

# VGAM Family Functions for Categorical Data

Beta Version 0.5-15

© Thomas W. Yee

Department of Statistics,  
University of Auckland,  
New Zealand

yee@stat.auckland.ac.nz  
<http://www.stat.auckland.ac.nz/~yee>

January 16, 2004

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Nominal Responses</b>	<b>2</b>
2.1	Multinomial logit model . . . . .	3
2.2	Stereotype Model . . . . .	3
<b>3</b>	<b>Ordinal Responses</b>	<b>4</b>
3.1	Models Involving Cumulative Probabilities . . . . .	4
3.2	Models Involving Stopping-ratios and Continuation-ratios . . . . .	5
3.3	Models Involving Adjacent Categories . . . . .	6
<b>4</b>	<b>Other Topics</b>	<b>6</b>
4.1	Input . . . . .	6
4.2	Output . . . . .	7
4.3	Constraints . . . . .	7
4.4	Implementation Details . . . . .	7
4.5	Convergence . . . . .	7

4.6	Over-dispersion . . . . .	8
4.7	Relationship with <code>binomialff()</code> . . . . .	8
4.8	The $x_{ij}$ Argument . . . . .	8
4.9	Reduced-rank Regression . . . . .	9
<b>5</b>	<b>Tutorial Examples</b>	<b>9</b>
5.1	Multinomial logit model . . . . .	9
5.2	Stereotype model . . . . .	10
5.3	Proportional odds model . . . . .	11
5.4	Stopping Ratio Model . . . . .	12
<b>6</b>	<b>Other Software</b>	<b>13</b>
<b>7</b>	<b>Yet To Do</b>	<b>14</b>
	<b>Exercises</b>	<b>14</b>
	<b>References</b>	<b>14</b>

[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  $Y = 1, 2, \dots, M + 1$ . Table 2 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 1 where the stages of a disease are  $Y = 1$  for none,  $Y = 2$  for mild, and  $Y = 3$  for severe symptoms.

General references for categorical data include Agresti (1990), Leonard (2000), Lloyd (1999) and McCullagh and Nelder (1989). 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 4.1.

## 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

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

where

$$\eta_j(\mathbf{x}) = \beta_j^T \mathbf{x}.$$

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

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

VGAM fits the multinomial logit model using the family function `multinomial()`. It uses the last column of the response matrix as baseline, 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).

## 2.2 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

Table 1: *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

$M \times p$  matrix of regression coefficients by a lower rank matrix. In detail, the reduced-rank concept replaces  $\mathbf{B} = (\beta_1, \dots, \beta_M)^T$  (without the intercepts) by

$$\mathbf{B} = \mathbf{C}\mathbf{A}^T \quad (1)$$

where  $\mathbf{C} = (\mathbf{c}_1 \mathbf{c}_2 \dots \mathbf{c}_r)$  is  $p \times r$ ,  $\mathbf{A} = (\mathbf{a}_1 \mathbf{a}_2 \dots \mathbf{a}_r)$  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

$$\eta_i = \eta_0 + \mathbf{A}\mathbf{C}^T \mathbf{x}_i = \eta_0 + \mathbf{A}\nu_i.$$

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 (1) is not unique as  $\eta_i = \eta_0 + \mathbf{A}\mathbf{M}\mathbf{M}^{-1}\nu_i$  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

$$\mathbf{A} = \begin{pmatrix} \mathbf{I}_r \\ \tilde{\mathbf{A}} \end{pmatrix},$$

where  $\tilde{\mathbf{A}}$  is a  $(M - r) \times r$  matrix. In fact, any  $r$  rows of  $\mathbf{A}$  may be chosen to represent  $\mathbf{I}_r$ <sup>1</sup>. 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  $P(Y \leq j | \mathbf{x})$  (see McCullagh (1980)), in particular, the *proportional odds model* is

$$\text{logit } P(Y \leq j | \mathbf{x}) = \beta_{(j)0} + \beta^T \mathbf{x}, \quad j = 1, \dots, M.$$

<sup>1</sup>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.

Table 2: Quantities defined in VGAM for an ordinal categorical response  $Y$  taking values  $1, \dots, M + 1$ . Covariates  $\mathbf{x}$  have been omitted for clarity.

Quantity	Notation	Range of $j$	VGAM Family function
$P(Y = j + 1)/P(Y = j)$	$\zeta_j$	$1, \dots, M$	<code>acat()</code>
$P(Y = j)/P(Y = j + 1)$	$\zeta_j^R$	$2, \dots, M + 1$	<code>acat(reverse=T)</code>
$P(Y > j   Y \geq j)$	$\delta_j^*$	$1, \dots, M$	<code>cratio()</code>
$P(Y < j   Y \leq j)$	$\delta_j^{*R}$	$2, \dots, M + 1$	<code>cratio(reverse=T)</code>
$P(Y \leq j)$	$\gamma_j$	$1, \dots, M$	<code>cumulative()</code>
$P(Y \geq j)$	$\gamma_j^R$	$2, \dots, M + 1$	<code>cumulative(reverse=T)</code>
$\log\{P(Y = j)/P(Y = M + 1)\}$		$1, \dots, M$	<code>multinomial()</code>
$P(Y = j   Y \geq j)$	$\delta_j$	$1, \dots, M$	<code>sratio()</code>
$P(Y = j   Y \leq j)$	$\delta_j^R$	$2, \dots, M + 1$	<code>sratio(reverse=T)</code>

We call models of the form

$$\text{logit } P(Y \leq j|\mathbf{x}) = \eta_j \quad j = 1, \dots, M$$

*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  $\beta_1 = \dots = \beta_M$ . 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

$$\text{logit } P(Y \leq j|\mathbf{x}) = \beta_{(j)0} + \eta, \quad j = 1, \dots, M. \quad (2)$$

If a complementary log-log link is chosen in (2) the result is known as the *proportional hazards model*. In theory, any other link function used for binomial data such as the probit link can be applied to cumulative probability models.

VGAM fits models based on cumulative probabilities using the family function `cumulative()`. It supports a variety of link functions and the parallelism assumption  $\beta_1 = \dots = \beta_M$  (or more generally,  $f_{(1)k}(x_k) = \dots = f_{(M)k}(x_k)$ ). For example,

```
vglm(y ~ x1, cumulative(link=probit, reverse=T, parallel=T), mydataframe)
```

fits the model

$$\Phi^{-1}\{P(Y \geq j|\mathbf{x})\} = \beta_{(j-1)0} + \beta_1 x_1, \quad j = 2, \dots, M + 1.$$

### 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 the probability of *stopping* at  $Y = j$  or *continuing* past

Table 3: Summary of whether `parallel=T` ever applies to the intercepts.

family=	Default: parallel=	Apply to intercepts?
acat	F	No
cratio	F	No
cumulative	F	No
multinomial	F	Yes
sratio	F	No

$Y = j$ , 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 2 summarizes the VGAM definition of a continuation-ratio, and also lists a quantity termed the *stopping-ratio* to help distinguish between the two types. Note that, regardless of type, one sometimes wishes to reverse  $Y$  so that it is enumerated from  $M + 1, M, \dots, 1$  instead; this is handled by prefixing their name by “reverse,” e.g.,  $\delta_j^R \equiv P(Y = j | Y \leq j)$ ,  $j = 2, \dots, M + 1$  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 2, VGAM supports a variety of link functions, the parallelism assumption  $\beta_1 = \dots = \beta_M$  (or more generally,  $f_{(1)k}(x_k) = \dots = f_{(M)k}(x_k)$ ), and the `zero=` argument to constrain  $\eta_j$  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 2 are all interrelated. The following formulae give some of these relationships.

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

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

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

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

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

## 4 Other Topics

### 4.1 Input

The response in `vg1m()/vgam()` can be

1. a  $n \times (M + 1)$  matrix of counts. The columns are best labelled, and the  $j$ th column denotes  $Y = j$ .
2. a vector. The unique values, when sorted (or levels if a factor), denote the  $M + 1$  levels from  $1, \dots, M + 1$ . If a factor, it may be ordered or unordered. The functions `factor()`, `ordered()`, `levels()`, `codes()` are useful; see the S-PLUS 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.

## 4.2 Output

Suppose `fit` is a categorical VGAM object. Like `binomialff()`, the fitted values (in `fitted(fit)`) are probabilities and `fit@prior.weights` 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.

## 4.3 Constraints

All categorical data family functions have the `parallel` and `zero` arguments. By default, `parallel=F` 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=T` is applied to the intercepts. Also, all categorical data family functions have a `reverse=F` argument.

## 4.4 Implementation Details

The S-PLUS expression `process.categorical.data.vgam` provides a unified way of handling the response variable. The variable `delete.zero.cols` should be set to `TRUE` or `FALSE` prior to evaluating `process.categorical.data` in `@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.

## 4.5 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 guaranteed 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.

## 4.6 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

$$\tilde{\sigma}^2 = X^2 / \{nM - p\} = X^2 / \{\text{residual d.f.}\}, \quad (3)$$

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

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 (3) for some models.

## 4.7 Relationship with `binomialff()`

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

If `y` is a vector of 0's and 1's 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(rev=T), ...)`.

These hold because the final level is used as the baseline category for the MLM.

## 4.8 The `xij` Argument

The `xij` argument allows for covariate values that are specific to each  $\eta_j$ . 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:

$$\begin{aligned} \text{logit } P(Y_{ij} = 1 | x_{ij}) &= \beta_{(1)0} + \beta_{(1)1} x_{ij}, \quad j = 1, 2; \quad i = 1, \dots, n; \\ \log \psi &= \beta_{(3)0}, \end{aligned} \quad (4)$$

where the dependency between the responses is modelled through the odds ratio  $\psi$ . 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:  $X_1 = \text{cost}$ , and  $X_2 = \text{journey duration}$ . 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  $M = 3$ .

$$\log \{P(Y = j)/P(Y = M + 1)\} = \eta_j = \beta_{(j)0} + \beta_1 x_{i1j} + \beta_2 x_{i2j}, \quad j = 1, \dots, M,$$

where  $x_{i1j}$  is the cost for the  $j$ th transport means for the  $i$ th person, and  $x_{i2j}$  is the journey duration of the  $j$ th transport means for the  $i$ th person. Note again that the regression coefficients  $\beta_1$  and  $\beta_2$  are the same.

To fit such models in VGAM one needs to use the `xij` argument. See the documentation elsewhere concerning this important feature.

## 4.9 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. For more details, see Yee and Hastie (2003).

Examples supplied included are `rrvam()`, `rrcumulative()`, and `rracat()`. The alternating algorithm is performed using the function `valt()`.

## 5 Tutorial Examples

In this section we illustrate some of the models. We will make use of the pneumoconiosis data in Table 1 (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  $x = \log(\text{exposure time})$ .

### 5.1 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$let <- log(pneumo$exposure.time)
> fit.mlm <- vglm(cbind(normal,mild,severe) ~ let, multinomial, pneumo)
> coef(fit.mlm, matrix=T)
              (Intercept)          let
log(mu[,1]/mu[,3])  11.975092 -3.0674665
log(mu[,2]/mu[,3])   3.039062 -0.9020936
> summary(fit.mlm)
```

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

```
Coefficients:
              Value Std. Error  t value
```

```

(Intercept):1 11.9750920 2.0004445 5.986216
(Intercept):2 3.0390622 2.3760683 1.279030
      let:1 -3.0674665 0.5652065 -5.427161
      let:2 -0.9020936 0.6689816 -1.348458

Number of linear predictors: 2

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

(Dispersion Parameter for Multinomial logit model family taken to be 1)

Residual Deviance: 5.347382 on 12 degrees of freedom

Number of Iterations: 4

```

Note that  $\mu[j,]$  is the  $j$ th 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)`.

## 5.2 Stereotype model

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

```

> fit1 = rrvglm(cbind(normal,mild,severe) ~ let, multinomial, Rank=1, pneumo)
> fit1@constraints
$(Intercept)":
      [,1] [,2]
[1,]    1    0
[2,]    0    1

$let:
      [,1]
[1,] 1.0000000
[2,] 0.2940857

> coef(fit1)
(Intercept):1 (Intercept):2      let
      11.9751      3.03908 -3.067469
> coef(fit1, matrix=T)
      (Intercept)      let
log(mu[,1]/mu[,3])    11.97510 -3.0674690
log(mu[,2]/mu[,3])    3.03908 -0.9020987

```

That is,  $\mathbf{A} = \text{fit1@constraints}\$let = (1, 0.2940857)^T$  and  $\mathbf{C} = (-3.067469)$ . Note that `coef(fit1, matrix=T)` 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.

### 5.3 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

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

We can fit this by

```
fit.pom <- vglm(cbind(normal,mild,severe) ~ let, cumulative(par=T), pneumo)
```

Then `coef(fit.pom, mat=T)` gives the matrix of regression coefficients

```
              (Intercept)      let
logit(P[Y<=1])    9.676124 -2.596816
logit(P[Y<=2])   10.581756 -2.596816
```

which agrees to that of McCullagh & Nelder. `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*.

Typing `summary(fit.pom)` gives

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

Coefficients:

```
              Value Std. Error  t value
(Intercept):1  9.676124  1.3241658  7.307336
(Intercept):2 10.581756  1.3455201  7.864435
           let  -2.596816  0.3811267 -6.813525
```

Number of linear predictors: 2

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

(Dispersion Parameter for Cumulative logit model family taken to be 1)

Residual Deviance: 5.026826 on 13 degrees of freedom

Number of Iterations: 3

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)
> 1-pchisq(deviance(fit.pom)-deviance(fit.npom),
           df=df.residual(fit.pom)-df.residual(fit.npom))
[1] 0.7058849
```

There is no evidence against the proportional odds assumption.

## 5.4 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 4, 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  $1$ .

```
> y = matrix(c(19, 29, 24, 497, 560, 269), 2, 3, byrow=T)
> dimnames(y) = list(NULL,
+   c("present but not enlarged", "enlarged", "greatly enlarged"))
> carrier = c(1, -1)
> y
      present but not enlarged enlarged greatly enlarged
[1,]                19         29         24
[2,]               497        560        269
> fit = vglm(y ~ carrier, sratio(par=T, rev=F), tr=T, crit="c")
VGLM  linear loop 1 : coefficients =
-0.77522, 0.46797, -0.26421
VGLM  linear loop 2 : coefficients =
-0.77525, 0.46795, -0.26423
VGLM  linear loop 3 : coefficients =
-0.77525, 0.46795, -0.26423
VGLM  linear loop 4 : coefficients =
-0.77525, 0.46795, -0.26423
> fit
Call:
vglm(formula = y ~ carrier, family = sratio(par = T, rev = F),
      tr = T, crit = "c")

Coefficients:
(Intercept):1 (Intercept):2      carrier
   -0.77525      0.46795      -0.26423

Degrees of Freedom: 4 Total; 1 Residual
Residual Deviance: 0.0056573
Log-likelihood: -1477.6
> coef(fit, matrix=T)
      logit(P[Y=1|Y>=1]) logit(P[Y=2|Y>=2])
(Intercept)      -0.77525      0.46795
```

Table 4: *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

```

carrier          -0.26423          -0.26423
> summary(fit)

Call:
vglm(formula = y ~ carrier, family = sratio(par = T, rev = F),
      tr = T, crit = "c")

Pearson Residuals:
  logit(P[Y=1|Y>=1]) logit(P[Y=2|Y>=2])
1           0.0510          -0.0524
2          -0.0108           0.0141

Coefficients:
              Value Std. Error t value
(Intercept):1 -0.775    0.1061  -7.31
(Intercept):2  0.468    0.1116   4.19
carrier        -0.264    0.0989  -2.67

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: 0.006 on 1 degrees of freedom

Log-likelihood: -1477.6 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          -1.03948          0.20372
2          -0.51102          0.73218
> fitted(fit)
  present but not enlarged enlarged greatly enlarged
1           0.26125  0.40687          0.33188
2           0.37495  0.42208          0.20296
> fitted(fit) * fit@prior.weights
  present but not enlarged enlarged greatly enlarged
1           18.81  29.295          23.895
2          497.19 559.682          269.128

```

## 6 Other Software

There are several other R/S-PLUS implementations for fitting categorical regression models. `polr` in library MASS fits cumulative logits, but as it is written in S it can easily be altered to allow for a probit link (change

pnorm to plogis, dnorm to dlogis, and alter the glm call to use probits.)

J. Lange has written software, and there is the DESIGN/HMISC library of F. Harrell.

## 7 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:

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

where  $G$  is the cdf of some continuous distribution such as the logistic distribution. Write a VGAM family function to fit this model. It should allow the logit, probit, complementary log-log and identity links, as well as the usual `parallel` and `reverse` and `zero` options.

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

## References

- Agresti, A., 1990. *Categorical Data Analysis*. Wiley, New York.
- Anderson, J. A., 1984. Regression and ordered categorical variables (with discussion). *Journal of the Royal Statistical Society, Series B, Methodological* 46, 1–30.
- Armstrong, B. G., Sloan, M., 1989. Ordinal regression models for epidemiologic data. *American Journal of Epidemiology* 129, 191–204.
- Chambers, J. M., Hastie, T. J. (Eds.), 1993. *Statistical Models in S*. Chapman & Hall, New York.
- Fahrmeir, L., Tutz, G., 1994. *Multivariate Statistical Modelling Based on Generalized Linear Models*. Springer-Verlag.
- Holmes, M. C., Williams, R., 1954. The distribution of carriers of *Streptococcus Pyogenes* amongst 2413 healthy children. *J. Hyg. Camb.* 52, 165–179.
- Leonard, T., 2000. *A Course in Categorical Data Analysis*. Chapman & Hall/CRC, Boca Raton, FL.
- Lloyd, C. J., 1999. *Statistical Analysis of Categorical Data*. Wiley, New York.
- McCullagh, P., 1980. Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B, Methodological* 42, 109–142.
- McCullagh, P., Nelder, J. A., 1989. *Generalized Linear Models*, 2nd Edition. Chapman & Hall, London.
- Peterson, B., 1990. Letter to the editor: Ordinal regression models for epidemiologic data. *American Journal of Epidemiology* 131, 745–746.
- Ripley, B. D., 1996. *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge.
- Yee, T. W., Hastie, T. J., 2003. Reduced-rank vector generalized linear models. *Statistical Modelling* 3, 15–41.