

# VGAM Family Functions for Log-linear Models

T. W. Yee

October 30, 2006

Beta Version 0.6-5

© Thomas W. Yee

Department of Statistics,  
University of Auckland,  
New Zealand  
yee@stat.auckland.ac.nz  
<http://www.stat.auckland.ac.nz/~yee>

## Contents

<b>1</b>	<b>Introduction</b>	<b>2</b>
1.1	VGLMs and VGAMs . . . . .	2
<b>2</b>	<b>VGAM Family Functions</b>	<b>2</b>
2.1	Some Background . . . . .	3
<b>3</b>	<b>Other Topics</b>	<b>5</b>
3.1	Input . . . . .	5
3.2	Output . . . . .	5
3.3	Constraints . . . . .	5
3.4	Advice . . . . .	5
3.5	Generic Functions . . . . .	5
<b>4</b>	<b>Tutorial Example</b>	<b>6</b>
4.1	Coalminers' Data . . . . .	6
4.2	Hunua Forest Data . . . . .	7

<b>Exercises</b>	<b>9</b>
<b>Acknowledgements</b>	<b>9</b>
<b>References</b>	<b>9</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 log-linear models for binary responses. Currently, only a limited selection of models are available. See the document on categorical data for related work.

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, see Yee and Wild (2001), and the VGAM *User Manual* should be consulted for general instructions about the software. General books with log-linear model content include Bishop et al. (1975), Christensen (1997), Lindsey (1995), and McCullagh and Nelder (1989).

### 1.1 VGLMs and VGAMs

Recall that we originally defined the VGAM class as any model for which the conditional distribution of  $\mathbf{y}$  (which may be multivariate) given  $\mathbf{x}$  is of the form

$$f(\mathbf{y}|\mathbf{x};\boldsymbol{\eta}) = h(\mathbf{y}, \eta_1, \dots, \eta_M) \quad (1)$$

where  $h(\cdot)$  is some known function and

$$\eta_j(\mathbf{x}) = \beta_{(j)0} + \sum_{k=1}^p f_{(j)k}(x_k), \quad (2)$$

are additive predictors. VGLMs constrain  $\eta_j = \boldsymbol{\beta}_j^T \mathbf{x}$ .

## 2 VGAM Family Functions

VGAM family functions called `loglinb2()` and `loglinb3()` have been written for bi-/tri-variate binary responses. The first fits

$$\log P(Y_1 = y_1, Y_2 = y_2|\mathbf{x}) = u_0(\mathbf{x}) + u_1(\mathbf{x})y_1 + u_2(\mathbf{x})y_2 + u_{12}(\mathbf{x})y_1y_2 \quad (3)$$

where  $y_j = 0$  or  $1$ ,

$$\begin{pmatrix} u_1(\mathbf{x}) \\ u_2(\mathbf{x}) \\ u_{12}(\mathbf{x}) \end{pmatrix} = \boldsymbol{\eta}(\mathbf{x}) = \begin{pmatrix} \eta_1(\mathbf{x}) \\ \eta_2(\mathbf{x}) \\ \eta_3(\mathbf{x}) \end{pmatrix},$$

and the  $\eta_j$  are additive predictors (2). The function `loglinb3()` fits

$$\log P(Y_1 = y_1, Y_2 = y_2, Y_3 = y_3 | \mathbf{x}) = u_0(\mathbf{x}) + u_1(\mathbf{x}) y_1 + u_2(\mathbf{x}) y_2 + u_3(\mathbf{x}) y_3 + u_{12}(\mathbf{x}) y_1 y_2 + u_{13}(\mathbf{x}) y_1 y_3 + u_{23}(\mathbf{x}) y_2 y_3 \quad (4)$$

where

$$\begin{pmatrix} u_1(\mathbf{x}) \\ u_2(\mathbf{x}) \\ u_3(\mathbf{x}) \\ u_{12}(\mathbf{x}) \\ u_{13}(\mathbf{x}) \\ u_{23}(\mathbf{x}) \end{pmatrix} = \boldsymbol{\eta}(\mathbf{x}) = \begin{pmatrix} \eta_1(\mathbf{x}) \\ \eta_2(\mathbf{x}) \\ \eta_3(\mathbf{x}) \\ \eta_4(\mathbf{x}) \\ \eta_5(\mathbf{x}) \\ \eta_6(\mathbf{x}) \end{pmatrix}.$$

As an example, if `ymatrix` is a  $n$  by 2 matrix of ones and zeros, then

```
fit = vgam(ymatrix ~ s(x, df=c(4,2)), loglinb2(exchangeable=TRUE))
```

would fit

$$\log P(Y_1 = y_1, Y_2 = y_2 | x) = u_0(x) + u_1(x) y_1 + u_2(x) y_2 + u_{12}(x) y_1 y_2,$$

subject to  $u_1 = u_2$ , using vector (smoothing) splines. Here,  $u_1$  and  $u_{12}$  are assigned 4 and 2 degrees of freedom respectively (1 = linear fit).

As another example,

```
fit = vgam(ymatrix ~ s(x, df=c(4,2)), loglinb3(exchangeable=TRUE))
```

fits (4) subject to  $u_1 = u_2 = u_3$  and  $u_{12} = u_{13} = u_{23}$ . It is harder to think of applications of exchangeable models with 3 binary responses, whereas there are plentiful examples with 2 binary responses, e.g., with eye and ear data.

## 2.1 Some Background

Suppose generally that the data are  $(y_{i1}, \dots, y_{iS}, \mathbf{x}_i)$ ,  $i = 1, \dots, n$ , where each  $y_{ij}$  is a binary response. For log-linear models it is often a good idea to force  $u_{jkl} \equiv 0$  and similarly for other higher order associations. Unless the data contain all fitted combinations, such assumptions are often necessary because the estimates become unbounded. It also reduces the complexity of the problem. Furthermore, higher order associations become increasingly more difficult to interpret.

In general we consider the log-linear model

$$\log P(Y_1 = y_1, \dots, Y_S = y_S | \mathbf{x}) = u_0(\mathbf{x}) + \sum_{j=1}^S u_j(\mathbf{x}) y_j + \sum_{j < k} u_{jk}(\mathbf{x}) y_j y_k. \quad (5)$$

The normalizing parameter  $u_0$  satisfies

$$e^{-u_0} = 1 + \sum_{j=1}^S e^{u_j} + \sum_{j < k} e^{u_j + u_k + u_{jk}} + \sum_{j < k < \ell} e^{u_j + u_k + u_\ell + u_{jk} + u_{j\ell} + u_{k\ell}} + \dots + \exp\left(\sum_{j=1}^S u_j + \sum_{j < k} u_{jk}\right).$$

One has  $\boldsymbol{\eta} = (\eta_1, \dots, \eta_M)^T = (u_1, \dots, u_S, u_{12}, \dots, u_{S-1,S})^T$  where  $M = S(S+1)/2$ . (An identity link for each of the  $u$ 's is chosen because the parameter space is unconstrained.)

With IRLS, it may be shown that Newton-Raphson coincides with Fisher scoring. Although the iterative proportional fitting procedure (IPFP; Bishop et al. (1975)) is the usual algorithm for log-linear models, our limited experience has indicated that IRLS works well.

## 3 Other Topics

### 3.1 Input

The response input for `loglinb2()` and `loglinb3()` must be a 2 and 3 column matrix of 1's and 0's, respectively. The `weights` argument may be used to input more than one observation per row.

### 3.2 Output

`loglinb2()` and `loglinb3()` return in the `fitted` component of the `vglm()/vgam()` object the joint probabilities. These should be extracted using, e.g., `fitted(fit)`. As an example, for  $S = 2$ , the order is  $\hat{P}(Y_{i1} = 0, Y_{i2} = 0)$ ,  $\hat{P}(Y_{i1} = 0, Y_{i2} = 1)$ ,  $\hat{P}(Y_{i1} = 1, Y_{i2} = 0)$ ,  $\hat{P}(Y_{i1} = 1, Y_{i2} = 1)$ ; see Section 4.

### 3.3 Constraints

Log-linear VGAM family functions have the `exchangeable` and `zero` arguments. The latter can be assigned a vector taking values in  $\{1, 2, \dots, S\}$ . If the vector contains a value  $j$ , then  $\eta_j$  to be modelled as an intercept only. This is often a good idea for the association parameters, as allowing them too much flexibility can lead to problems during estimation.

The argument `exchangeable=FALSE` by default, and can be set to `TRUE` (for all terms in the model), or a formula such as `exchangeable = TRUE ~ x1 + x3` (meaning all terms except the intercept, `x1` and `x3` are not exchangeable.)

### 3.4 Advice

Our experience has shown that association parameters should be more smooth than the marginals, e.g., linear or an intercept only. If not, the  $\mathbf{W}_i$  may become non-positive-definite during estimation, leading to problems.

### 3.5 Generic Functions

There are a number of methods functions that support objects of class "vgam", for example, `resid(fit, "working")` returns the working residuals, and `fitted(fit)` returns the  $n$  by  $2^S$  matrix of joint probabilities. The methods function `plot.vgam()` is available for `vglm()/vglm()` and `vgam()` objects.

## 4 Tutorial Example

We illustrate some of the software by fitting a simple model or two.

### 4.1 Coalminers' Data

We try to mimic the results of §6.6 of McCullagh and Nelder (1989). At the moment, `loglinb2()` only takes a  $n \times 2$  matrix of 0's and 1's as input. Below, `counts` is  $n \times 4$  and `yy` is  $n \times 2$ .

```
> data(coalminers)
> coalminers = transform(coalminers, age = (age - 42)/5)
> temp = vglm(cbind(nBnW, nBW, BnW, BW) ~ age, binom2.or,
+   coalminers)
> counts = round(c(weights(temp, type = "prior")) * temp@y)
> fred = matrix(c(0, 0, 0, 1, 1, 0, 1, 1), 4, 2, byrow = TRUE)
> yy = kronecker(matrix(1, nrow(counts), 1), fred)
> wt = c(t(counts))
> age = rep(coalminers$age, rep(4, length(coalminers$age)))
> yy = yy[wt > 0, ]
> age = age[wt > 0]
> wt = wt[wt > 0]
> fit = vglm(yy ~ age, loglinb2, wei = wt)
> coef(fit, mat = TRUE)
```

	u1	u2	u12
(Intercept)	-3.4777878	-2.0089779	3.0594787
age	0.5154003	0.2006049	-0.1661525

As an exercise, reconcile this with p.234 of McCullagh and Nelder (1989) who obtain

$$\begin{aligned}\text{logit } P(Y_1 = 1|Y_2 = 1, x) &= -0.418 + 0.349x \\ \text{logit } P(Y_2 = 1|Y_1 = 1, x) &= 1.051 + 0.034x \\ \text{log odds - ratio} &= 3.059 - 0.166x.\end{aligned}$$

## 4.2 Hunua Forest Data

The data consists of the presence/absence of 17 plant species collected in the Hunua forest near Auckland, New Zealand. The variable altitude (m) is explanatory. The first species is *Agathis australis*, or Kauri, a famous New Zealand evergreen.

```
> data(hunua)
> names(hunua)

 [1] "agaaus"  "beitaw"  "corlae"  "cyadea"  "cyamed"  "daccup"
 [7] "dacdac"  "eladen"  "hedarb"  "hohpop"  "kniexc"  "kuneri"
[13] "lepsco"  "metro"   "neslan"  "rhosap"  "vitluc"  "altitude"

> fit = vglm(cbind(dacdac, metro) ~ altitude, loglinb2,
+           hunua)
> fitted(fit)[1:4, ]

           00           01           10           11
1 0.7789428 0.02125847 0.1934821 0.006316637
2 0.7737730 0.02059477 0.1984841 0.007148179
3 0.7570811 0.01869120 0.2138875 0.010340201
4 0.7510931 0.01808447 0.2191360 0.011686431

> coef(fit, matrix = TRUE)

                u1                u2                u12
(Intercept) -1.103104270 -3.826711087  1.28803383
altitude     -0.003218316  0.002505874 -0.01232055
```

The coefficients may be interpreted in terms of odds ratios.

Is the linearity assumption justified? Let's fit a VGAM. Then Figure 1 was produced by

```
> fits = vgam(cbind(dacdac, metrob) ~ s(altitude, df = c(4,  
+ 4, 1.5)), loglinb2, hunua)  
> par(mfrow = c(2, 2), las = 1, mar = c(5, 5, 1, 1))  
> plot(fits, se = TRUE, cex = 0.9, lcol = "blue")
```

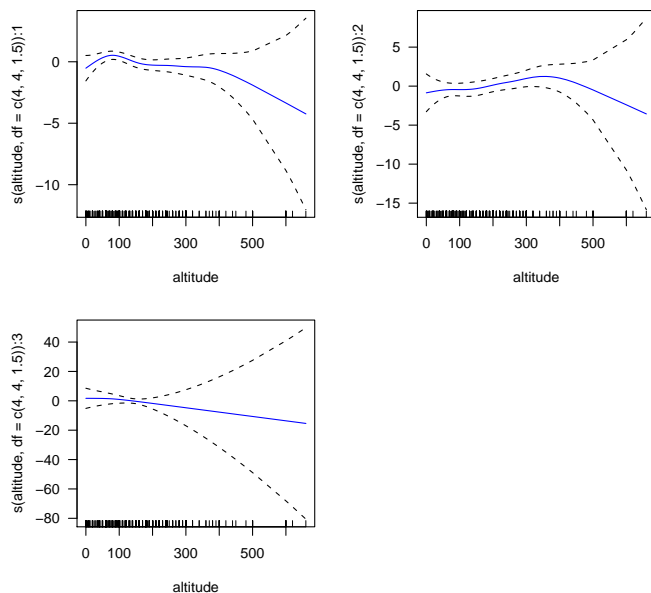


Figure 1: VGAM fitted to two species in the Hunua data set.

The massive standard errors of  $\hat{u}_{12}$  at high altitudes is partly due to the few data there. Plotting without `se=TRUE` suggests replacing it by an intercept only.

## Exercises

1. [Time consuming] Write a VGAM family function for four binary responses. Call it `loglinb4()`. Enumerate  $\boldsymbol{\eta}$  as  $(u_1, \dots, u_4, u_{12}, \dots, u_{14}, u_{23}, \dots, u_{34})^T$ , and allow for the exchangeable and zero arguments.

## ACKNOWLEDGEMENTS

The author wishes to thank Dr Neil Mitchell for the Hunua data.

## References

- Bishop, Y. M. M., Fienberg, S. E., Holland, P. W., 1975. Discrete Multivariate Analyses: Theory and Practice. MIT Press.
- Chambers, J. M., Hastie, T. J. (Eds.), 1993. Statistical Models in S. Chapman & Hall, New York.
- Christensen, R., 1997. Log-linear Models and Logistic Regression, 2nd Edition. Springer-Verlag.
- Lindsey, J. K., 1995. Modelling Frequency and Count Data. Clarendon Press.
- McCullagh, P., Nelder, J. A., 1989. Generalized Linear Models, 2nd Edition. Chapman & Hall, London.