the *elasticities* are

$$\frac{\partial\, p_j(\boldsymbol{x}_i)}{\partial\, x_{ik}} \frac{x_{ik}}{p_j(\boldsymbol{x}_i)} \;=\; x_{ik} \left\{ \beta_{(j)k} - \sum_{s=1}^{M+1} p_s(\boldsymbol{x}_i)\, \beta_{(s)k} \right\}. \tag{4}$$

For the `"cumulative"` family with `reverse = FALSE` we have $p_j = \gamma_j - \gamma_{j-1} = h(\eta_j) - h(\eta_{j-1})$ where $h = g^{-1}$ is the inverse of the link function. Then

$$\frac{\partial\, p_j(\boldsymbol{x})}{\partial\, \boldsymbol{x}} \;=\; h'(\eta_j)\, \boldsymbol{\beta}_j - h'(\eta_{j-1})\, \boldsymbol{\beta}_{j-1}. \tag{5}$$

The function `margeff()` will also accept a `cumulative()` VGLM.

Table 2: Period of exposure (years) and severity of pneumoconiosis amongst a group of coalminers.

Exposure Time	Normal	Mild	Severe
5.8	98	0	0
15.0	51	2	1
21.5	34	6	3
27.5	35	5	8
33.5	32	10	9
39.5	23	7	8
46.0	12	6	10
51.5	4	2	5

2.2 Stereotype Model

This model, which we sometimes refer to as the *reduced-rank multinomial logit model*, was proposed by Anderson (1984) who described it as being suitable for *all* (ordered or unordered) categorical response variables. The basic idea is that if M and p are even moderately large then the total number of regression coefficients in the multinomial logit model will be large. One method of parsimony is to approximate the $M \times p$ matrix of regression coefficients by a

lower rank matrix. In detail, the reduced-rank concept replaces $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_M)^T$ (without the intercepts) by

$$\mathbf{B} \;=\; \mathbf{C}\mathbf{A}^T \tag{6}$$

where $\mathbf{C} = (\boldsymbol{c}_1\, \boldsymbol{c}_2 \cdots \boldsymbol{c}_r)$ is $p \times r$, $\mathbf{A} = (\boldsymbol{a}_1\, \boldsymbol{a}_2 \cdots \boldsymbol{a}_r)$ is $M \times r$ and r (usually $\ll \min(M, p)$) is the rank of $\mathbf{A}$ and $\mathbf{C}$. It is convenient to write

$$\boldsymbol{\eta}_i \;=\; \boldsymbol{\eta}_0 + \mathbf{A}\,\mathbf{C}^T\boldsymbol{x}_i \;=\; \boldsymbol{\eta}_0 + \mathbf{A}\,\boldsymbol{\nu}_i \;.$$

The stereotype model is a special case of a *Reduced-rank VGLM* (RR-VGLM). It may be fitted using `rrvglm()` and `multinomial()`. See the other documentation regarding RR-VGLMs for more details.

It is well known that the factorization (6) is not unique as $\boldsymbol{\eta}_i = \boldsymbol{\eta}_0 + \mathbf{A}\,\mathbf{M}\,\mathbf{M}^{-1}\,\boldsymbol{\nu}_i$ for any nonsingular matrix $\mathbf{M}$. A common form of constraint which ensures $\mathbf{A}$ is of rank r and unique is to restrict it to the form

$$\mathbf{A} \;=\; \left(\begin{array}{c} \mathbf{I}_r \\ \widetilde{\mathbf{A}} \end{array} \right),$$

where $\widetilde{\mathbf{A}}$ is a $(M - r) \times r$ matrix. In fact, any r rows of $\mathbf{A}$ may be chosen to represent $\mathbf{I}_r$[1]. This method of identifiability is implemented in VGAM.

3 Ordinal Responses

3.1 Models Involving Cumulative Probabilities

When the response is ordered the most common models involve the cumulative probabilities $P(Y \leq j|\boldsymbol{x})$ (see McCullagh (1980)), in particular, the *proportional odds model* is

$$\text{logit}\, P(Y \leq j|\boldsymbol{x}) \;=\; \beta_{(j)1} + \boldsymbol{\beta}^T\boldsymbol{x}_{(-1)}, \qquad j = 1, \ldots, M.$$

We call models of the form

$$\text{logit}\, P(Y \leq j|\boldsymbol{x}) \;=\; \eta_j \qquad j = 1, \ldots, M$$

cumulative logit models as they involve the logit link function and cumulative probabilities. The proportional odds model is a cumulative logit model with the *parallelism* assumption $\boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_M$. It is well-known that the parallelism assumption applied to a cumulative logit model results in the effect of the covariates on the odds ratio being the same regardless of the division point j, hence the name *proportional odds model* (this property is called *strict stochastic ordering* (McCullagh, 1980)). In practice, the parallelism assumption should be checked; see, e.g., Armstrong and Sloan (1989), Peterson (1990). In general, the proportional odds model is

$$\text{logit}\, P(Y \leq j|\boldsymbol{x}) \;=\; \beta_{(j)1} + \eta, \qquad j = 1, \ldots, M. \tag{7}$$

If a complementary log-log link is chosen in (7) the result is known as the *proportional hazards model*. In theory, any other link function used for binomial data such as the probit link (which

[1] Actually, if the 'wrong' rows of $\mathbf{A}$ are chosen to represent $\mathbf{I}_r$, then $\mathbf{A}$ may be ill-conditioned at the solution, or even failing to exist.

is often referred to as the *ordinal probit model* or *cumulative probit model*) can be applied to cumulative probability models.

One reason for the proportional-odds cumulative-logit model's popularity is its connection to the idea of a continuous latent response. It can be shown that proportional-odds model can be motivated by the categorical outcome Y being a categorized version of an unobservable (latent) continuous variable, Z, say. Then $Y = k$ if $c_{k-1} < Z \leq c_k$ where the c_k are known as *cutpoints*. It is assumed that the regression of Z on $\boldsymbol{x}$ has the form

$$Z \;=\; \boldsymbol{\beta}^T \boldsymbol{x} + \varepsilon,$$

where $\boldsymbol{x}$ has an intercept and ε is a random error from a logistic distribution with mean zero and constant variance. Then the coarsened version Y will be related to the $\boldsymbol{x}$ by a proportional-odds cumulative logit model. Note that the logistic distribution has a bell-shaped density similar to a normal curve, and it may be fitted to raw data with VGAM family functions `logistic1()` and `logistic2()`. If ε had normal errors rather than logistic errors then the cumulative logit equations would change to have a probit link—this is known as a cumulative probit model. For more information see McCullagh and Nelder (1989).

VGAM fits models based on cumulative probabilities using the family function `cumulative()`. It has a limited shortcut called `propodds()`. Liu and Agresti (2005) call this class of models *cumulative link models*. Of course, `cumulative()` allows other link functions, and the parallelism assumption $\boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_M$ (or more generally, $f_{(1)k}(x_k) = \cdots = f_{(M)k}(x_k)$). For example,

```
> vglm(y ~ x2, cumulative(link = probit, reverse = TRUE,
+     parallel = TRUE), mydataframe)
```

fits the model

$$\Phi^{-1}\{P(Y \geq j|\boldsymbol{x})\} \;=\; \beta_{(j-1)1} + \beta_2 x_2, \qquad j = 2, \ldots, M+1.$$

Similarly,

```
> vgam(y ~ s(x2), cumulative(link = probit, reverse = TRUE,
+     parallel = TRUE), mydataframe)
```

fits the model

$$\Phi^{-1}\{P(Y \geq j|\boldsymbol{x})\} \;=\; \beta_{(j-1)1} + f_{(1)2}(x_2), \qquad j = 2, \ldots, M+1$$

for some smooth function $f_{(1)2}$.

If the proportional odds assumption is inadequate then one can try to use a different link function or add extra terms such as interaction terms into the linear predictor. Another strategy is the so-called partial proportional odds model (Peterson and Harrell, 1990) which VGAM may fit.

3.2 Models Involving Stopping-ratios and Continuation-ratios

Quantities known as *continuation-ratios* are useful for the analysis of a sequential process, e.g., to ascertain the effect of a covariate on the number of children a couple choose to have ($Y = 1$ (no children), 2 (1 child), 3 (2 child), 4 (3+ children)), or whether a risk factor is related to the progression of a disease ($Y = 1$ (no disease), 2 (localized), 3 (widespread), 4 (terminal)). For such data, there are two ways of modelling the situation—these are deciding whether to look at

Table 3: Summary of the default values of the `parallel` argument, and whether `parallel = TRUE` ever applies to the intercept.

`family =`	Default: `parallel =`	Apply to intercepts?
`acat`	`FALSE`	No
`cratio`	`FALSE`	No
`cumulative`	`FALSE`	No
`multinomial`	`FALSE`	Yes
`propodds`	`TRUE` (fixed value)	No
`sratio`	`FALSE`	No

the probability of *stopping* at $Y = j$ or *continuing* past $Y = j$, given that Y has reached level j in the first place. Which of the two is more natural depends on the particular application.

Because there are differences in the literature Table 1 summarizes the VGAM definition of a continuation-ratio, and also lists a quantity termed the *stopping-ratio* to help distinguish between the two types. The stopping ratio matches the definition of the continuation-ratio in Agresti (2002). Note that, regardless of type, one sometimes wishes to reverse Y so that it is enumerated from $M+1, M, \ldots, 1$ instead; this is handled by prefixing their name by "reverse," e.g., $\delta_j^R \equiv P(Y = j|Y \leq j),\ j = 2, \ldots, M+1$ is called a reverse stopping ratio. Some authors use "backward" instead of "reverse" and "forward" if it is not reversed.

For all the models listed in Table 1, VGAM allows other link functions to be used, as well as allowing the parallelism assumption $\boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_M$ (or more generally, $f_{(1)k}(x_k) = \cdots = f_{(M)k}(x_k)$), and the `zero =` argument to constrain η_j be be a intercept term only. To see the default values of these arguments use the `args()` function, e.g., `args(sratio)`.

The quantities in Table 1 are all interrelated. The following formulae give some of these relationships.

$$
\begin{aligned}
\gamma_j &= 1 - \gamma_{j-1}^R, \quad \gamma_0 = 0, \quad \gamma_{M+1} = 1,\ j = 1, \ldots, M, \\
\delta_j &= \frac{p_j}{1 - \gamma_{j-1}},\ j = 1, \ldots, M, \\
\delta_j^R &= \frac{p_j}{\gamma_j},\ j = 2, \ldots, M+1, \quad \delta_1^R = 1, \\
\delta_j^* &= 1 - \delta_j = \frac{1 - \gamma_j}{1 - \gamma_{j-1}},\ j = 1, \ldots, M, \\
\delta_1^* &= 1 - p_1, \quad \delta_{M+1}^* = 0, \\
\delta_j^{*R} &= 1 - \delta_j^R = \frac{\gamma_{j-1}}{\gamma_j},\ j = 2, \ldots, M+1, \\
\delta_1 &= p_1, \quad \delta_{M+1} = 1, \\
\delta_1^{*R} &= 0, \quad \delta_{M+1}^{*R} = 1 - p_{M+1}.
\end{aligned}
$$

For further information on these models see, e.g., Armstrong and Sloan (1989).

3.3 Models Involving Adjacent Categories

Adjacent-category models are models for categorical data in the form

$$
g(p_j/p_{j+1}) = \eta_j, \qquad j = 1, \ldots, M,
$$

for some link function g (log or identity because p_j/p_{j+1} does not necessarily lie in $[0,1]$.) Reverse adjacent-category models use

$$g(p_j/p_{j-1}) \;=\; \eta_{j-1}, \qquad j = 2, \ldots, M+1.$$

4 The Bradley-Terry Model

Consider an experiment consisting of n_{ij} judges who compare pairs of items T_i, $i = 1, \ldots, M+1$. They express their preferences between T_i and T_j. Let $N = \sum \sum_{i<j} n_{ij}$ be the total number of pairwise comparisons, and assume independence for ratings of the same pair by different judges and for ratings of different pairs by the same judge.

A model describing this experiment was proposed by Bradley and Terry (1952) and Zermelo (1929). Let π_i be the *worth* of item T_i,

$$P[T_i > T_j] \;=\; p_{i/ij} \;=\; \frac{\pi_i}{\pi_i + \pi_j},$$

$i \neq j$, where "$T_i > T_j$" means i is preferred over j. Suppose that $\pi_i > 0$. Let Y_{ij} be the number of times that T_i is preferred over T_j in the n_{ij} comparisons of the pairs. Then $Y_{ij} \sim \text{Bin}(n_{ij}, p_{i/ij})$. Maximum likelihood estimation of the parameters $\pi_1, \ldots, \pi_{M+1}$ involves maximizing

$$\prod_{i<j}^{M+1} \binom{n_{ij}}{y_{ij}} \left(\frac{\pi_i}{\pi_i + \pi_j}\right)^{y_{ij}} \left(\frac{\pi_j}{\pi_i + \pi_j}\right)^{n_{ij} - y_{ij}}.$$

By default, $\pi_{M+1} \equiv 1$ is used for identifiability, however, this can be changed very easily. Note that one can define linear predictors η_{ij} of the form

$$\text{logit}\left(\frac{\pi_i}{\pi_i + \pi_j}\right) \;=\; \log\left(\frac{\pi_i}{\pi_j}\right) \;=\; \lambda_i - \lambda_j. \tag{8}$$

The VGAM framework can handle the Bradley-Terry model only for intercept models. It has

$$\lambda_j \;=\; \eta_j \;=\; \log \pi_j = \beta_{(1)j}, \qquad j = 1, \ldots, M. \tag{9}$$

As well as having many applications in the field of preferences, the Bradley-Terry model has many uses in modelling 'contests' between teams i and j, where only one of the teams can win in each contest (ties are not allowed under the classical model).

The R package `BradleyTerry` by D. Firth can fit the Bradley-Terry model; see Firth (2005) for details.

4.1 Example

Consider the effect of the food-enhancer monosodium glutamate (MSG) on the flavour of apple sauce in the data given in Table 4. Treatments 1, 2 and 3 are increasing amounts of the substance, and Treatment 4 is a control with no MSG. Four independent comparisons were made of each of the six pairs.

We apply the VGAM family function `brat()`, which implements the Bradley-Terry model, to the apple sauce data.

```
> amsg = matrix(c(NA, 3, 3, 3, 1, NA, 3, 4, 1, 1, NA, 0,
+     1, 0, 4, NA), 4, 4, byrow = TRUE)
> dimnames(amsg) = list(winner = as.character(1:4), loser = as.character(1:4))
> fit = vglm(Brat(amsg) ~ 1, brat)
```

This gives

```
> summary(fit)

Call:
vglm(formula = Brat(amsg) ~ 1, family = brat)

Pearson Residuals:

Coefficients:
                Value Std. Error t value
(Intercept):1  1.2112    0.84083  1.4405
(Intercept):2  0.8943    0.80747  1.1075
(Intercept):3 -0.9961    0.87068 -1.1441

Number of linear predictors:  3

Names of linear predictors: log(alpha1), log(alpha2), log(alpha3)

Dispersion Parameter for brat family:   1

Log-likelihood: -25.989 on 0 degrees of freedom

Number of Iterations: 4

> Coef(fit)

   alpha1    alpha2    alpha3
3.3576125 2.4456117 0.3693147
```

By default, the last reference group is baseline, so that $\lambda_4 \equiv 1$. We have $\widehat{\lambda}_1 \approx 3.358$, $\widehat{\lambda}_2 \approx 2.446$, $\widehat{\lambda}_3 \approx 0.369$, therefore we conclude that Treatment 1 is the most preferred, followed by Treatments 2 and 4, and lastly Treatment 3. It appears that the more MSG, the worse the taste, however, the control treatment tastes the second worse.

Finally,

```
> InverseBrat(fitted(fit))

           1         2         3         4
1         NA 0.5785771 0.9009064 0.7705165
```

Table 4: Apple sauce data and model. The i and j index the items. The C_k denote columns of a contrast matrix with respect to the λs.

i	j	y_{ij}	n_{ij}	η_{ij}	C_1	C_2	C_3	C_4
1	2	3	4	$\lambda_1 - \lambda_2$	1	–1		
1	3	3	4	$\lambda_1 - \lambda_3$	1		–1	
1	4	3	4	$\lambda_1 - \lambda_4$	1			–1
2	3	3	4	$\lambda_2 - \lambda_3$		1	–1	
2	4	4	4	$\lambda_2 - \lambda_4$		1		–1
3	4	0	4	$\lambda_3 - \lambda_4$			1	–1

```
2 0.42142293        NA 0.8688013 0.7097758
3 0.09909362 0.1311987        NA 0.2697077
4 0.22948346 0.2902242 0.7302923        NA
```

gives the estimated probabilities of Treatments i "beating" Treatments j, $\widehat{P}[i > j]$.

4.2 Bradley-Terry Model with Ties

Consider a Bradley-Terry model with ties (no preference). This is useful because it is common for people to say that both T_i and T_j are equally good or bad. There are at least two ways of modelling ties:

(i)

$$\begin{aligned} P(T_i > T_j) &= \frac{\pi_i}{\pi_i + \pi_j + \pi_0}, \\ P(T_i < T_j) &= \frac{\pi_j}{\pi_i + \pi_j + \pi_0}, \\ P(T_i = T_j) &= \frac{\pi_0}{\pi_i + \pi_j + \pi_0}. \end{aligned}$$

Here, $\pi_0 > 0$ is an extra parameter.

(ii) Davidson (1970) proposed

$$\begin{aligned} P(T_i > T_j) &= \frac{\pi_i}{\pi_i + \pi_j + q\sqrt{\pi_i\,\pi_j}}, \\ P(T_i < T_j) &= \frac{\pi_j}{\pi_i + \pi_j + q\sqrt{\pi_i\,\pi_j}}, \\ P(T_i = T_j) &= \frac{q\sqrt{\pi_i\,\pi_j}}{\pi_i + \pi_j + q\sqrt{\pi_i\,\pi_j}}, \end{aligned}$$

where $q > 1$.

The first model, (i), has been implemented in the VGAM family function `bratt()`. It has

$$\boldsymbol{\eta} = (\log \pi_1, \ldots, \log \pi_{M-1}, \log \pi_0)^T$$

by default, where there are M competitors and $\pi_M \equiv 1$. Like `brat()`, one can choose a different reference group and reference value.

For futher information about ties, see Kuk (1995) and Torsney (2004).

5 Other Topics

5.1 Input

The response in `vglm()`/`vgam()` can be

1. an $n \times (M+1)$ matrix of counts. The columns are best labelled, and the jth column denotes $Y = j$.

2. a vector. The unique values, when sorted (or levels if a factor), denote the $M+1$ levels from $1, \ldots, M+1$. If a factor, it may be ordered or unordered. The functions `factor()`, `ordered()`, `levels()`, are useful; see the R online help.

Note: if `weights` is used as input, then any zero values should be deleted first. For example, if `n` is a vector containing zeros, then something like

```
> vglm(..., weights = n, subset = n > 0)
```

should be used.

5.2 Output

Suppose `fit` is a categorical VGAM object. Like `binomialff()`, the fitted values (in `fitted(fit)`) are probabilities and `weights(fit, type="prior")` contain the $n_i = \sum_{j=1}^{M+1} y_{ij}$. However, `fitted(fit)` is a $n \times (M+1)$ matrix, whose rows sum to unity.

5.3 Constraints

All categorical data family functions have the `parallel` and `zero` arguments. By default, `parallel = FALSE` and `zero = NULL` for all models. This means that to make the parallelism assumption one must explicitly invoke it as such. The reason for this is that the parallelism assumption must be checked, and the software discourages users from making assumptions without thinking. Unfortunately, the default values may fail on some data, or lead to a large number of parameters.

Table 3 summarizes whether `parallel = TRUE` is applied to the intercepts. Also, all categorical data family functions have a `reverse = FALSE` argument.

5.4 Implementation Details

The S expression `process.categorical.data.vgam` provides a unified way of handling the response variable. The variable `delete.zero.colns` should be set to `TRUE` or `FALSE` prior to evaluating `process.categorical.data.vgam` in the slot `@initialize` to handle columns of the response matrix that contain all 0's. In some situations these must be deleted.

Similarly, the expression `deviance.categorical.data.vgam` computes the deviance for all the models in this document.

5.5 Convergence

The VGAM family functions described in this document use the type of algorithm described in McCullagh (1980). He showed that a unique maximum of the likelihood is guarenteed for sufficiently large sample sizes, though infinite parameter values can arise with sparse data sets containing certain patterns of zeros. Usually, one obtains rapid convergence to the MLEs.

5.6 Over-dispersion

Over-dispersion for polytomous responses can occur just like it does for binary responses (see documentation on GLM and GAM VGAM family functions). Sec. 5.5 of McCullagh and Nelder (1989) use

$$\tilde{\sigma}^2 \;=\; X^2/\{nM-p\} \;=\; X^2/\{\text{residual d.f}\}, \tag{10}$$

where X^2 is Pearson's statistic. They state that this is approximately unbiased for σ^2, is consistent for large n regardless of whether the data are sparse, and is approximately independent of the estimated $\hat{\boldsymbol{\beta}}$.

Following GLM and GAM-type VGAM family functions, categorical family functions might one day have a `dispersion` argument. The default for all will be `1`, and if set to `0`, it will be estimated using some formula. Note that `summary()` uses a generalization of the Pearson statistic, which reduces to (10) for some models.

5.7 Relationship with `binomialff()`

VGAM can fit to binomial responses using `binomialff()`, which is incompatible with `glm()` and `gam()`.

If `y` is a vector of 0s and 1s then simple logistic regression can be performed as follows:

1. `glm(y ~ ..., family = binomial, ...)`
2. `vglm(y ~ ..., family = binomialff, ...)`
3. `vglm(1-y ~ ..., multinomial, ...)`,
4. `vglm(cbind(y,1-y) ~ ..., multinomial, ...)`,
5. `vglm(y ~ ..., cumulative(reverse = TRUE), ...)`.

These hold because the final level is used as the baseline category for the MLM. However, this reference category can be changed using an argument of `multinomial()`.

5.8 The `xij` Argument

The `xij` argument allows for covariate values that are specific to each η_j. For example, suppose we have two binary responses, $Y_j = 1$ or 0 for presence/absence of a cataract, where $j = 1, 2$ for the left and right eyes respectively. We have a single covariate, called ocular pressure, which measures the internal fluid pressure within each eye. With data from n people, it would be natural to fit an exchangeable bivariate logistic model:

$$\begin{aligned} \text{logit}\, P(Y_{ij} = 1) &= \beta_{(1)1} + \beta_{(1)2}\, x_{ij}, \quad j = 1, 2; \quad i = 1, \ldots, n; \\ \log\, \psi &= \beta_{(3)1}, \end{aligned} \tag{11}$$

where the dependency between the responses is modelled through the odds ratio ψ. Note that the regression coefficient for x_{i1} and x_{i2} is the same, and $x_{i1} \neq x_{i2}$ in general.

Another example is if you were an econometrician interested in peoples' choice of transport for travelling to work. Suppose $Y = 1$ means bus, $Y = 2$ means train, $Y = 3$ means car, and $Y = 4$ means walking to work, and that we have two covariates: $X_2 =$ cost, and $X_3 =$ journey duration. Suppose we collect data from a random sample of n people from some population, and that each person has access to all these transport modes. For such data, a natural regression model would be a multinomial logit model with $M = 3$: for $j = 1, \ldots, M$,

$$\log \frac{P(Y = j)}{P(Y = M + 1)} \;=\; \eta_j \;=\; \beta_{(j)1} + \beta_2 (x_{i2j} - x_{i2,M+1}) + \beta_3 (x_{i3j} - x_{i3,M+1}),$$

where x_{i2j} is the cost for the jth transport means for the ith person, and x_{i3j} is the journey duration of the jth transport means for the ith person. Note that, apart from the intercepts, there are two regression coefficients, β_2 and β_3.

To fit such models in VGAM one needs to use the `xij` and `form2` arguments. See Yee (2010) and the online documentation concerning this important feature, e.g., `fill()`, `vglm.control()`.

5.9 Reduced-rank Regression

A RR-VGLM (reduced-rank VGLM) replaces the (large) $M \times p$ coefficient matrix $\mathbf{B}^T$ by $\mathbf{A}\,\mathbf{C}^T$, where $\mathbf{A}$ is $M \times r$ and $\mathbf{C}$ is $p \times r$, $r \ll \min(M, p)$. The application to the multinomial logit model is only one special case (called the stereotype model). For more details, see Yee and Hastie (2003).

6 Tutorial Examples

In this section we illustrate some of the models. We will make use of the pneumoconiosis data in Table 2 (see McCullagh and Nelder (1989)). Recall the ordinal response is the severity of pneumoconiosis in coalface workers (categorized as 1 = normal, 2 = mild pneumoconiosis, 3 = severe pneumoconiosis) and the covariate is the number of years of exposure. In fact, in the following, we use $x_2 = \log$(exposure time).

6.1 Multinomial logit model

Although Y is ordinal for the pneumoconiosis data, we fit the multinomial logit model for the sake of illustration.

```
> data(pneumo)
> pneumo = transform(pneumo, let = log(exposure.time))
> fit.mlm <- vglm(cbind(normal, mild, severe) ~ let, multinomial,
+     pneumo)
> coef(fit.mlm, matrix = TRUE)

            log(mu[,1]/mu[,3]) log(mu[,2]/mu[,3])
(Intercept)          11.975092          3.0390622
let                  -3.067466         -0.9020936

> summary(fit.mlm)

Call:
vglm(formula = cbind(normal, mild, severe) ~ let, family = multinomial,
    data = pneumo)

Pearson Residuals:
                        Min       1Q    Median      3Q     Max
log(mu[,1]/mu[,3]) -0.75473 -0.20032 -0.043255 0.43644 0.92283
log(mu[,2]/mu[,3]) -0.81385 -0.43818 -0.121281 0.15716 1.03229

Coefficients:
               Value Std. Error t value
(Intercept):1 11.9751    2.00044  5.9862
(Intercept):2  3.0391    2.37607  1.2790
let:1         -3.0675    0.56521 -5.4272
let:2         -0.9021    0.66898 -1.3485

Number of linear predictors:  2
```

```
Names of linear predictors: log(mu[,1]/mu[,3]), log(mu[,2]/mu[,3])

Dispersion Parameter for multinomial family:   1

Residual Deviance: 5.34738 on 12 degrees of freedom

Log-likelihood: -25.25054 on 12 degrees of freedom

Number of Iterations: 4
```

Note that `mu[j,]` is the jth column of the matrix of fitted values. The estimated variance-covariance matrix of the regression coefficients is `summary(fit.mlm)@cov.unscaled` (if the dispersion parameter is unity), or more elegantly, `vcov(fit.mlm)`.

6.2 Stereotype model

In the following, `fit1` fits a rank-1 stereotype model.

```
> data(pneumo)
> pneumo = transform(pneumo, let = log(exposure.time))
> fit1 = rrvglm(cbind(normal, mild, severe) ~ let, multinomial,
+     Rank = 1, pneumo)
> constraints(fit1)

$`(Intercept)`
     [,1] [,2]
[1,]    1    0
[2,]    0    1

$let
          [,1]
[1,] 1.0000000
[2,] 0.2940843

> coef(fit1)

(Intercept):1 (Intercept):2           let
    11.975092      3.039062     -3.067466

> coef(fit1, matrix = TRUE)

            log(mu[,1]/mu[,3]) log(mu[,2]/mu[,3])
(Intercept)          11.975092          3.0390622
let                  -3.067466         -0.9020936
```

That is, $\mathbf{A}$ = `(constraints(fit1))[["let"]]` $= (1, 0.2940857)^T$ and $\mathbf{C} = (-3.067469)$. Note that `coef(fit1, matrix = TRUE)` looks very similar to `fit.mlm` above—it should be! This is because this RR-MLM is exactly the same model as the MLM because the rank equals the number of covariates which is one. It is like a saturated reduced-rank model.

6.3 Proportional odds model

Page 179 of McCullagh and Nelder (1989) fit a proportional odds model to the pneumoconiosis data. They concluded that the logarithm of exposure time was strongly preferred to simply exposure time, and fitted the model

$$\text{logit } \hat{\gamma}_j(x_i) \;=\; \hat{\alpha}_j - \hat{\beta}\, \log(\text{exposure time}_i), \quad i = 1, \ldots, 8, \quad j = 1, 2.$$

We can fit this (not exactly with respect to the signs) by

```
> fit.pom <- vglm(cbind(normal, mild, severe) ~ let, fam = propodds,
+     pneumo)
```

or

```
> fit.pom <- vglm(cbind(normal, mild, severe) ~ let, fam = cumulative(parallel = TRUE),
+     pneumo)
```

Then the matrix of regression coefficients can be obtained

```
> coef(fit.pom, matrix = TRUE)
```

```
            logit(P[Y<=1]) logit(P[Y<=2])
(Intercept)       9.676092      10.581724
let              -2.596806      -2.596806
```

which agrees to that of McCullagh & Nelder. Note that `coef(fit.pom)` is recommended over using `fit.pom@coefficients`, which, for this example, is the vector $(\hat{\alpha}_1, \hat{\alpha}_2, \hat{\beta})^T$. The general enumeration of the elements in `@coefficients` is given in the VGAM *User Manual*.

Also,

```
> summary(fit.pom)
```

```
Call:
vglm(formula = cbind(normal, mild, severe) ~ let, family = cumulative(parallel = TRUE),
    data = pneumo)

Pearson Residuals:
                   Min        1Q  Median      3Q     Max
logit(P[Y<=1]) -1.2479 -0.071638 0.14410 0.30859 0.77144
logit(P[Y<=2]) -1.0444 -0.184147 0.30926 0.33528 0.50477

Coefficients:
               Value Std. Error t value
(Intercept):1  9.6761     1.3241  7.3078
(Intercept):2 10.5817     1.3454  7.8649
let           -2.5968     0.3811 -6.8140

Number of linear predictors:  2

Names of linear predictors: logit(P[Y<=1]), logit(P[Y<=2])

Dispersion Parameter for cumulative family:   1
```

```
Residual Deviance: 5.02683 on 13 degrees of freedom

Log-likelihood: -25.09026 on 13 degrees of freedom

Number of Iterations: 6
```

Now one can formally test the proportional odds assumption by a likelihood ratio test. This can be done by

```
> fit.npom <- vglm(cbind(normal, mild, severe) ~ let, cumulative,
+     pneumo)
> pchisq(deviance(fit.pom) - deviance(fit.npom), df = df.residual(fit.pom) -
+     df.residual(fit.npom), lower.tail = FALSE)

[1] 0.7058849
```

There is no evidence against the proportional odds assumption.

The fitted probabilities can be obtained by

```
> fitted(fit.npom)

     normal        mild      severe
1 0.9937774 0.004356443 0.001866134
2 0.9327738 0.042506379 0.024719812
3 0.8461118 0.090169357 0.063718886
4 0.7448902 0.137177345 0.117932441
5 0.6373943 0.175768145 0.186837528
6 0.5350518 0.199663151 0.265285079
7 0.4375094 0.208312388 0.354178218
8 0.3677969 0.204411220 0.427791905
```

6.4 Stopping Ratio Model

Page 88 of Fahrmeir and Tutz (1994) fit a "sequential logit" model to some tonsil data. The data is reproduced in Table 5, and has been examined by, e.g., McCullagh (1980). It is assumed that tonsil size starts off in the "present but not enlarged" form, and then can become enlarged, and then possibly "greatly enlarged." Thus a sequential model such as the stopping ratio model maybe appropriate.

We reproduce the results of column 2 of Table 3.6 of Fahrmeir and Tutz (1994) by the following. Note that the variable `carrier` here takes on values -1 and 1, rather than the usual 0 and 1.

```
> ymat = matrix(c(19, 29, 24, 497, 560, 269), 2, 3, byrow = TRUE)
> dimnames(ymat) = list(NULL, c("present but not enlarged",
+     "enlarged", "greatly enlarged"))
> tonsils = data.frame(carrier = c(1, -1))
> ymat

     present but not enlarged enlarged greatly enlarged
[1,]                       19       29              24
[2,]                      497      560             269
```

```
> fit = vglm(ymat ~ carrier, sratio(par = TRUE, rev = FALSE),
+     tonsils, trace = TRUE, crit = "c")

VGLM    linear loop  1 :  coefficients =
-0.775224575,  0.467970163, -0.264207381
VGLM    linear loop  2 :  coefficients =
-0.775249403,  0.467949487, -0.264230636
VGLM    linear loop  3 :  coefficients =
-0.775249437,  0.467949444, -0.264230673
VGLM    linear loop  4 :  coefficients =
-0.775249437,  0.467949444, -0.264230673

> fit

Call:
vglm(formula = ymat ~ carrier, family = sratio(par = TRUE, rev = FALSE),
    data = tonsils, trace = TRUE, crit = "c")

Coefficients:
(Intercept):1 (Intercept):2       carrier
   -0.7752494     0.4679494    -0.2642307

Degrees of Freedom: 4 Total; 1 Residual
Residual Deviance: 0.005657256
Log-likelihood: -11.76594

> coef(fit, matrix = TRUE)

            logit(P[Y=1|Y>=1]) logit(P[Y=2|Y>=2])
(Intercept)         -0.7752494          0.4679494
carrier             -0.2642307         -0.2642307

> summary(fit)

Call:
vglm(formula = ymat ~ carrier, family = sratio(par = TRUE, rev = FALSE),
    data = tonsils, trace = TRUE, crit = "c")

Pearson Residuals:
  logit(P[Y=1|Y>=1]) logit(P[Y=2|Y>=2])
1           0.050963          -0.052368
2          -0.010777           0.014092

Coefficients:
                Value Std. Error t value
(Intercept):1 -0.77525   0.106085 -7.3078
(Intercept):2  0.46795   0.111556  4.1948
carrier       -0.26423   0.098873 -2.6724

Number of linear predictors:  2
```

```
Names of linear predictors: logit(P[Y=1|Y>=1]), logit(P[Y=2|Y>=2])

Dispersion Parameter for sratio family:   1

Residual Deviance: 0.00566 on 1 degrees of freedom

Log-likelihood: -11.76594 on 1 degrees of freedom

Number of Iterations: 4
```

The linear predictors and fitted values can be obtained as follows.

```
> predict(fit)

  logit(P[Y=1|Y>=1]) logit(P[Y=2|Y>=2])
1         -1.0394801          0.2037188
2         -0.5110188          0.7321801

> fitted(fit)

  present but not enlarged  enlarged greatly enlarged
1                0.2612503 0.4068696        0.3318801
2                0.3749547 0.4220828        0.2029625

> fitted(fit) * c(weights(fit, type = "prior"))

  present but not enlarged  enlarged greatly enlarged
1                 18.81002  29.29461         23.89537
2                497.18998 559.68173        269.12829
```

7 FAQ

Models for categorical responses (aka categorical data analysis, or CDA) are a very frequent form of data, and the modelling functions of this article seem to be heavily used relative to other VGAM family functions. Here is a list of some common FAQs. We will assume the fitted model is called `fit`.

1. Q: *How I can extract the standard errors for the model coefficients from the output?*

   A: A (bad) way is to try

   ```
   > slotNames(fit)
   > summary(fit)@coef3
   > summary(fit)@coef3[, "Std. Error"]
   ```

   But better is

   ```
   > coef(summary(fit))[, "Std. Error"]
   ```

   This is also recommended:

   ```
   > sqrt(diag(vcov(fit)))
   ```

Note that the slot name "coef3" may change in the future so it shouldn't be used directly.

2. Q: *Why aren't the Wald statistics p-values printed out in the summary()?*

   A: VGAM fits so many models from a broad range of statistical fields and subjects that it is potentially dangerous to print out a p-value. The methods function summaryvglm() ought to be accurate for all VGAM family functions, therefore it was thought that it was best to leave out p-values. Possibly an argument pvalues could be added in the future.

8 Other Software

There are several other R implementations for fitting categorical regression models. Some of them are described in Yee (2010), who argues that these implementations tend to be piecemeal in nature, and do not provide a unified framework that is as understandable or comprehensive as VGLMs/VGAMs.

9 Yet To Do

Things yet to be done include

1. Allow for a dispersion parameter. Currently this is possible for GLM-type family functions. A unified and uniform way of handling these is to have dispersion=0 or dispersion=1. It is easy to modify existing VGAM family functions, but this hasn't been done yet—the theory is undeveloped.

2. Improvement to biplots of RR-VGLMs are needed.

3. RR-VGAMs are yet to be implemented. This involves vector projection pursuit regression.

4. Investigate vector smoothing subject to the constraint that the component functions never intersect. This would mean the non-parametric cumulative logit model would not have problems with negative probabilities etc.

Exercises

1. McCullagh (1980) proposed the following model which incorporates dispersion effects:

$$P(Y \leq j|\boldsymbol{x}) \;=\; G\left(\frac{\beta_{(j)0} - \boldsymbol{\beta}^T\boldsymbol{x}}{\tau_x}\right),$$

   where G is the cumulative distribution function of some continuous distribution such as the logistic distribution. Write a VGAM family function to fit this model. It should have the usual parallel and reverse and zero options.

Table 5: Tonsil data. The size of the tonsil is cross-classified as to whether the child was a carrier of Streptococcus pyogenes. The data was collected from 2413 healthy children (Holmes and Williams, 1954).

Carrier?	Present but not enlarged	Enlarged	Greatly enlarged
Yes	19	29	24
No	497	560	269

2. For the multinomial logit model show that Fisher scoring is equivalent to Newton-Raphson. To do this, show that the likelihood score vector $\boldsymbol{d}_i = n_i(\boldsymbol{y}_i^* - \boldsymbol{p}_i^*)$, and $\mathbf{W}_i = n_i(\mathrm{diag}(\boldsymbol{p}_i^*) - \boldsymbol{p}_i^*\boldsymbol{p}_i^{*T})$, where $\boldsymbol{y}_i^* = (y_{i1}, \ldots, y_{iM})^T$ are sample proportions, $n_i = \sum_{j=1}^{M+1} y_{ij}$ is the number of counts with $\boldsymbol{x}_i$, and $\boldsymbol{p}_i = (p_{i1}, \ldots, p_{iM})^T = E(\boldsymbol{y}_i^*)$ are fitted probabilities. The asterix here is used to denote that $\boldsymbol{y}_i^*$ is $\boldsymbol{y}_i$ with $y_{i,M+1}$ dropped. Here, $\boldsymbol{d}_i = \partial\ell_i/\partial\boldsymbol{\eta}_i$, $\ell = \sum_{i=1}^n \ell_i$ is the log-likelihood, and $\mathbf{W}_i = -\partial^2\ell_i/(\partial\boldsymbol{\eta}\,\partial\boldsymbol{\eta}^T)$.

3. Obtain an expression for the marginal effects of a cumulative logit model, cf. (3) for the MLM. Modify `margeff()` to compute this quantity for a `cumulative()` VGLM.

4. Motivate the Bradley-Terry model using a latent variable argument that uses the logistic distribution. Replacing the logistic distribution by a normal distribution results in the Thurstone model (i.e., replace the logit link in (8) by a probit link). Adapt `brat()` to handle both distributions. Is it feasible to handle the `cloglog()` and `tanl()` links as well? If so, implement those too.

5. Another way of modelling paired comparison data is via the model (Rao and Kupper, 1967)

$$\begin{aligned} P(T_i > T_j) &= \pi_i/(\pi_i + q\pi_j), \\ P(T_i < T_j) &= \pi_j/(\pi_j + q\pi_i), \end{aligned}$$

   where $q > 1$. This model can be motivated by latent variables and a logistic distribution since $P(T_i > T_j) = G(\lambda_i - \lambda_j - \tau)$, $\tau \geq 0$, where $G(\cdot)$ is the logistic distribution function and $\pi_i = \exp(\lambda_i)$, $q = \exp(\tau)$.

6. Write a VGAM family function to implement the model (ii) for ties in Section 4.2. It is based on Davidson (1970).

7. Write a VGAM family function to implement triple comparisons, i.e., let

$$\theta_{ijk} = P(T_i > T_j > T_k)$$

   be the probability that T_i is preferred to T_j and T_j is preferred to T_k. One way is to choose

$$\theta_{ijk} = \frac{\pi_i\pi_j}{(\pi_i + \pi_j + \pi_k)(\pi_j + \pi_k)}.$$

   An alternative is

$$\theta_{ijk} = \frac{\pi_i^2\pi_j}{\pi_i^2\pi_j + \pi_j^2\pi_i + \pi_i^2\pi_k + \pi_k^2\pi_i + \pi_j^2\pi_k + \pi_k^2\pi_j}.$$

   Write a VGAM family function for each of these models.

8. Run the help-file example of the family function `brat()`. Show that the working residuals are all zero (or zero to working precision, depending on the convergence criterion—can you get the model to iterate even closer to the MLE?). Explain the theory behind why this is to be expected. Are Pearson residuals defined when the working residuals are all zero? Should `summary.vglm()` print out the working residuals if the Pearson residuals are undefined?

9. (a) Consider the standard *random utility maximization* model (RUM), with an indirect utility function given by

$$U_{ij} \;=\; \eta_{ij} + \varepsilon_{ij} \tag{12}$$

with $i = 1, \ldots, n$ and $j = 1, \ldots, J$. Then U_{ij} is the *utility* that individual i gains from outcome j. Let the observed individual characteristics be $\boldsymbol{x}_i$ and the characteristics of the outcomes be $\boldsymbol{z}_{ij}$, then

$$\eta_{ij} \;=\; \boldsymbol{x}_i^T \boldsymbol{\beta}_j + \boldsymbol{z}_{ij}^T \boldsymbol{\alpha}. \tag{13}$$

Finally, ε_{ij} in (12) is assumed to follow an independent extreme value distribution. The probability that individual i selects outcome j is given by

$$P_{ij} \;=\; P\left[U_{ij} = \max(U_{i1}, \ldots, U_{iJ})\right]$$

with choice set $C = \{1, 2, \ldots, J\}$, and associated multinomial logit model probabilities given by

$$P_{ij}^m \;=\; \frac{\exp(\eta_{ij})}{\sum_{k=1}^J \exp(\eta_{ik})}. \tag{14}$$

(b) Write a VGAM family function called `dogit()` to implement the dogit model. The following description comes from Fry and Harris (2005). The multinomial logit model imposes some strong restrictions on the model, most notably that of the independence of irrelevant alternatives (IIA). The IIA property of the MLM says that the odds ratio P_{ij}/P_{ik}, $j \neq k$, is independent of all other alternatives, and independent of additions to, and deletions from, the full choice set. This can be an unrealistic assumption.

Of the several non-IIA alternatives to the MLM proposed, the dogit model of Gaudry and Dagenais (1979) is attractive. The dogit model expands on (14) by having

$$P_{ij}^d \;=\; \frac{\exp(\eta_{ij}) + \theta_j \sum_{k=1}^J \exp(\eta_{ik})}{\left(1 + \sum_{k=1}^J \theta_k\right) \sum_{k=1}^J \exp(\eta_{ik})}. \tag{15}$$

Here, $\theta_j \geq 0$ for $j = 1, \ldots, J$. This can be motivated by a two-part choice process (Fry and Harris, 1996).

Nb. Gaudry and Dagenais (1979) say that the IIA property of a model requires that the relative probabilities of any two alternatives be independent of the attributes of the remaining alternatives whether the latter are present or absent from the set of available choices. This property, introduced by Luce (1959) to explain the psychology of choice, was used by McFadden (1968) to derive the multinomial logit model within a different theoretical framework. Gaudry and Dagenais (1979) call it the "dogit" model because "the model avoids or dodges the researcher's dilemma of choosing *a priori* between a format which commits to IIA restrictions and one which excludes them—whence its name."

(c) Write a VGAM family function called `ogev()` to implement the standard ordered GEV model, whose probabilities are given by

$$P_{ij}^o \;=\; \frac{\xi_{ij}\left((\xi_{i,j-1} + \xi_{ij})^{\rho-1} + (\xi_{ij} + \xi_{i,j+1})^{\rho-1}\right)}{\sum_{r=1}^{J+1} (\xi_{i,r-1} + \xi_{ir})^{\rho}} \tag{16}$$

where $\xi_{ij} = \exp(\eta_{ij}/\rho)$ with the η_{ij} given by (13), $\xi_{i0} = \xi_{i,J+1} = 0$ and $0 < \rho \leq 1$.

(d) Write a VGAM family function called `dogev()` to implement the DOGEV (dogit ordered GEV) model whose probabilities are

$$P_{ij}^D \;=\; \frac{\theta_j}{1+\sum_{k=1}^M \theta_k} + \frac{P_{ij}^o}{1+\sum_{k=1}^M \theta_k}. \tag{17}$$

Note that in all models, the log-likelihood function is

$$\ell^k(\boldsymbol{\phi}) \;=\; \sum_{i=1}^n \sum_{j=1}^J d_{ij} \log P_{ij}^k$$

with $k = m, d, o, D$ and model parameter vector $\boldsymbol{\phi} = (\boldsymbol{\beta}, \boldsymbol{\alpha})$, $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta})$, $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \rho)$, and $(\boldsymbol{\beta}, \boldsymbol{\alpha}, \boldsymbol{\theta}, \rho)$ respectively, where $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_J)$. Here, $d_{ij} = 1$ if individual i chooses outcome j, else $d_{ij} = 0$. Of course, one should use the `zero` argument to make ρ intercept-only by default.

(e) Show that

$$P_{ij}^d \;=\; \frac{\theta_j}{1+\sum_{k=1}^J \theta_k} + \frac{P_{ij}^m}{1+\sum_{k=1}^J \theta_k}.$$

References

Agresti, A., 2002. Categorical Data Analysis, 2nd Edition. Wiley, New York, USA.

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

Armstrong, B. G., Sloan, M., 1989. Ordinal regression models for epidemiologic data. American Journal of Epidemiology 129, 191–204.

Bradley, R. A., Terry, M. E., 1952. Rank analysis of incomplete block designs. I. The method of paired comparisons. Biometrika 39, 324–345.

Chambers, J. M., Hastie, T. J. (Eds.), 1993. Statistical Models in S. Chapman & Hall, New York, USA.

Davidson, R. R., 1970. On extending the Bradley Terry model to accommodate ties in paired comparison experiments. J. Amer. Statist. Assoc. 65, 317–328.

Fahrmeir, L., Tutz, G., 1994. Multivariate Statistical Modelling Based on Generalized Linear Models. Springer-Verlag.

Fahrmeir, L., Tutz, G., 2001. Multivariate Statistical Modelling Based on Generalized Linear Models, 2nd Edition. Springer-Verlag, New York, USA.

Firth, D., 2005. Bradley-Terry models in R. Journal of Statistical Software 12 (1), 1–12.

Fry, T. R. L., Harris, M., 1996. A Monte Carlo study of tests for the independence of irrelevant alternatives property. Transportation Research, Part B 30B (1), 19–30.

Fry, T. R. L., Harris, M. N., 2005. The dogit ordered generalized extreme value model. The Australian and New Zealand Journal of Statistics 47 (4), 531–542.

Gaudry, M., Dagenais, M., 1979. The dogit model. Transportation Research, Part B 13B (2), 105–112.

Holmes, M. C., Williams, R., 1954. The distribution of carriers of Streptococcus Pyogenes amongst 2413 healthy children. J. Hyg. Camb. 52, 165–179.

Kuk, A. Y. C., 1995. Modelling paired comparison data with large numbers of draws and large variability of draw percentage among players. The Statistician 44, 523–528.

Leonard, T., 2000. A Course in Categorical Data Analysis. Chapman & Hall/CRC, Boca Raton, FL, USA.

Liu, I., Agresti, A., 2005. The analysis of ordered categorical data: An overview and a survey of recent developments. Sociedad Estadística e Investigación Operativa Test 14 (1), 1–73.

Lloyd, C. J., 1999. Statistical Analysis of Categorical Data. Wiley, New York, USA.

Long, J. S., 1997. Regression Models for Categorical and Limited Dependent Variables. Sage Publications, Thousand Oaks, CA, USA.

Luce, R. D., 1959. Individual Choice Behavior. Wiley, New York, USA.

McCullagh, P., 1980. Regression models for ordinal data. J. Roy. Statist. Soc. Ser. B 42 (2), 109–142.

McCullagh, P., Nelder, J. A., 1989. Generalized Linear Models, 2nd Edition. Chapman & Hall, London.

McFadden, D., 1968. The revealed preferences of a government bureaucracy. Tech. Rep. 17, Department of Economics, University of California.

Peterson, B., 1990. Letter to the editor: Ordinal regression models for epidemiologic data. American Journal of Epidemiology 131, 745–746.

Peterson, B., Harrell, F. E., 1990. Partial proportional odds models for ordinal response variables. Applied Statistics 39 (2), 205–217.

Rao, P. V., Kupper, L. L., 1967. Ties in paired comparison experiments: a generalisation of the Bradley Terry. J. Amer. Statist. Assoc. 62, 192–204.

Ripley, B. D., 1996. Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge.

Simonoff, J. S., 2003. Analyzing Categorical Data. Springer-Verlag, New York, USA.

Thompson, L. A., 2009. R (and S-PLUS) Manual to Accompany Agresti's *Categorical Data Analysis* (2002), 2^{nd} edition.
URL `https://home.comcast.net/~lthompson221/Splusdiscrete2.pdf`

Torsney, B., 2004. Fitting Bradley Terry models using a multiplicative algorithm. In: Antoch, J. (Ed.), Proceedings in Computational Statistics COMPSTAT 2004. Physica-Verlag, Heidelberg, pp. 513–526.

Yee, T. W., 2008. The VGAM package. R News 8 (2), 28–39.
URL `http://CRAN.R-project.org/doc/Rnews/`

Yee, T. W., 2010. The VGAM package for categorical data analysis. Journal of Statistical Software 32 (10), 1–34.
URL http://www.jstatsoft.org/v32/i10/

Yee, T. W., Hastie, T. J., 2003. Reduced-rank vector generalized linear models. Statistical Modelling 3 (1), 15–41.

Zermelo, E., 1929. Die berechnung turnier-ergebnisse als ein maximumproblem der wahrscheinlichkeitsrechnung. Math. Zeit. 29, 436–460.