

Lecture 21

1. Conditional logistic regression
2. Logistic ordinal regression
3. Proportional Odds Model
4. Generalized logits model

1

Matched-pair logistic regression

In a clinical study to compare a new treatment for a skin disease against placebo, investigators enrolled 2 patients at each of 79 clinics and randomly assigned one to each treatment (T or P).

Covariates: baseline condition, age and gender.

At the end of the study, each patient was rated as improved or not improved.

Obs	center	treatment	gender	age	improve	baseline_ score
1	1	T	F	27	0	1
2	1	P	F	32	0	2
3	41	T	F	13	1	2
4	41	P	M	22	0	3
5	2	T	F	41	1	3

(Source: Ch 10 of Stokes, Davis, Koch: *Categorical Data Analysis Using the SAS System, 2nd ed.*)

2

This is a stratified randomization, with clinic as stratum.

If the response improved were continuous (and normally distributed) then we could fit:

```
Proc GLM;  
  class clinic treatment gender;  
  model improved = clinic baseline_score age gender treatment;
```

This doesn't work in logistic regression, because the difference of the responses within pairs is either -1, 0, or 1.

As a result, we cannot estimate $79 - 1 = 78$ parameters for the clinics.

3

Conditional logistic regression

This adaptation of logistic regression uses the pairs where the responses were different, and avoids estimating the parameters for strata.

```
Proc Logistic descending ;  
  class center treatment(ref="P") gender(ref="F");  
  model improve = baseline_score treatment;  
  strata center;
```

The strata variable must be in the class statement.

4

The LOGISTIC Procedure

Conditional Analysis

Model Information

Data Set	WORK.STOKES_TRIAL
Response Variable	improve
Number of Response Levels	2
Number of Strata	79
Number of Uninformative Strata	25
Frequency Uninformative	50
Model	binary logit
Optimization Technique	Newton-Raphson ridge

Number of Observations Read	158
Number of Observations Used	158

5

Conditional Analysis

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
baseline_score	2.937	1.564	5.514
treatment T vs P	2.037	1.028	4.034

Those receiving the new treatment had twice the odds of improvement compared to placebo.

CLodds=PL not allowed with strata statement.

6

Recall the lymphedema trial: women in each group (exercise, control) were rated as symptomatic (yes/no) at baseline and study end.

Why can't we use conditional logistic regression, with woman as strata?

Because treatment is the same within women (same treatment at baseline and end), so we would not be able to compare treatments.

7

Ordinal logistic regression

In the NHANES 2004 example, we looked at the relation between the rate of obesity and age and gender. Obesity is a binary response, defined by $\text{BMI} \geq 30$.

However, there are other categories:

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

If we wish to examine rates of obesity and overweight in relation to age and gender, then we have three ordered categories:

Normal weight < Overweight < Obese

and our response takes one of three values.

8

This is an example of an **ordinal response**, in which the response variable is one of 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:

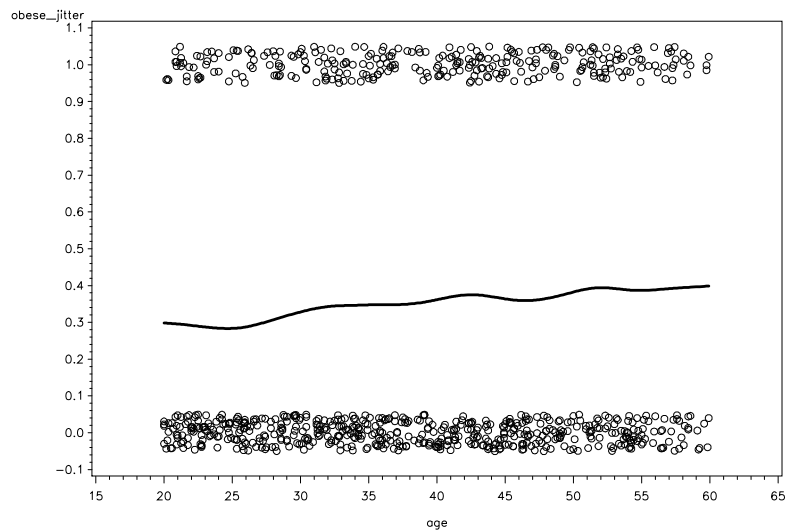
Worse < No Change < Recovered

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

Ordinal models use only the ranks of the categories.

9

In the NHANES 2004 logistic regression example, we looked at the relation between the rate of obesity and age and gender. There is a slow rise in obesity rate with age.



Make age into a predictor with categories:

"young" for those aged 20–39 years old, and "old" for those aged 40–59.

Proportional odds model for ordinal responses

The **proportional odds model** deals with ordinal categories by modeling **cumulative odds**:

1. Odds of being in the top category vs the rest: obese vs normal + overweight
2. Odds of being in the top two categories vs the rest: overweight or obese vs normal
3. Odds of being in the top three categories vs the rest, if there are > 3 categories, etc.

To define these odds we define cumulative probabilities:

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

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

11

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

We want to find the odds ratios for old to young of:

obesity

overweight + obesity

12

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

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

13

The responses in the proportional odds model are log odds, as they were in logistic regression:

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

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

$$\text{logit}(\theta_{h3}) = \log \left[\frac{\theta_{h3}}{1 - \theta_{h3}} \right]$$

$$\text{logit}(\theta_{h2}) = \log \left[\frac{\theta_{h2}}{1 - \theta_{h2}} \right]$$

14

The **Proportional Odds Model** fits one group effect with separate category intercepts:

$$\text{logit}(\theta_{hk}) = \alpha_k + x_h\beta$$

β estimates the covariate effect and $\{\alpha_k\}$ are category intercepts.

We are estimating an “average” effect (odds ratio) of age for both BMI cut-points: the ratio of the odds for someone 40+ of “being in a heavier category” to the odds for someone 20–39.

Does it make sense to average the observed odds ratios?

15

The proportional odds model assumes that the effect of the covariate (the odds ratio) is the same for all the cumulative odds:

old/young odds ratio for being *obese* = odds ratio for being *overweight or obese*

Odds ratios on the log scale:

$$\text{logit}(\theta_{23}) - \text{logit}(\theta_{13}) = (\alpha_3 + \beta) - \alpha_3 = \beta$$

$$\text{logit}(\theta_{22}) - \text{logit}(\theta_{12}) = (\alpha_2 + \beta) - \alpha_2 = \beta$$

Proc Logistic tests this assumption.

16

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, and the default for Proc Logistic when the response has > 2 levels is the proportional odds model.

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.

17

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.

SAS tests the null hypothesis that 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

18

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

The two intercepts are α_2 and α_3 : nuisance parameters.

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.

19

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

20

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

21

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?

Generalized logits model

This model fits log odds between a reference category and each other category:

$$\text{logit}_{hjk} = \log \left[\frac{\pi_{hjk}}{\pi_{hjr}} \right]$$

where π_{hjr} is the probability of the reference category.

The model fits separate effects of each predictor on each odds ratio:

$$\text{logit}_{hk} = \alpha_k + x_{hj}\beta_k$$

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

23

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.

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

40+ has 2 times greater odds of being obese and 1.6 times greater odds of being overweight. Women have about half the men's odds of being overweight. No difference for obesity.

24

References on logistic and ordinal regression:

Stokes, Davis, Koch (2000) *Categorical Data Analysis Using the SAS System, Second Edition*

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

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