glm における *.L, *.Q, *.C などについて

青木繁伸

2014 年 1 月 22 日

1 VGAM パッケージの vglm 関数

1.1 累積ロジスティックモデル parallel=FALSE

以下のようなデータを用い，Sex, Age で Reaction を予測する場合を考えよう。

```
, , Reaction = unimportant

        Age
Sex      18-23 24-40 >40
  female    26     9   5
  male      40    17   8

, , Reaction = normal

        Age
Sex      18-23 24-40 >40
  female    12    21  14
  male      17    15  15

, , Reaction = important

        Age
Sex      18-23 24-40 >40
  female     7    15  41
  male       8    12  18
```

Reaction は，unimportant < normal < important のように順序がついている。

累積ロジスティックモデルを当てはめるには，vglm において parallel=FALSE を指定する。

```
> library(VGAM)
> ans1 <- vglm(Reaction ~ Sex + Age, family=cumulative(parallel=FALSE), data=d)
> summary(ans1)
Call:
vglm(formula = Reaction ~ Sex + Age, family = cumulative(parallel = FALSE),
    data = d)

Pearson residuals:
                    Min       1Q   Median      3Q    Max
logit(P[Y<=1]) -1.6770 -0.70505 -0.26361 0.68698 2.9509
logit(P[Y<=2]) -2.7983 -0.71940  0.27233 0.76441 1.5713

Coefficients:
              Estimate Std. Error  z value
(Intercept):1 -1.104499    0.20395 -5.41553
(Intercept):2  0.508895    0.18407  2.76461
Sexmale:1      0.590708    0.26736  2.20944
```

```
Sexmale:2      0.572251    0.26851  2.13123
Age.L:1       -1.598005    0.25360 -6.30135
Age.L:2       -1.486816    0.24401 -6.09332
Age.Q:1        0.095079    0.24077  0.39490
Age.Q:2       -0.083539    0.23651 -0.35322

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: 580.5941 on 592 degrees of freedom

Log-likelihood: -290.297 on 592 degrees of freedom

Number of iterations: 6
```

係数は coefficients によって取り出すことができる。

```
> coeff <- matrix(coefficients(ans1), byrow=TRUE, ncol=2)
> dimnames(coeff) <- list(c("Intercept", "Sex.male", "Age.L", "Age.Q"),
+                         paste("logit", 1:2))
> coeff
              logit 1     logit 2
Intercept -1.10449871  0.50889501
Sex.male   0.59070807  0.57225102
Age.L     -1.59800468 -1.48681617
Age.Q      0.09507914 -0.08353905
```

glm でのカテゴリー変数の使われ方に注意が必要である。上に示したデータの Sex のようにカテゴリーが 2 つしかない場合には通常のダミー変数として扱われる。すなわち，最初のカテゴリー（今の場合は female）が基準とされ，2 番目のカテゴリーが変数名に合成される（Sexmale ）。つまり，female は 0，male は 1 として使用される。3 つ以上カテゴリーを持つ場合には，直交多項式の値が使われる。Age はカテゴリーが 3 つなので，1:3 の数値ベクトルを 2 次までの直交多項式で表した値が使われる。

```
> val <- poly(1:3, degree=2)
> print(round(val, 5))
            1        2
[1,] -0.70711  0.40825
[2,]  0.00000 -0.81650
[3,]  0.70711  0.40825
attr(,"degree")
[1] 1 2
attr(,"coefs")
attr(,"coefs")$alpha
[1] 2 2

attr(,"coefs")$norm2
[1] 1.0000000 3.0000000 2.0000000 0.6666667

attr(,"class")
[1] "poly"   "matrix"
```

上の例では，"18-23" に対しては 1 次，2 次のデータに対してはそれぞれ −0.70711，0.40825，"24-40" に対しては 0，−0.8165，">40" に対しては 0.70711，0.40825 が使われる。変数名としては元の変数名に .L : Linear，.Q : Quadratic，.C : Cubic などが付け加えられる。

元のデータは，性と年齢で 6 通りある。

```
> d2 <- data.frame(Sex=gl(2, 3), Age=gl(3, 1, 6))
> d2$Sex <- factor(d2$Sex, 1:2, labels=c("female", "male"), ordered=TRUE)
> d2$Age <- factor(d2$Age, 1:3, labels=c("18-23", "24-40", ">40"), ordered=TRUE)
> d2
     Sex   Age
1 female 18-23
2 female 24-40
3 female   >40
4   male 18-23
5   male 24-40
6   male   >40
```

それぞれのロジットは次のようになる。

```
> (logit <- predict(ans1, newdata=d2))
  logit(P[Y<=1]) logit(P[Y<=2])
1     0.06427713    1.526128123
2    -1.18213050    0.577104358
3    -2.19564276   -0.576547463
4     0.65498520    2.098379142
5    -0.59142243    1.149355378
6    -1.60493469   -0.004296444
```

データフレーム d2 は，計算に使われるデータ行列として，d3 のようになる。

```
> d3 <- matrix(0, 6, 3)
> d3[, 1] <- as.numeric(d2$Sex == "male")
> d3[, 2:3] <- val[as.integer(d2$Age),]
> d3
     [,1]          [,2]       [,3]
[1,]    0 -7.071068e-01  0.4082483
[2,]    0 -7.850462e-17 -0.8164966
[3,]    0  7.071068e-01  0.4082483
[4,]    1 -7.071068e-01  0.4082483
[5,]    1 -7.850462e-17 -0.8164966
[6,]    1  7.071068e-01  0.4082483
```

logit は以下のように計算されている。

```
> t(t(d3%*%coeff[2:4,])+coeff[1,]) # = logit
         logit 1      logit 2
[1,]  0.06427713  1.526128123
[2,] -1.18213050  0.577104358
[3,] -2.19564276 -0.576547463
[4,]  0.65498520  2.098379142
[5,] -0.59142243  1.149355378
[6,] -1.60493469 -0.004296444
```

例えば，female, "18-23" では

$$\text{logit}(P[Y <= 1]) = -1.104487+0.590691 \times 0+-1.597990 \times -7.071068e-01+0.095098 \times 0.4082483 = 0.06428619 \quad (1)$$

$$\text{logit}(P[Y <= 2]) = 0.508880+0.572275 \times 0+-1.486795 \times -7.071068e-01+-0.083561 \times 0.4082483 = 1.526089 \quad (2)$$

male, ">40" では

$$\text{logit}(P[Y <= 1]) = -1.104487+0.590691 \times 1+-1.597990 \times 7.071068e-01+0.095098 \times 0.4082483 = -1.604922 \quad (3)$$

$$\text{logit}(P[Y <= 2]) = 0.508880+0.572275 \times 1+-1.486795 \times 7.071068e-01+-0.083561 \times 0.4082483 = -0.004281491 \quad (4)$$

などとなる（丸めの誤差の範囲内で一致）。

従属変数のそれぞれのカテゴリーに対応する確率を求めるには，predict で type="response" を指定すればよい。

```
> predict(ans1, newdata=d2, type="response")
  unimportant    normal important
1   0.5160638 0.3053754 0.1785609
2   0.2346693 0.4057315 0.3595992
3   0.1001425 0.2595850 0.6402726
4   0.6581330 0.2326126 0.1092545
5   0.3563085 0.4030846 0.2406068
6   0.1672931 0.3316328 0.5010741
```

これは，次のように計算されている。確率 P[Y<=1] ， P[Y<=2] は $1/(\{1+\exp(-logit)\}$ で計算される。

```
> P <- 1/(1+exp(-logit))
> colnames(P) <- c("P[Y<=1]", "P[Y<=2]")
> P
    P[Y<=1]   P[Y<=2]
1 0.5160638 0.8214391
2 0.2346693 0.6404008
3 0.1001425 0.3597274
4 0.6581330 0.8907455
5 0.3563085 0.7593932
6 0.1672931 0.4989259
```

これに基づいて，確率 P[Y=1]，P[Y=2]，P[Y=3] は以下のように計算される。

```
> P2 <- cbind(P[,1], P[,2]-P[,1], 1-P[,2])
> colnames(P2) <- paste("P[Y=", 1:3, "]", sep="")
> P2
     P[Y=1]    P[Y=2]    P[Y=3]
1 0.5160638 0.3053754 0.1785609
2 0.2346693 0.4057315 0.3595992
3 0.1001425 0.2595850 0.6402726
4 0.6581330 0.2326126 0.1092545
5 0.3563085 0.4030846 0.2406068
6 0.1672931 0.3316328 0.5010741
```

また，元のデータでの確率は fitted で求めることができる。

```
> head(fitted(ans1), 10)
   unimportant    normal important
1    0.5160638 0.3053754 0.1785609
2    0.5160638 0.3053754 0.1785609
3    0.5160638 0.3053754 0.1785609
4    0.5160638 0.3053754 0.1785609
5    0.5160638 0.3053754 0.1785609
6    0.5160638 0.3053754 0.1785609
7    0.5160638 0.3053754 0.1785609
8    0.5160638 0.3053754 0.1785609
9    0.5160638 0.3053754 0.1785609
10   0.5160638 0.3053754 0.1785609
```