Examples of Using R for Modeling Ordinal Data

Alan Agresti

Department of Statistics, University of Florida

Supplement for the book *Analysis of Ordinal Categorical Data*, 2nd ed., 2010 (Wiley), abbreviated below as *OrdCDA*

Summary of R (and S-Plus)

- A detailed discussion of the use of R for models for categorical data is available on-line in the free manual prepared by Laura Thompson to accompany Agresti (2002). A link to this manual is at www.stat.ufl.edu/∼aa/cda/software.html.
- Specialized R functions available from various R libraries. Prof. Thomas Yee at Univ. of Auckland provides VGAM for vector generalized linear and additive models (www.stat.auckland.ac.nz/∼yee/VGAM).
- In VGAM, the *vglm* function fits wide variety of models. Possible models include the cumulative logit model (family function *cumulative*) with proportional odds or partial proportional odds or nonproportional odds, cumulative link models (family function *cumulative*) with or without common effects for each cutpoint, adjacent-categories logit models (family function *acat*), and continuation-ratio logit models (family functions *cratio* and *sratio*).

Example of Cumulative Logit Modeling with and Without Proportional Odds: Detecting trend in dose response

Effect of intravenous medication doses on patients with subarachnoid hemorrhage trauma (p. 207, *OrdCDA*)

	Glasgow Outcome Scale (y)				
Treatment Group (*x*)	Death	Veget. State	Major Disab.	Minor Disab.	Good Recov.
Placebo	59	25	46	48	32
Low dose	48	21	44	47	30
Med dose	44	14	54	64	31
High dose	43	4	49	58	41

Model with linear effect of dose (scores x = 1, 2, 3, 4) on cumulative logits for outcome,

$$\text{logit}[P(y \leq j)] = \alpha_j + \beta x$$

has ML estimate $\hat{\beta} = -0.176$ $(SE = 0.056)$

R for modeling dose-response data, using Thomas Yee's vglm() function in his VGAM library

```
> trauma <- read.table("trauma.dat", header=TRUE)
> trauma
  dose y1 y2 y3 y4 y5
1    1 59 25 46 48 32
2    2 48 21 44 47 30
3    3 44 14 54 64 31
4    4 43  4 49 58 41
> library(VGAM)
> fit <- vglm(cbind(y1,y2,y3,y4,y5) ~ dose,
        family=cumulative(parallel=TRUE), data=trauma)
> summary(fit)

Coefficients:
                 Value Std. Error t value
(Intercept):1 -0.71917    0.15881 -4.5285
(Intercept):2 -0.31860    0.15642 -2.0368
(Intercept):3  0.69165    0.15793  4.3796
(Intercept):4  2.05701    0.17369 11.8429
dose          -0.17549    0.05632 -3.1159

Residual Deviance: 18.18245 on 11 degrees of freedom
Log-likelihood: -48.87282 on 11 degrees of freedom
Number of Iterations: 4
> fitted(fit)
         y1         y2        y3        y4        y5
1 0.2901506 0.08878053 0.2473198 0.2415349 0.1322142
2 0.2553767 0.08321565 0.2457635 0.2619656 0.1536786
3 0.2234585 0.07701184 0.2407347 0.2808818 0.1779132
4 0.1944876 0.07043366 0.2325060 0.2975291 0.2050436
```

The fitted values shown here are the 5 estimated multinomial response probabilities for each of the 4 treatment groups.

Note: propodds() is another possible family for vglm; it defaults to cumulative(reverse = TRUE, link = "logit", parallel = TRUE)

R for modeling dose-response data using polr() in MASS library, for which response must be an ordered factor

```
> trauma2 <- read.table("trauma2.dat", header=TRUE)
> trauma2
   dose response count
1     1        1    59
2     1        2    25
3     1        3    46
4     1        4    48
5     1        5    32
6     2        1    48
...
20    4        5    41
> y <- factor(trauma2$response)
> fit.clogit <- polr(y ~ dose, data=trauma2, weight=count)
> summary(fit.clogit)

Re-fitting to get Hessian

Coefficients:
         Value Std. Error  t value
dose 0.1754816 0.05671224 3.094245

Intercepts:
    Value   Std. Error t value
1|2 -0.7192  0.1589    -4.5256
2|3 -0.3186  0.1569    -2.0308
3|4  0.6917  0.1597     4.3323
4|5  2.0570  0.1751    11.7493

Residual Deviance: 2461.349

> fitted(fit.clogit)
           1          2         3         4         5
1  0.2901467 0.08878330 0.2473217 0.2415357 0.1322126
2  0.2901467 0.08878330 0.2473217 0.2415357 0.1322126
...
20 0.1944866 0.07043618 0.2325084 0.2975294 0.2050394
```

Note this uses the model formula based on the latent variable approach, for which $\hat{\beta}$ has different sign.

R for modeling dose-response data without proportional odds, using vglm() in VGAM library without parallel=TRUE option

```
> trauma <- read.table("trauma.dat", header=TRUE)
> trauma
  dose y1 y2 y3 y4 y5
1    1 59 25 46 48 32
2    2 48 21 44 47 30
3    3 44 14 54 64 31
4    4 43  4 49 58 41
> library(VGAM)
> fit2 <- vglm(cbind(y1,y2,y3,y4,y5) ~ dose, family=cumulative, data=trauma)
> summary(fit2)

Coefficients:
                  Value Std. Error  t value
(Intercept):1 -0.864585   0.194230 -4.45133
(Intercept):2 -0.093747   0.178494 -0.52521
(Intercept):3  0.706251   0.175576  4.02248
(Intercept):4  1.908668   0.238380  8.00684
dose:1        -0.112912   0.072881 -1.54926
dose:2        -0.268895   0.068319 -3.93585
dose:3        -0.182341   0.063855 -2.85555
dose:4        -0.119255   0.084702 -1.40793

Residual Deviance: 3.85163 on 8 degrees of freedom
Log-likelihood: -41.70741 on 8 degrees of freedom

> pchisq(deviance(fit)-deviance(fit2),
df=df.residual(fit)-df.residual(fit2), lower.tail=FALSE)
[1] 0.002487748
```

The improvement in fit is statistically significant, but perhaps not substantively significant; effect of dose is moderately negative for each cumulative probability.

R for modeling *mental impairment* data with partial proportional odds (life events but not SES), using vglm() in VGAM library.

```
> fit3 < -vglm(impair ~ ses + life, family=cumulative(parallel=FALSE~ses))

> summary(fit3)

Coefficients:
              Estimate Std. Error  z value
(Intercept):1 -0.17660    0.69506 -0.25408
(Intercept):2  1.00567    0.66327  1.51623
(Intercept):3  2.39555    0.77894  3.07539
ses:1          0.98237    0.76430  1.28531
ses:2          1.54149    0.73732  2.09066
ses:3          0.73623    0.81213  0.90655
life          -0.32413    0.12017 -2.69736

Names of linear predictors: logit(P[Y<=1]), logit(P[Y<=2]), logit(P[Y<=3])

Residual deviance: 97.36467 on 113 degrees of freedom
Log-likelihood: -48.68234 on 113 degrees of freedom
```

Deviance (97.36, df=113) similar to obtained with more complex non-proportional odds model (96.75, df=111) and the simpler proportional odds model (99.10, df=115).

Example: Religious fundamentalism by region (2006 GSS data)

	y = Religious Beliefs		
x = Region	Fundamentalist	Moderate	Liberal
Northeast	92 (14%)	352 (52%)	234 (34%)
Midwest	274 (27%)	399 (40%)	326 (33%)
South	739 (44%)	536 (32%)	412 (24%)
West/Mountain	192 (20%)	423 (44%)	355 (37%)

Create indicator variables $\{r_i\}$ for region and consider model

$$\text{logit}[P(y \leq j)] = \alpha_j + \beta_1 r_1 + \beta_2 r_2 + \beta_3 r_3$$

Score test of proportional odds assumption compares with model having separate $\{\beta_i\}$ for each logit, that is, 3 extra parameters. SAS (PROC LOGISTIC) reports:

```
-------------------------------------------------------------------------
          Score Test for the Proportional Odds Assumption

                Chi-Square       DF     Pr > ChiSq
                   93.0162        3         <.0001
-------------------------------------------------------------------------
```

R for religion and region data, using vglm() in VGAM library

```
> religion <- read.table("religion_region.dat",header=TRUE)
> religion
  region  y1  y2  y3
1      1  92 352 234
2      2 274 399 326
3      3 739 536 412
4      4 192 423 355
> z1 < -ifelse(region==1,1,0); z2 <-ifelse(region==2,1,0); z3 <-ifelse(region==3,1,0)
> cbind(z1,z2,z3)
     z1 z2 z3
[1,]  1  0  0
[2,]  0  1  0
[3,]  0  0  1
[4,]  0  0  0
> library(VGAM)
> fit.po <- vglm(cbind(y1,y2,y3) ~ z1+z2+z3, family=cumulative(parallel=TRUE),
         data=religion)
> summary(fit.po)
Coefficients:
                  Value Std. Error   t value
(Intercept):1 -1.261818   0.064033 -19.70584
(Intercept):2  0.472851   0.061096   7.73948
z1            -0.069842   0.093035  -0.75071
z2             0.268777   0.083536   3.21750
z3             0.889677   0.075704  11.75211
Residual Deviance: 98.0238 on 3 degrees of freedom
Log-likelihood: -77.1583 on 3 degrees of freedom
> fit.npo <- vglm(cbind(y1,y2,y3) ~ z1+z2+z3, family=cumulative,data=religion)
> summary(fit.npo)
Coefficients:
                  Value Std. Error   t value
(Intercept):1 -1.399231   0.080583 -17.36377
(Intercept):2  0.549504   0.066655   8.24398
z1:1          -0.452300   0.138093  -3.27532
z1:2           0.090999   0.104731   0.86888
z2:1           0.426188   0.107343   3.97032
z2:2           0.175343   0.094849   1.84866
z3:1           1.150175   0.094349  12.19065
z3:2           0.580174   0.087490   6.63135
Residual Deviance: -5.1681e-13 on 0 degrees of freedom
Log-likelihood: -28.1464 on 0 degrees of freedom
> pchisq(deviance(fit.po)-deviance(fit.npo),
      df=df.residual(fit.po)-df.residual(fit.npo),lower.tail=FALSE)
[1] 4.134028e-21
```

Example of Adjacent-Categories Logit Model:
Stem Cell Research and Religious Fundamentalism

		Stem Cell Research			
	Religious	Definitely	Probably	Probably	Definitely
Gender	Beliefs	Fund	Fund	Not Fund	Not Fund
Female	Fundamentalist	34 (22%)	67 (43%)	30 (19%)	25 (16%)
	Moderate	41 (25%)	83 (52%)	23 (14%)	14 (9%)
	Liberal	58 (39%)	63 (43%)	15 (10%)	12 (8%)
Male	Fundamentalist	21 (19%)	52 (46%)	24 (21%)	15 (13%)
	Moderate	30 (27%)	52 (47%)	18 (16%)	11 (10%)
	Liberal	64 (45%)	50 (36%)	16 (11%)	11 (8%)

For gender g (1 = females, 0 = males) and religious beliefs treated quantitatively with x = (1, 2, 3), consider adjacent-categories logit (ACL) model

$$\log(\pi_j/\pi_{j+1}) = \alpha_j + \beta_1 x + \beta_2 g$$

R: It’s easy with the vglm() function in VGAM library, as adjacent-categories logit model is a model option.

```
> stemcell <- read.table("scresrch.dat",header=TRUE)
> stemcell
 religion gender y1 y2 y3 y4
1    1         0 21 52 24 15
2    1         1 34 67 30 25
3    2         0 30 52 18 11
4    2         1 41 83 23 14
5    3         0 64 50 16 11
6    3         1 58 63 15 12
> fit.adj <- vglm(cbind(y1,y2,y3,y4) ~ religion + gender,
 family=acat(reverse=TRUE, parallel=TRUE), data=stemcell)
> summary(fit.adj)

Coefficients:
                  Value Std. Error  t value
(Intercept):1 -0.95090   0.142589 -6.66880
(Intercept):2  0.55734   0.145084  3.84147
(Intercept):3 -0.10656   0.164748 -0.64680
religion       0.26681   0.047866  5.57410
gender        -0.01412   0.076706 -0.18408

Number of linear predictors:  3

Residual Deviance: 11.99721 on 13 degrees of freedom
Log-likelihood: -48.07707 on 13 degrees of freedom

> fitted(fit.adj)
         y1        y2        y3         y4
1 0.2177773 0.4316255 0.1893146 0.16128261
2 0.2138134 0.4297953 0.1911925 0.16519872
3 0.2975956 0.4516958 0.1517219 0.09898673
4 0.2931825 0.4513256 0.1537533 0.10173853
5 0.3830297 0.4452227 0.1145262 0.05722143
6 0.3784551 0.4461609 0.1163995 0.05898444
```

Example of Continuation-Ratio Logit Model:

Tonsil Size and Streptococcus

	Tonsil Size		
Carrier	Not enlarged	Enlarged	Greatly Enlarged
Yes	19 (26%)	29 (40%)	24 (33%)
No	497 (37%)	560 (42%)	269 (20%)

Let x = whether carrier of Streptococcus pyogenes (1 = yes, 0 = no)

Continuation-ratio logit model

$$\log\left[\frac{\pi_1}{\pi_2+\pi_3}\right] = \alpha_1 + \beta x, \quad \log\left[\frac{\pi_2}{\pi_3}\right] = \alpha_2 + \beta x$$

estimates an assumed common value for cumulative odds ratio from first part of model and for local odds ratio from second part.

R: VGAM library has continuation-ratio logit model option in vglm() function

```
> tonsils <- read.table("tonsils.dat",header=TRUE)
> tonsils
  carrier  y1  y2  y3
1       1  19  29  24
2       0 497 560 269
> library(VGAM)
> fit.cratio <- vglm(cbind(y1,y2,y3) ~ carrier,
           family=cratio(reverse=FALSE, parallel=TRUE), data=tonsils)
> summary(fit.cratio)

Coefficients:
                 Value Std. Error t value
(Intercept):1  0.51102   0.056141   9.1025
(Intercept):2 -0.73218   0.072864 -10.0486
carrier        0.52846   0.197747   2.6724

Residual Deviance: 0.00566 on 1 degrees of freedom
Log-likelihood: -11.76594 on 1 degrees of freedom

> fitted(fit.cratio)
         y1        y2        y3
1 0.2612503 0.4068696 0.3318801
2 0.3749547 0.4220828 0.2029625
```

Adjacent Categories Logit and Continuation Ratio Logit Models with Nonproportional Odds

- As in cumulative logit case, model of proportional odds form fits poorly when there are substantive dispersion effects,
- Each model has a more general non-proportional odds form, the ACL version being

$$\log[P(y_i = j)/P(y_i = j + 1)] = \alpha_j + \boldsymbol{\beta}_j' \mathbf{x}_i$$

- Unlike cumulative logit model, these models do not have structural problem that cumulative probabilities may be out of order.
- Models lose ordinal advantage of parsimony, but effects still have ordinal nature, unlike baseline-category logit (BCL) models.
- Can fit general ACL model with software for BCL model, converting its $\{\hat{\beta}_j^*\}$ estimates to $\hat{\beta}_j = \hat{\beta}_j^* - \hat{\beta}_{j+1}^*$, since

$$\log\left(\frac{\pi_j}{\pi_{j+1}}\right) = \log\left(\frac{\pi_j}{\pi_c}\right) - \log\left(\frac{\pi_{j+1}}{\pi_c}\right),$$

  or using specialized software such as vglm function in R without “PARALLEL = TRUE” option.

Example: Data on stemcell research that had been fitted with ACL model of proportional odds form

```
> vglm(cbind(y1,y2,y3,y4) ~ religion + gender,
+ family=acat(reverse=TRUE, parallel=FALSE), data=stemcell)

         y1        y2        y3         y4
1 0.1875000 0.4642857 0.2142857 0.13392857
2 0.2179487 0.4294872 0.1923077 0.16025641
3 0.2702703 0.4684685 0.1621622 0.09909910
4 0.2546584 0.5155280 0.1428571 0.08695652
5 0.4539007 0.3546099 0.1134752 0.07801418
6 0.3918919 0.4256757 0.1013514 0.08108108

Call:
vglm(formula = cbind(y1, y2, y3, y4) ~ religion + gender,
family = acat(reverse = TRUE, parallel = FALSE), data = stemcell)

Coefficients:
(Intercept):1 (Intercept):2 (Intercept):3    religion:1    religion:2
  -1.24835878    0.47098433    0.42740812    0.43819661    0.25962043
   religion:3      gender:1      gender:2      gender:3
   0.01192302   -0.13683357    0.18706754   -0.16093003

Degrees of Freedom: 18 Total; 9 Residual
Residual Deviance: 5.675836
Log-likelihood: -44.91638
```

We then get separate effects of religion and of gender for each logit. The change in the deviance is 11.997 - 5.676 = 6.32 based on df = 13 - 9 = 4 (P = 0.18), so simpler model is adequate.

Stereotype model: Multiplicative paired-category logits

ACL model with separate effects for each pair of adjacent categories is equivalent to standard BCL model

$$\log\left[\frac{\pi_j}{\pi_c}\right] = \alpha_j + \boldsymbol{\beta}_j^{'}\boldsymbol{x}, \quad j = 1, \ldots, c-1.$$

- Disadvantage: lack of parsimony (treats response as *nominal*)
- $c-1$ parameters for each predictor instead of a single parameter
- No. parameters large when either c or no. of predictors large

Anderson (1984) proposed alternative model nested between ACL model with proportional odds structure and the general ACL or BCL model with separate effects $\{\beta_j\}$ for each logit.

Stereotype model:

$$\log\left[\frac{\pi_j}{\pi_c}\right] = \alpha_j + \phi_j \boldsymbol{\beta}' \boldsymbol{x}, \quad j = 1, \ldots, c-1.$$

- For predictor x_k, $\phi_j \beta_k$ represents log odds ratio for categories j and c for a unit increase in x_k. By contrast, general BCL model has log odds ratio β_{jk} for this effect, which requires many more parameters
- $\{\phi_j\}$ are parameters, treated as "scores" for categories of y.
- Like proportional odds models, stereotype model has advantage of single parameter β_k for effect of predictor x_k (for given scores $\{\phi_j\}$).
- Stereotype model achieves parsimony by using same scores for each predictor, which may or may not be realistic.
- Identifiability requires location and scale constraints on $\{\phi_j\}$, such as $(\phi_1 = 1, \phi_c = 0)$ or $(\phi_1 = 0, \phi_c = 1)$.
- Corresponding model for category probabilities is

$$P(y_i = j) = \frac{\exp(\alpha_j + \phi_j \boldsymbol{\beta}' \boldsymbol{x}_i)}{1 + \sum_{k=1}^{c-1} \exp(\alpha_k + \phi_k \boldsymbol{\beta}' \boldsymbol{x}_i)}$$

- Model is *multiplicative* in parameters, which makes model fitting awkward (*gnm* add-on function to R fits this and other nonlinear models; also *rrvglm* function in VGAM library).

Example: Boys' Disturbed Dreams by Age (Anderson)

	Degree of Suffering (ordinal)			
Age	Not severe (1)	(2)	(3)	Very severe (4)
5-7	7	4	3	7
8-9	10	15	11	13
10-11	23	9	11	7
12-13	28	9	12	10
14-15	32	5	4	3

Let x_i = age for row i, using mid-point scores (6, 8.5, 10.5, 12.5, 14.5). Consider model

$$\log\left[\frac{\pi_j}{\pi_4}\right] = \alpha_j + \phi_j \beta x_i$$

With identifiability constraints $(\phi_1 = 1, \phi_4 = 0)$, Anderson reported (with SE values in parentheses)

$$\hat{\phi}_1 = 1.0,\ \hat{\phi}_2 = 0.19\ (0.25),\ \hat{\phi}_3 = 0.36\ (0.24),\ \hat{\phi}_4 = 0.0.$$

In this model $\hat{\beta} = 0.31$. Estimates suggest possibly $\phi_2 = \phi_3 = \phi_4$, an ordinary logit model collapsing categories 2–4.

Can fit stereotype model in Yee's VGAM package, as reduced-rank multinomial logit model with function rrvglm().

```
> dreams <- read.table("dreams.dat", header=T)
> dreams
   age y1 y2 y3 y4
1  6.0  7  4  3  7
2  8.5 10 15 11 13
3 10.5 23  9 11  7
4 12.5 28  9 12 10
5 14.5 32  5  4  3

> library(VGAM)
> fit.stereo <- rrvglm(cbind(y1,y2,y3,y4) ~ age, multinomial, data=dreams)
> summary(fit.stereo)

Pearson Residuals:
  log(mu[,1]/mu[,4]) log(mu[,2]/mu[,4]) log(mu[,3]/mu[,4])
1            1.58695           -0.74923           -0.70538
2           -1.31353            0.87080            0.12071
3            0.80427           -0.05333            0.64526
4           -1.11202           -0.46304            0.22031
5            0.91070            0.16150           -0.65056

Coefficients:
                 Value Std. Error  t value
I(lv.mat):1    0.19346   0.251754  0.76847
I(lv.mat):2    0.36221   0.238527  1.51853
(Intercept):1 -2.45444   0.845365 -2.90341
(Intercept):2 -0.55464   0.890800 -0.62263
(Intercept):3 -1.12464   0.916067 -1.22768
age            0.30999   0.078019  3.97325

Names of linear predictors:
log(mu[,1]/mu[,4]), log(mu[,2]/mu[,4]), log(mu[,3]/mu[,4])

Residual Deviance: 9.74913 on 9 degrees of freedom
Log-likelihood: -31.47252 on 9 degrees of freedom

Number of Iterations: 4
```

Example of Cumulative Probit Model:

Religious fundamentalism by highest educational degree

(GSS data from 1972 to 2006)

	Religious Beliefs		
Highest Degree	Fundamentalist	Moderate	Liberal
Less than high school	4913 (43%)	4684 (41%)	1905 (17%)
High school	8189 (32%)	11196 (44%)	6045 (24%)
Junior college	728 (29%)	1072 (43%)	679 (27%)
Bachelor	1304 (20%)	2800 (43%)	2464 (38%)
Graduate	495 (16%)	1193 (39%)	1369 (45%)

For cumulative link model

$$\text{link}[P(y \leq j)] = \alpha_j + \beta x_i$$

using scores $\{x_i = i\}$ for highest degree,

$\hat{\beta} = -0.206$ $(SE = 0.0045)$ for probit link

$\hat{\beta} = -0.345$ $(SE = 0.0075)$ for logit link

R: vglm() function in VGAM library has cumulative probit model option

```
> fundamentalism <- read.table("fundamentalism.dat",header=TRUE)
> fundamentalism
  degree   y1    y2   y3
1      0 4913  4684 1905
2      1 8189 11196 6045
3      2  728  1072  679
4      3 1304  2800 2468
5      4  495  1193 1369
> library(VGAM)
> fit.cprobit <- vglm(cbind(y1,y2,y3) ~ degree,
   family=cumulative(link=probit, parallel=TRUE), data=fundamentalism)

> summary(fit.cprobit)

Call:
vglm(formula = cbind(y1, y2, y3) ~ degree, family = cumulative(link = probit,
    parallel = TRUE), data = fundamentalism)

Coefficients:
                 Value Std. Error t value
(Intercept):1 -0.22398  0.0079908 -28.030
(Intercept):2  0.94001  0.0086768 108.336
degree        -0.20594  0.0044727 -46.044

Number of linear predictors:  2

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

Residual Deviance: 48.70723 on 7 degrees of freedom
```

Example of Complementary Log-Log Link:

Life table for gender and race (percent)

	Males		Females	
Life Length	White	Black	White	Black
0-20	1.3	2.6	0.9	1.8
20-40	2.8	4.9	1.3	2.4
40-50	3.2	5.6	1.9	3.7
50-65	12.2	20.1	8.0	12.9
Over 65	80.5	66.8	87.9	79.2

Source: 2008 Statistical Abstract of the United States

For gender g (1 = female; 0 = male), race r (1 = black; 0 = white), and life length y, consider model

$$\log\{-\log[1 - P(y \leq j)]\} = \alpha_j + \beta_1 g + \beta_2 r$$

Good fit with this model or a cumulative logit model or a cumulative probit model

R: vglm() function in VGAM library has cumulative complementary log-log model option

```
> life <- read.table("lifetable.dat",header=TRUE)
> life
  gender race y1 y2 y3  y4  y5
1      0    0 13 28 32 122 805
2      0    1 26 49 56 201 668
3      1    0  9 13 19  80 879
4      1    1 18 24 37 129 792

> library(VGAM)
> fit.cloglog <- vglm(cbind(y1,y2,y3,y4,y5) ~ gender+race,
         family=cumulative(link=cloglog, parallel=TRUE),data=life)

> summary(fit.cloglog)

Call:
vglm(formula = cbind(y1, y2, y3, y4, y5) ~ gender + race,
    family = cumulative(link = cloglog, parallel = TRUE), data = life)

Coefficients:
                 Value Std. Error  t value
(Intercept):1 -4.21274   0.133834 -31.4773
(Intercept):2 -3.19223   0.091148 -35.0225
(Intercept):3 -2.58210   0.076360 -33.8147
(Intercept):4 -1.52163   0.062317 -24.4176
gender        -0.53827   0.070332  -7.6533
race           0.61071   0.070898   8.6139
```

$\beta_1 = -0.538, \;\; \beta_2 = 0.611$

Gender effect:

$$P(y > j \mid g = 0, r) = [P(y > j \mid g = 1, r)]^{\exp(0.538)}$$

Given race, proportion of men living longer than a fixed time equals proportion for women raised to $\exp(0.538) = 1.71$ power.

Given gender, proportion of blacks living longer than a fixed time equals proportion for whites raised to $\exp(0.611) = 1.84$ power.

Cumulative logit model has gender effect = -0.604, race effect = 0.685.

If Ω denotes odds of living longer than some fixed time for white women, then estimated odds of living longer than that time are

$\exp(-0.604)\Omega = 0.55\Omega$ for white men

$\exp(-0.685)\Omega = 0.50\Omega$ for black women

$\exp(-0.604 - 0.685)\Omega = 0.28\Omega$ for black men

Example of Partial Proportional Odds Model:

Smoking Status and Degree of Heart Disease

Smoking	Degree of Heart Disease				
status	1	2	3	4	5
Smoker	350 (23%)	307 (20%)	345 (22%)	481 (31%)	67 (4%)
Non-smoker	334 (45%)	99 (13%)	117 (16%)	159 (22%)	30 (4%)

y ordinal: 1 = No disease, ..., 5 = Very severe disease

x binary: 1 = smoker, 0 = non-smoker

Sample cumulative log odds ratios:

$$-1.04,\ -0.65,\ -0.46,\ -0.07.$$

Consider model letting effect of x depend on j,

$$\text{logit}[P(Y \leq j)] = \alpha_j + \beta_1 x + (j-1)\beta_2 x.$$

Cumulative log odds ratios are

$$\log \theta_{11}^C = \beta_1,\ \ \log \theta_{12}^C = \beta_1 + \beta_2,\ \ \log \theta_{13}^C = \beta_1 + 2\beta_2,\ \ \log \theta_{14}^C = \beta_1 + 3\beta_2.$$

Fit obtained using Joe Lang's mph.fit function in R (analysis 2 at www.stat.ufl.edu/∼aa/ordinal/mph.html).

$$\hat{\beta}_1 = -1.017\ (SE = 0.094),\ \ \hat{\beta}_2 = 0.298\ (SE = 0.047)$$

give estimated cumulative log odds ratios

$$\log \hat{\theta}_{11}^C = -1.02,\ \ \log \hat{\theta}_{12}^C = -0.72,\ \ \log \hat{\theta}_{13}^C = -0.42,\ \ \log \hat{\theta}_{14}^C = -0.12.$$

Some Models that Lang's mph.fit R Function Can Fit by ML:

- mph stands for *multinomial Poisson homogeneous* models, which have general form

$$\mathbf{L}(\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$$

  for probabilities or expected frequencies $\boldsymbol{\mu}$ in a contingency table, where $\mathbf{L}$ is a general link function (Lang 2005).

- Important special case is *generalized loglinear model*

$$\mathbf{C}\log(\mathbf{A}\boldsymbol{\mu}) = \mathbf{X}\boldsymbol{\beta}$$

  for matrices $\mathbf{C}$ and $\mathbf{A}$ and vector of parameters $\boldsymbol{\beta}$.

- This includes ordinal logit models, such as cumulative logit; e.g., $\mathbf{A}$ forms cumulative prob's and their complements at each setting of explanatory var's (each row has 0's and 1's), and $\mathbf{C}$ forms contrasts of log prob's to generate logits (each row contains 1, -1, and otherwise 0's).

- Includes models for ordinal odds ratios, such as model where all global log odds ratios take common value β. (*OrdCDA, Sec. 6.6*)

- Another special case has form $\mathbf{A}\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, which includes multinomial mean response model that mimics ordinary regression (scores in each row of $\mathbf{A}$).

Example of Mean Response Model:

Political views by political party and gender

		Political Views							
Gender	Party ID	1	2	3	4	5	6	7	Total
Females	Democrat	42	201	136	320	83	63	18	863
	Independent	33	87	107	459	123	92	19	920
	Republican	5	19	29	177	121	183	52	586
Males	Democrat	28	120	89	202	51	37	10	537
	Independent	20	79	124	362	120	90	18	813
	Republican	3	12	26	128	107	211	47	534

Political views:

1 = extremely liberal, 4 = moderate, 7 = extremely conservative

Let g = 1 (females), g = 0 (males), and p_1 and p_2 indicator variables for Democrats and Independents.

ML fit of mean response model (using R function mph.fit, analysis 3 at www.stat.ufl.edu/$\sim$aa/ordinal/mph.html)

$$\hat{\mu} = 5.081 - 0.063g - 1.513p_1 - 1.049p_2,$$

with SE = 0.040 for g, 0.052 for p_1, and 0.048 for p_2.

Mean political ideology estimated to be about one category more conservative for Republicans than for Independents, about 1.5 categories more conservative for Republicans than for Democrats.

Example of GEE for Repeated Measurement:

Randomized Clinical Trial for Treating Insomnia

Randomized, double-blind clinical trial compared hypnotic drug with placebo in patients with insomnia problems

	Time to Falling Asleep				
	Initial	Follow-up			
Treatment		<20	20–30	30–60	>60
Active	<20	7	4	1	0
	20–30	11	5	2	2
	30–60	13	23	3	1
	>60	9	17	13	8
Placebo	<20	7	4	2	1
	20–30	14	5	1	0
	30–60	6	9	18	2
	>60	4	11	14	22

y_t = time to fall asleep

x = treatment (0 = placebo, 1 = active)

t = occasion (0 = initial, 1 = follow-up after 2 weeks)

Model

$$\text{logit}[P(y_t \leq j)] = \alpha_j + \beta_1 t + \beta_2 x + \beta_3 (t \times x), \quad j = 1, 2, 3$$

GEE estimates:

$\hat{\beta}_1 = 1.04$ ($SE = 0.17$), placebo occasion effect
$\hat{\beta}_2 = 0.03$ ($SE = 0.24$), treatment effect initially
$\hat{\beta}_3 = 0.71$ ($SE = 0.24$), interaction

Considerable evidence that distribution of time to fall asleep decreased more for treatment than placebo group

Occasion effect = 1.04 for placebo, 1.04 + 0.71 = 1.75 for active

Odds ratios $e^{1.04} = 2.8,\ e^{1.75} = 5.7$

Treatment effect $e^{0.03} = 1.03$ initially,
$e^{0.03+0.71} = 2.1$ follow-up

R analysis using repolr() function in repolr library, here using the independence working correlation structure

(Thanks to Anestis Touloumis)

```
> insomnia<-read.table("insomnia.dat",header=TRUE)
> insomnia<-as.data.frame(insomnia)
> insomnia
       case treat occasion outcome
        1       1     0        1
        1       1     1        1
        2       1     0        1
        2       1     1        1
        3       1     0        1
        3       1     1        1
        4       1     0        1
        4       1     1        1
        5       1     0        1
...
      239       0     0        4
      239       0     1        4

> library(repolr)
> fit <- repolr(formula = outcome ~ treat + occasion + treat * occasion,
       + subjects="case", data=insomnia, times=c(1,2), categories=4,
        corstr = "independence")
> summary(fit$gee)

Coefficients:
                  Estimate Naive S.E.     Naive z Robust S.E.    Robust z
factor(cuts)1  -2.26708899  0.2027367 -11.1824294   0.2187606 -10.3633343
factor(cuts)2  -0.95146176  0.1784822  -5.3308499   0.1809172  -5.2591017
factor(cuts)3   0.35173977  0.1726860   2.0368745   0.1784232   1.9713794
treat           0.03361002  0.2368973   0.1418759   0.2384374   0.1409595
occasion        1.03807641  0.2375992   4.3690229   0.1675855   6.1943093
treat:occasion  0.70775891  0.3341759   2.1179234   0.2435197   2.9063728
```

- The package geepack contains a function *ordgee* for ordinal GEE analyses, but a PhD student of mine (Anestis Touloumis) and I have found it to be very unreliable. Touloumis is preparing his own package, based on modeling associations using local odds ratios.
- The glmmAK package contains a function *cumlogitRE* for using MCMC (Bayesian approach) to fit cumulative logit models with random effects. The ordinal package has a function *clmm* that can fit cumulative link models with random effects; it uses only a Laplace approximation, but promises adaptive Gauss-Hermite quadrature in the future.
- Can fit nonlinear models such as stereotype model using *gnm* add-on function to R by Firth and Turner: www2.warwick.ac.uk/fac/sci/statistics/staff/research/turner/gnm/

  Can also fit stereotype model within VGAM , as a reduced-rank multinomial logit model; e.g., rrvglm(y x1 + x2, multinomial).
- R function *mph.fit* prepared by Joe Lang at Univ. of Iowa can fit many other models for contingency tables that are difficult to fit with ML, such as global odds ratio models, marginal models.