Modeling Ordinal Categorical Data

Alan Agresti

Distinguished Professor Emeritus

Department of Statistics

University of Florida, USA

Presented for Vienna University of Economics and Business

May 21, 2013

Ordinal categorical responses

- Patient quality of life (excellent, good, fair, poor)
- Political philosophy (very liberal, slightly liberal, moderate, slightly conservative, very conservative)
- Government spending (too low, about right, too high)
- Categorization of an inherently continuous variable, such as body mass index, BMI = weight(kg)/[height(m)]^2,
  measured as ($<$ 18.5, 18.5-25, 25-30, $>$ 30)
  for (underweight, normal weight, overweight, obese)

For ordinal response variable y with c categories, our focus is on modeling how

$$P(y = j), \quad j = 1, 2, \ldots, c,$$

depends on explanatory variables $\boldsymbol{x}$, which can be categorical and/or quantitative.

The models treat observations on y at fixed $\boldsymbol{x}$ as *multinomial*.

Outline

1: Logistic Regression Using Cumulative Logits

(“proportional odds” model, non-proportional odds)

2: Other Ordinal Models

(adjacent-category logits, continuation-ratio logits, cumulative probits and complementary log-log)

These notes are extracted from a two-day short course that I’ve presented at Padova, Firenze, and Groningen.

Focus of tutorial

- The primary methods for modeling ordinal categorical responses
- Emphasis on concepts, examples of use, complicating issues, rather than theory, derivations, or technical details
- Examples included of how to fit models using SAS, R, Stata (thanks, Kat Chzhen for Stata), but output is provided to enhance interpretation, not to teach software.
- For R for ordinal models, Thomas Yee's VGAM library is especially useful; see `www.stat.auckland.ac.nz/~yee/VGAM`. Joseph Lang's R function mph.fit (link at `www.stat.ufl.edu/~aa/ordinal/ord.html`) fits some nonstandard models, must be requested from him at U. of Iowa (jblang@iowa.uiowa.edu). Also useful is detailed R tutorial by Laura Thompson to accompany my book *Categorical Data Analysis*, linked at R section of `www.stat.ufl.edu/~aa/cda/cda.html`.
- This lecture assumes some familiarity with basic categorical data methods (contingency tables, logistic regression).
- Lecture based on material in *Analysis of Ordinal Categorical Data*, 2nd ed., Wiley, 2010

1. Logistic Regression Using Cumulative Logits

y an ordinal response (c categories)

x an explanatory variable

Model $P(y \leq j),\ \ j = 1, 2, \cdots, c-1,$ using logits

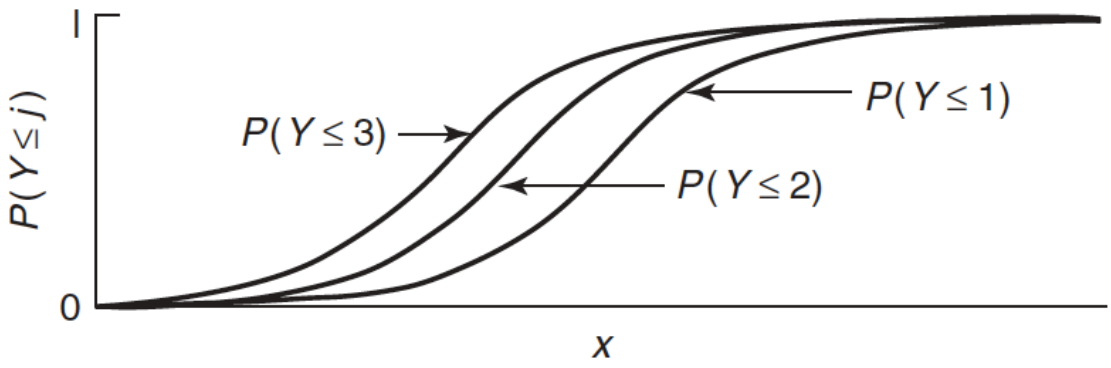
$$\begin{aligned} \text{logit}[P(y \leq j)] &= \log[P(y \leq j)/P(y > j)] \\ &= \alpha_j + \beta x, \quad j = 1, \ldots, c-1 \end{aligned}$$

This is called a *cumulative logit* model.

As in ordinary logistic regression, effects described by odds ratios. Here, we compare odds of being below vs. above any point on the response scale (*cumulative odds ratios*).

For fixed j, looks like ordinary logistic regression for binary response (below j, above j).

See figure on next page for $c = 4$ categories.



Model satisfies

$$\log\left[\frac{P(y \le j \mid x_1)/P(y > j \mid x_1)}{P(y \le j \mid x_2)/P(y > j \mid x_2)}\right] = \beta(x_1 - x_2)$$

for all j (*Proportional odds* property)

- β = *cumulative log odds ratio* for 1-unit increase in predictor
- Model assumes effect β is identical for every "cutpoint" for cumulative probability, $j = 1, \cdots, c-1$
- Logistic regression is special case $c = 2$
- Software for maximum likelihood (ML) fitting includes R functions *vglm* in VGAM library and *polr* (proportional odds logistic regression) in MASS library, SAS (PROC LOGISTIC, PROC GENMOD), Stata programs *ologit, oglm*, SPSS program *plum*.

Example: Detecting trend in dose response

Effect of intravenous medication doses on patients with subarachnoid hemorrhage trauma

	Glasgow Outcome Scale (y)				
Treatment Group (x)	Death	Veget. State	Major Disab.	Minor Disab.	Good Recov.
Placebo	59 (28%)	25	46	48	32 (15%)
Low dose	48 (25%)	21	44	47	30 (16%)
Med dose	44 (21%)	14	54	64	31 (15%)
High dose	43 (22%)	4	49	58	41 (21%)

Some indication that chance of death decreases as dose increases.

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

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

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

Likelihood-ratio test of H_0 $\beta = 0$ has test statistic = 9.6 (df = 1, P = 0.002), based on twice difference in maximized log likelihoods compared to simpler model with $\beta = 0$.

R for modeling dose-response data, using vglm() in 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)  #  estimated multinomial response prob's
         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

> vglm(cbind(y1,y2,y3,y4,y5) ~ 1,  #  null model
       family=cumulative(parallel=TRUE), data=trauma)
Coefficients:
(Intercept):1 (Intercept):2 (Intercept):3 (Intercept):4
   -1.1423167    -0.7459897     0.2506811     1.6064484
Degrees of Freedom: 16 Total; 12 Residual
Residual Deviance: 27.79488
Log-likelihood: -53.67903

> 1 - pchisq(2*(53.67903 - 48.87282) , df=1)
[1] 0.001932658 # P-value for likelihood-ratio test of no dose effect
```

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 logit$[P(y \leq j)] = \alpha_j - \boldsymbol{\beta}'\boldsymbol{x}$ based on a latent variable model (p. 18 of these notes), for which $\hat{\beta}$ has opposite sign.

SAS for cumulative logit modeling of dose-response data

```
data trauma;
input dose outcome count @@;
datalines;
1 1 59  1 2 25  1 3 46  1 4 48  1 5 32
2 1 48  2 2 21  2 3 44  2 4 47  2 5 30
3 1 44  3 2 14  3 3 54  3 4 64  3 5 31
4 1 43  4 2  4  4 3 49  4 4 58  4 5 41
;
proc logistic; freq count; * proportional odds cumulative logit model;
   model outcome = dose / aggregate scale=none;
run;
```

SOME OUTPUT:

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	18.1825	11	1.6530	0.0774
Pearson	15.8472	11	1.4407	0.1469

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	9.6124	1	0.0019
Score	9.4288	1	0.0021
Wald	9.7079	1	0.0018

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept 1	1	-0.7192	0.1588	20.5080	<.0001
Intercept 2	1	-0.3186	0.1564	4.1490	0.0417
Intercept 3	1	0.6916	0.1579	19.1795	<.0001
Intercept 4	1	2.0570	0.1737	140.2518	<.0001
dose	1	-0.1755	0.0563	9.7079	0.0018

Stata for modeling trauma data

Note: This uses parameterization

logit$[P(y \leq j)] = \alpha_j - \boldsymbol{\beta}'\boldsymbol{x}$

generated by latent variable model. For some details about the use of the *ologit* function, see

www.ats.ucla.edu/stat/stata/output/stata_ologit_output.htm

and

www.stata.com/help.cgi?ologit

```
-------------------------------------------------------------------
. *using grouped count data
.
. infile dose y1 y2 y3 y4 y5 using trauma.txt in 2/5, clear
(eof not at end of obs)
(4 observations read)

. gen groupid=_n
.
. reshape long y, i(groupid)
(note: j = 1 2 3 4 5)

Data                               wide   ->   long
-----------------------------------------------------------------------------
Number of obs.                        4   ->      20
Number of variables                   7   ->       4
j variable (5 values)                     ->   _j
xij variables:
                           y1 y2 ... y5   ->   y
-----------------------------------------------------------------------------

. rename y count
. rename _j y
.
. list

     +----------------------------+
     | groupid   y   dose   count |
     |----------------------------|
  1. |       1   1      1      59 |
  2. |       1   2      1      25 |
  3. |       1   3      1      46 |
  4. |       1   4      1      48 |
  5. |       1   5      1      32 |
     |----------------------------|
```

```
  6. |       2   1      2      48 |
  7. |       2   2      2      21 |
  8. |       2   3      2      44 |
  9. |       2   4      2      47 |
 10. |       2   5      2      30 |
     |-----------------------------|
 11. |       3   1      3      44 |
 12. |       3   2      3      14 |
 13. |       3   3      3      54 |
 14. |       3   4      3      64 |
 15. |       3   5      3      31 |
     |-----------------------------|
 16. |       4   1      4      43 |
 17. |       4   2      4       4 |
 18. |       4   3      4      49 |
 19. |       4   4      4      58 |
 20. |       4   5      4      41 |
     +-----------------------------+
.
. ologit y dose [fw=count]  // counts are used as frequency weights

Ordered logistic regression                     Number of obs   =        802
                                                LR chi2(1)      =       9.61
                                                Prob > chi2     =     0.0019
Log likelihood = -1230.6744                     Pseudo R2       =     0.0039

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        dose |   .1754861   .0567122     3.09   0.002     .0643322      .28664
-------------+----------------------------------------------------------------
       /cut1 |  -.7191664   .1589164                     -1.030637    -.407696
       /cut2 |  -.3186011   .1568861                     -.6260921   -.0111101
       /cut3 |   .6916531   .1596505                      .378744    1.004562
       /cut4 |   2.057009   .1750751                     1.713868     2.40015
------------------------------------------------------------------------------
.
```

Goodness-of-fit statistics:

Pearson X^2 = 15.8

deviance G^2 = 18.2

(df = 16 $-$ 5 = 11)

P-values = 0.15 and 0.18

Model seems to fit adequately

Odds ratio interpretation: For dose $i+1$, estimated odds of outcome $\leq j$ (instead of $> j$) equal $\exp(-0.176) = 0.84$ times estimated odds for dose i; equivalently, for dose $i+1$, estimated odds of outcome $\geq j$ (instead of $< j$) equal $\exp(0.176) = 1.19$ times estimated odds for dose i.

95% confidence interval for $\exp(-\beta)$ is

$$e^{0.176 \pm 1.96(0.056)} = (1.07, 1.33).$$

- Cumulative odds ratio for dose levels (rows) 1 and 4 equals

$$e^{(4-1)0.176} = 1.69$$

- Any equally-spaced scores (e.g. 0, 10, 20, 30) for dose provide same fitted values and same test statistics (different $\hat{\beta}$, SE).
- Unequally-spaced scores more natural in many cases (e.g., doses may be 0, 125, 250, 500). "Sensitivity analysis" usually shows substantive results don't depend much on that choice, unless data highly unbalanced (e.g., Graubard and Korn 1987).
- The cumulative logit model uses ordinality of y without assigning category scores.
- Alternative analysis treats dose as factor, using indicator variables. Double the log-likelihood increases only 0.13, df = 2. With β_4 = 0:
  $\hat{\beta}_1$ = 0.52, $\hat{\beta}_2$ = 0.40, $\hat{\beta}_3$ = 0.20 (SE = 0.18 each)

  Testing H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4$ gives likelihood-ratio (LR) stat. = 9.8 (df = 3, P = 0.02).

  Using ordinality often increases power (focused on df = 1).

R for modeling dose-response data, with dose as a factor, using the *vglm* function in the *VGAM* library:

```
-----------------------------------------------------------------
> attach(trauma)

> library(VGAM)

> fit2 <- vglm(cbind(y1,y2,y3,y4,y5) ~ factor(dose),
+   family=cumulative(parallel=TRUE), data=trauma)

> summary(fit2)
Coefficients:
              Estimate Std. Error  z value
(Intercept):1 -0.91880    0.13204 -6.95875
(Intercept):2 -0.51826    0.12856 -4.03122
(Intercept):3  0.49215    0.12841  3.83255
(Intercept):4  1.85785    0.14527 12.78927
factor(dose)2 -0.11756    0.17843 -0.65885
factor(dose)3 -0.31740    0.17473 -1.81649
factor(dose)4 -0.52077    0.17795 -2.92657

Residual deviance: 18.04959 on 9 degrees of freedom
Log-likelihood: -48.80638 on 9 degrees of freedom
Number of iterations: 4

> 1 - pchisq(2*(53.67903 - 48.80638), df=3)
[1] 0.02086  # P-value for likelihood-ratio test of no dose effect
-----------------------------------------------------------------
```

SAS for modeling dose-response data, with dose as a factor using a CLASS statement to create indicator predictors for first three categories

```
data trauma;
input dose outcome count @@;
datalines;
1 1 59  1 2 25  1 3 46  1 4 48  1 5 32
2 1 48  2 2 21  2 3 44  2 4 47  2 5 30
3 1 44  3 2 14  3 3 54  3 4 64  3 5 31
4 1 43  4 2  4  4 3 49  4 4 58  4 5 41
;
proc logistic; freq count; class dose / param=ref; * treat dose as factor;
   model outcome = dose / aggregate scale=none;
run;
```

SOME OUTPUT WITH DOSE AS A FACTOR:

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	18.0496	9	2.0055	0.0346
Pearson	15.7881	9	1.7542	0.0714

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	9.7453	3	0.0209
Score	9.5583	3	0.0227
Wald	9.8440	3	0.0199

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept 1	1	-1.4396	0.1416	103.3943	<.0001
Intercept 2	1	-1.0390	0.1369	57.6363	<.0001
Intercept 3	1	-0.0286	0.1317	0.0472	0.8280
Intercept 4	1	1.3371	0.1428	87.7207	<.0001
dose 1	1	0.5208	0.1779	8.5641	0.0034
dose 2	1	0.4032	0.1820	4.9072	0.0267
dose 3	1	0.2034	0.1779	1.3071	0.2529

Checking goodness of fit for contingency tables

- With nonsparse contingency table data, can check goodness of fit using Pearson X^2, deviance G^2 comparing observed cell counts to expected frequency estimates.

- At setting i of predictor with $n_i = \sum_{j=1}^{c} n_{ij}$ multinomial observations, expected frequency estimates equal

$$\hat{\mu}_{ij} = n_i \hat{P}(y = j), \quad j = 1, \ldots, c.$$

- Pearson test statistic is

$$X^2 = \sum_{i,j} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}.$$

  Deviance (likelihood-ratio test statistic for testing that model holds against unrestricted alternative) is

$$G^2 = 2 \sum_{i,j} n_{ij} \log \left(\frac{n_{ij}}{\hat{\mu}_{ij}} \right).$$

  df = No. multinomial parameters $-$ no. model parameters

- With sparse data, continuous predictors, can use such measures to compare nested models.

Other properties of cumulative logit models

- Model extends to multiple explanatory variables,

$$\text{logit}[P(y \leq j)] = \alpha_j + \beta_1 x_1 + \cdots + \beta_k x_k$$

that can be qualitative (i.e., factors) or quantitative
(use indicator variables for factors)

- For subject i with values $\boldsymbol{x}_i$ on a set of explanatory variables, estimated conditional distribution function is

$$\hat{P}(y_i \leq j) = \frac{\exp(\hat{\alpha}_j + \hat{\boldsymbol{\beta}}' \boldsymbol{x}_i)}{1 + \exp(\hat{\alpha}_j + \hat{\boldsymbol{\beta}}' \boldsymbol{x}_i)}$$

Estimated probability of outcome j is

$$\hat{P}(y_i = j) = \hat{P}(y_i \leq j) - \hat{P}(y_i \leq j - 1)$$

- Can motivate proportional odds structure by a regression model for underlying continuous *latent variable*
(Anderson and Philips 1981, McKelvey and Zavoina 1975)

y = observed ordinal response

y^* = underlying continuous latent variable,

$y^* = \boldsymbol{\beta}'\boldsymbol{x} + \epsilon$ where ϵ has cdf G with mean 0. Thresholds (cutpoints) $-\infty = \alpha_0 < \alpha_1 < \ldots < \alpha_c = \infty$ such that

$$y = j \text{ if } \alpha_{j-1} < y^* \leq \alpha_j$$

Then, at fixed $\boldsymbol{x}$ (see figure on next page)

$$P(y \leq j) = P(y^* \leq \alpha_j) = P(y^* - \boldsymbol{\beta}'\boldsymbol{x} \leq \alpha_j - \boldsymbol{\beta}'\boldsymbol{x})$$

$$= P(\epsilon \leq \alpha_j - \boldsymbol{\beta}'\boldsymbol{x}) = G(\alpha_j - \boldsymbol{\beta}'\boldsymbol{x})$$

$\rightarrow$ Model $G^{-1}[P(y \leq j \mid \boldsymbol{x})] = \alpha_j - \boldsymbol{\beta}'\boldsymbol{x}$

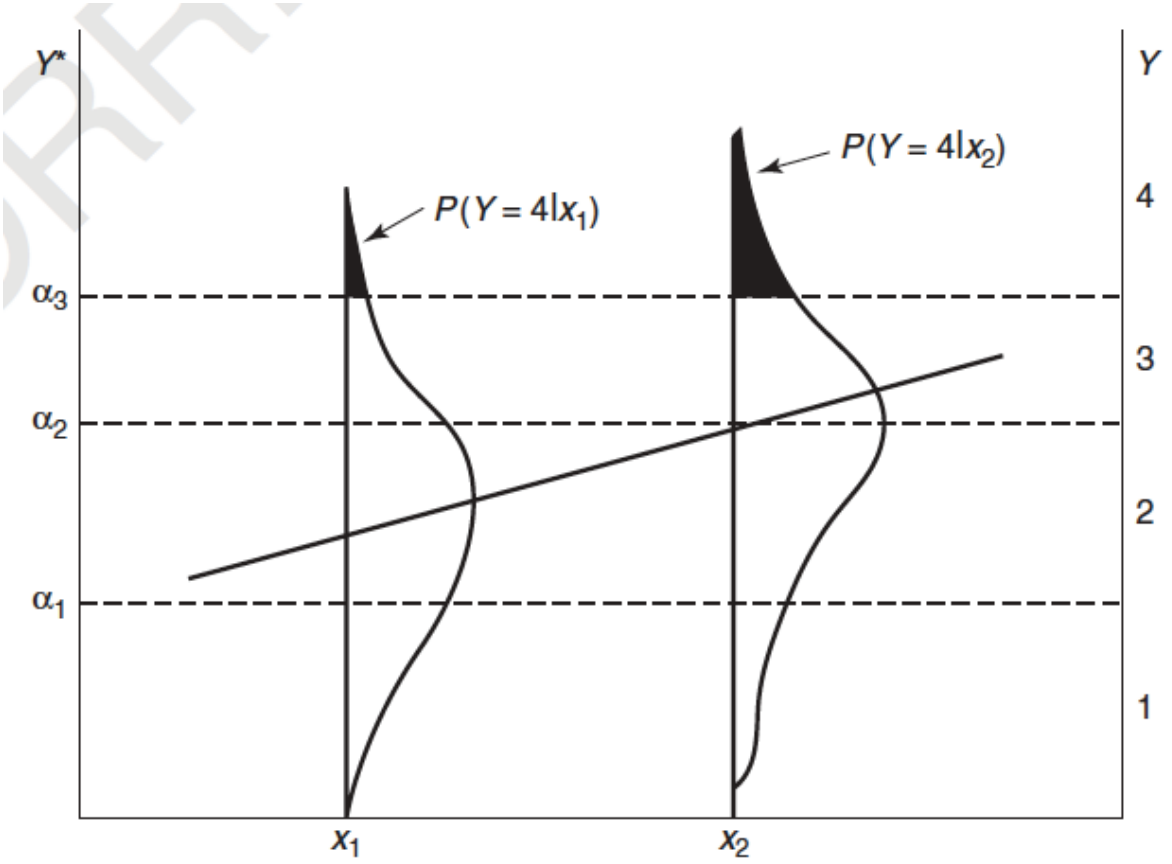
with G^{-1} a *link function*. Get cumulative logit model when G = logistic cdf $(G^{-1} =$ logit). So, cumulative logit model fits well when regression model holds for underlying logistic response.

Note: The model is often expressed as

$$\text{logit}[P(y \leq j)] = \alpha_j - \boldsymbol{\beta}'\boldsymbol{x}.$$

Then, $\beta_j > 0$ has usual interpretation of 'positive' effect

(Stata *ologit* and SPSS use this parameterization. Same fit, estimates, as using $\alpha_j + \boldsymbol{\beta}'\boldsymbol{x}$, except sign of $\boldsymbol{\beta}$)



Note: This derivation suggests such models are designed to detect shifts in *location* (center), not dispersion (spread), at different settings of explanatory variables.

This model and most others in this tutorial imply that conditional distributions of y at different settings of explanatory variables are *stochastically ordered*; i.e., the cdf at one setting is always above or always below the cdf at another level.

Other properties of cumulative logit models (continued)

- Can use similar model with alternative "cumulative link"

$$\text{link}[P(y_i \leq j)] = \alpha_j - \boldsymbol{\beta}'\boldsymbol{x}_i$$

  of cumulative prob.'s (McCullagh 1980); e.g., *cumulative probit* model (link fn. = inverse of standard normal cdf) applies naturally when underlying regression model has normal y^*.

- Effects $\boldsymbol{\beta}$ invariant to choice and number of response categories (If model holds for given response categories, holds with same $\boldsymbol{\beta}$ when response scale collapsed in any way).

- For subject i, let $(y_{i1}, \ldots, y_{ic})$ be binary indicators of the response, where $y_{ij} = 1$ when response in category j. For independent multinomial observations at values $\boldsymbol{x}_i$ of the explanatory variables for subject i, the likelihood function is

$$\prod_{i=1}^{n}\left\{\prod_{j=1}^{c}\left[P(Y_i = j \mid \boldsymbol{x}_i)\right]^{y_{ij}}\right\} =$$

$$\prod_{i=1}^{n}\left\{\prod_{j=1}^{c}\left[P(Y_i \leq j \mid \boldsymbol{x}_i) - P(Y_i \leq j-1 \mid \boldsymbol{x}_i)\right]^{y_{ij}}\right\} =$$

$$\prod_{i=1}^{n}\left\{\prod_{j=1}^{c}\left[\frac{\exp(\alpha_j + \boldsymbol{\beta}'\boldsymbol{x}_i)}{1+\exp(\alpha_j + \boldsymbol{\beta}'\boldsymbol{x}_i)} - \frac{\exp(\alpha_{j-1} + \boldsymbol{\beta}'\boldsymbol{x}_i)}{1+\exp(\alpha_{j-1} + \boldsymbol{\beta}'\boldsymbol{x}_i)}\right]^{y_{ij}}\right\}$$

Model fitting and inference

- Model fitting requires iterative methods. Log likelihood is concave (Pratt 1981). To get standard errors,
  Newton-Raphson inverts *observed* information matrix $-\partial^2 L(\boldsymbol{\beta})/\partial\beta_a\partial\beta_b$ (e.g., SAS PROC GENMOD)
  Fisher scoring inverts *expected* information matrix $E(-\partial^2 L(\boldsymbol{\beta})/\partial\beta_a\partial\beta_b)$ (e.g., R *vglm* function, SAS PROC LOGISTIC).

- McCullagh (1980) provided Fisher scoring algorithm for cumulative link models.

- Inference uses standard methods for testing H_0: $\beta_j = 0$ (likelihood-ratio, Wald, score tests) and inverting tests of H_0: $\beta_j = \beta_{j0}$ to get confidence intervals for β_j.

  Wald: $z = \frac{\hat{\beta}_j - \beta_{j0}}{SE}$, or $z^2 \sim \chi^2$ poorest method for small n or extremely large estimates (infinite being a special case)

  Likelihood-ratio: $-2([L(\hat{\boldsymbol{\beta}}_0) - L(\hat{\boldsymbol{\beta}})] \sim \chi^2$

Alternative ways of summarizing effects

- Some researchers find odds ratios difficult to interpret.
- Can compare probabilities or cumulative prob's for y directly, such as comparing $\hat{P}(y = 1)$ or $\hat{P}(y = c)$ at maximum and minimum values of a predictor (at means of other predictors).

- Summary measures of predictive power include

  (1) R^2 for regression model for underlying latent response variable (McKelvey and Zavoina 1975, provided by Stata)

  (2) *correlation* between y and estimated mean of conditional dist. of y from model fit, based on scores $\{v_j\}$ for y (mimics multiple correlation).

  (3) *concordance index* (probability that observations with different outcomes are concordant with predictions)

Checking fit (general case) and selecting a model

- Lack of fit may result from omitted predictors (e.g., interaction between predictors), violation of proportional odds assumption, wrong link function. Often, lack of fit results when there are dispersion as well as location effects.
- Can check particular aspects of fit using likelihood-ratio test to compare to more complex models (test statistic = change in deviance).
- Some software (e.g., PROC LOGISTIC) provides score test of proportional odds assumption, by comparing model to more general "non-proportional odds model" with effects $\{\boldsymbol{\beta}_j\}$. This test applicable also when X^2, G^2 don't apply, but is liberal (i.e., P(Type I error) too high). LR test also possible, except when more general model has cumulative probabilities out-of-order.
- When model with proportional odds structure fails, we can use estimated effects in non-proportional odds model (e.g., using *vglm* function in R or by fitting binary logistic model to each collapsing) to describe effects more fully.
- Even if proportional odds model has lack of fit, it may usefully summarize "first-order effects" and have good power for testing H_0: no effect, because of its parsimony

Cumulative logit models without proportional odds

Generalized model permits effects of explanatory variables to differ for different cumulative logits,

$$\text{logit}[P(y_i \leq j)] = \alpha_j + \boldsymbol{\beta}_j^{'}\boldsymbol{x}_i, \quad j = 1, \ldots, c-1.$$

Each predictor has $c-1$ parameters, allowing different effects for logit$[P(y_i \leq 1)$, logit$[P(y_i \leq 2)], \ldots,$ logit$[P(y_i \leq c-1)]$.

Even if this model fits better, for reasons of parsimony a simple model with proportional odds structure is sometimes preferable.

- Effects of explanatory variables easier to summarize and interpret.
- With large n, small P-value in test of proportional odds assumption may reflect statistical, not practical, significance.
- Effect estimators using simple model are biased but may have smaller MSE than estimators from more complex model, and tests may have greater power, especially when more complex model has many more parameters.
- Is variability in effects great enough to make it worthwhile to use more complex model?

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

> 1 - pchisq(deviance(fit)-deviance(fit2),
df=df.residual(fit)-df.residual(fit2))
[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.

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
-----------------------------------------------------------------
```

SAS for cumulative logit modeling, assuming proportional odds, of GSS religion and region data

```
data religion;
input region fund count;
    datalines;
        1 1 92
        1 2 352
        1 3 234
        2 1 274
        2 2 399
        2 3 326
        3 1 739
        3 2 536
        3 3 412
        4 1 192
        4 2 423
        4 3 355
        ;
proc genmod; weight count; class region;
    model fund = region / dist=multinomial link=clogit lrci type3 ;
run;
proc logistic; weight count; class region / param=ref;
    model fund = region / aggregate scale=none;
run;
```

GENMOD output:

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Likelihood Ratio 95% Confidence Limits		Chi-Square
Intercept1		1	-1.2618	0.0632	-1.3863	-1.1383	398.10
Intercept2		1	0.4729	0.0603	0.3548	0.5910	61.56
region	1	1	-0.0698	0.0901	-0.2466	0.1068	0.60
region	2	1	0.2688	0.0830	0.1061	0.4316	10.48
region	3	1	0.8897	0.0758	0.7414	1.0385	137.89
region	4	0	0.0000	0.0000	0.0000	0.0000	.

R for religion and region data, using vglm() for cumulative logit modeling with and without proportional odds structure

```
> 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
> r1 < -ifelse(region==1,1,0); r2 <-ifelse(region==2,1,0); r3 <-ifelse(region==3,1,0)
> cbind(r1,r2,r3)
     r1 r2 r3
[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) ~ r1+r2+r3,
         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
r1            -0.069842   0.093035  -0.75071
r2             0.268777   0.083536   3.21750
r3             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) ~ r1+r2+r3, family=cumulative,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
r1:1          -0.452300   0.138093  -3.27532
r1:2           0.090999   0.104731   0.86888
r2:1           0.426188   0.107343   3.97032
r2:2           0.175343   0.094849   1.84866
r3:1           1.150175   0.094349  12.19065
r3:2           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
> 1 - pchisq(deviance(fit.po)-deviance(fit.npo),
      df=df.residual(fit.po)-df.residual(fit.npo))
[1] 4.134028e-21
```

Stata for modeling religion and region data, for cumulative logit modeling with and without proportional odds

```
-------------------------------------------------------------------------------
. infile region y1 y2 y3 using region.txt in 2/5, clear
(eof not at end of obs)
(4 observations read)

. list
     +--------------------------+
     | region    y1    y2    y3 |
     |--------------------------|
  1. |      1    92   352   234 |
  2. |      2   274   399   326 |
  3. |      3   739   536   412 |
  4. |      4   192   423   355 |
     +--------------------------+

. gen groupid=_n
. reshape long y, i(groupid)
(note: j = 1 2 3)

Data                               wide   ->   long
-------------------------------------------------------------------------------
Number of obs.                        4   ->      12
Number of variables                   5   ->       4
j variable (3 values)                     ->   _j
xij variables:
                                y1 y2 y3   ->   y
-------------------------------------------------------------------------------

. rename y count
. rename _j y
. list
     +------------------------------+
     | groupid   y   region   count |
     |------------------------------|
  1. |       1   1        1      92 |
  2. |       1   2        1     352 |
  3. |       1   3        1     234 |
  4. |       2   1        2     274 |
  5. |       2   2        2     399 |
     |------------------------------|
  6. |       2   3        2     326 |
  7. |       3   1        3     739 |
  8. |       3   2        3     536 |
```

```
  9. |       3   3        3     412 |
 10. |       4   1        4     192 |
     |-------------------------------|
 11. |       4   2        4     423 |
 12. |       4   3        4     355 |
     +-------------------------------+
.
. tab region, gen(reg) // create dummy indicators for region

     region |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |          3       25.00       25.00
          2 |          3       25.00       50.00
          3 |          3       25.00       75.00
          4 |          3       25.00      100.00
------------+-----------------------------------
      Total |         12      100.00
.
. *check the proportional odds assumption
. omodel logit y reg1 reg2 reg3 reg4 [fw=count]

Ordered logit estimates                           Number of obs   =       4334
                                                  LR chi2(3)      =     206.48
                                                  Prob > chi2     =     0.0000
Log likelihood = -4622.4007                       Pseudo R2       =     0.0218

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        reg1 |   .0698393   .0901259     0.77   0.438    -.1068042    .2464828
        reg2 |  -.2687773   .0830439    -3.24   0.001    -.4315402   -.1060143
        reg3 |  -.8896776   .0757644   -11.74   0.000    -1.038173    -.741182
-------------+----------------------------------------------------------------
       _cut1 |  -1.261818   .0632411          (Ancillary parameters)
       _cut2 |   .4728514   .0602666
------------------------------------------------------------------------------
Approximate likelihood-ratio test of proportionality of odds
across response categories:
         chi2(3) =     98.78
       Prob > chi2 =    0.0000
```

```
. *model WITHOUT PROPORTIONAL ODDS ASSUMPTION
>
. gologit2 y reg1 reg2 reg3 reg4 [fw=count]

Generalized Ordered Logit Estimates               Number of obs   =       4334
                                                  LR chi2(6)      =     304.51
                                                  Prob > chi2     =     0.0000
Log likelihood = -4573.3888                       Pseudo R2       =     0.0322
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
        reg1 |   .4523001   .1380932     3.28   0.001     .1816424    .7229577
        reg2 |  -.4261876   .1073435    -3.97   0.000     -.636577   -.2157982
        reg3 |  -1.150175   .0943489   -12.19   0.000    -1.335095   -.9652542
       _cons |   1.399231   .0805834    17.36   0.000     1.241291    1.557172
-------------+----------------------------------------------------------------
2            |
        reg1 |   -.090999   .1047314    -0.87   0.385    -.2962688    .1142709
        reg2 |  -.1753435   .0948488    -1.85   0.065    -.3612436    .0105567
        reg3 |  -.5801736   .0874895    -6.63   0.000      -.75165   -.4086973
       _cons |  -.5495045   .0666552    -8.24   0.000    -.6801463   -.4188627
------------------------------------------------------------------------------
```

Model assuming proportional odds has (with $\beta_4 = 0$)

$(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$ = $(-0.07, 0.27, 0.89)$

For more general model,

$(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$ = $(-0.45, 0.43, 1.15)$ for logit$[P(Y \leq 1)]$

$(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$ = $(0.09, 0.18, 0.58)$ for logit$[P(Y \leq 2)]$.

Change in sign of $\hat{\beta}_1$ reflects lack of stochastic ordering of regions 1 and 4; their cdf's don't always have same order.

Compared to resident of West, a Northeast resident is less likely to be fundamentalist (see $\hat{\beta}_1 = -0.45 < 0$ for logit$[P(Y \leq 1)]$) but slightly more likely to be fundamentalist or moderate and slightly less likely to be liberal (see $\hat{\beta}_1 = 0.09 > 0$ for logit$[P(Y \leq 2)]$).

Peterson and Harrell (1990) proposed *partial proportional odds model* falling between proportional odds model and more general model,

$$\text{logit}[P(y_i \leq j)] = \alpha_j + \boldsymbol{\beta}' \boldsymbol{x}_i + \boldsymbol{\gamma}_j' \mathbf{u}_i, \quad j = 1, \ldots, c-1.$$

2. Other Ordinal Models

a. Models using *adjacent-category logits* (ACL)

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

- Odds uses adjacent categories, whereas in cumulative logit model it uses entire response scale, so interpretations use *local* odds ratios instead of *cumulative* odds ratios.
- Model also has proportional odds structure, for these logits (effect $\boldsymbol{\beta}$ same for each cutpoint j).
- Effects in paired-category logit models such as ACL are estimable with retrospective studies (e.g., case-control) that sample $\boldsymbol{x}$ conditional on y, but not with models such as cumulative logit that group categories together (Mukherjee and Liu 2008).

- Anderson (1984) noted that if

$$(\boldsymbol{x} \mid y = j) \sim N(\boldsymbol{\mu}_j, \boldsymbol{\Sigma})$$

then

$$\log\left[\frac{P(y = j \mid \boldsymbol{x})}{P(y = j+1 \mid \boldsymbol{x})}\right] = \alpha_j + \boldsymbol{\beta}_j^{'}\boldsymbol{x}$$

with

$$\boldsymbol{\beta}_j = \boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_j - \boldsymbol{\mu}_{j+1})$$

Equally-spaced means imply ACL model holds with same effects for each logit.

- ACL and cumulative logit models with proportional odds structure fit well in similar situations and provide similar substantive results (both imply stochastic orderings of conditional distributions of y at different predictor values)

- Which to use? Cumulative logit extends inference to underlying continuum and is invariant with respect to choice of response categories. ACL gives effects in terms of fixed categories, which is preferable to provide interpretations for given categories rather than underlying continuum, and those effects are estimable with retrospective studies.

ACL model effects for any pair of response categories

Since for $j < k$,

$$\log\left(\frac{\pi_j}{\pi_k}\right) = \log\left(\frac{\pi_j}{\pi_{j+1}}\right) + \log\left(\frac{\pi_{j+1}}{\pi_{j+2}}\right) + \cdots + \log\left(\frac{\pi_{k-1}}{\pi_k}\right),$$

ACL model $\log\left[\frac{\pi_j(\boldsymbol{x})}{\pi_{j+1}(\boldsymbol{x})}\right] = \alpha_j + \boldsymbol{\beta}'\boldsymbol{x}$

implies paired-category logistic model

$$\log\left[\frac{\pi_j(\boldsymbol{x})}{\pi_k(\boldsymbol{x})}\right] = \sum_{i=j}^{k-1} \alpha_i + \boldsymbol{\beta}'(k-j)\boldsymbol{x}$$

so log odds ratios multiplied by $(k-j)$.

Model equivalently can be expressed in terms of baseline-category logits (BCL), which with baseline c are

$$\log\left(\frac{\pi_1}{\pi_c}\right),\ \log\left(\frac{\pi_2}{\pi_c}\right),\ \ldots,\ \log\left(\frac{\pi_{c-1}}{\pi_c}\right).$$

ACL model in terms of category probabilities is

$$P(Y=j) = \frac{\exp[\sum_{k=j}^{c-1}\alpha_k + (c-j)\boldsymbol{\beta}'\mathbf{x}]}{1 + \sum_{k=1}^{c-1}\exp[\sum_{k'=j}^{c-1}\alpha_{k'} + (c-j)\boldsymbol{\beta}'\mathbf{x}]}.$$

Example: Stem Cell Research and Religious Fundamentalism

(from 2006 General Social Survey)

		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), ACL model

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

is equivalent to BCL model

$$\log(\pi_j/\pi_4) = \alpha_j^* + \beta_1(4-j)x + \beta_2(4-j)g$$

R: vglm() function in VGAM library has adjacent-categories logit model as 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
```

SAS: Can fit with PROC NLMIXED, which permits specifying the log-likelihood to be maximized, here *ll* statement and expressing model as baseline-category logit model.

```
data stemcell;
input religion gender y1 y2 y3 y4;
datalines;
1 0 21 52 24 15
1 1 34 67 30 25
2 0 30 52 18 11
2 1 41 83 23 14
3 0 64 50 16 11
3 1 58 63 15 12
;
/* Adjacent-categories logit model with proportional odds */
proc nlmixed data=stemcell;
eta1 = alpha1 + alpha2 + alpha3 + 3*beta1*religion + 3*beta2*gender;
eta2 = alpha2 + alpha3 + 2*beta1*religion + 2*beta2*gender;
eta3 = alpha3 + beta1*religion + beta2*gender;
p4 = 1 / (1 + exp(eta1) + exp(eta2) + exp(eta3));
p1 = exp(eta1)*p4;
p2 = exp(eta2)*p4;
p3 = exp(eta3)*p4;
ll = y1*log(p1) + y2*log(p2) + y3*log(p3) + y4*log(p4);
model y1 ~ general(ll);
run;
```

Parameter Estimates

Parameter	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
alpha1	-0.9509	0.1426	6	-6.67	0.0006	0.05	-1.2998	-0.6020
alpha2	0.5573	0.1451	6	3.84	0.0085	0.05	0.2023	0.9123
alpha3	-0.1066	0.1648	6	-0.65	0.5417	0.05	-0.5097	0.2966
beta1	0.2668	0.04787	6	5.57	0.0014	0.05	0.1497	0.3839
beta2	-0.01412	0.07671	6	-0.18	0.8600	0.05	-0.2018	0.1736

- For moderates, estimated odds of (definitely fund) vs. (probably fund) are $\exp(0.2668) = 1.31$ times estimated odds for fundamentalists, whereas estimated odds of (definitely fund) vs. (definitely not fund) are $\exp[3(0.2668)] = 2.23$ times the estimated odds for fundamentalists, for each gender.

- Ordinal models with trend in location display strongest association with most extreme categories. e.g., for liberals, estimated odds of (definitely fund) vs. (definitely not) are $\exp[2(3)(0.2668)] = 4.96$ times estimated odds for fundamentalists, for each gender.

- Model describes 18 multinomial probabilities (3 for each religion$\times$gender combination) using 5 parameters. Deviance $G^2 = 12.00$, $df = 18 - 5 = 13$ (P-value = 0.53).

- Similar substantive results with cumulative logit model.

  Religious beliefs effect larger ($\hat{\beta}_1 = 0.488$, $SE = 0.080$), since refers to entire response scale. However, statistical significance similar, with $(\hat{\beta}_1/SE) > 5$ for each model.

Adjacent-Categories Logit Models with Nonproportional Odds

- As in cumulative logit case, model of proportional odds form fits poorly when there are substantive dispersion effects,
- The more general non-proportional odds form is

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

- Unlike cumulative logit model, this model does 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 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 (and simpler to interpret).

b. Models using *continuation-ratio* logits

$\log[P(y_i = j)/P(y_i \geq j+1)], \;\; j = 1, ..., c-1,$ or

$\log[P(y_i = j+1)/P(y_i \leq j)], \;\; j = 1, ..., c-1$

Let $\omega_j = P(y = j \mid y \geq j) = \frac{\pi_j}{\pi_j + \cdots + \pi_c}$

Then

$$\log\left(\frac{\pi_j}{\pi_{j+1} + \cdots + \pi_c}\right) = \log[\omega_j/(1-\omega_j)],$$

- Of interest when a *sequential* mechanism determines the response outcome (Tutz 1991) or for grouped survival data
- Simple model with proportional odds structure is

$$\text{logit}[\omega_j(\boldsymbol{x})] = \alpha_j + \boldsymbol{\beta}'\boldsymbol{x}, \quad j = 1, \ldots, c-1,$$

- More general model $\text{logit}[\omega_j(\boldsymbol{x})] = \alpha_j + \boldsymbol{\beta}_j'\boldsymbol{x}$ has fit equivalent to fit of $c-1$ separate binary logit models, because multinomial factors into binomials,

$$p(y_{i1}, \ldots, y_{ic}) = p(y_{i1})p(y_{i2} \mid y_{i1}) \cdots p(y_{ic} \mid y_{i1}, \ldots, y_{i,c-1}) =$$

$$bin[1, \; y_{i1}; \; \omega_1(\boldsymbol{x}_i)] \cdots bin[1 - y_{i1} - \cdots - y_{i,c-2}, \; y_{i,c-1}; \; \omega_{c-1}(\boldsymbol{x}_i)].$$

Example: 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 fits well (deviance 0.01, $df = 1$):

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

has $\hat{\beta} = -0.528$ ($SE = 0.196$). Model estimates an assumed common value $\exp(-0.528) = 0.59$ for cumulative odds ratio from first part of model and for local odds ratio from second part.

e.g., given that tonsils were enlarged, for carriers, estimated odds of response enlarged rather than greatly enlarged were 0.59 times estimated odds for non-carriers.

By contrast, cumulative logit model estimates $\exp(-0.6025) = 0.55$ for each cumulative odds ratio, and ACL model estimates $\exp(-0.429) = 0.65$ for each local odds ratio. (Both these models also fit well: Deviances 0.30, 0.24, $df = 1$.)

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

> fit2.cratio <- vglm(cbind(y1,y2,y3) ~ carrier,
           family=sratio(parallel=TRUE), data=tonsils)
```

Note: family=cratio parameterizes as reciprocal, so $\hat{\beta}$ has opposite sign; will get correct sign using family=sratio as shown at end of code.

SAS: Fit continuation-ratio logit models using procedures for binary logistic regression

```
------------------------------------------------------------------------------
data tonsils; * look at data as indep. binomials;
input stratum carrier success failure;  n = success + failure;
datalines;
1 1 19 53
1 0 497 829
2 1 29 24
2 0 560 269
;
proc genmod data=tonsils; class stratum;
model success/n = stratum carrier / dist=binomial link=logit lrci type3;
------------------------------------------------------------------------
```

Parameter		DF	Estimate	Standard Error	Likelihood Ratio 95% Confidence Limits		Chi-Square
Intercept		1	0.7322	0.0729	0.5905	0.8762	100.99
stratum	1	1	-1.2432	0.0907	-1.4220	-1.0662	187.69
stratum	2	0	0.0000	0.0000	0.0000	0.0000	.
carrier		1	-0.5285	0.1979	-0.9218	-0.1444	7.13

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
carrier	1	7.32	0.0068

```
------------------------------------------------------------------------
```

Or, can fit directly using PROC NLMIXED

```
------------------------------------------------------------------------------
data tonsil;
input carrier y1 y2 y3;
datalines;
1 19 29 24
0 497 560 269
;
proc nlmixed data=tonsil;
eta1 = alpha1 + beta*carrier;  eta2 = alpha2 + beta*carrier;
p1 = exp(eta1)/(1+exp(eta1));
p2 = exp(eta2)/((1+exp(eta1))*(1+exp(eta2)));
p3 = 1-p1-p2;
ll = y1*log(p1) + y2*log(p2) + y3*log(p3);
model y1 ~ general(ll);
run;
------------------------------------------------------------------------
```

Stata for modeling tonsil inflammation data

```
. infile carrier y1 y2 y3 using tonsils.txt in 2/3, clear
(eof not at end of obs)
(2 observations read)

. list

     +---------------------------+
     | carrier    y1    y2    y3 |
     |---------------------------|
  1. |       1    19    29    24 |
  2. |       0   497   560   269 |
     +---------------------------+

. gen groupid=_n

.
. reshape long y, i(groupid)
(note: j = 1 2 3)

Data                               wide   ->   long
-----------------------------------------------------------------------------
Number of obs.                        2   ->       6
Number of variables                   5   ->       4
j variable (3 values)                     ->   _j
xij variables:
                              y1 y2 y3   ->   y
-----------------------------------------------------------------------------

. rename y count
. rename _j y
.
. list

     +-------------------------------+
     | groupid   y   carrier   count |
     |-------------------------------|
  1. |       1   1         1      19 |
  2. |       1   2         1      29 |
  3. |       1   3         1      24 |
  4. |       2   1         0     497 |
  5. |       2   2         0     560 |
     |-------------------------------|
  6. |       2   3         0     269 |
     +-------------------------------+
```

```
. tab carrier y [fw=count], row

           |                y
   carrier |         1          2          3 |     Total
-----------+---------------------------------+----------
         0 |       497        560        269 |     1,326
           |     37.48      42.23      20.29 |    100.00
-----------+---------------------------------+----------
         1 |        19         29         24 |        72
           |     26.39      40.28      33.33 |    100.00
-----------+---------------------------------+----------
     Total |       516        589        293 |     1,398
           |     36.91      42.13      20.96 |    100.00

.
. ologit y carrier [fw=count] // ordered logit model

Ordered logistic regression                     Number of obs   =       1398
                                                LR chi2(1)      =       7.02
                                                Prob > chi2     =     0.0081
Log likelihood = -1477.7474                     Pseudo R2       =     0.0024

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     carrier |   .6026492   .2274158     2.65   0.008     .1569224    1.048376
-------------+----------------------------------------------------------------
       /cut1 |  -.5085091   .0563953                     -.6190418   -.3979763
       /cut2 |    1.36272   .0673406                      1.230735    1.494705
------------------------------------------------------------------------------

.
. ocratio y carrier [fw=count]  // continuation ratio model

Continuation-ratio logit Estimates                  Number of obs =    2280
                                                    chi2(1)       =    7.32
                                                    Prob > chi2   =  0.0068
Log Likelihood = -1477.599                          Pseudo R2     =  0.0025

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     carrier |   .5284613    .197904     2.67   0.008     .1405766     .916346
-------------+----------------------------------------------------------------
 _cut1       |  -.5110188   .0561416             (Ancillary parameters)
 _cut2       |   .7321801   .0728583
------------------------------------------------------------------------------
```

c. Cumulative Probit Models

Denote *cdf* of standard normal by Φ.

Cumulative probit model is

$$\Phi^{-1}[P(y \leq j)] = \alpha_j + \boldsymbol{\beta}'\boldsymbol{x}, \quad j = 1, \ldots, c-1$$

Recall that in binary response case with single predictor and $\beta > 0$, this says that as a function of x, $P(y = 1)$ looks like a normal cdf for some μ, σ.

As in proportional odds models (logit link), effect $\boldsymbol{\beta}$ is same for each cumulative probability.

(Here, not appropriate to call this a “proportional odds” model, because interpretations do not apply to odds or odds ratios.)

Properties

- Motivated by underlying normal regression model for latent variable y^* with constant σ.
  (σ = 1 gives standard normal for link function).
- Then, coefficient β_k of x_k has interpretation that a unit increase in x_k corresponds to change in $E(y^*)$ of β_k standard deviations, keeping fixed other predictor values.
- Logistic and normal *cdf*s having same mean and standard deviation look similar, so cumulative probit models and cumulative logit models fit well in similar situations.
- Standard logistic distribution $G(y) = e^y/(1 + e^y)$ has mean 0 and standard deviation $\pi/\sqrt{3} = 1.8$. The ML estimates from cumulative logit models tend to be about 1.6 to 1.8 times ML estimates from cumulative probit models.
- Quality of fit and statistical significance essentially same for cumulative probit and cumulative logit models. Both imply stochastic orderings at different $\boldsymbol{x}$ levels and are designed to detect location rather than dispersion effects.

Example: Religious fundamentalism by highest educational degree

(GSS data from 1972 to 2006, huge n, example chosen to show difficulty of discriminating between logit and probit even with enormous sample sizes.)

	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

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

Residual Deviance: 48.70723 on 7 degrees of freedom

> vglm(cbind(y1,y2,y3) ~ degree,
   family=cumulative(link=logit, parallel=TRUE), data=fundamentalism)

Coefficients:
(Intercept):1 (Intercept):2        degree
   -0.3520540     1.5498053    -0.3446603

Degrees of Freedom: 10 Total; 7 Residual
Residual Deviance: 45.3965
```

SAS: PROC GENMOD and LOGISTIC fit cumulative probit

```
-----------------------------------------------------------------------
data religion;
input degree religion count;
    datalines;
    0 1 4913
    0 2 4684
...
    4 3 1369
        ;
proc logistic; weight count;
    model religion = degree  / link=probit aggregate scale=none;
proc logistic; weight count; class degree / param=ref;
    model religion = degree  / link=probit aggregate scale=none;
-----------------------------------------------------------------------
             Score Test for the Equal Slopes Assumption
                  Chi-Square       DF     Pr > ChiSq
                      0.2452        1         0.6205

           Deviance and Pearson Goodness-of-Fit Statistics
    Criterion          Value       DF     Value/DF     Pr > ChiSq
    Deviance         48.7072        7       6.9582         <.0001
    Pearson          48.9704        7       6.9958         <.0001

               Criterion    Intercept Only  Intercept and Covariates
               -2 Log L      105528.77           103389.09

                                  Standard          Wald
Parameter      DF    Estimate       Error    Chi-Square    Pr > ChiSq
Intercept 1     1     -0.2240     0.00799      785.6659        <.0001
Intercept 2     1      0.9400     0.00868    11736.5822        <.0001
degree          1     -0.2059     0.00447     2120.0908        <.0001
-----------------------------------------------------------------------
    Criterion          Value       DF     Value/DF     Pr > ChiSq
    Deviance          5.1606        4       1.2902         0.2712
    Pearson           5.1616        4       1.2904         0.2711

                                  Standard          Wald
Parameter      DF    Estimate       Error    Chi-Square    Pr > ChiSq
Intercept 1     1     -1.0169      0.0210     2355.5732        <.0001
Intercept 2     1      0.1478      0.0206       51.3520        <.0001
degree    0     1      0.8298      0.0231     1289.2450        <.0001
degree    1     1      0.5599      0.0217      666.9138        <.0001
degree    2     1      0.4639      0.0303      234.1537        <.0001
degree    3     1      0.1695      0.0247       47.0787        <.0001
```

Stata for cumulative logit and probit modeling of religious beliefs

```
-------------------------------------------------------------------------------
. infile degree y1 y2 y3 using religion.txt in 2/6, clear
(eof not at end of obs)
(5 observations read)
.
. list
     +-----------------------------+
     | 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 |
     +-----------------------------+

. gen groupid=_n
.
. reshape long y, i(groupid)
(note: j = 1 2 3)

Data                               wide   ->   long
-----------------------------------------------------------------------------
Number of obs.                        5   ->      15
Number of variables                   5   ->       4
j variable (3 values)                     ->   _j
xij variables:
                                y1 y2 y3   ->   y
-----------------------------------------------------------------------------

. rename y count
. rename _j y
.
. list

     +-------------------------------+
     | groupid   y   degree   count |
     |-------------------------------|
  1. |       1   1        0    4913 |
  2. |       1   2        0    4684 |
  3. |       1   3        0    1905 |
  4. |       2   1        1    8189 |
  5. |       2   2        1   11196 |
     |-------------------------------|
  6. |       2   3        1    6045 |
```

```
  7. |       3   1        2     728 |
  8. |       3   2        2    1072 |
  9. |       3   3        2     679 |
 10. |       4   1        3    1304 |
     |------------------------------|
 11. |       4   2        3    2800 |
 12. |       4   3        3    2468 |
 13. |       5   1        4     495 |
 14. |       5   2        4    1193 |
 15. |       5   3        4    1369 |
     +------------------------------+
.
. ologit y degree [fw=count] // ordered logit

Ordered logistic regression                     Number of obs   =      49040
                                                LR chi2(1)      =    2142.99
                                                Prob > chi2     =     0.0000
Log likelihood = -51692.888                     Pseudo R2       =     0.0203

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      degree |   .3446603   .0075309    45.77   0.000        .3299    .3594205
-------------+----------------------------------------------------------------
       /cut1 |   -.352054   .0130676                     -.3776659   -.3264421
       /cut2 |   1.549805   .0149954                      1.520415    1.579196
------------------------------------------------------------------------------
.
. oprobit y degree [fw=count] // ordered probit

Ordered probit regression                       Number of obs   =      49040
                                                LR chi2(1)      =    2139.68
                                                Prob > chi2     =     0.0000
Log likelihood = -51694.544                     Pseudo R2       =     0.0203

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      degree |   .2059429   .0044745    46.03   0.000     .1971731    .2147128
-------------+----------------------------------------------------------------
       /cut1 |  -.2239807   .0079956                     -.2396517   -.2083096
       /cut2 |   .9400106   .0086789                      .9230003     .957021
------------------------------------------------------------------------------
```

- From probit $\hat{\beta} = -0.206$, for category increase in highest degree, mean of underlying response on religious beliefs estimated to decrease by 0.21 standard deviations.
- From logit $\hat{\beta} = -0.345$, estimated odds of response in fundamentalist rather than liberal direction multiply by $\exp(-0.345) = 0.71$ for each category increase in degree. e.g., estimated odds of fundamentalist rather than moderate or liberal for those with less high school education are $1/\exp[4(-0.345)] = 4.0$ times estimated odds for those with graduate degree.

  For each category increase in highest degree, mean of underlying response on religious beliefs estimated to decrease by $0.345/(\pi/\sqrt{3}) = 0.19$ standard deviations.

Goodness of fit?

Cumulative probit: Deviance = 48.7 ($df = 7$)

Cumulative logit: Deviance = 45.4 ($df = 7$)

Either link treating education as *factor* passes goodness-of-fit test, but fit not practically different than with simpler linear trend model.

e.g., Probit deviance = 5.2, logit deviance = 2.4 ($df = 4$)

Probit $\hat{\beta}_1 = 0.83, \hat{\beta}_2 = 0.56, \hat{\beta}_3 = 0.46, \hat{\beta}_4 = 0.17, \hat{\beta}_5 = 0$

d. Cumulative Log-Log Links

Logit and probit links have symmetric S shape, in sense that $P(y \leq j)$ approaches 1.0 at same rate as it approaches 0.0.

Model with *complementary log-log link*

$$\log\{-\log[1 - P(y \leq j)]\} = \alpha_j + \boldsymbol{\beta}'\boldsymbol{x}$$

approaches 1.0 at *faster* rate than approaches 0.0. It and corresponding *log-log link*,

$$\log\{-\log[P(y \leq j)]\},$$

based on underlying skewed distributions (extreme value) with *cdf* of form $G(y) = \exp\{-\exp[-(y-a)/b]\}$.

- Model with complementary log-log link has interpretation that

  $P(y > j \mid \boldsymbol{x} \text{ with } x_k = x+1) = P(y > j \mid \boldsymbol{x} \text{ with } x_k = x)^{\exp(\beta_k)}$

- Most software provides complementary log-log link, but can fit model with log-log link by reversing order of categories and using complementary log-log link.

Example: Life table for gender and race

(These are population percentages, not counts, so we use model for description but not inference)

	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 (SE values irrelevant)

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
```

SAS: Use PROC GENMOD or LOGISTIC for complementary log-log link

```
data lifetab;
input sex $ race $ age count;
    datalines;
        m w 20  13
        f w 20   9
        m b 20  26
        f b 20  18
...
        m w 100 805
        f w 100 879
        m b 100 668
        f b 100 792
        ;
proc logistic; freq count; class sex race / param=ref;
    model age = sex race  / link=cloglog aggregate scale=none;
run;
proc genmod; freq count; class sex race;
    model  age = sex race  / dist=multinomial link=CCLL lrci type3 obstats;
run;
```

The GENMOD Procedure

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Standard Error	Likelihood Ratio 95% Confidence Limits		Chi-Square
Intercept1		1	-4.2127	0.1338	-4.4840	-3.9587	991.04
Intercept2		1	-3.1922	0.0911	-3.3741	-3.0168	1226.85
Intercept3		1	-2.5821	0.0764	-2.7340	-2.4347	1143.60
Intercept4		1	-1.5216	0.0623	-1.6458	-1.4015	596.43
sex	f	1	-0.5383	0.0703	-0.6769	-0.4011	58.57
sex	m	0	0.0000	0.0000	0.0000	0.0000	.
race	b	1	0.6107	0.0709	0.4725	0.7506	74.20
race	w	0	0.0000	0.0000	0.0000	0.0000	.

Stata for comp. log-log link modeling of life table data

```
-------------------------------------------------------------------------------.
infile gender race y1 y2 y3 y4 y5 using ltable.txt in 2/5, clear
(eof not at end of obs)
(4 observations read)

.
. gen groupid=_n

.
. reshape long y, i(groupid)
(note: j = 1 2 3 4 5)

Data                               wide   ->   long
-----------------------------------------------------------------------------
Number of obs.                        4   ->      20
Number of variables                   8   ->       5
j variable (5 values)                     ->   _j
xij variables:
                           y1 y2 ... y5   ->   y
-----------------------------------------------------------------------------

. rename y percent
. rename _j y
.
. gen count=percent*10
.
. tab gender y if race==0 [fw=count], row


           |                              y
    gender |         1          2          3          4          5 |     Total
-----------+--------------------------------------------------------+----------
         0 |        13         28         32        122        805 |     1,000
           |      1.30       2.80       3.20      12.20      80.50 |    100.00
-----------+--------------------------------------------------------+----------
         1 |         9         13         19         80        879 |     1,000
           |      0.90       1.30       1.90       8.00      87.90 |    100.00
-----------+--------------------------------------------------------+----------
     Total |        22         41         51        202      1,684 |     2,000
           |      1.10       2.05       2.55      10.10      84.20 |    100.00


. tab gender y if race==1 [fw=count], row

+----------------+
```

```
| Key            |
|----------------|
|   frequency    |
| row percentage |
+----------------+
```

gender	y = 1	2	3	4	5	Total
0	26	49	56	201	668	1,000
	2.60	4.90	5.60	20.10	66.80	100.00
1	18	24	37	129	792	1,000
	1.80	2.40	3.70	12.90	79.20	100.00
Total	44	73	93	330	1,460	2,000
	2.20	3.65	4.65	16.50	73.00	100.00

```
.
. *cumulative complementary log log model
. ocratio y gender race [fw=count], link(cloglog) cumulative

Ordered cloglog Estimates                         Number of obs =   15430
                                                  chi2(2)       =  136.22
                                                  Prob > chi2   =  0.0000
Log Likelihood = -2917.397                        Pseudo R2     =  0.0228

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      gender |   .5382685   .0703365     7.65   0.000     .4004115    .6761254
        race |  -.6107102   .0708956    -8.61   0.000     -.749663   -.4717574
-------------+----------------------------------------------------------------
 _cut1       |   -4.21274   .1338366             (Ancillary parameters)
 _cut2       |   -3.19223   .1228357
 _cut3       |  -2.582102   .1161383
 _cut4       |  -1.521633   .0906746
------------------------------------------------------------------------------
```

gender effect estimate $\beta_1 = -0.538$

race effect estimate $\beta_2 = 0.611$

Gender effect described by :

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

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 with proportional odds structure:
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

Extensions to Clustered and Multivariate Data

- **Marginal models**: Generalized estimating equations (GEE) methods extend to ordinal responses, such as for cumulative logit models (Lipsitz et al. 1994, Touloumis et al. 2013).

  R: *multgee* package has *ordLORgee* function that can fit cumulative link and adjacent-categories logit models, based on using local odds ratios to describe working association structure. Also, can use *repolr* function in *repolr* library for proportional odds version of cumulative logit model.

  SAS: PROC GENMOD, but only with "independence working correlation structure."

- **Random effects models**: Can include random effects in the various types of ordinal logit models (Hedeker and Gibbons 1994, Tutz and Hennevogl 1996, Agresti and Natarajan 2001).

  R: *clmm* function in *ordinal* package fits cumulative logit models with random effects, using Laplace approximation.

  SAS: PROC NLMIXED uses Gauss-Hermite quadrature for ML fitting of random effects models, extending PROC MIXED to handle non-normal response and link functions of GLMs.

Summary of Ordinal Modeling

- Logistic regression for binary responses extends in various ways to handle ordinal responses: Use logits for cumulative probabilities, adjacent-response categories, or a mix (continuation-ratio logits).
- Other ordinal multinomial models include cumulative link models (e.g., probit).
- Which model to use? Apart from certain types of data in which grouped response models are invalid (e.g., cumulative logits with case-control data or effects varying among logits), we may consider

  (1) how we want to summarize effects (e.g., cumulative prob's with cumulative logit, individual category prob's with ACL)

  and

  (2) do we want a connection with an underlying latent variable model (natural with cumulative logit and other cumulative link models)?

Software for Modeling Ordinal Data

SAS

- PROC FREQ provides large-sample and small-sample tests of independence in two-way tables, measures of association and their estimated SEs.
- PROC GENMOD fits multinomial cumulative link models and Poisson loglinear models , and it can perform GEE analyses for marginal models as well as Bayesian model fitting for binomial and Poisson data.
- PROC LOGISTIC fits cumulative link models.
- PROC NLMIXED and PROC GLIMMIX fit models with random effects. PROC NLMIXED can also fit other generalized nonlinear models.
- PROC CATMOD can fit baseline-category logit models by ML, and hence adjacent-category logit models.
- See *Categorical Data Analysis Using SAS*, 3rd ed., by M. Stokes, C. S. Davis, and G. G. Koch (2012) for more details about using SAS for categorical data analyses.

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*).

- Many other R functions can fit cumulative logit and other cumulative link models. Thompson's manual (p. 121) describes the *polr* function from the MASS library, used in these notes for the dose-response data (p. 19).

- *multgee* package has *ordLORgee* function that can fit cumulative link and adjacent-categories logit models, based on using local odds ratios to describe working association structure. The package repolr contains a function *repolr* for repeated proportional odds logistic regression. The package geepack contains a function *ordgee* for ordinal GEE analyses, but a PhD student of mine and I have found it to be very unreliable (often gives incorrect results, such as for example in Thompson manual).

- The *clmm* function in the *ordinal* package can fit cumulative logit models with random effects. The package glmmAK contains a function *cumlogitRE* for using MCMC to fit such models.

- R function *mph.fit* prepared by Joe Lang at Univ. of Iowa can fit many models for contingency tables that are difficult to fit with ML, such as mean response models, global odds ratio models, marginal models for contingency tables.

Stata

- The *ologit* program (`www.stata.com/help.cgi?ologit`) fits cumulative logit models, also using GEE.
- The *oprobit* program (`www.stata.com/help.cgi?oprobit`) fits cumulative probit models.
- Continuation-ratio logit models can be fitted with the *ocratio* module (`www.stata.com/search.cgi?query=ocratio`) and with the *seqlogit* module. The *ocratio* module also fits models with complementary log-log link.
- The GLLAMM module (`www.gllamm.org`) can fit a very wide variety of models, including cumulative logit models with random effects. See `www.stata.com/search.cgi?query=gllamm`.

SPSS

- On ANALYZE menu, choose REGRESSION option and ORDINAL suboption to get ORDINAL REGRESSION menu for fitting cumulative link model. Clicking on *Options*, you can request link functions such as logit, probit, complementary log-log. Clicking on *Output*, you can request test of parallelism (i.e., proportional odds for logit link).
- GENLOG function in SPSS can fit adjacent-categories logit models.
- For GEE methods, on ANALYZE menu, select GENERALIZED LINEAR MODELS option and GENERALIZED ESTIMATING EQUATIONS (GEE) suboption. On GEE window, click on *Repeated* and select form for working correlation model, and click on *Type of Model* to specify model for ordinal logistic or probit response.

Partial Bibliography: Analysis of Ordinal Categorical Data

Some Books

Agresti, A. 2013. *Categorical Data Analysis*, Wiley, 3rd ed.

Agresti, A. 2010. *Analysis of Ordinal Categorical Data*, Wiley, 2nd ed.

Clogg and Shihadeh (1994). *Statistical Models for Ordinal Variables*, Sage.

Greene, W. H., and D. A. Hensher. 2010. *Modeling Ordered Choices*. Cambridge U. Press.

Some Survey Articles

Agresti, A. 1999. Modelling ordered categorical data: Recent advances and future challenges. *Statist. Medic.* **18**: 2191–2207.

Agresti, A., and R. Natarajan. 2001. Modeling clustered ordered categorical data: A survey. *Intern. Statist. Rev.*, 69: 345-371.

Liu, I., and A. Agresti. 2005. The analysis of ordered categorical data: An overview and a survey of recent developments (with discussion). *Test* **14**: 1–73.

McCullagh, P. 1980. Regression models for ordinal data. **42**: *J. Royal. Stat. Society, B*, 109–142.