

Lecture 18

1. Odds ratio and relative risk
 2. Log-binomial regression
 3. Ordinal regression: cumulative categories—proportional odds
 4. Ordinal regression: comparison to single reference—generalized logits
-

Review of log-binomial regression: *Am J Epidemiol* 2005; 162:199–200

Stokes, et al. (2000) *Categorical Data Analysis Using the SAS System, 2nd Edition*

McCullagh, Nelder (1989) *Generalized Linear Models, Second Edition*

Harrell (2001) *Regression Modeling Strategies* (Springer)

1

Odds and risk for rare events

Logistic regression gives odds ratios from regression coefficients:

$$\exp(\hat{\beta}_X) = \text{odds ratio for 1-unit increase in } X$$

Odds ratio sometimes used as an estimate of risk ratio = relative risk

Suppose number of events is A , number of non-events is B ,

$$\text{odds} = \frac{\text{number with event}}{\text{number without event}} = \frac{A}{B}$$

$$\text{risk} = \frac{\text{number with event}}{\text{number with event} + \text{number without event}} = \frac{A}{B + A}$$

2

Relationship between odds ratio and relative risk

$$\text{odds} = \frac{\text{number with event}}{\text{number without event}} = A/B, \quad \text{risk} = \frac{\text{number with event}}{\text{total number}} = A/(A+B)$$

\Rightarrow odds > risk.

Let π_1 and π_2 be risk in groups 1 and 2, respectively.

$$\text{relative risk} = \frac{\pi_1}{\pi_2}, \quad \text{odds ratio} = \frac{\frac{\pi_1}{1-\pi_1}}{\frac{\pi_2}{1-\pi_2}} = \left(\frac{\pi_1}{1-\pi_1}\right)\left(\frac{1-\pi_2}{\pi_2}\right)$$

$$\frac{\text{odds ratio}}{\text{relative risk}} = \left(\frac{\pi_1}{1-\pi_1}\right)\left(\frac{1-\pi_2}{\pi_2}\right)\left(\frac{\pi_2}{\pi_1}\right) = \frac{1-\pi_2}{1-\pi_1}$$

3

$$\frac{\text{odds ratio}}{\text{relative risk}} = \frac{1-\pi_2}{1-\pi_1}$$

Fix relative risk as constant θ (theta):

$$\text{relative risk} = \theta = \frac{\pi_1}{\pi_2} \quad \Rightarrow \quad \pi_1 = \theta\pi_2$$

Substituting gives

$$\frac{\text{odds ratio}}{\text{relative risk}} = \frac{1-\pi_2}{1-\pi_1} = \frac{1-\pi_2}{1-\theta\pi_2}$$

For relative risk = $\theta = 1$, this ratio is one \Rightarrow OR = RR

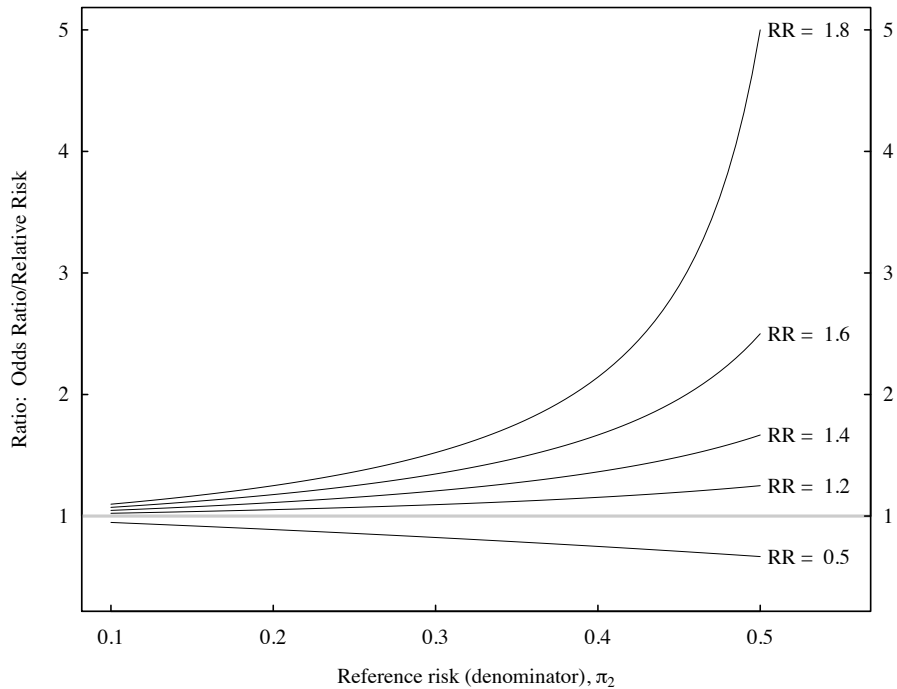
For relative risk = $\theta > 1$, this ratio is greater than one \Rightarrow OR > RR

For relative risk = $\theta < 1$, this ratio is less than one \Rightarrow OR < RR

Odds ratio is farther than relative risk from null value 1.

4

Graph of Odds Ratio/Relative Risk = $(1 - \pi_2)/(1 - \theta\pi_2)$ for range of relative risk.



rare event

common event

5

Rare events: the number of events A is very small relative to the number of non-events B ,

$$B \approx B + A, \quad \text{so} \quad \frac{A}{B} \approx \frac{A}{B + A}$$

that is, odds \approx risk (rate).

When event is rare in both groups, odds ratio \approx relative risk.

But when events are not rare, odds ratios are not good estimates of relative risk.

Alternative: estimate relative risk directly.

6

Log-binomial regression: estimating relative risks directly

Logistic regression: binomial response y , mean chance of event is $\pi(x)$ and

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x$$

Log-binomial regression: binomial response y , mean chance of event is $\pi(x)$ and

$$\log(\pi(x)) = \beta_0 + \beta_1 x$$

Change the function that links the mean π to the linear regression $\beta_0 + \beta_1 x$

7

With log-binomial model, we estimate relative risk, not odds ratio.

Regression coefficient for predictor x is

$$\begin{aligned}\beta_1 &= (\beta_0 + \beta_1(x+1)) - (\beta_0 + \beta_1(x)) \\ &= \log(p(x+1)) - \log(p(x)) \\ &= \log\left(\frac{p(x+1)}{p(x)}\right) \\ &= \log(\mathbf{relative\ risk\ for\ unit\ increase\ in\ }x)\end{aligned}$$

$\Rightarrow \exp(\beta_1) = \mathbf{relative\ risk\ for\ 1\text{-unit\ increase\ in\ }X}$

8

Non-rare event: OR vs RR

Example from Greenland (*Am J Epidemiol* 2004; 160:301–305)

Cohort study followed 192 women with breast cancer, classified by breast-cancer stage (I, II, III) and receptor level (low, high).

Survival = alive 5 years after diagnosis

	Stage I		Stage II		Stage III	
	Low	High	Low	High	Low	High
Deaths	2	5	9	17	12	9
Survivors	10	50	13	57	2	6
Total	12	55	22	74	14	15

Event is death, not rare in this example.

(Data source: Newman SC. *Biostatistical methods in epidemiology*. New York, NY: Wiley, 2001.)

9

Model: death rate = receptor stage

Compare results from

1. Logistic regression odds ratios
2. Log binomial relative risks

First: get data into SAS

	Stage I		Stage II		Stage III	
	Low	High	Low	High	Low	High
Deaths	2	5	9	17	12	9
Survivors	10	50	13	57	2	6
Total	12	55	22	74	14	15

```

data greenland; * from Am J Epidemiol (2004) 160:301-305;
  input receptor$ stage deaths survivors;
  total = deaths + survivors; compute, don't type
  cards;
low 1 2 10
high 1 5 50
low 2 9 13
high 2 17 57
low 3 12 2
high 3 9 6
;

```

11

Grouped binary responses (binomial)

```

Proc Logistic ;
  model number of events / number of subjects = predictors ;

```

NO descending option since events are specified in the model statement.

```

Proc Logistic data=greenland;
  class receptor (ref="high") stage (ref="1") ;
  model deaths/total = receptor stage / CLodds=PL;

```

Set reference level in the class statement.

12

Number of Observations Read	6
Number of Observations Used	6
Sum of Frequencies Read	192
Sum of Frequencies Used	192

Response Profile

Ordered Value	Binary Outcome	Total Frequency
1	Event	54
2	Nonevent	138

13

Start by checking for interaction:

```
proc logistic data=greenland;
  class receptor(ref="high") stage(ref="1");
  model deaths/total = receptor stage receptor*stage / CLodds=PL;
```

Type 3 Analysis of Effects

Effect	DF	Wald	
		Chi-Square	Pr > ChiSq
receptor	1	4.3873	0.0362
stage	2	22.8100	<.0001
receptor*stage	2	0.3381	0.8445

14

```
Proc Logistic data=greenland;
  class receptor (ref="high") stage (ref="1") ;
  model deaths/total = receptor stage / CLodds=PL;
```

Regression coefficients:

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.5503	0.2212	6.1906	0.0128
receptor low	1	0.4598	0.1977	5.4080	0.0200
stage 2	1	-0.2223	0.2499	0.7914	0.3737
stage 3	1	1.5791	0.3231	23.8796	<.0001

15

Odds ratio estimates from logistic regression.

Profile Likelihood Confidence Interval for Adjusted Odds Ratios

Effect	Unit	Estimate	95% Confidence Limits	
receptor low vs high	1.0000	2.508	1.148	5.454
stage 2 vs 1	1.0000	3.110	1.306	8.303
stage 3 vs 1	1.0000	18.839	6.299	63.727

Next: relative risk estimates from log-binomial regression.

16

Proc Genmod: log-binomial regression

Use Proc Genmod, specify binomial distribution and **log** link:

```
Proc Genmod data=greenland;
  class receptor stage ;
  model deaths/total = receptor stage
    / type3 dist=binomial link= log ;
```

To fit logistic regression (same results as previous page), specify **logit** link:

```
Proc Genmod data=greenland;
  class receptor stage ;
  model deaths/total = receptor stage
    / type3 dist=binomial link= logit ;
```

17

Reference level is supposed to work in Proc Genmod but doesn't. Either recode levels, or use ESTIMATE statement.

Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square
Intercept		1	-0.1390	0.0955	-0.3262	0.0482	2.12
receptor	high	1	-0.4436	0.2021	-0.8396	-0.0476	4.82
receptor	low	0	0.0000	0.0000	0.0000	0.0000	.
stage	1	1	-1.7695	0.3875	-2.5290	-1.0101	20.85
stage	2	1	-0.8381	0.2154	-1.2604	-0.4158	15.13
stage	3	0	0.0000	0.0000	0.0000	0.0000	.
Scale		0	1.0000	0.0000	1.0000	1.0000	

Parameter		Pr > ChiSq
Intercept		0.1457
receptor	high	0.0281
receptor	low	.
stage	1	<.0001
stage	2	0.0001
stage	3	.
Scale		

18

```

Proc Genmod data=greenland;
  class receptor(ref=first) stage(ref=first); * ref doesn't work;
  model deaths/total = receptor stage / type3 dist=bin link=log;

  estimate "receptor low vs. high" receptor -1 1 / exp ;
  estimate "stage2 vs stage1" stage -1 1 0 / exp;
  estimate "stage 3 vs stage1" stage -1 0 1/exp;

  ODS output ParameterEstimates = reg_coef;

```

Genmod doesn't calculate relative risks for you.

Use ODS to save regression coefficients to data set, then back-transform.

19

Exponentiating the estimated difference gives the relative risk:

Contrast Estimate Results

Label	Estimate	Standard Error	Alpha	Confidence Limits	
receptor low vs. high	0.4436	0.2021	0.05	0.0476	0.8396
Exp(receptor low vs. high)	1.5583	0.3149	0.05	1.0487	2.3155
stage2 vs stage1	0.9314	0.3937	0.05	0.1599	1.7030
Exp(stage2 vs stage1)	2.5382	0.9991	0.05	1.1734	5.4903
stage 3 vs stage1	1.7695	0.3875	0.05	1.0101	2.5290
Exp(stage 3 vs stage1)	5.8680	2.2738	0.05	2.7458	12.5406

20

Relative risks from log-binomial model

Label	Estimate	Confidence Limits	
Exp(receptor low vs. high)	1.5583	1.0487	2.3155
Exp(stage2 vs stage1)	2.5382	1.1734	5.4903
Exp(stage 3 vs stage1)	5.8680	2.7458	12.5406

Odds ratios from logistic regression

Effect	Estimate	95% Confidence Limits	
receptor low vs high	2.508	1.148	5.454
stage 2 vs 1	3.110	1.306	8.303
stage 3 vs 1	18.839	6.299	63.727

For non-rare events, odds ratio is farther than relative risk from null value 1.

21

Why doesn't everyone use log-binomial instead of logistic regression?

1. Logistic is numerically more stable: log-binomial does not always converge to produce an answer.
2. Logistic is conventional approach, software more developed.

22

Ordinal logistic regression

In the logistic regression example, we looked at how rate of obesity related to age and gender in NHANES 2004. Obesity is a binary response, defined by $BMI \geq 30$.

However, there is an intermediate category:

- Obese: $BMI \geq 30$.
- Overweight: $25 \leq BMI < 30$
- Normal weight: $18 \leq BMI < 25$

Examine how rates of obesity and overweight relate to age and gender, then three ordered categories:

Normal weight < Overweight < Obese

23

Ordinal response

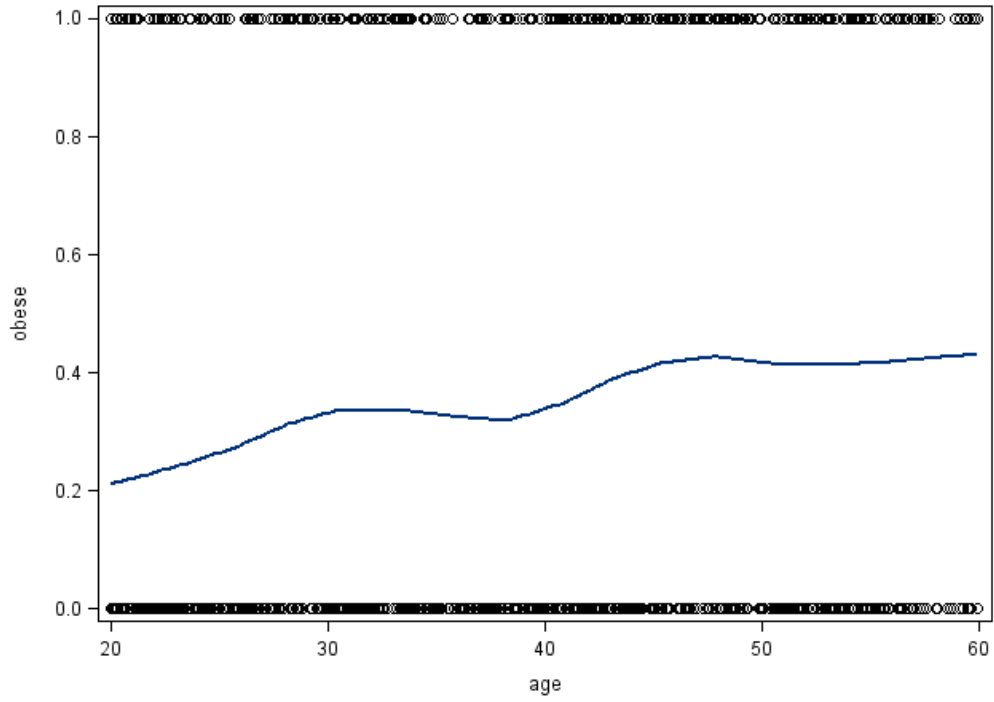
Response variable has three or more ordered categories.

Ordinal response categories may be defined by a continuous measurement scale, as obesity and overweight are defined with reference to the BMI scale. Or they may just be ordered:

Worse < No Change < Recovered

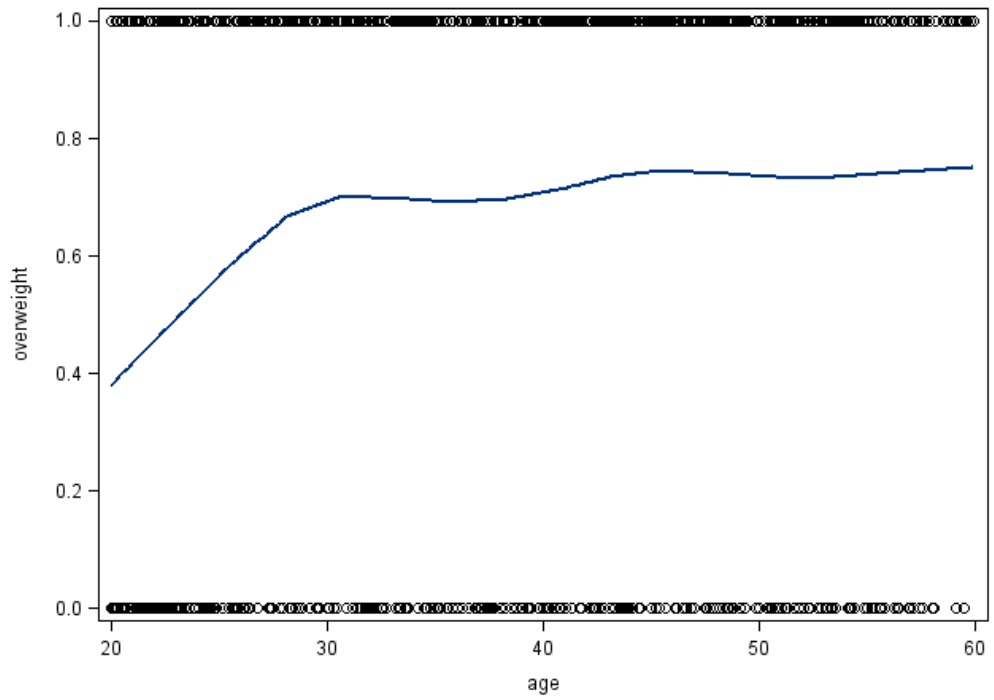
where it does not make sense to ask about the distance between categories.

Ordinal models use only the ranks of the categories.



Probability of being obese, as function of age.

25



Probability of being overweight or obese, as function of age.

26

Divide age into two categories:

young aged 20–39 years old, and **old** aged 40–60.

Frequency				
Row Pct	1_normal	2_overwt	3_obese	Total
old	118	153	183	454
	25.99	33.70	40.31	
young	190	155	142	487
	39.01	31.83	29.16	
Total	308	308	325	941

Percent in each weight category for old: $\hat{\pi}_{11}, \hat{\pi}_{12}, \hat{\pi}_{13}$.

Percent in each weight category for young: $\hat{\pi}_{21}, \hat{\pi}_{22}, \hat{\pi}_{23}$.

27

Multinomial distribution (generalizes binomial distribution):

parameters are the probabilities (π_{ij}) of being in each category.

Logistic regression estimates the difference of odds (on the log scale).

Ordinal regression will fit two logistic regression simultaneously:

- Odds of being in the top category vs the rest:
obese vs (overweight + normal)
- Odds of being in the top two categories vs the rest:
(overweight + obese) vs normal

Difference in odds between young and old assumed to be same for both.

28

Proportional odds model for ordinal responses

Proportional odds model forces > 2 ordinal categories into binary comparisons by combining categories in sequence from the top. Gives **cumulative odds**:

1. Odds of top category vs the rest: obese vs (normal + overweight)
2. Odds of top two categories vs the rest: (obese + overweight) vs normal
3. Odds of being in the top three categories vs the rest, etc.

To define these odds we define cumulative probabilities:

$$\theta_3 = \text{chance of obesity} = \pi_3$$

$$\theta_2 = \text{chance of obesity or overweight} = \pi_2 + \pi_3,$$

29

Frequency					
Row Pct	1_normal	2_overwt	3_obese		Total
old	118	153	183		454
young	190	155	142		487
Total	308	308	325		941

Find odds ratios for old to young of:

obesity

overweight + obesity

30

	Normal +			
	Obese	Overweight	odds	odds ratio
Age 40+	183	271	0.6753	
Age 20–39	142	345	0.4116	1.64

	Obese +			
	Overweight	Normal	odds	odds ratio
Age 40+	336	118	2.85	
Age 20–39	297	190	1.56	1.84

31

Proportional odds model combines two logistic regression models:

$\text{logit}(\theta_{h3}) = \log \text{ odds of being in the top category vs the rest, for group } h$

$\text{logit}(\theta_{h2}) = \log \text{ odds of being in the top two category vs the rest, for group } h$

$$\left\{ \begin{array}{l} \text{logit}(\theta_{h3}) = \log \left[\frac{\theta_{h3}}{1 - \theta_{h3}} \right] = \alpha_3 + x_h \beta \\ \text{logit}(\theta_{h2}) = \log \left[\frac{\theta_{h2}}{1 - \theta_{h2}} \right] = \alpha_2 + x_h \beta \end{array} \right.$$

β estimates the same covariate effect in both models: an “average” effect (odds ratio) of age for both BMI cut-points, ratio of the odds for someone 40+ of “being in a heavier category” to the odds for someone 20–39.

Proc Logistic tests this assumption.

32

Fitting the proportional odds model in Proc Logistic

```
Proc Logistic descending data=mayod327.age_bmi_sample;  
  class age_category /param=glm;  
  model bmi_cat = age_category ;
```

bmi_cat has 3 levels.

Default when the response has > 2 levels is the *proportional odds model*.

33

It is critical to check that SAS is combining categories in the right direction: use the `descending` option to reverse the order.

Response Profile		
Ordered		Total
Value	bmi_cat	Frequency
1	3_obese	325
2	2_overwt	308
3	1_normal	308

Probabilities modeled are cumulated over the lower Ordered Values.

From the log file:

NOTE: PROC LOGISTIC is fitting the cumulative logit model. The probabilities modeled are summed over the responses having the lower Ordered Values in the Response Profile table.

34

Test of proportional odds assumption

SAS tests H_0 : odds are proportional, against a larger model with separate effects of age for each category-comparison.

In our example, this means two β values instead of one, so this larger model has 1 extra parameter and the test has 1 degree of freedom.

Score Test for the Proportional Odds Assumption

Chi-Square	DF	Pr > ChiSq
0.5562	1	0.4558

35

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	3_obese	1	-0.9162	0.0926	97.7973	<.0001
Intercept	2_overwt	1	0.4680	0.0886	27.8689	<.0001
age_category	old	1	0.5451	0.1210	20.2876	<.0001
age_category	young	0	0	.	.	.

Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits
age_category	old vs young	1.725	1.361 2.187

Age effect: the old have 1.7 times the odds of being obese compared to the young, and 1.7 times the odds of being overweight or obese.

Better: those 40–60 have 1.7 times higher odds of being in a heavier category than those 20–39.

Intercepts are α_2 and α_3 : not informative.

36

Why not just do separate logistic regressions for each cut-point?

Ordinal logistic regression also usually gives greater precision (more power): the standard error for the regression coefficient is smaller.

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
obese vs rest					
age_category old	1	0.4951	0.1382	12.8353	0.0003
obese+overweight vs rest					
age_category old	1	0.5996	0.1417	17.9064	<.0001
Proportional Odds					
age_category old	1	0.5451	0.1210	20.2876	<.0001

Often a good model check. Regression coefficient from proportional odds is essentially an average of the regression coefficients in the 2 logistic regression models, but they are quite close.

Proportional odds model with age group and gender

```
Proc Logistic descending data=mayod327.age_bmi_sample;
  class age_category gender /param=glm;
  model bmi_cat = age_category gender ;
```

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept 3_obese	1	-0.9528	0.1097	75.4200	<.0001
Intercept 2_overwt	1	0.4318	0.1058	16.6533	<.0001
age_category old	1	0.5440	0.1211	20.1875	<.0001
age_category young	0	0	.	.	.
gender female	1	0.0775	0.1204	0.4148	0.5195
gender male	0	0	.	.	.

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
age_category old vs young	1.723	1.359	2.184
gender female vs male	1.081	0.854	1.368

With 2 age categories and 2 genders, we have 4 subgroups. We assume that the odds ratios between any two are the same for all cumulative comparisons of categories.

Score Test for the Proportional Odds Assumption

Chi-Square	DF	Pr > ChiSq
13.2908	2	0.0013

It appears that this assumption fails for this data. What now?

39

Generalized logits model: unranked categories

Proportional odds makes odds from adjoining *ordered* categories.

If categories are not ordered, then proportional odds cannot be applied.

Generalized logits makes odds between one reference category and all the other categories.

Handles categories without order, eg. vanilla, strawberry, chocolate

```
Proc Logistic descending data=mayod327.age_bmi_sample;  
  class age_category gender / param=glm;  
  model bmi_cat = age_category gender / link=glogit ;
```

40

Default is to use the highest category as reference:

Ordered Value	bmi_cat	Total Frequency
1	3_obese	325
2	2_overwt	308
3	1_normal	308

Logits modeled use bmi_cat='1_normal' as the reference category.

Notice that degrees of freedom are twice as large as they should be:

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
age_category	2	20.6261	<.0001
gender	2	12.8702	0.0016

We are essentially fitting two separate models (normal vs overweight, normal vs obese). H_0 : reg coef for both models = 0

41

Here are the regression coefficients: note the doubling

Parameter	bmi_cat	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	3_obese	1	-0.3337	0.1394	5.7279	0.0167
Intercept	2_overwt	1	0.00496	0.1314	0.0014	0.9699
age_category old	3_obese	1	0.7278	0.1621	20.1676	<.0001
age_category old	2_overwt	1	0.4754	0.1643	8.3743	0.0038
age_category young	3_obese	0	0	.	.	.
age_category young	2_overwt	0	0	.	.	.
gender female	3_obese	1	0.0814	0.1613	0.2545	0.6139
gender female	2_overwt	1	-0.4576	0.1633	7.8491	0.0051
gender male	3_obese	0	0	.	.	.
gender male	2_overwt	0	0	.	.	.

If c response categories, then usual degrees of freedom are multiplied by $(c - 1)$.

42

Odds Ratio Estimates

Effect	bmi_cat	Point Estimate	95% Wald Confidence Limits
age_category old vs young	3_obese	2.071	1.507 2.845
age_category old vs young	2_overwt	1.609	1.166 2.220
gender female vs male	3_obese	1.085	0.791 1.488
gender female vs male	2_overwt	0.633	0.459 0.872

Averaging across genders, those over 40 have 2 times greater odds of being obese and 1.6 times greater odds of being overweight.

Averaging across ages, women have about half the men's odds of being overweight, but about the same odds for obesity.