The Analysis of Ordinal Data with Graphs and Odds Ratios

Robin High
Department of Biostatistics
University of Nebraska Medical Center
Omaha, NE

University of Nebraska Medical Center

Outline

Ordinal Logistic Regression Models

- Cumulative Logit
- Checking the Proportional Odds Assumption
- Adjacent Logit
- Odds Ratios
- Predicted Probabilities
- Visual Displays

Power Analysis with Ordered Categories

Ordinal Data
Categorical Data with an Inherent Order

Responses that reflect an ordered progression from the lowest to highest (or highest to lowest) levels in succession without reversing trend

Evaluate responses difficult or impossible to quantify, qualitative or "subjective" endpoints, such as:

- **Levels of agreement or disagreement**
- **Behavior (frequency of activity)**
- **Perceived difficulty of a task**
- **Sensation of pain**
- **Medical tests**
  - **Gleason: <= 6, 7, 8, 9, 10 (range is 2-10, typically recorded 6-10 or grade divided into three groups: <= 6, 7, 8-10)**
  - **PI-RADS: 1,2,3,4,5 (measures how likely cancer is present)**

Coding actual numbers into ordinal categories is generally NOT recommended for data analysis purposes

Coding Ordinal Data

Order Matters

Default settings of SAS procedures interpret character data or variables in a CLASS statement (discrete) are sorted in increasing order

- Alphabetical: a b c d e ..
- Numerical: 1 2 3 .. k

Where to start: ascending order

- Code ordinal data response level of "greatest" interest as 1 with increasing integers for remaining levels
- Contrast with binary data coded as 0/1

Assign actual values of numerical data codes with a format

```
PROC FORMAT;
VALUE rsp  1='Never' 2='Rarely' 3='Sometimes' 4='Often' 5='Always';
VALUE rspA 1='a Never' 2='b Rarely' 3='c Sometimes' 4='d Often' 5='e Always';
RUN;
```

Apply order= option when displaying data:

order = internal (with rsp)

order = formatted (with rspA)

Coding Data Preferences for Analysis (especially important with ordinal data)

All Data

- Define with short variable names (acronymns) and assign labels

Categorical Data

- Code with integers or short character of same length and case (some macros require all numbers)
- Numerical or alphabetical order matters
- Assign actual values with formats

General recommendations for data analysis:

- Do not assign formats to data in the DATA step
- When making data sets with ODS options, analyze data in procedures without FORMAT statements
- Apply FORMATs within the various SAS procedures that print or display the desired printed results

Analysis of Ordinal Data

Ordinal nature of outcome data much too often ignored

Ordinal data as a subjective measurement of an attribute, e.g., 2 is larger than 1, 3 is larger than 2, etc.,
yet how much numerically is not well-defined

Reduce k ordinal levels to 2 or 3 levels

- Apply only if small cell counts present

Analyze as continuous/interval data (treat as numbers)

- Resort to t-tests, ANOVA, regression methods often a matter of convenience for a "quicker, simpler" interpretation

Preferred approach is to summarize with odds ratios and predicted probabilities displayed in tables and graphs

SAS Software Procedures

Statistical Procedures with MODEL statement options

PROC LOGISTIC / link = clogit < alogit > unequalslopes

Procedures not discussed, available as needed

PROC GENMOD / dist=multinomial link = cumlogit

PROC GLIMMIX / dist=multinomial link = cumlogit

PROC NLMIXED

PROC CATMOD

Make Visual Displays

PROC SGPLOT

PROC SGPANEL

PROC PLOT (rough graphs for diagnostic purposes)

Utility Procedures

PROC TABULATE

PROC FREQ

PROC FORMAT

PROC TRANSPOSE

PROC MODEL (ETS)

Binary Logistic Regression

	a Yes	b No	Odds	Odds Ratio
a Treatment	p11	p12	odds1 = p11 / p12	OR = odds1/odds2
b Control	p21	p22	odds2 = p21 / p22	

Data summarized as row percents in a 2x2 table

p11 > p21 implies odds ratio will be > 1

Columns: outcome

Primary level of interest placed in column 1

Rows: treatment

Reference category (Control) placed in row 2

Row probabilities sum to 1

In SAS enter:

```
PROC FREQ DATA= inpdat order=formatted;
TABLE x * y / nocol nopercent cmh2;
FORMAT x trt. y ysn. ;
RUN;
```

Ordinal Logistic Regression with k=4 levels

	1=Very Good	2=Good	3=Moderate	4=Poor	Sum
a Treatment	p11	p12	p13	p14	1.0
b Control	p21	p22	p23	p24	1.0

```
Response has k discrete categories (usually 3 to 5 levels)

Make a 2 x k table (k=4 in table above)

Columns: ordinal outcome
         level of primary interest in column 1

Rows: Treatment vs Control
      Reference category in bottom row (row 2)
```

Row probabilities, p_{ij}, sum to 1 for i = 1,2

Ordinal Logistic Regression with k=4 levels

Choose ascending or descending order of the response

ascending

descending

	1=Very Good	2=Good	3=Moderate	4=Poor
a Treatment	p11	p12	p13	p14
b Control	p21	p22	p23	p24

The objective is to compute the odds ratio(s) and predicted probabilities based on one of two methods of grouping the ordinal responses (ascending or descending) with two types of link functions

Cumulative logit
Adjacent logit

Response Profile: Ascending

```
ODS select responseprofile;
PROC LOGISTIC DATA=indat ;  * default is ascending levels for the response;
MODEL y = x1 x2;
RUN;

Response Profile

 Ordered                        Total
   Value            y       Frequency

       1            1              60
       2            2              42
       3            3              21
       4            4              14

Probabilities modeled are cumulated over the lower Ordered Values.


Computed over lower order values in this setting means having these odds:

 [ Prob(y=1)                           ] / [ Prob(y=2) + Prob(y=3) + Prob(y=4) ]
 [ Prob(y=1) + Prob(y=2)               ] / [             Prob(y=3) + Prob(y=4) ]
 [ Prob(y=1) + Prob(y=2) + Prob(y=3)   ] / [                         Prob(y=4) ]
```

Response Profile: Descending

```
ODS select responseprofile;
PROC LOGISTIC DATA=indat descending;
MODEL y = x1 x2;
RUN;

Response Profile

 Ordered                         Total
   Value            y        Frequency

       1            4            14
       2            3            21
       3            2            42
       4            1            60

Probabilities modeled are cumulated over the lower Ordered Values.


Computed over lower order values in this setting means having these odds:

 [ Prob(y=4)                         ] / [ Prob(y=3) + Prob(y=2) + Prob(y=1) ]
 [ Prob(y=4) + Prob(y=3)             ] / [             Prob(y=2) + Prob(y=1) ]
 [ Prob(y=4) + Prob(y=3) + Prob(y=2) ] / [                         Prob(y=1) ]
```

Ordinal Logistic Regression: Cumulative Logit

	1 Very Good	2 Good	3 Moderate	4 Poor
Treatment	p11	p12	p13	p14
Control	p21	p22	p23	p24

```
Cumulative logit
Probabilities summed across all four response levels

Row T:  oddsT  =  A  /  B
Row C:  oddsC  =  C  /  D

Where components A, B, C, D defined for the first comparison
```

```
Compare (Very Good) to ( Good, Moderate, Poor)

T: odds11 = (p11                )  / (p12 + p13 + p14)
C: odds21 = (p21                )  / (p22 + p23 + p24)
```

Ordinal Logistic Regression: Cumulative Logit

```
Cumulative logit
Probabilities summed across all four response levels

Treatment:  odds T  =    A      /     B
  Control:  odds C  =    C      /     D

With k=4 have three ways to write pairs of equations for the odds

T: odds11 = (p11              )  / (p12 + p13 + p14)
C: odds21 = (p21              )  / (p22 + p23 + p24)

T: odds12 = (p11 + p12        )  / (p13 + p14)
C: odds22 = (p21 + p22        )  / (p23 + p24)

T: odds13 = (p11 + p12 + p13)  / (p14)
C: odds23 = (p21 + p22 + p23)  / (p24)


Cumulative Proportional Odds Model

Three equations of predicted values all compute one odds ratio
```

$$\text{Odds ratio} = \frac{\text{Odds T}}{\text{Odds C}} = \frac{(A\ /\ B)}{(C\ /\ D)}$$

Checking the Proportional Odds Assumption
A recent SAS resource with examples

Usage Note 37944

Plotting empirical (observed) logits for binary and ordinal response data

https://support.sas.com/kb/37/944.html

Page also has a link to download the

CtoN.sas macro

(character to numeric)

Ordinal Logistic Regression
Checking the proportional odds condition

PROC LOGISTIC

Score Test for the Proportional Odds Assumption

Visual aids

SAS Global Forum 2013
Paper 446-2013

Ordinal Response Modeling with the LOGISTIC Procedure
Bob Derr, SAS Institute Inc.

www.lexjansen.com

Search for "proportional odds"

Download pdf file for 446-2013

Direct link to the macros is found on page 11

http://support.sas.com/rnd/app/stat/papers/aastaffcode2013.html

Graphical Assessment of proportional odds

SUPPORT

RESOURCES / **FOCUS AREAS**

FOCUS AREAS

- Base SAS
- Graphics
- Enterprise Management Integration
- Migration
- Scalability & Performance
- SAS for Containers
- Statistics & Operations Research
  - SAS/ETS
  - SAS/IML
  - SAS/OR
  - SAS/QC
  - SAS/STAT

Programs Referenced in 2013 SAS Global Forum Papers

Ordinal Response Modeling with the LOGISTIC Procedure

By Bob Derr

- LRTest Macro
- LRTestCycle Macro

Does the addition of variables improve model? (unequalslopes option included)

- OneClassPlot Macro
- OneContPlot Macro
- EmpiricalLogitPlot Macro
- EXPlot Macro

Macros which produce graphical displays of the proportional odds assumption

Hints to Run the Macros

- **All explanatory variables must be numeric (early version need to dummy code categorical data before submitting macros, except the two LRT macros)**
- **Ordinal response coded from 1 to k [1,2, .., k]**
- **Submit file with individual observations (frequency indicated by a "count" variable not available)**
- **May want to modify the SGPLOT and SGPANEL options, especially graph dimensions and display options, and what folder to store it**
- **Data sets merged without a BY statement; macros will not work with this option:**

```
OPTIONS mergeNoBy=error;
```

Change to

```
OPTIONS mergeNoBy=warn;
```

Plots evaluate proportional odds in two ways

Empirical Logit Plot

ELP: Plots the empirical logits against observed X values (all coded as numeric), one at a time, unadjusted models

Are the plotted lines parallel?

EXPlot

EXP: Compares the mean of the X variable within each level of Y, to the expected value of X|Y=j given that the proportional odds assumption holds, E(X|Y=j), j = 1, 2, .. k
From Harrell (2001, 2015).

Do the two lines follow the same trajectory?

Proportional Odds Assumption
Which table can be evaluated with one odds ratio?

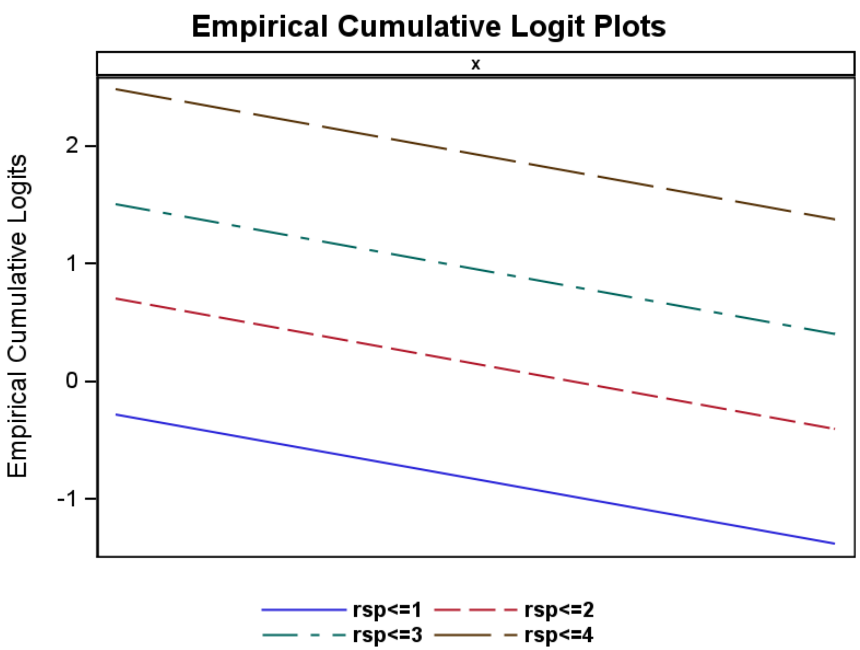
Table 1	1	2	3	4	5	Total
Treatment	86	48	30	21	15	200
	.43	.24	.15	.105	.075	1.0
Control	40	40	40	40	40	200
	.2	.2	.2	.2	.2	1.0

Table 2	1	2	3	4	5	Total
Treatment	86	55	9	24	26	200
	.43	.275	.045	.12	.13	1.0
Control	40	40	40	40	40	200
	.2	.2	.2	.2	.2	1.0

Table 1 Summary

Score Test for the Proportional Odds Assumption

Chi-Square	DF	Pr > ChiSq
0.0059	3	0.9999



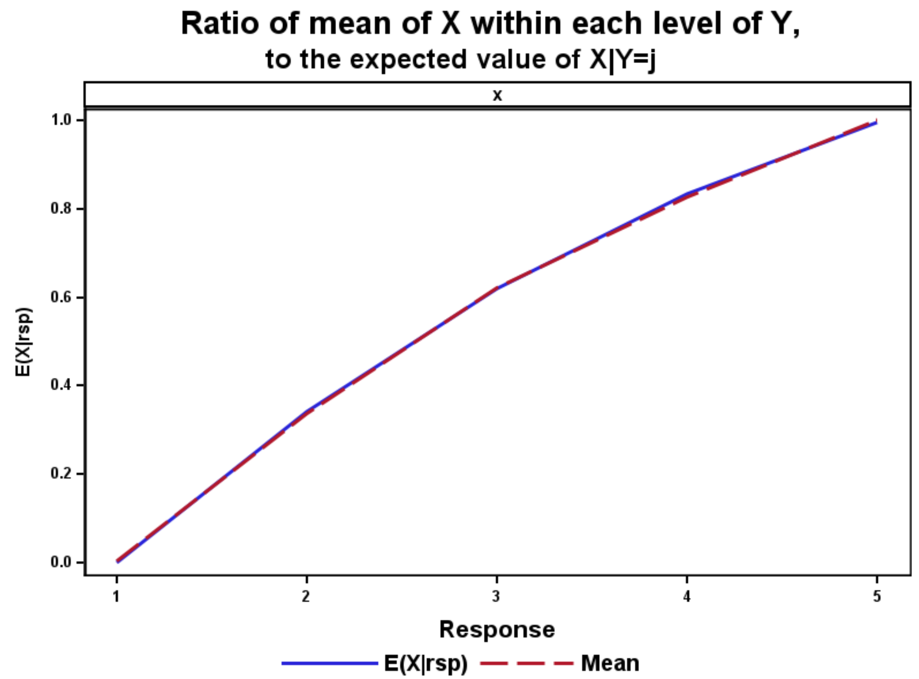
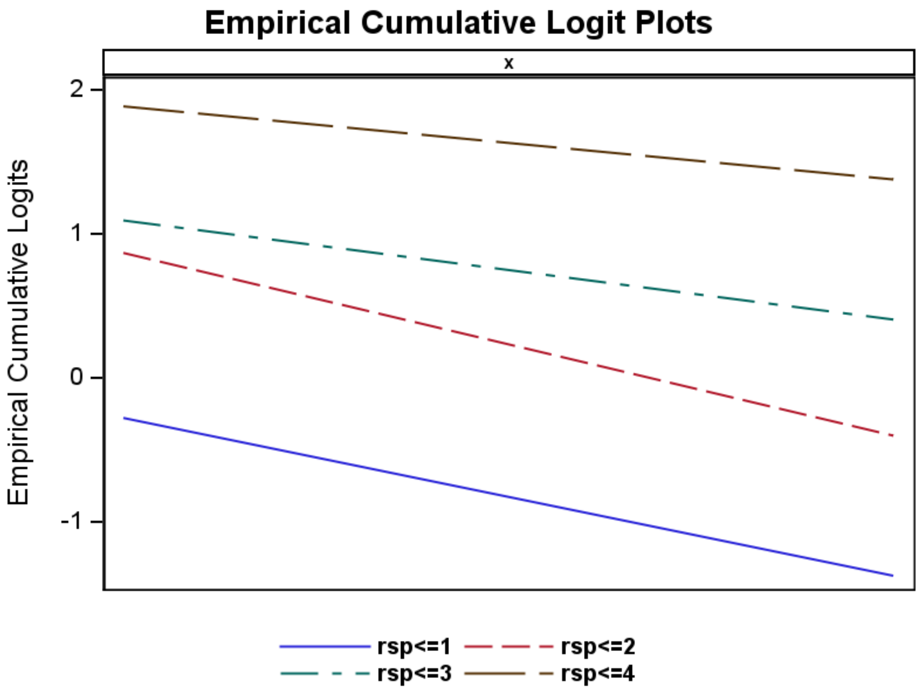
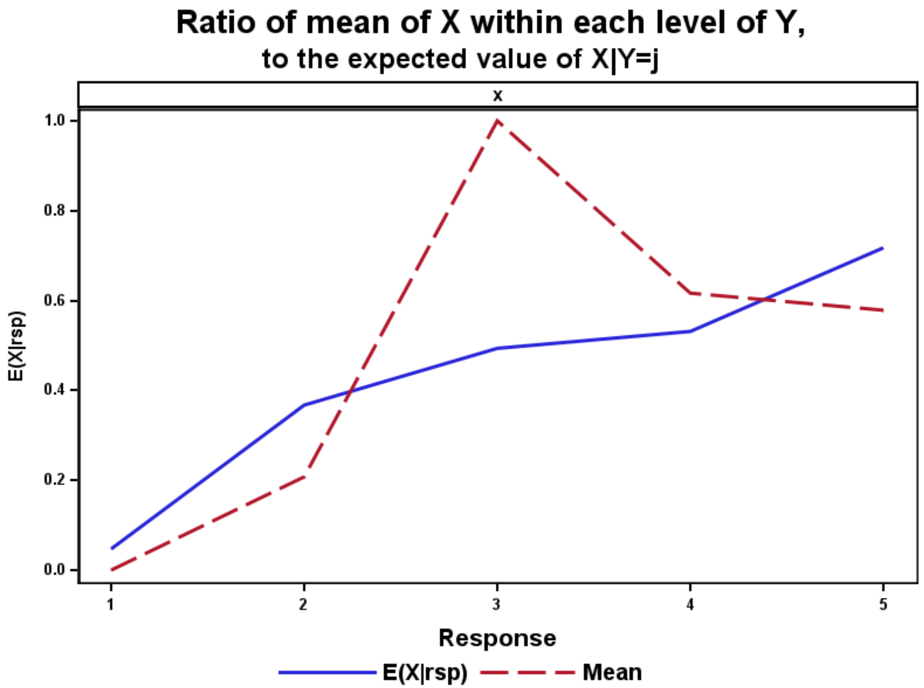


Table 2 Summary

Score Test for the Proportional Odds Assumption

Chi-Square	DF	Pr > ChiSq
24.53	3	<.0001





Ordinal Logistic Regression: Adjacent Logits

	a Very Good	b Good	c Moderate	d Poor
a Treatment	p11	p12	p13	p14
b Control	p21	p22	p23	p24

Adjacent logit

- **compare probabilities from adjacent cells**

```
T: oddsT =  A   /   B
C: oddsC =  C   /   D

T: odds11 = p11 / p12
C: odds21 = p21 / p22

T: odds12 = p12 / p13
C: odds22 = p22 / p23

T: odds13 = p13 / p14
C: odds23 = p23 / p24
```

Both models produce one odds ratio for the entire table

The computed odds ratio derived from three ratios:

```
     = odds11 / odds21
  OR = odds12 / odds22
     = odds13 / odds23
```

Ordinal Logistic Regression
Adjacent Logits

```
PROC LOGISTIC DATA=indat;
CLASS x1 / param=ref;
MODEL y = x1 x2 / link = alogit ;
SCORE data=scr out=prd;
RUN;
```

Adacent logit

```
With k=4, have three linear predictors

eta1 = LOG( P(y=1)/P(y=2)) = LOG(p1/p2)
eta2 = LOG( P(y=2)/P(y=3)) = LOG(p2/p3)
eta3 = LOG( P(y=3)/P(y=4)) = LOG(p3/p4)

Solve equations for p1, p2, p3, p4 in terms of eta1, eta2, eta3

p1 = 1 / total;
p2 = exp(eta1)*p1;
p3 = exp(eta2)*p2;
p4 = exp(eta3)*p3;

where
total = 1 + exp(eta1)
          + exp(eta1+eta2)
          + exp(eta1+eta2+eta3);
```

Computations of Ordinal Logistic Regression Models in PROC NLMIXED

SAS Global Forum, 2013
Paper 445-2013
Models for Ordinal Response Data

Robin High
University of Nebraska Medical Center

Ordinal Logistic Regression
Display results: Effect plots

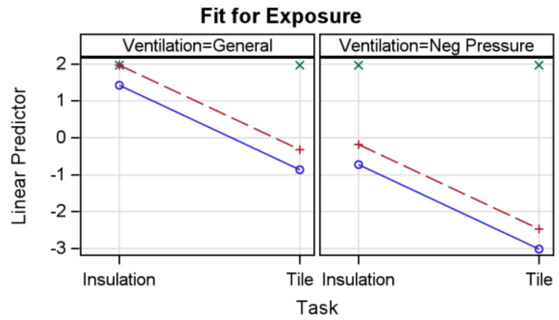
```
PROC LOGISTIC DATA=indat;
CLASS x1 x2 / param=ref ;
MODEL response = x1 x2 / link=clogit aggregate;

EFFECTPLOT interaction(x=x1 plotby=x2 sliceby=response) / link;
EFFECTPLOT interaction(x=x1 plotby=x2 sliceby=response) / cluster;
EFFECTPLOT interaction(x=x1 plotby=x2 sliceby=response) / polybar;
EFFECTPLOT interaction(x=x1 plotby=x2 sliceby=response) / individual;
EFFECTPLOT interaction(x=x1 plotby=x2 sliceby=response) / individual Noconnect;

Run;
```

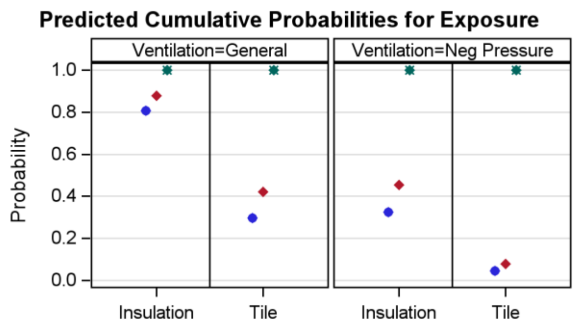
A Few Examples of Effect Plots

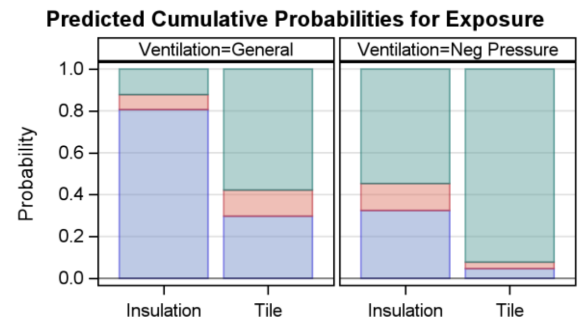
LINK



CLUSTER

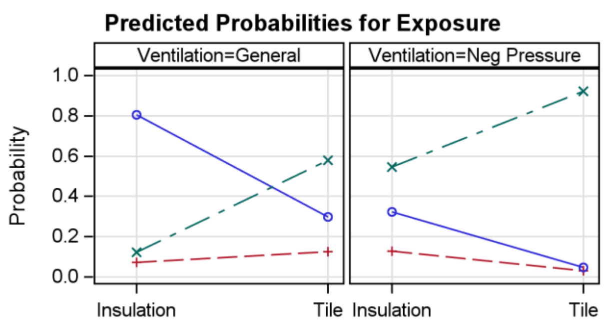
Cumulative Probabilities

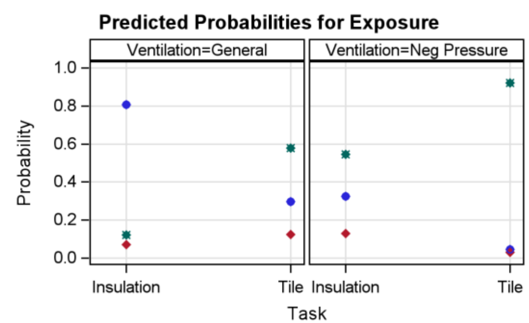




POLYBAR

Individual Probabilities





INDIVIDUAL NOCONNECT

Example 48.4: PROC GENMOD
Ice Cream Testing

Counts

	Very Good	Good	Middle	Bad	Very Bad
Brand 1	70	71	151	30	46
Brand 2	20	36	130	74	70
Brand 3	50	55	140	52	50

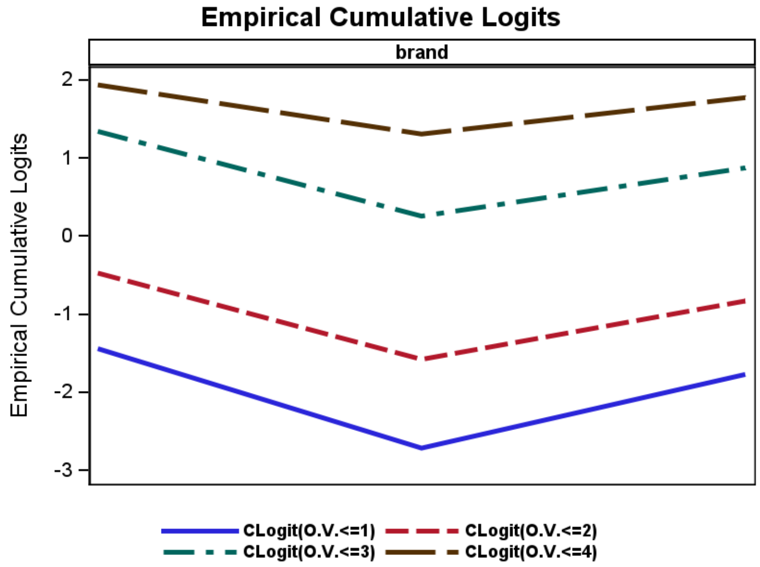
Row Percents

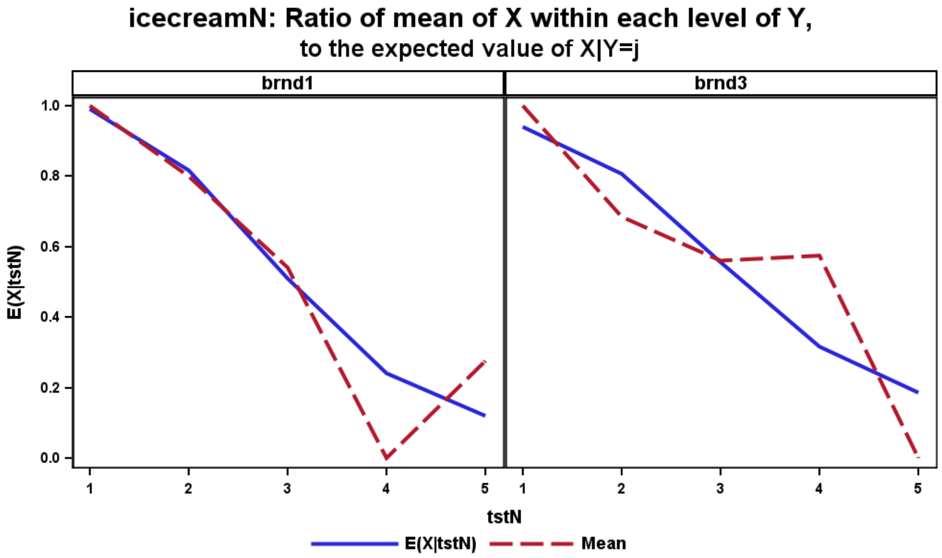
	Very Good	Good	Middle	Bad	Very Bad
Brand 1	19.0	19.3	41.0	8.2	12.5
Brand 2	6.1	10.9	39.4	22.4	21.2
Brand 3	14.4	15.9	40.3	15.0	14.4

Checking the proportional odds assumption

Score Test for the Proportional Odds Assumption

Chi-Square	DF	Pr > ChiSq
11.1216	6	0.0847





LRT: Model 2 (unequal slopes) vs Model 1 (equal slopes)

DF	ChiSq	P	
6	9.865	0.13	pvalue prefers equal slopes

Collecting Predicted Values SCORE statement

```
DATA iceScr;
brand = 'ice1'; output;
brand = 'ice2'; output;
brand = 'ice3'; output;
run;

PROC LOGISTIC DATA=icecream ;
FREQ count;
CLASS brand / param=ref;
MODEL tstN = brand / link=alogit aggregate ;
SCORE data=iceScr out=prdA;
TITLE 'LOGISTIC: Adjacent Logit';
run;

PROC PRINT DATA=prdA; RUN;
```

brand	P_1	P_2	P_3	P_4	P_5
ice1	0.18744	0.18692	0.40526	0.12082	0.09956
ice2	0.07057	0.11233	0.38876	0.18500	0.24333
ice3	0.13756	0.16180	0.41375	0.14549	0.14140

Ice Cream Testing
Predicted Cell Probabilities

Cumulative logits

	Very Good	Good	Middle	Bad	Very Bad
Brand 1	0.186	0.196	0.405	0.112	0.100
Brand 2	0.076	0.105	0.388	0.193	0.238
Brand 3	0.135	0.161	0.419	0.143	0.141

Odds Ratios

1 vs 3: 1.47
2 vs 3: 0.52

Adjacent logits

	Very Good	Good	Middle	Bad	Very Bad
Brand 1	0.187	0.187	0.405	0.121	0.100
Brand 2	0.071	0.112	0.389	0.185	0.243
Brand 3	0.138	0.162	0.414	0.145	0.141

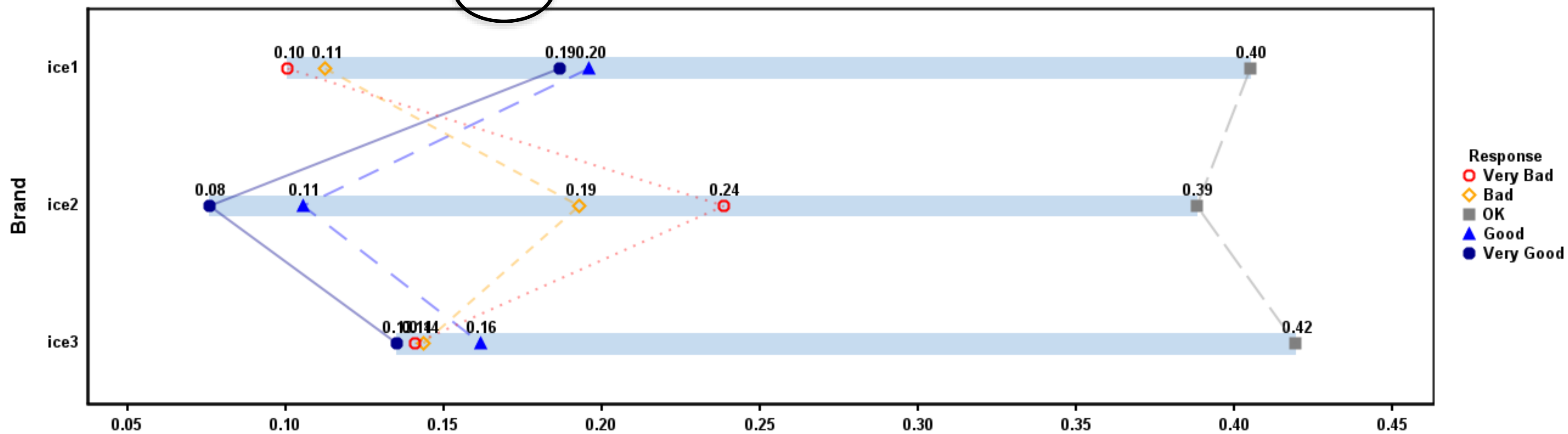
Odds Ratios

1 vs 3: 1.18
2 vs 3: 0.74

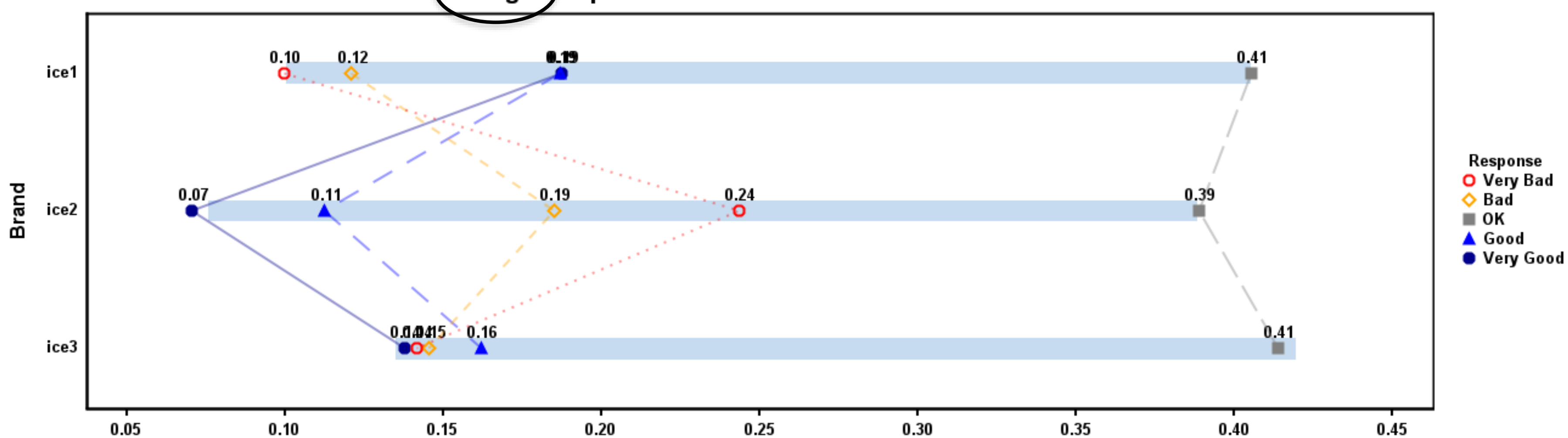
Cell probabilities very close for all three brands and five responses

Visual Comparison of Probabilities

Logit: Response Probabilities based on ice cream brand



Alogit: Response Probabilities based on ice cream brand



Cumulative Logits: Ice Cream Testing
Odds Ratios from Predicted Cell Probabilities

Based on cumulative logits

	Very Good	Good	Middle	Bad	Very Bad
Brand 1	0.186	0.196	0.405	0.112	0.100
Brand 2	0.076	0.105	0.388	0.193	0.238
Brand 3	0.135	0.161	0.419	0.143	0.141

Odds Ratio

1 vs 3: 1.47

Verify odds ratios from rows 1 and 3:

$$\text{OR 1 vs 3} = \frac{0.186 \;/\; (0.196 + 0.405 + 0.112 + 0.100)}{0.135 \;/\; (0.161 + 0.419 + 0.143 + 0.141)} = 1.47$$

$$\text{OR 1 vs 3} = \frac{(0.186 + 0.196) \;/\; (0.405 + 0.112 + 0.100)}{(0.135 + 0.161) \;/\; (0.419 + 0.143 + 0.141)} = 1.47$$

Odds Ratio=0.52 to compare Brand 2 with Brand 3 derived from probabilities in rows 2 and 3

Adjacent Logits: Ice Cream Testing
Odds Ratios from Predicted Cell Probabilities

Based on adjacent logits

	Very Good	Good	Middle	Bad	Very Bad
Brand 1	0.187	0.187	0.405	0.121	0.100
Brand 2	0.071	0.112	0.389	0.185	0.243
Brand 3	0.138	0.162	0.414	0.145	0.141

Odds Ratios

1 vs 3: 1.18

Verify Odds Ratios with probabilities from rows 1 and 3

$$\text{OR } 1 \text{ vs } 3 = \frac{0.187 / 0.187}{0.138 / 0.162} = 1.18$$

$$\text{OR } 1 \text{ vs } 3 = \frac{0.187 / 0.405}{0.162 / 0.414} = 1.18$$

Odds Ratio=0.74 to compare Brand 2 to Brand 3 derived in the same manner with rows 2 and 3

Ordinal Logistic Regression

Which model to apply?

Both approaches fit well in similar situations and provide similar substantive results (with large cell counts)

<u>Cumulative logit</u> extends inference to underlying continuum (all k levels)

<u>Adjacent logit</u> gives effects in terms of the categories preferable to interpret categories next to each other rather then the entire continuum

Logistic Regression
Partial Proportional Odds

When diagnostics indicate non-proportional outcomes, apply the “unequalslopes” option in PROC LOGISTIC

```
PROC LOGISTIC DATA=indat;
CLASS x1 / param=ref;
MODEL y = x1 x2 / link=clogit unequalslopes;
RUN;
```

Or specify the specific variable name(s) to apply unequalslopes

```
MODEL y = x1 x2 / link=clogit unequalslopes=x1;
```

Ordinal Logistic Regression with NLMIXED
Odds Ratios for both Cumulative Logits

Proportional Odds: one odds ratio

```
proc nlmixed data=indat(rename=(outcome = y));
parms Int1 -2 Int2 0 Int3 1.5 bx .07 ;

eta1 = Int1 + bx * (trt=1)  ; * trt=0 ref category;
eta2 = Int2 + bx * (trt=1)  ;
eta3 = Int3 + bx * (trt=1)  ;

cp1= 1 / (1 + exp(-eta1));
cp2= 1 / (1 + exp(-eta2));
cp3= 1 / (1 + exp(-eta3));

p1 = cp1;           * calculate single cell probs ;
p2 = cp2 - cp1;     * from the cumulative probs;
p3 = cp3 - cp2;
p4 =   1 - cp3;

lk =   (p1**(y EQ 1)) * (p2**(y EQ 2))
     * (p3**(y EQ 3)) * (p4**(y EQ 4)); * ascending;

  lk = max(min(lk,1-1E-9),1E-9);
lglk = log(lk);

ESTIMATE 'Odds Ratio' EXP(bx);

MODEL y ~ general(lglk);
TITLE "NLMIXED: ordinal Logr Regr with link = clogit";
run;
```

Partial Proportional odds: three odds ratios

```
parms Int1 -2 Int2 0 Int3 1.5 bx .07 da .1 db .1 ;

eta1 = Int1 + (bx    ) * (trt=1) ;
eta2 = Int2 + (bx+da) * (trt=1) ;
eta3 = Int3 + (bx+db) * (trt=1) ;

ESTIMATE 'Odds Ratio 1 vs 234' EXP(bx     );
ESTIMATE 'Odds Ratio 12 vs 34' EXP(bx + da);
ESTIMATE 'Odds Ratio 123 vs 4' EXP(bx + db);
```

Example

Backache Severity Data

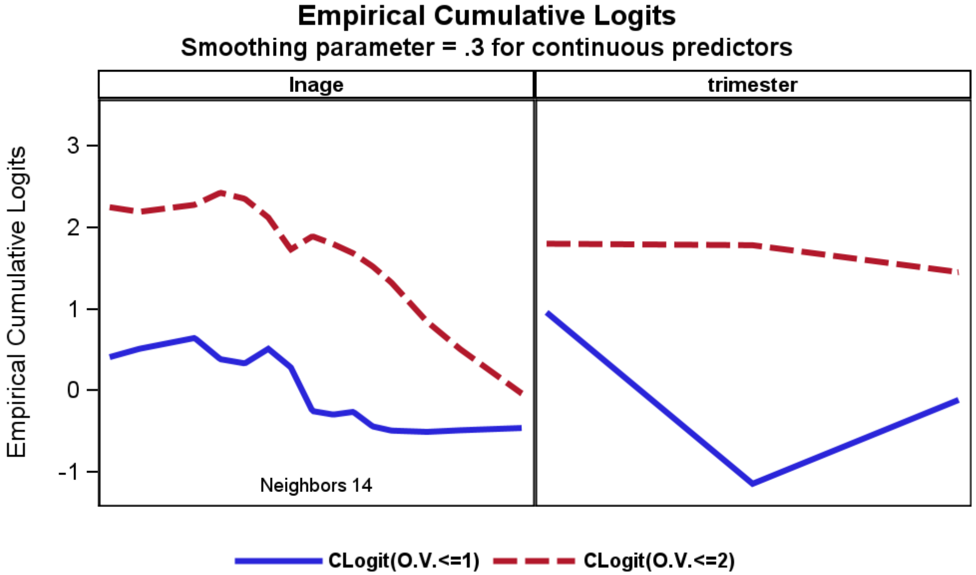
Problem Solving: A Statistician's Guide
by Chris Chatfield (1995)

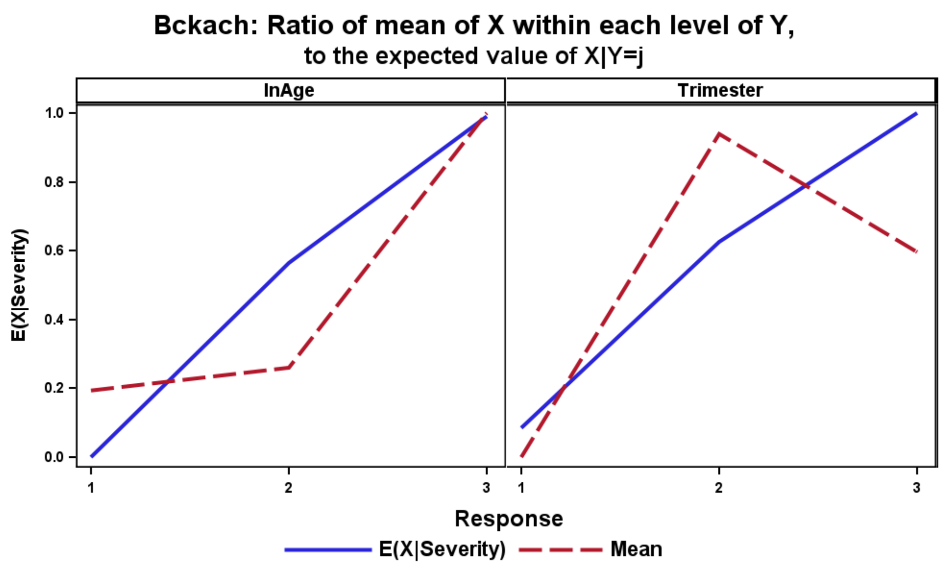
data available in a downloadble text file at:

https://lib.stat.cmu.edu/datasets

Evaluate: women who gave birth in a London hospital, is severity of back pain (three ordinal levels) related to trimester pain began (discrete) and age (continuous in years)

Evaluate Proportional Odds





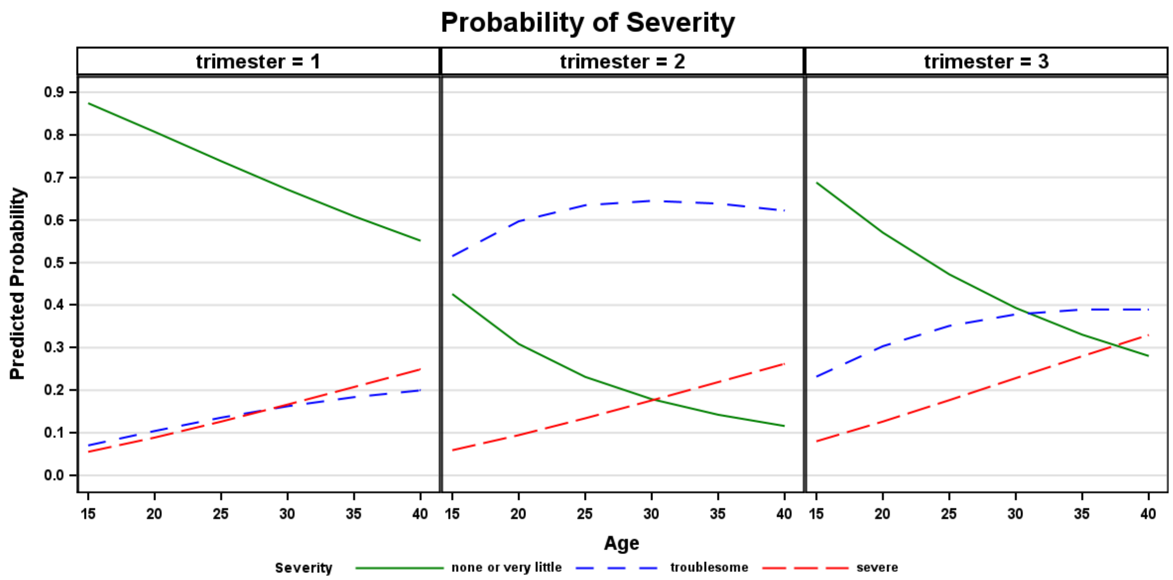
Score Test

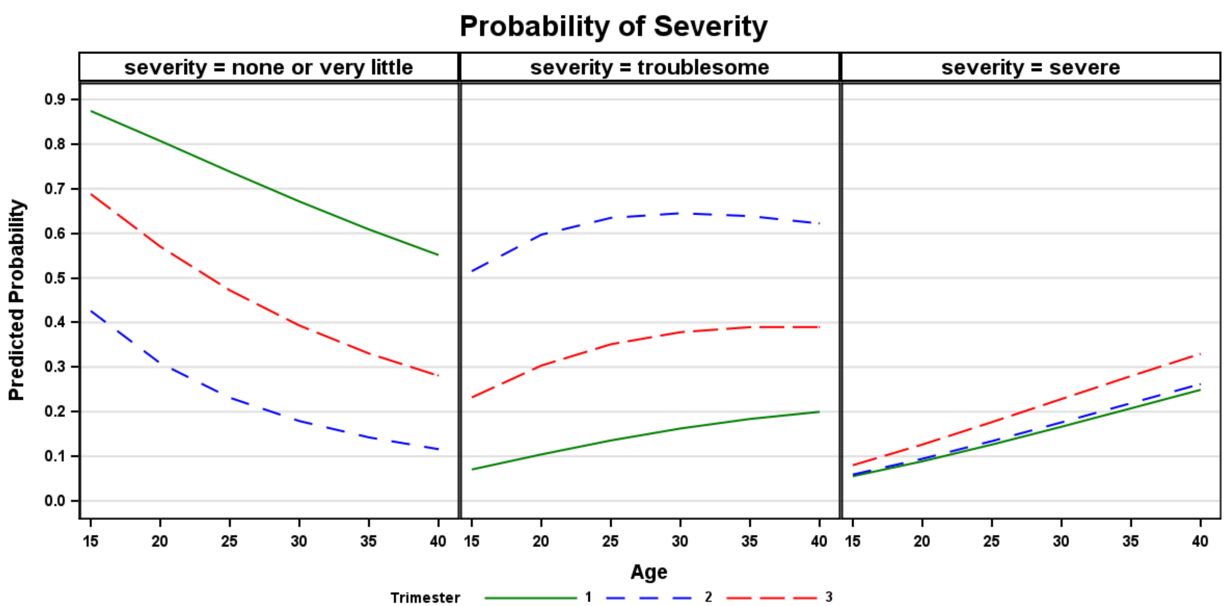
Chi-Square	DF	Pr > ChiSq
22.28	3	<.0001

Trimester (panel on right in both plots) shows signs of non-proportional odds

```
proc logistic
      data=Bckach descending;
class Trimester / param=ref;
model Severity =lnAge Trimester
/ unequalslopes=Trimester;
Run;
```

Graph Predicted Probabilities

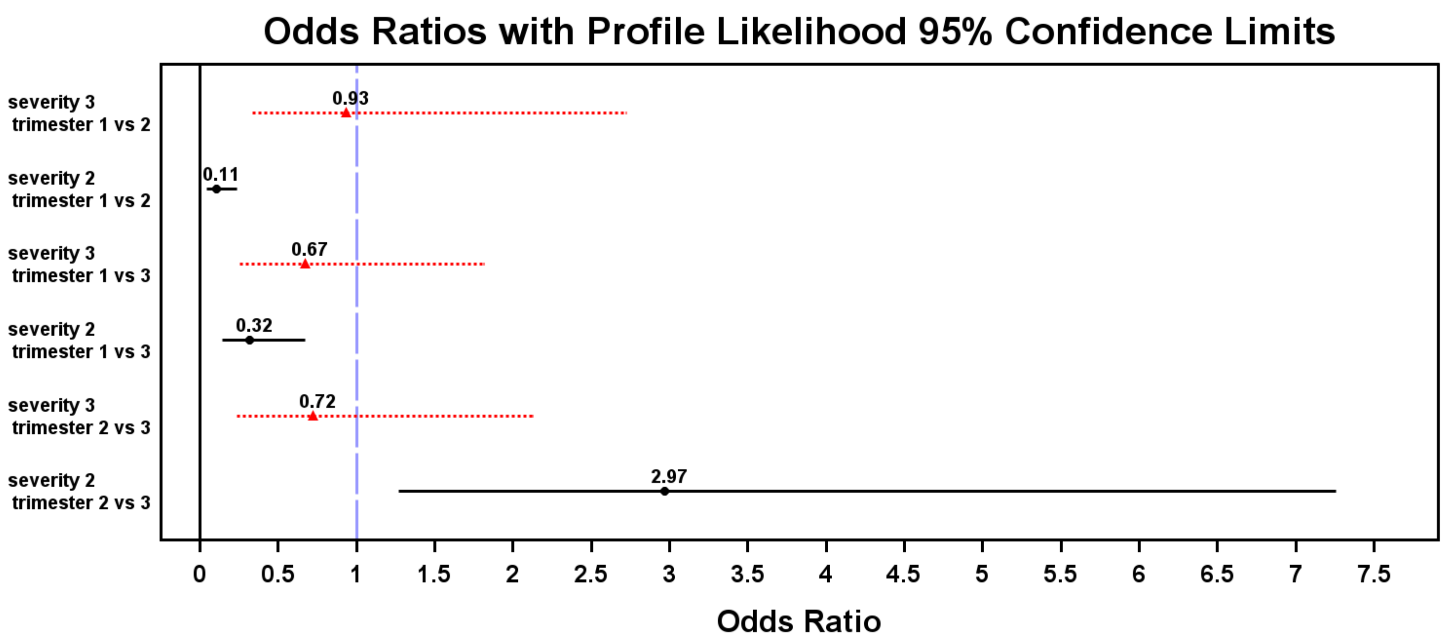




For any of the three trimesters at a specific age, the three probabilities for level of severity sum to 1

Odds Ratios: Descending option

Effect	Odds Ratio	Lower CL	Upper CL
Sev 3: Tri 1 vs 2	0.93	0.34	2.73
Sev 2: Tri 1 vs 2	0.11	0.045	0.24
Sev 3: Tri 1 vs 3	0.67	0.25	1.82
Sev 2: Tri 1 vs 3	0.32	0.15	0.67
Sev 3: Tri 2 vs 3	0.72	0.24	2.13
Sev 2: Tri 2 vs 3	2.97	1.27	7.26



Compute Odds Ratios from Probabilities

descending

Age = 20	Severity		
Trimester	1=none/very little	2=troublesome	3=severe
1	0.80691	0.10425	0.08884
2	0.30863	0.59682	0.09455
3	0.57017	0.30342	0.12641

Severity 3: trimester 1 vs 2

$$\text{OR} = \frac{0.08884 \;/\; (0.80691 + 0.10425)}{0.09455 \;/\; (0.30863 + 0.59682)} = 0.93$$

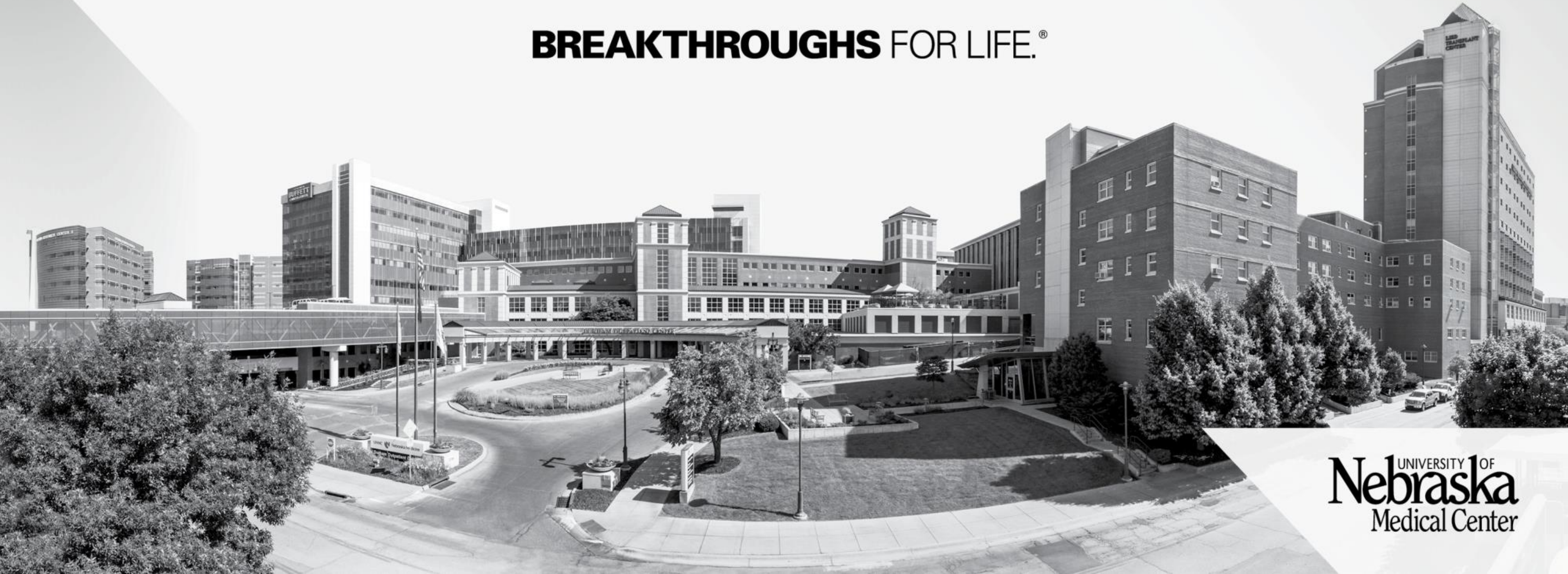
Severity 2: trimester 1 vs 2

$$\text{OR} = \frac{(0.08884 + 0.10425) \;/\; 0.80691}{(0.09455 + 0.59682) \;/\; 0.30863} = 0.11$$

Remaining odds ratios computed in like manner
from rows 1 and 3 (Trimesters 1 and 3)
rows 2 and 3 (Trimesters 2 and 3)

University of Nebraska Medical Center℠

BREAKTHROUGHS FOR LIFE.®



University of Nebraska Medical Center