Modeling Ordered Categorical Data

Lin Chen

Department of Public Health Sciences
The University of Chicago

Ordered Categorical Data

- An ordered categorical variable (also called an ordinal variable) is a categorical variable (with more than two levels) where there is a natural ordering of the categories.
- Examples:
  1. In a clinical trial on pain relievers, the degree of pain control may be described as totally ineffective, poor control, moderate control, or good control.
  2. Stage of cancer are ordered by extent of disease: stage I (localized), II, III, IV (metastatic).
  3. Agreement level on a survey question: strongly disagree, disagree, neutral, agree, strongly agree.
- The quantitative distance between levels may not be known and may not be the same. increasing order

Ordinal Response: notation

- Let $C_1, C_2, ..., C_k,\ k \geq 2$ denote the k ordered categories for the response (in increasing order). # of categories
- Let Y_i be the response variable for the i^{th} individual, with Y_i taking the value j if the response is in category $C_j,\ j = 1, 2, ..., k$.
  category
- Define $p_{ij} = P(Y_i = j) = P[$*individual* i *responds in category* $C_j]$
- The *cumulative probability* for Y_i is denoted as $\gamma_{ij} = P(Y_i \leq j)$ Hence $\gamma_{ij} = p_{i1} + p_{i2} + \cdots + p_{ij}$ and $\gamma_{ik} = \sum_{j=1}^{k} p_{ij} = 1$
- We now introduce an *unobservable / latent* continuous random variable Z_i which is such that

$$Y_i = j,\ \textit{if}\ d_{j-1} < Z_i \leq d_j$$

where $-\infty = d_0 < d_1 < \cdots < d_k = \infty$. We refer to $d_1, d_2, ..., d_{k-1}$ as the *cut points*. Response threshold

Latent Variable Idea

- Thus, $\gamma_{ij} = P(Y_i \le j) = P(Z_i \le d_j),\ \ j = 1, 2, ..., k$
- Assume that Z_i have a logistic distribution with mean μ_i and unit standard deviation, then

cumulative probabilities

$$\gamma_{ij} = P(Z_i \le d_j) = \frac{e^{d_j - \mu_i}}{1 + e^{d_j - \mu_i}} \tag{1}$$

It then follows that

$$\log(\frac{\gamma_{ij}}{1-\gamma_{ij}}) = d_j - \mu_i \tag{2}$$

We further suppose that μ_i is a linear combination of the explanatory variables for the i^{th} individual, and set $\mu_i = \boldsymbol{\beta}' \boldsymbol{x_i}$. The ordered logit model assuming proportional odds is given by

Assuming same quantitative distance

$$\log(\frac{\gamma_{ij}}{1-\gamma_{ij}}) = d_j - \boldsymbol{\beta}' \boldsymbol{x_i} \tag{3}$$

Inversely affects the probability of achieving better outcomes

log(cumulative probabilities)

↑ odds ↓ odds

- Why negative β? Response categories are ordered from worst to best outcomes. ↑ in explanatory variables : ↓ in log-odds in a higher category v.s. lower category
  ex) : death for stages of cancer is ordered from End Stage ~ Stage I.

Proportional Odds Model: $\log(\frac{\gamma_{ij}}{1-\gamma_{ij}}) = d_j \boxed{-\boldsymbol{\beta}' \boldsymbol{x_i}}$

- This model uses cumulative probabilities up to a response threshold, thereby making the whole range of ordinal categories **binary** at that threshold.
- Intercept d_j is the log-odds of falling into or below category j when $\boldsymbol{x}_i = \boldsymbol{0}$
- Two individuals with covariates $\boldsymbol{x_1}$, and $\boldsymbol{x_2}$ respectively, $\log\left(\frac{\gamma_{1j}}{1-\gamma_{1j}}\right) = d_j - \boldsymbol{\beta}'\boldsymbol{x_1}$, $\log\left(\frac{\gamma_{2j}}{1-\gamma_{2j}}\right) = d_j - \boldsymbol{\beta}'\boldsymbol{x_2}$, we have

$$\log\left(\frac{\gamma_{1j}/(1-\gamma_{1j})}{\gamma_{2j}/(1-\gamma_{2j})}\right) = -\boldsymbol{\beta}'(\boldsymbol{x_1} - \boldsymbol{x_2}) \tag{4}$$

- The log odds ratio $-\boldsymbol{\beta}'(\boldsymbol{x_1} - \boldsymbol{x_2})$ in (4) does not depend on the category j. The log odds ratio of being in category C_j or worse is proportional to the difference between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ where $-\boldsymbol{\beta}$ is the constant of proportionality (same for $j = 1, 2, ..., k-1$). The model is a "proportional odds model".
- When $k = 2$, the model is $\log(\frac{\gamma_{i1}}{1-\gamma_{i1}}) = d_1 - \boldsymbol{\beta}'\boldsymbol{x_i}$, where $\gamma_{i1} = p_{i1}$ with only one cut point. This reduces to the standard logistic model for binary data.

Example: *Small Cell Lung Cancer*

In a clinical trial evaluating treatment of small cell lung cancer described by Holtbrugge and Schumacher (1991), two treatment strategies were compared: *sequential therapy* (same combination of chemotherapeutic agents were administered in each treatment cycle) vs. *alternating therapy* (three different combinations were given, alternating between cycles). Data were obtained from 299 patients.

Table 1: Tumor response by sex and chemotherapy strategy.

worst — best

Sex of patient	Therapy strategy	Progressive disease (PD)	Stable disease (SD) (no change)	Partial remission (PR)	Complete remission (CR)
0(Male)	0(Sequential)	28	45	29	26
1(Female)	0	4	12	5	2
0	1(Alternating)	41	44	20	20
1	1	12	7	3	1
		85	108	57	49

\+ cancer — − cancer

↑ odds of outcome — ↓ odds of outcome

Response is naturally ordered from worst to best.

Lung Cancer: dataset organization for ordered analysis (wide form)

```
. use "Small_cell_lung_cancer_wide.dta", clear

. list

     +------------------------------------------------------+
     |  sex   therapy   count1   count2   count3   count4   |
     |------------------------------------------------------|
  1. |    0         0       28       45       29       26   |
  2. |    0         1       41       44       20       20   |
  3. |    1         0        4       12        5        2   |
  4. |    1         1       12        7        3        1   |
     +------------------------------------------------------+
```

Each row with multiple observations

Lung Cancer: dataset organization for ordered analysis (wide $\longrightarrow$ long form)

- The analysis commands require “long form", so we reshape the data to long form.

```
. reshape long count, i(sex therapy) j(category)
(note: j = 1 2 3 4)

Data                                wide   ->   long
-----------------------------------------------------------------------------
Number of obs.                         4   ->     16
Number of variables                    6   ->      4
j variable (4 values)                      ->   category
xij variables:
               count1 count2 ... count4   ->   count
-----------------------------------------------------------------------------
```

Lung Cancer: dataset organization for ordered analysis (long form)

The data is in long form now and is ready for analysis.

```
. list
```

	sex	therapy	category	count
1.	0	0	1	28
2.	0	0	2	45
3.	0	0	3	29
4.	0	0	4	26
5.	0	1	1	41
6.	0	1	2	44
7.	0	1	3	20
8.	0	1	4	20
9.	1	0	1	4
10.	1	0	2	12
11.	1	0	3	5
12.	1	0	4	2
13.	1	1	1	12
14.	1	1	2	7
15.	1	1	3	3
16.	1	1	4	1

Each row with a single observation

Lung Cancer: dataset organization for ordered analysis (long ⟶ wide form)

What if we want to change it back to wide format?

```
. reshape wide count, i(sex therapy) j(category)
(note: j = 1 2 3 4)

Data                                   long   ->   wide
-----------------------------------------------------------------------------
Number of obs.                           16   ->      4
Number of variables                       4   ->      6
j variable (4 values)              category   ->   (dropped)
xij variables:
                                      count   ->   count1 count2 ... count4
-----------------------------------------------------------------------------

. list

     +-----------------------------------------------------+
     | sex   therapy   count1   count2   count3   count4 |
     |-----------------------------------------------------|
  1. |   0         0       28       45       29       26 |
  2. |   0         1       41       44       20       20 |
  3. |   1         0        4       12        5        2 |
  4. |   1         1       12        7        3        1 |
     +-----------------------------------------------------+
```

This matches the table data summary.

Lung Cancer: $\log\frac{\gamma_{ij}}{1-\gamma_{ij}} = d_j, \ j = 1,2,3$

cutpoint/threshold: $\boxed{j-1}$

Next, take the long data form and run an ordered logit model without predictors (the null model, with only intercepts for each cutoff j). The Stata function for ordered logit model is `ologit`.

```
. reshape long count, i(sex therapy) j(category)
. ologit category [fweight=count], nolog

Ordered logistic regression                     Number of obs     =        299
Log likelihood = -399.98398                     Pseudo R2         =     0.0000

------------------------------------------------------------------------------
    category |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       /cut1 |  -.9233248   .1282092                      -1.17461   -.6720393
       /cut2 |   .5992511   .1208938                      .3623036    .8361986
       /cut3 |   1.629641   .1562311                      1.323433    1.935848
------------------------------------------------------------------------------

. estimates store Null
```

"j" "frequency weight"

The *estimates store filename* **command provides storage of model info for contrasting later**

Lung Cancer: $\log\frac{\gamma_{ij}}{1-\gamma_{ij}} = d_j,\ j = 1,2,3$

What do these parameters represent?

Recall Table 1.

Sex of patient	Therapy strategy	Progressive disease (PD) $j=1$	Stable disease (SD) (no change) $j=2$	Partial remission (PR) $j=3$	Complete remission (CR)
0(Male)	0(Sequential)	28	45	29	26
1(Female)	0	4	12	5	2
0	1(Alternating)	41	44	20	20
1	1	12	7	3	1
		85	108	57	49

why no cut 4? Assumed any value above cut 3 = 4^{th} category

Param	estimating what	raw odds	log odds	Cumul. prob.	Prob.
cut1	log(odds PD/higher response category) PD v.s. SD,PR,CR	0.397 $= \frac{85}{108+57+49}$ v.s.	-0.9233 =log(0.397)	.28 $= \frac{0.397}{1+0.397}$	.28
cut2	log(odds PD or SD/higher category) PD,SD v.s. PR,CR	1.821 $= \frac{85+108}{57+49}$	0.5992 $= \log(1.821)$	.65 includes cut 1 $= \frac{1.821}{1+1.821}$	.37 =.65-.28
cut3	log(odds PD or SD or PR/higher category) PD,SD,PR v.s. CR	5.10 $= \frac{85+108+57}{49}$	1.6296 $= \log(5.10)$	.84 $= \frac{5.10}{1+5.10}$	.19=.84-.65

Model predicts odds (probability) of **being in a given category or lower** (worse) **vs higher categories**

Lung Cancer: $\log\frac{\gamma_{ij}}{1-\gamma_{ij}} = d_j - \beta_1 sex,\ j = 1,2,3$

Next we consider sex as a predictor in the model:

```
. ologit category sex [fweight=count], nolog

Ordered logistic regression                       Number of obs     =        299
                                                  LR chi2(1)        =       3.34
                                                  Prob > chi2       =     0.0676
Log likelihood = -398.31341                       Pseudo R2         =     0.0042

------------------------------------------------------------------------------
    category |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |  -.5218702     .28707    -1.82   0.069    -1.084517    .0407767
-------------+----------------------------------------------------------------
       /cut1 |   -1.01504    .138661                     -1.286811   -.7432694
       /cut2 |   .5188191   .1285166                      .2669312     .770707
       /cut3 |   1.557141   .1609139                      1.241756    1.872527
------------------------------------------------------------------------------

. estimates store S
```

Lung Cancer: $\log\frac{\gamma_{ij}}{1-\gamma_{ij}} = d_j - \beta_1$ *therapy*, $j = 1,2,3$

We now consider the predictor of interest, therapy type (sequential versus alternating):

```
. ologit category therapy [fweight=count], nolog

Ordered logistic regression                       Number of obs   =        299
                                                  LR chi2(1)      =       7.31
                                                  Prob > chi2     =     0.0068
Log likelihood = -396.32657                       Pseudo R2       =     0.0091

------------------------------------------------------------------------------
    category |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     therapy |  -.5699142   .2117716    -2.69   0.007    -.9849789   -.1548495
-------------+----------------------------------------------------------------
       /cut1 |   -1.21673   .1704333                     -1.550773   -.8826866
       /cut2 |   .3382206   .1542139                       .035967    .6404743
       /cut3 |   1.380296   .1801627                      1.027184    1.733409
------------------------------------------------------------------------------

. estimates store T
```

What does this model say?

Lung Cancer: $\log\frac{\gamma_{ij}}{1-\gamma_{ij}} = d_j - \beta_1$ *therapy*, $j = 1,2,3$

Let's focus on the comparison of PR,SD or PR verus CR. That is, `/cut3`.

j=4

```
. tab therapy category [fweight=count], row col
```

therapy	1	2	3	4	Total
0	32	57	34	28	151
	21.19	37.75	22.52	18.54	100.00
	37.65	52.78	59.65	57.14	50.50
1	53	51	23	21	148
	35.81	34.46	15.54	14.19	100.00
	62.35	47.22	40.35	42.86	49.50
Total	85	108	57	49	299
	28.43	36.12	19.06	16.39	100.00
	100.00	100.00	100.00	100.00	100.00

cut 3

- For therapy 0, odds of progression, stable, or partial response versus complete response is $123/28 = 4.39$. 32+57+34/28 For therapy 1, odds of progression, stable, or partial response versus complete response is $6.04$5. 53+51+23/21
- The proportional odds model predicts $\exp(1.38) = 3.97$ for therapy 0, and $\exp(\overset{d}{1.38} \overset{-}{-} (\overset{\beta_1}{-.570})) = 7.02$ for therapy 1. Its estimates differ from the estimates based on logistic regression considering only cut 3.

Don't forget (−) not (+)

Lung Cancer: $\log\frac{\gamma_{ij}}{1-\gamma_{ij}} = d_j - \beta_1 \textit{sex} - \beta_2 \textit{therapy},\ j = 1,2,3$

Let's consider more models. This one includes therapy as a predictor and also adjusts for sex.

X_1 X_2

```
. ologit category sex therapy [fweight=count], nolog

Ordered logistic regression                     Number of obs   =        299
                                                LR chi2(2)      =      10.91
                                                Prob > chi2     =     0.0043
Log likelihood = -394.52832                     Pseudo R2       =     0.0136

------------------------------------------------------------------------------
    category |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |  -.5413938   .2871816    -1.89   0.059    -1.104259    .0214717
     therapy |   -.580685   .2121478    -2.74   0.006     -.996487    -.164883
-------------+----------------------------------------------------------------
       /cut1 |  -1.318043   .1797769                     -1.670399   -.9656869
       /cut2 |   .2492335   .1613881                     -.0670813    .5655484
       /cut3 |   1.300056   .1849928                      .9374766    1.662635
------------------------------------------------------------------------------

. estimates store ST
```

Lung Cancer: $\log \frac{\gamma_{ij}}{1-\gamma_{ij}} = d_j - \beta_1 \textit{sex} - \beta_2 \textit{therapy} - \beta_3 \textit{sex} \times \textit{therapy}, \;\; j = 1, 2, 3$

One last sex-by-therapy interaction model:

```
. gen st = sex*therapy
. ologit category sex therapy st [fweight=count], nolog
```

x_1 x_2 x_1x_2

```
Ordered logistic regression                       Number of obs   =        299
                                                  LR chi2(3)      =      11.96
                                                  Prob > chi2     =     0.0075
Log likelihood = -394.00492                       Pseudo R2       =     0.0149
```

category	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
sex	-.2741906	.3873497	-0.71	0.479	-1.033382	.4850008
therapy	-.488071	.2305167	-2.12	0.034	-.9398754	-.0362666
st	-.5904159	.5791605	-1.02	0.308	-1.72555	.5447177
/cut1	-1.275657	.184367			-1.63701	-.9143045
/cut2	.2957159	.1678283			-.0332216	.6246534
/cut3	1.345164	.1905977			.9715991	1.718728

```
. estimates store SXT
```

Lung Cancer: Model Comparison via LR Tests

These tests are nested, and we can use likelihood ratio tests `lrtest` to compare them.

```
. lrtest S Null
Likelihood-ratio test                          LR chi2(1)  =    3.34
(Assumption: Null nested in S)                 Prob > chi2 =  0.0676

. lrtest T Null
Likelihood-ratio test                          LR chi2(1)  =    7.31
(Assumption: Null nested in T)                 Prob > chi2 =  0.0068

. lrtest ST T
Likelihood-ratio test                          LR chi2(1)  =    3.60
(Assumption: T nested in ST)                   Prob > chi2 =  0.0579

. lrtest ST S
Likelihood-ratio test                          LR chi2(1)  =    7.57
(Assumption: S nested in ST)                   Prob > chi2 =  0.0059

. lrtest SXT ST
Likelihood-ratio test                          LR chi2(1)  =    1.05
(Assumption: ST nested in SXT)                 Prob > chi2 =  0.3062
```

Adding S — ~0.05 — keep S

Adding T — keep T

Adding S — ~0.05 — keep S

Adding T — keep T

Adding S×T — omit interaction — no diff. from "perfect"

Based on these proportional odds models, we conclude that both sex and therapy affect tumor response, but there is not evidence that the interaction between sex and therapy is an important predictor for tumor response. We choose Model ST. ✓

Lung Cancer: Writing the Fitted Model

```
                             S     T
. ologit category sex therapy [fweight=count], nolog
------------------------------------------------------------------------------
    category |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |  -.5413938   .2871816    -1.89   0.059    -1.104259    .0214717
     therapy |   -.580685   .2121478    -2.74   0.006    -.996487    -.164883
-------------+----------------------------------------------------------------
       /cut1 |  -1.318043   .1797769                     -1.670399   -.9656869
       /cut2 |   .2492335   .1613881                     -.0670813    .5655484
       /cut3 |   1.300056   .1849928                      .9374766    1.662635
------------------------------------------------------------------------------
```

- The fitted model based on $\log\frac{\gamma_{ij}}{1-\gamma_{ij}} = d_j - \beta_1 \textit{sex} - \beta_2 \textit{therapy}$, $j = 1, 2, 3$ can be written as:

$$j=1 \qquad \log\frac{\hat{\gamma}_{i1}}{1-\hat{\gamma}_{i1}} = -1.318 + \overset{-(-0.541)}{0.541} \cdot \textit{sex} + \overset{-(-0.581)}{0.581} \cdot \textit{therapy}$$

$$j=2 \qquad \log\frac{\hat{\gamma}_{i2}}{1-\hat{\gamma}_{i2}} = .249 + 0.541 \cdot \textit{sex} + 0.581 \cdot \textit{therapy}$$

$$j=3 \qquad \log\frac{\hat{\gamma}_{i3}}{1-\hat{\gamma}_{i3}} = 1.300 + 0.541 \cdot \textit{sex} + 0.581 \cdot \textit{therapy}$$

constant slope

Lung Cancer: $\log\frac{\gamma_{ij}}{1-\gamma_{ij}} = d_j - \beta_1 \textit{sex} - \beta_2 \textit{therapy},\ j = 1,2,3$ Interpreting the estimated coefficients

The parameter estimates $\hat{d}_j, j = 1,2,3$ are **the estimated log odds of falling into or below category j** when Sex=0 (male) and Therapy = 0 (sequential therapy)

The parameter estimates $\hat{\beta}_1, \hat{\beta}_2$ associated with *Sex* and *Therapy* can be interpreted in terms of **log odds ratios**.

Example: $\overset{\text{worse}}{-\hat{\beta}_2}(= 0.5806..)$ gives the estimated log odds ratio of the probability of a response in category C_j or worse, j = 1,2,3, comparing the alternating therapy (therapy = 1) with sequential therapy (therapy = 0) adjusting for sex. Odds ratio is 1.79 $=\exp(-\hat{\beta}_2)$ (so alternating therapy is a bit worse). Interpret in $-\beta$

negative

Lung Cancer: Predicted Category Probabilities

We would like to have the predicted probabilities for each category under each condition. The results are based on the additive model $\log\frac{\gamma_{ij}}{1-\gamma_{ij}} = d_j - \beta_1 \textit{sex} - \beta_2 \textit{therapy},\ j = 1,2,3.$

"`f(%5.2f)`" is for formatting and rounding the numbers.

without displaying output

```
. quietly ologit category sex therapy [fweight=count], nolog
. predict p1 p2 p3 p4
(option pr assumed; predicted probabilities)
. table sex therapy, c(mean p1 mean p2 mean p3 mean p4) f(%5.2f)

----------------------
          |   therapy
      sex |    0      1
----------+-----------
        0 | 0.21   0.32
          | 0.35   0.37
          | 0.22   0.17
          | 0.21   0.13
          |
        1 | 0.32   0.45
          | 0.37   0.35
          | 0.18   0.12
          | 0.14   0.08
----------------------
```

Lung Cancer: Predicted Cumulative Probabilities

```
. gen cp2 = p1+p2

. gen cp3 = cp2+p3

. gen cp4 = cp3+p4

. table sex therapy, c(mean p1 mean cp2 mean cp3 mean cp4) f(%5.2f)

----------------------
          |  therapy
      sex |     0      1
----------+-----------
        0 |  0.21   0.32
          |  0.56   0.70
          |  0.79   0.87
          |  1.00   1.00
          |
        1 |  0.32   0.45
          |  0.69   0.80
          |  0.86   0.92
          |  1.00   1.00
----------------------
```

ex): $\underbrace{0.21 + 0.35}_{p_1 + p_2} = 0.56$

Lung Cancer: Predicted Probabilities

Note: the *lincom* command can also be used to make predictions for specific covariate combinations (pay attention to - signs) with precision estimate. The `_b[varName]` function gives the coefficient estimate of the variable.

```
                                          Cut 1                         sex=1                     therapy=1
. * estimate probability for PD vs better for female on alternating therapy
. lincom  _b[/cut1] - 1*sex - 1*therapy

 ( 1)  - [category]sex - [category]therapy + [cut1]_cons = 0
------------------------------------------------------------------------------
    category |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |  -.1959642   .2892904    -0.68   0.498    -.7629629    .3710345
------------------------------------------------------------------------------

. display invlogit(-.1959642)
.45116513
```

$$\frac{e^{-0.196}}{1+e^{-0.196}} = 0.451$$

```
. * lower bound
. display invlogit( -.7629629)
.31800333

. * upper bound
. display invlogit(  .3710345)
.59170893
```

Testing the Proportional Odds Assumption

- The final model from the proportional odds model

$$\log \frac{\gamma_{ij}}{1-\gamma_{ij}} = d_j - \beta_1 \textit{sex} - \beta_2 \textit{therapy}, \;\; j = 1,2,3 \tag{5}$$

assumes the effect of therapy and of sex do not depend on which cutpoint between response categories we are considering.

$$\log \frac{\hat{\gamma}_{i1}}{1-\hat{\gamma}_{i1}} = -1.318 + 0.541 \cdot \textit{sex} + 0.581 \cdot \textit{therapy}$$

$$\log \frac{\hat{\gamma}_{i2}}{1-\hat{\gamma}_{i2}} = 0.249 + 0.541 \cdot \textit{sex} + 0.581 \cdot \textit{therapy}$$

$$\log \frac{\hat{\gamma}_{i3}}{1-\hat{\gamma}_{i3}} = 1.300 + 0.541 \cdot \textit{sex} + 0.581 \cdot \textit{therapy}$$

Testing the proportional odds assumption (cont.)

- We can test this proportional odds assumption by allowing for a different covariate effect at each cutpoint, which motivates the following generalized ordered logit model.
- *Generalized ordered logit model*:

$$\boxed{\log \frac{\gamma_{ij}}{1-\gamma_{ij}}} = d_j - \beta_{j1} \cdot \textit{sex} - \beta_{j2} \cdot \textit{therapy}, \quad j = 1,2,3 \tag{6}$$

Predictor now depends on category

- The proportional odds model/ordered logit model is *nested within* the generalized ordered logit model. Under the null,

$$H_0 : \beta_{11} = \beta_{21} = \beta_{31} \textit{ and } \beta_{12} = \beta_{22} = \beta_{32}$$

so we can use difference in deviance, i.e., the likelihood-ratio test to perform model comparison to check this assumption.

The generalized ordered logit model: (continued)

- In Stata, we need a user-written program (gologit or gologit2) to perform the analysis using this model. You may use "`ssc install`"
- In contrast to the way that the proportional odds model is parameterized as in **ologit**, the generalized ordered logit model is parameterized in **gologit2** as follows:

$$\boxed{\log \frac{1-\gamma_{ij}}{\gamma_{ij}}} = d_j + \beta_{j1} \cdot \mathit{sex} + \beta_{j2} \cdot \mathit{therapy}, \;\; j = 1, 2, 3 \tag{7}$$

- In Stata, due to different model parameterization, intercept estimates d_j from **gologit2** will be of opposite sign as cuts from proportional odds model from **ologit**

Lung Cancer: gologit2 model

```
. ssc install gologit2
. gologit2 category sex therapy [fweight=count]

Generalized Ordered Logit Estimates               Number of obs   =        299
                                                  LR chi2(6)      =      14.10
                                                  Prob > chi2     =     0.0285
Log likelihood = -392.93348                       Pseudo R2       =     0.0176

------------------------------------------------------------------------------
    category |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
1            |
         sex |  -.3645641   .3443519    -1.06   0.290    -1.039481    .3103533
     therapy |  -.7317005   .2633247    -2.78   0.005    -1.247807   -.2155936
       _cons |   1.373901   .2091236     6.57   0.000     .9640261    1.783775
-------------+----------------------------------------------------------------
2            |
         sex |  -.6700663   .3720467    -1.80   0.072    -1.399264    .0591318
     therapy |  -.5229065   .2458232    -2.13   0.033    -1.004711   -.0411019
       _cons |  -.2566465   .1745179    -1.47   0.141    -.5986952    .0854023
-------------+----------------------------------------------------------------
3            |
         sex |  -1.171411    .619793    -1.89   0.059    -2.386183     .043361
     therapy |  -.3439213   .3156354    -1.09   0.276    -.9625554    .2747128
       _cons |  -1.341935   .2158886    -6.22   0.000    -1.765069   -.9188013
------------------------------------------------------------------------------
```

Lung Cancer: gologit2 model vs. ologit model

- The fitted generalized ordered logit model gologit2
  ($\log\frac{\gamma_{ij}}{1-\gamma_{ij}} = -d_j - \beta_{j1}\,sex - \beta_{j2}\,therapy,\ \ j = 1,2,3$)
  flipped sign to get − for parameters

$$\log\frac{\hat{\gamma}_{i1}}{1-\hat{\gamma}_{i1}} = -1.374 + 0.365\cdot sex + 0.732\cdot therapy$$

$$\log\frac{\hat{\gamma}_{i2}}{1-\hat{\gamma}_{i2}} = 0.257 + 0.670\cdot sex + 0.523\cdot therapy$$

$$\log\frac{\hat{\gamma}_{i3}}{1-\hat{\gamma}_{i3}} = 1.342 + 1.171\cdot sex + 0.344\cdot therapy$$

- The previous fitted proportional odds model ologit
  ($\log\frac{\gamma_{ij}}{1-\gamma_{ij}} = d'_j - \beta'_1\,sex - \beta'_2\,therapy,\ \ j = 1,2,3$)

$$\log\frac{\hat{\gamma}_{i1}}{1-\hat{\gamma}_{i1}} = -1.318 + 0.541\cdot sex + 0.581\cdot therapy$$

$$\log\frac{\hat{\gamma}_{i2}}{1-\hat{\gamma}_{i2}} = .249 + 0.541\cdot sex + 0.581\cdot therapy$$

$$\log\frac{\hat{\gamma}_{i3}}{1-\hat{\gamma}_{i3}} = 1.300 + 0.541\cdot sex + 0.581\cdot therapy$$

Testing the proportional odds assumption: *Lung Cancer data*

- Log likelihood = -394.52832 under the proportional odds ologit model/ordered logit model
- Log Likelihood = -392.93348 under the generalized ordered logit model glogit2
- The two models are nested, we could calculate **the likelihood ratio test statistic**

$$\Lambda = -2(\ell_{\text{ordered}}^{\text{current}} - \ell_{\text{general}}^{\text{full}}) \sim \chi^2_{\text{df}} \text{ under } H_0$$

Here $LR = -2 \times (-394.52832 - (-392.93348)) = 3.18968$ with 4 d.f. This is not close to statistical significance. good fit for current model

- This comparison suggests that the proportional odds assumption is plausible for the *small cell lung cancer* study.

Summary

Ordinal Logistic Regression

- When outcome variable has multiple ordered categories, ordinal logit model (**ordinal logistic regression**) is a useful model extending the logistic regression model
- There are several choices for defining the outcome metric - here we examined cumulative logits - odds of a given category or below vs. categories above
- An important assumption is the proportional odds assumption. This assumption needs to be checked for ordinal logistic model.
- The **likelihood ratio test** is helpful in comparing nested generalized linear models.
- Working with the ordinal model and coefficient interpretation can be more difficult (review Slide 19-20).
- Nonetheless, these models effectively support evaluation of multiple covariates in relation to ordered discrete outcomes.