<u>Topics involved with these notes</u>:

- *Ordinal logistic regression*
- *Generalized linear model with cumulative logit*
- *GzLMM for OLR data*
- *Fitting models with SAS and R*

<u>Associated reading</u>: related topics in the BIOS6643 course notes (see 'Modeling independent or correlated non-normal data' chapter), and in particular the section on ordinal logistic regression.

Ordinal logistic regression

- Sometimes outcomes are not completely continuous and yet are not nominal.

- If there is an intrinsic ordering to the levels but it still makes sense to treat it as more of a categorical variable, then we can use ordinal logistic regression as the modeling tool.

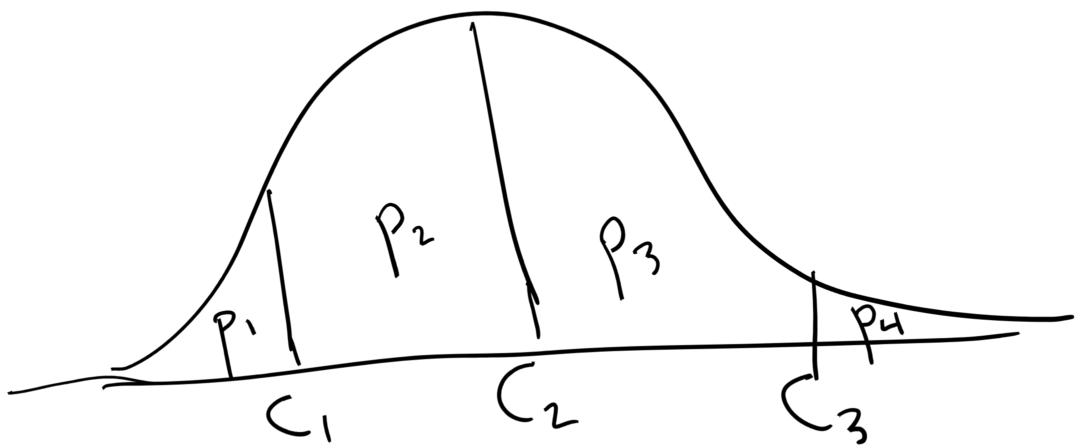
- Sometimes count or continuous variables can be categorized into ordinal levels.

- Ordinal logistic regression is a generalization of standard logistic regression for outcomes that have more than 2 levels.

- In this case, we can use the *cumulative logit* as the link in the generalized linear model.

Cumulative logit model

- Consider the distribution, cut-points and probabilities:



- The cumulative odds for the j^{th} category is

$$\frac{P(Y_i \le C_j)}{P(Y_i > C_j)} = \frac{p_{i1} + p_{i2} + \cdots + p_{ij}}{p_{i,j+1} + \cdots + p_{iJ}}$$

where $p_{ij}=\mathrm{P}(C_{j-1} \le Y_i < C_j)$

- The cumulative logit model is

$$log\left(\frac{p_{i1}+p_{i2}+\cdots+p_{ij}}{p_{i,j+1}+\cdots+p_{iJ}}\right)=\mathbf{x}_{ij}^{r}\boldsymbol{\beta}_j$$

- The proportional odds model is

$$log\left(\frac{p_{i1}+p_{i2}+\cdots+p_{ij}}{p_{i,j+1}+\cdots+p_{iJ}}\right)=\beta_{0j}+\beta_1 x_{i1}+\cdots+\beta_k x_{ik}$$

- In the proportional odds model (commonly used), the Beta's for covariates (other than intercept) do not depend on j. This is a somewhat restrictive assumption but it can be checked.

- A couple of examples

  - In modeling exacerbations, we can use categories of 0 versus 1 to 4 versus 5 or more (or 0 versus 1 or 2 or more). Without other covariates the model is

    0
    1-4
    5 or more

  - For DoD data, grading exposure of military personnel by 'low', 'medium' or 'high' based on job type.

  - If there are no covariates, for these applications, we can use

$$\log\left[\frac{P(Y_i \le C_j)}{P(Y_i > C_j)}\right] = \alpha_j$$

for j=1,2 (2 cut-points, 3 probabilities that sum to 1).

Analysis using exacerbation data

- Understanding exacerbation frequencies for COPD subjects is important in knowing who needs to be treated or more closely monitored.
- Below is the frequency distribution of GOLD 2 subjects in the COPDGene study using baseline data (i.e., total exacerbations in the year prior to baseline).

Exacerbations	Frequency	Percent
0	646	53.52
1-4	365	30.24
5 or more	196	16.24

Modeling the data using SAS

```
proc genmod data=simpler plots=(reschi); where finalgold_baseline=2 and
visitnum=1; class triexac;
 model triexac= / dist=multinomial link=cumlogit pscale; run;
```

```
The GENMOD Procedure
Model Information
Distribution              Multinomial
Link Function             Cumulative Logit
Dependent Variable        triexac
Number of Observations Used     1207

Criteria For Assessing Goodness Of Fit
Criterion     DF   Value          Value/DF
Log Likelihood     -1196.6336
Full Log Likel.    -1196.6336
AIC                2397.2673
```

```
Response Profile
Ordered
Value    triexac  Total Frequency
1        0        646
2        1        365
3        2        196

PROC GENMOD is modeling the
probabilities of levels of triexac
having LOWER Ordered Values in the
response profile table.
```

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	SE	Wald 95% Conf Limits		Wald Chi-Sq	Pr>ChiSq
Intercept1	1	0.1411	0.0577	0.0280	0.2542	5.98	0.0145
Intercept2	1	1.6406	0.0780	1.4876	1.7935	441.87	<.0001
Scale (fixed)	0	1.0000	0.0000	1.0000	1.0000		

0 vs. 1 or more

0-4 vs 5 or more

Approach in R:

Int 1 > 0
1-4
Int 2 > 5+

```
library(MASS)

m=polr(factor(triexac)~1,data=dat,Hess=TRUE)

> summary(m)
Call:
polr(formula = factor(triexac) ~ 1, data = dat, Hess = TRUE)

No coefficients

Intercepts:
    Value   Std. Error t value
0|2  0.1411  0.0577     2.4446
2|3  1.6406  0.0780    21.0207

Residual Deviance: 2393.267
AIC: 2397.267
```

Including FEV1pp (lung function measure) as a predictor:

```
proc genmod data=simpler plots=(reschi);
 where finalgold_baseline=2 and visitnum=1;
 class triexac;
 model triexac=fev1pp/ dist=multinomial link=cumlogit; run;
```

0 vs 1-4 vs 5+

lung function

The GENMOD Procedure

Model Information

Data Set	WORK.SIMPLER
Distribution	Multinomial
Link Function	Cumulative Logit
Dependent Variable	triexac
Number of Obs Used	1207

Class Level Information

Class	Levels	Values
triexac	3	0 1 2

Response Profile

Ordered Value	**triexac**	**Total Freq**
1	0	646
2	1	365
3	2	196

PROC GENMOD is modeling the probabilities of levels of triexac having LOWER Ordered Values in the response profile table. One way to change this to model the probabilities of HIGHER Ordered Values is to specify the DESCENDING option in the PROC statement.

Criteria For Assessing Goodness Of Fit

Criterion	**DF**	**Value**	**Value/DF**
Log Likelihood		-1177.0568	
Full Log Likelihood		-1177.0568	
AIC (smaller is better)		2360.1137	

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	**DF**	**Est.**	**SE**	**Wald 95% Conf Limits**		**Wald Chi-Sq**	**Pr>ChiSq**
Intercept1	1	-2.5421	0.4352	-3.3950	-1.6891	34.12	<.0001
Intercept2	1	-1.0058	0.4307	-1.8499	-0.1617	5.45	0.0195
FEV1pp_utah	1	0.0412	0.0066	0.0281	0.0542	38.44	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

Note: The scale parameter was held fixed.

Mean-correcting data to make intercepts more interpretable

When you have a predictor in the model, the intercepts are the odds of lower versus higher categories (Intercept1 is lowest versus two highest and Intercept2 is lowest and middle versus highest) when the predictors are 0. Since many variables like FEV1 are not often 0, you can mean correct the variable to make the intercepts more interpretable. For this application, the mean FEV1pp is 65, so if we create a new variable, FEV1pp_meancor=FEV1pp_utah-65 and fit this, the intercepts will then be evaluated at the mean level of FEV1pp:

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	SE	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept1	1	0.1332	0.0585	0.0186	0.2479	5.19	0.0227
Intercept2	1	1.6695	0.0790	1.5147	1.8243	446.80	<.0001
fev1pp_meancor	1	0.0412	0.0066	0.0281	0.0542	38.44	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		

The odds of lower versus middle or high exp(0.1332)=1.14 at the average FEV1, while the odds of lower or middle versus high is exp(1.67)=5.31. Note that these are very similar to the model with FEV1.

Interpreting coefficients of predictors in models

By default, ordinal logistic regression model compares lower levels relative to higher ones, but when you add a continuous predictor like FEV1, it is more intuitive to reverse this. For example, we would expect FEV1 and exacerbations to be inversely related, since those with higher lung function would be expected to have fewer exacerbations. You can address this one of 2 ways: (i) for the continuous predictor, just flip the sign on the estimated coefficient and CI endpoints before exponentiating for odds interpretation (also, flip the CI endpoints); (ii) adjust the direction in the software that you are using. For example, in SAS, reverse the direction by including the *desc* option with the outcome [i.e., *y* (*desc*)] or add *descending* in the PROC GENMOD statement [i.e., PROC GENMOD data=dat descending].

As an example: for FEV1 we had 0.0412, 0.0281 and 0.0542 as the mean and CI endpoints. So the modified results would be exp(-0.0412)=0.960 and 95% CI exp(-0.0542)=0.947 to exp(-0.0281)=0.972. You could make this in terms of a 10% FEV1 change (more natural variability) by multiplying before 10 before exponentiating.

Using R:

```
#Ordinal logistic regression
library(MASS)
m2=polr(factor(triexac)~FEV1pp,data=dat,Hess=TRUE)
summary(m2)
```

```
> summary(m2)
Call:
polr(formula = factor(triexac) ~ FEV1pp, data = dat, Hess = TRUE)

Coefficients:
          Value Std. Error t value
FEV1pp -0.04116    0.00664  -6.199

Intercepts:
    Value   Std. Error t value
0|1 -2.5421  0.4352    -5.8411
1|2 -1.0058  0.4307    -2.3351

Residual Deviance: 2354.114
AIC: 2360.114
```

Ordinary logistic regression, adding random effects

Case study:

"Since 2001, 3 million soldiers have deployed to Southwest Asia (SWA), with exposure to inhalants that cause respiratory disease. Department of Defense uses standard occupational codes, termed Military Occupational Specialty (MOS), to classify military personnel by job/training. We characterized Marine MOS by estimated exposure to inhalational hazards. We developed an MOS-exposure matrix containing five major deployment inhalational hazards--sandstorms, burn pits, exhaust fumes, combat dust, occupational VDGF (vapor, dust, gas, fumes)--plus time worked outdoors. A 5 member expert panel of two physician deployment veterans and three occupational pulmonologists independently ranked 38 Marine MOS codes for estimated exposure intensity (3=high, 2=medium, 1=low) to each hazard." From Pepper et al., 2017.

The MOS occupational codes (or MOS_num) are numbered 1 through 38, for convenience, but they relate to specific job types. For example, 1=personnel and administration, 2=intelligence, 3=infantry, etc.

Our data follows this form, for a given inhalation hazard:

	MOS_num					
Rater	#1	#2	#3	#4	#5	…
1	1	1	3	2	1	
2	1	1	3	1	1	
3	1	2	3	2	1	
…						

The outcome is ordinal and given that there are only 3 levels (3 is high exposure, 2 is medium, 1 is low), a 3-level ordinal variable. Each judge rates each MOS_num (job type). Rater and MOS_num can be thought of as randomly sampled from a population (otherwise they could be considered fixed effects). We will add random intercepts for rater and MOS_num; they are *crossed random effects*.

A GzLMM that can be used to fit our data has the form

$$\lambda_{ijk} = \log\left[\frac{P(Y_{ij} \le k \mid b_i, b_j)}{1 - P(Y_{ij} \le k \mid b_i, b_j)}\right] = \alpha_k + b_i + b_j \quad ,$$

where i=MOS_num, j=rater, and k is outcome level; α_k, k=1, 2 are fixed intercepts; b_i and b_j are random intercepts for MOS_num and rater, respectively. $b_i \sim N(0, \sigma^2_{job})$ and , $b_j \sim N(0, \sigma^2_{rater})$.

Some questions of interest for our data:

(1) How do variances for raters compare with the variances over MOS types?
(2) Are there any raters that significantly differ from the group average?
(3) After adjusting for crossed random effects of MOS type and rater, what are the cumulative odds of low, medium, high exposure for a given inhalation hazard?
(4) What is the probability of a particular job of having a high exposure to a given exposure type?

To answer these questions, we can fit the ordinal logistic regression model shown on the last slide that accounts for multiple measures per MOS type (called *MOS_num* below), which is the experimental unit here (instead of subjects).

Descriptive approach to obtaining probabilities and odds ratios

- First, to get an understanding of the statistics we're dealing with, let's consider the data more descriptively.

- In the data, we have 115 MOS's assigned as 'low exposure' job types (62.5%), 56 as 'medium' (30.4%) and 13 as 'high' (7.1%).

- Without considering the correlation, the odds of a low classification for a randomly selected MOS is 0.625/(1–0.625)=1.67 and the odds of a low or medium classification is (0.929)/(1–0.929) = 13.08.

- When we fit the model, we account for the fact that rater's score every MOS; i.e., the random effects are crossed.

SAS Code for one inhalation exposure source, burn pits:

```
proc glimmix data=all2 method=laplace;
 class mos_num rater;
 model burn_pits=
   / solution dist=multinomial link=cumlogit;
   random mos_num rater / solution;  run;
```

Note that in the model statement, there are no effects; thus we only have the intercepts. Since there are three levels of the outome, we'll have 2 intercepts. Covariates can be included but here we just include the fixed intercepts and random intercepts for rater and MOS_num.

```
The GLIMMIX Procedure

Model Information
Response Variable            Burn_Pits
Response Distribution              Multinomial (ordered)
Link Function                      Cumulative Logit
Variance Function            Default
Estimation Technique               Maximum Likelihood
Likelihood Approximation           Laplace
Degrees of Freedom Method          Containment
Number of Observations Used        184


Response Profile

Ordered Value       Burn_Pits    n
1            1           115
2            2           56
3            3           13

The GLIMMIX procedure is modeling the probabilities of levels of Burn_Pits having lowered ordered values in the
Response Profile table.
```

The Laplace method approximates the true likelihood, and hence considered ML estimation.

The intercept for Burn_Pits=2 means that the associated odds ratio will be for level 1, relative to 2 or 3; the intercept for Burn_Pits=2 compares 1 or 2 versus 3.

Covariance Parameter Estimates

Cov Parm	Estimate	SE
MOS_num	2.9181	1.2889
rater	0.7259	0.6157

The variance estimates indicate that the variability of the exposure estimates among job types (MOS_num) is 4 times greater than for the raters, which is probably reassuring to the raters.

Solutions for Fixed Effects

Effect	Burn_Pits	Estimate	SE	DF	t Value	Pr > |t|
Intercept	**1**	0.7868	0.5219	4	1.51	0.2062
Intercept	**2**	3.8512	0.6760	4	5.70	0.0047

The odds of a rater ascribing a job type as having low exposure (relative to medium or high) is exp(0.7868)=2.20; the odds of low or medium versus high is exp(3.8512)=47.05

Solution for Random Effects

Effect	rater	MOS_num	Estimate	Std Err Pred	DF	t Value	Pr > |t|
MOS_num		**1**	0.5945	0.9462	142	0.63	0.5308
MOS_num		**2**	-0.1359	0.8699	142	-0.16	0.8761
MOS_num		**3**	-3.2523	1.0217	142	-3.18	0.0018
…							
MOS_num		**73**	-0.09849	0.8591	142	-0.11	0.9089
rater	**Gottschall**		1.1538	0.5745	142	2.01	0.0465
rater	**Kreft**		-0.3367	0.4953	142	-0.68	0.4977
rater	**Meehan**		-0.4260	0.4993	142	-0.85	0.3951
rater	**Pepper**		-0.9793	0.5202	142	-1.88	0.0618
rater	**Rose**		0.08430	0.4930	142	0.17	0.8645

MOS_num 3 is Infantry, thus we'd expect higher exposure rating. The signs appear flipped here since we're modeling probability of lower rating. For raters, Gottschall tends to rate lower exposure, while Pepper tends to rate higher exposure. A way to model higher exposure would be to add *(desc)* after the outcome in the model statement.

Using R:

r.e. for job type

r.e. for rater

```
library(ordinal)
results <- clmm(factor(Burn_Pits)~(1|MOS_num)+(1|rater), data = dat)
summary(results)

> summary(fmm1)
Cumulative Link Mixed Model fitted with the Laplace approximation

formula: factor(Burn_Pits) ~ (1 | MOS_num) + (1 | rater)
data:    dat

 link  threshold nobs logLik  AIC    niter    max.grad cond.H
 logit flexible  184  -139.39 286.77 114(460) 3.18e-06 1.1e+01

Random effects:
 Groups  Name        Variance Std.Dev.
 MOS_num (Intercept) 2.9181   1.708
 rater   (Intercept) 0.7259   0.852
Number of groups:  MOS_num 37,  rater 5

No Coefficients

Threshold coefficients:
    Estimate Std. Error z value
1|2   0.7868     0.5219   1.508
2|3   3.8512     0.6760   5.697
```

- The ordinal logistic regression model here is $P\left(Y_{ij} \le k \middle| b_i, b_j\right) = \frac{1}{1+e^{-\lambda_{ijk}}}$, and so $P\left(Y_{ij} \le k \middle| b_i = 0, b_j = 0\right) = \frac{1}{1+e^{-\alpha_k}}$. From the latter, we can estimate that for an average rater and MOS_num, the probability of low classification (k=1) as $1/(1+e^{-0.7868}) = 68.7\%$.

- Job and rater-specific probability estimates can be obtained by using the first formula. We can also compute for specific MOS_num or raters, holding the other at its mean, since random effects are crossed. For example, for an average MOS_num the probability of a low classification for Gottschall is $1/(1+e^{-(0.7868+1.15)}) = 87.4\%$, while for Pepper it is $1/(1+e^{-(0.7868-0.98)}) = 45.2\%$.

- We can get probabilities for any given level by computing the cumulative probabilities, and then taking differences [e.g., $P(Y=3)=P(Y\le3) - P(Y\le2)$.]

Using the mixed-effects ordinal logistic regression for longitudinal data

- We can generalize the formula for the mixed-effects ordinal logistic regression model so that it can be used for clustered / longitudinal data and include covariates. One such model that is useful for repeated measures within subjects (or subjects within clusters) is

$$\lambda_{ijk} = \log\left[\frac{P(Y_{ij} \le k \mid \mathbf{b}_i)}{1 - P(Y_{ij} \le k \mid \mathbf{b}_i)}\right] = \alpha_k + \mathbf{x}_{ij}^{r}\boldsymbol{\beta} + \mathbf{z}_{ij}^{r}\mathbf{b}_i$$

where *i* denotes subject, with measure *j* (or subject *j* in cluster *i*). Here, we have hierarchical data and so the random effects (as is usually done) are defined for the level 2 data (subjects).

- The previous model can be used for longitudinal ordinal logistic regression, although we only account for repeated measures via random effects. (Using pseudo-likelihood methods, you could consider models that account for random effects or serial correlation, or both.)

- Now we have what is called a proportional odds model (see McCullagh, 1980) that results from the fact that the relationship between the cumulative logit and the predictors does not depend on k.

- For example, say that the previous case study also had measurements over time (x=time). If we added this as a predictor, then the cumulative logits (and hence probabilities) would not change over time.

- We can generalize the model slightly so that for certain predictors, we do not require the proportional odds assumption.

- For example, Hedeker and Mermelstein (1998, 200) suggest the model

$$\lambda_{ijk} = \log\left[\frac{P(Y_{ij} \le k \mid \mathbf{b}_i)}{1 - P(Y_{ij} \le k \mid \mathbf{b}_i)}\right] = \alpha_k + \mathbf{x}_{ij}^{r}\boldsymbol{\beta} + \mathbf{s}_{ij}^{r}\boldsymbol{\gamma}_k + \mathbf{z}_{ij}^{r}\mathbf{b}_i$$

where the additional term involving $\boldsymbol{\gamma}_k$ allows the effects for the associated covariates to vary across the cumulative logits.

- For more detail, see the above references or Hedeker and Gibbons (2006). Hedeker does warn about use of this partial proportional odds model, with respect to inference for certain values of the covariates. For more detail, see Hedeker and Gibbons (2006).