Topics in Biostatistics Categorical Data Analysis and Logistic Regression, part 2

B. Rosner, 5/21/18

Outline

- 1. Testing for effect modification in logistic regression analyses
- 2. Conditional logistic regression
- 3. Other types of logistic regression

Aspirin and MI in the Physicians' Health Study

- We consider another dataset consisting of men in the Physicians' Health Study (PHS).

- The PHS was a randomized trial where approximately 22,000 men were randomized to either aspirin (ASA) or aspirin placebo in 1982 and were followed until 1990 for the development of myocardial infarction(MI) (a type of heart attack).

Aspirin and MI in the Physicians' Health Study (cont.)

- Blocking was used to balance treatment assignments within age groups.

- The results stratified by age are given in the following table:

Aspirin and MI in the Physicians' Health Study (cont.)

	Aspirin group		**Placebo group**	
Age (years)	**N**	**N with MI**	**N**	**N with MI**
40-49	4527	27	4524	24
50-59	3725	51	3725	87
60-69	2045	39	2045	84
70-84	740	22	740	44
Total	11037	139	11034	239

Effect modification- introduction

- One assumption made in the primary analyses for the Physicians' Health Study is that the effect of ASA is the same for all age groups.

- However, it seems from the table (see previous slide) that ASA was more effective in preventing MI for older men vs. younger men.

Aspirin and MI in the Physicians' Health Study (cont.)

- We can assess this by fitting the model:

Effect modification- Physicians' Health Study

$$\log it(p_i) = \alpha + \beta_1(x_{1i} - 62) + \beta_2 x_{2i} + \beta_3 x_{2i}(x_{1i} - 62)$$

The terms x_{1i}, x_{2i} are referred to as main effects.

x_{1i} = age, x_{2i} = 1 if ASA, = 0 else.

The term $x_{2i}(x_{1i} - 62)$ is referred to as an interaction effect.

Interpretation of Main Effects in a Logistic Regression Model with interaction terms

- In logistic regression, the main effects have a different interpretation once an interaction term is in the model.
- In the previous model,

$\beta_1 =$ effect of age when $x_{2i} = 0$ (i.e. in the placebo group),

$\beta_2 =$ effect of ASA when age = 62.

Effect modification – interpretation of parameters

- The effect of ASA on MI at age x_1 is estimated by $\beta_2 + \beta_3(x_1\text{-}62)$.
- Thus, if $\beta_3 \neq 0$, then the effect of MI varies with age.

- Specifically, β_2 = ln(odds) of ASA for MI when age = 62 (the approximate mean age).

Effect modification – interpretation of parameters (cont.)

- The reason for approximate mean-centering is to make the interpretation of the main effects meaningful.

- If we had not mean-centered age, then β_2 = ln(odds) of ASA for MI when age = 0, which is not very meaningful.

- We present the Stata analyses of effect modification in the next slide.

Stata analyses of effect modification in PHS study

```
. gen age_62 = age-62

. gen age_asa=asa*age_62

. logistic mi age_62 asa age_asa [fweight=freq]

Logistic regression                                     Number of obs   =      22071
                                                        LR chi2(3)      =     189.07
                                                        Prob > chi2     =     0.0000
Log likelihood = -1817.5845                             Pseudo R2       =     0.0494

------------------------------------------------------------------------------
          mi | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      age_62 |   1.071163   .0064502    11.42   0.000     1.058595     1.08388
         asa |    .554374   .0610914    -5.35   0.000     .4466851    .6880252
     age_asa |   .9803535   .0097263    -2.00   0.046     .9614745    .9996032
------------------------------------------------------------------------------
```

Stata analyses of effect modification in PHS study (cont.)

- We see that β_3 is statistically significant (OR = 0.98, 95% CI = 0.96-1.00, p = 0.046).

- Thus, there is significant evidence for interaction.

Stata analyses of effect modification in PHS study (cont.)

- The estimated OR for ASA is given by:
- $$\exp\left[\beta_2 + \beta_3(age - 62)\right]$$
  $$= \exp(\beta_2) \text{ x } \exp[\beta_3(age - 62)]$$
  $$= 0.554\left[\exp(\beta_3)\right]^{age-62}$$
  $$= 0.554 \text{ x } 0.980^{\text{age-62}}.$$

- We have evaluated this expression for age 45, 55, 65 and 75 as follows:

Stata analyses of effect modification in PHS study (cont.)

Age	Estimated OR for ASA
45	0.78
55	0.64
65	0.52
75	0.43

Stata analyses of effect modification in PHS study (cont.)

- Thus, the effect of ASA is stronger for older men than for younger men.

CONDITIONAL LOGISTIC REGRESSION

Nurses' Health Study Blood Sub-study

- A sub-study of approximately 32,826 women in the Nurses' Health Study (NHS) provided a blood sample in 1989/1990.

- The blood has been stored in freezers and used in nested case-control studies.

- One case-control study was conducted (Missmer, et al, JNCI,1998) relating serum estradiol to breast cancer risk.

Nurses' Health Study Blood Sub-study (cont.)

- 164 breast cancer cases (occurring from the time of the blood draw until 2000) and 346 controls (1 or 2 matched controls for each case, who did not have breast cancer at the time of diagnosis of the case) were matched to each case by age, fasting status and PMH use at the time of the blood draw.

Nurses' Health Study Blood Sub-study (cont.)

- In addition, the matched sets were analyzed in the same batch.

- Since there is often substantial batch-to-batch assay variability, it is important that the matching be taken into account in the analysis.

Logistic Regression for Matched Sets

- Suppose there are m matched sets.
- Suppose also that we have n_{1i} cases and n_{2i} controls in the ith matched set.

- Let p_{ij} = probability of disease for the jth subject in the ith matched set.

- We assume a logistic model of the form:

Logistic Regression for Matched Sets (cont.)

$$\text{logit}(p_{ij}) = \alpha_i + \sum_{k=1}^{K} \beta_k x_{ijk},$$

where

x_{ijk} = value of the kth covariate measured on the jth subject in the ith matched set.

Logistic Regression for Matched Sets (cont.)

- Since the matched sets are usually small, it is virtually impossible to directly estimate the α_i.
- Thus, it is impossible to estimate the absolute probability of disease.
- Instead, we use a conditional approach to estimate the other regression parameters (i.e., β_k, k=1,…, K).

Conditional Logistic Regression

- Suppose for the sake of simplicity that there is 1 case and 1 control for each matched set and that subject 1 is a case and subject 2 is a control.

- Let Y_{ij} = 1 if the jth woman in the ith matched pair is a case, = 0 if a control.

- From the logistic model we have:

Conditional Logistic Regression (cont.)

$$\Pr(Y_{i1} = 1 \text{ and } Y_{i2} = 0) =$$

$$\frac{\exp(\alpha_i + \sum_{k=1}^{K} \beta_k x_{i1k})}{[1 + \exp(\alpha_i + \sum_{k=1}^{K} \beta_k x_{i1k})]} \text{ x } \frac{1}{[1 + \exp(\alpha_i + \sum_{k=1}^{K} \beta_k x_{i2k})]}$$

$$\Pr(Y_{i1} = 0 \text{ and } Y_{i2} = 1) =$$

$$\frac{1}{[1 + \exp(\alpha_i + \sum_{k=1}^{K} \beta_k x_{i1k})]} \text{ x } \frac{\exp(\alpha_i + \sum_{k=1}^{K} \beta_k x_{i2k})}{[1 + \exp(\alpha_i + \sum_{k=1}^{K} \beta_k x_{i2k})]}$$

Conditional Logistic Regression (cont.)

- We now consider the conditional probability that subject 1 is a case given that exactly 1 out of 2 subjects in the matched set is a case given by:

Conditional Logistic Regression (cont.)

$$\Pr(Y_{i1}=1 \text{ and } \mathrm{Y_{i2}}=0 \text{ given } 1 \text{ case in a matched set})$$

$$\equiv \mathrm{L_i} = \frac{\Pr(Y_{i1}=1 \text{ and } \mathrm{Y_{i2}}=0)}{\Pr(Y_{i1}=1 \text{ and } \mathrm{Y_{i2}}=0)+\Pr(Y_{i1}=0 \text{ and } \mathrm{Y_{i2}}=1)}$$

$$= \frac{\exp(\sum_{k=1}^{K}\beta_k x_{i1k})}{\exp(\sum_{k=1}^{K}\beta_k x_{i1k})+\exp(\sum_{k=1}^{K}\beta_k x_{i2k})}$$

Conditional Logistic Regression (cont.)

- L_i is the contribution to the conditional likelihood for the ith matched set.

- Notice that the effect of the matched set (α_i) does not appear in L_i.

- Also, if the x's are the same for both members of the matched pair, then $L_i=1/2$. Thus, the matched pair is not informative and is not used in the analysis.

Conditional Logistic Regression (cont.)

- The overall conditional likelihood is given by:
- $L = L_1 \text{ x } L_2 \text{ x}\ldots \text{ x } L_m$.

- We then find the MLE's that maximize the conditional likelihood L.

Conditional Logistic Regression (cont.)

- This approach can be generalized to allow for any number of cases (n_{1i}) and controls(n_{2i}) in a matched set.

- Also, different matched sets do not have to have the same number of cases and controls or even the same total number of subjects.

- This model is referred to as a conditional logistic regression model.

Interpretation of Parameters in Conditional Logistic Regression

- Suppose we have two subjects 1 and 2 in a matched set who differ by 1 unit for the 1st covariate and are the same for all other covariates. From the logistic model we have:

Conditional Logistic Regression (cont.)

$$\text{logit}(p_{i1}) = \alpha_i + \beta_1(x_1 + 1) + \sum_{k=2}^{K} \beta_k x_{i1k},$$

$$\text{logit}(p_{i2}) = \alpha_i + \beta_1(x_1) + \sum_{k=2}^{K} \beta_k x_{i1k}.$$

Thus,

$$\text{logit}(p_{i1}) - \text{logit}(p_{i2}) = \beta_1,$$

or

$$OR_{1 \text{ vs. } 2} = \exp(\beta_1).$$

Conditional Logistic Regression (cont.)

- Thus, the odds ratios in conditional logistic regression have a conditional interpretation.
- They quantify the OR between disease and exposure for two subjects within a matched set.
- This is different from unconditional logistic regression, where all the subjects are assumed to be independent of each other.

Analysis of NHS Blood Study data

- Some descriptive data for the study are given in the next slide.

Analysis of NHS Blood Study data (cont.)

```
. summarize case currentpmh ageblood estradiol

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        case |       510    .3215686     .467537         0          1
  currentpmh |       510    .1509804    .3583813         0          1
    ageblood |       510    60.96863    4.989478        45         69
    estradiol |       510    8.907843    7.741739         2         85
```

```
. table case
----------------------
     case |      Freq.
----------+-----------
        0 |         346
        1 |         164
----------------------
```

Analysis of NHS Blood Study data (cont.)

We see that there were a total of 510 women in the study of whom 164 were cases and 346 were controls.

- A listing of part of the data are given on the next slide.

Analysis of NHS Blood Study data (cont.)

	id	matchid	case	curpmh	age	estradiol
19.	100241	107261	0	0	65	11
20.	212974	107261	0	0	64	8
21.	108215	108215	1	0	58	8
22.	106487	108946	0	0	62	6
23.	108946	108946	1	0	61	4
24.	116697	108946	0	0	58	9
25.	102266	109861	0	1	64	6
26.	103214	109861	0	0	65	5
27.	109861	109861	1	1	66	5
28.	100696	110294	0	1	66	3
29.	127187	110294	0	0	68	6

Analysis of NHS Blood Study data (cont.)

- We see that each subject has an individual id (id) and a matchid which identifies the matched set which the subject belongs to.

- A matched set is only useful for the analysis if there is at least one case and at least one control.

Analysis of NHS Blood Study data (cont.)

- Thus, matchid 107261 is not useful since there were no cases , while matchid 108215 is not useful because there were no controls.

- However, matchid 108946 is useful because it has 1 case and 2 controls.

- Some of the subjects originally in the blood study were not included due to missing data on other covariates, which were used in the primary analyses.

Analysis of NHS Blood Study data (cont.)

- We also see that the mean estradiol = 8.9 with sd = 7.7, which indicates that the distribution of estradiol is quite skewed.

- We chose to create a new variable (ln_estradiol) = ln(estradiol) for the purpose of analysis.

- We now will fit the conditional logistic regression model given by:

Analysis of NHS Blood Study data (cont.)

$$\text{logit}(p_{ij}) = \alpha_i + \beta_1 \text{ageblood}_{ij} + \beta_2 \text{currentpmh}_{ij} + \beta_3 \ln(\text{estradiol}_{ij}),$$

where

$p_{ij} = \Pr(\text{jth subject in the ith matched pair is a case})$,

$\text{ageblood}_{ij} = \text{age at the blood draw for the jth subject in the ith matched pair}$

$\text{currentpmh}_{ij} = 1$ if the jth subject in the ith matched pair used postmenopausal hormones at the time of the blood draw, $= 0$ otherwise

$\ln\_\text{estradiol}_{ij} = \ln(\text{estradiol for the jth subject in the ith matched pair})$.

Analysis of NHS Blood Study data (cont.)

- The results are obtained using the Stata clogit program as follows:

Analysis of NHS Blood Study data (cont.)

```
. clogit case currentpmh ageblood ln_estradiol, strata(matchid)
note: multiple positive outcomes within groups encountered.
note: 82 groups (126 obs) dropped due to all positive or
      all negative outcomes.

Iteration 0:   log likelihood = -132.67886
Iteration 1:   log likelihood = -132.50721
Iteration 2:   log likelihood = -132.50714
Iteration 3:   log likelihood = -132.50714

Conditional (fixed-effects) logistic regression   Number of obs   =        384
                                                  LR chi2(3)      =      12.57
                                                  Prob > chi2     =     0.0057
Log likelihood = -132.50714                       Pseudo R2       =     0.0453
```

Analysis of NHS Blood Study data (cont.)

case	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
currentpmh	.237519	.3007713	0.79	0.430	-.351982	.82702
ageblood	-.0248404	.1526609	-0.16	0.871	-.3240501	.2743694
ln_estradiol	.6826214	.2054062	3.32	0.001	.2800327	1.08521

Analysis of NHS Blood Study data (cont.)

- We see that 82 matched pairs(126 women) were not used in the analysis because there were either 0 cases or 0 controls in the matched pair.

- We also see that ln_estradiol is significant with p-value = 0.001, while the other 2 variables in the model are not significant.

Analysis of NHS Blood Study data (cont.)

- To assess the magnitude of effect of the variables in terms of odds ratios, we can specify
-
- `.clogit, or`

- with results as follows:

Analysis of NHS Blood Study data (cont.)

```
. clogit, or

Conditional (fixed-effects) logistic regression   Number of obs   =        384
                                                  LR chi2(3)      =      12.57
                                                  Prob > chi2     =     0.0057
Log likelihood = -132.50714                       Pseudo R2       =     0.0453

------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  currentpmh |   1.268099   .3814079     0.79   0.430     .7032928    2.286495
    ageblood |   .9754656   .1489154    -0.16   0.871      .723214    1.315701
ln_estradiol |   1.979059   .4065109     3.32   0.001     1.323173    2.960062
------------------------------------------------------------------------------
```

Analysis of NHS Blood Study data (cont.)

- We see that the OR for a 1 unit increase in ln(estradiol) is 1.98 with 95% CI = (1.32, 2.96).

Analysis of NHS Blood Study data (cont.)

- 1. Thus, suppose we have 2 women in the same matched set, one of whom has a ln(estradiol) 1 unit higher than the other (i.e., approximately 2.7 times as high).

- 2. The woman with the higher estradiol has a 2-fold higher odds of having breast cancer vs. the woman with the lower estradiol, holding the other variables (i.e., age, current pmh use) constant.

Analysis of NHS Blood Study data (cont.)

- For comparison, we have also used unconditional logistic regression on this dataset based on 510 women, ignoring the matching. The results are given on the next slide.

Analysis of NHS Blood Study data (cont.)

```
. logistic case currentpmh ageblood ln_estradiol

Logistic regression                                     Number of obs   =        510
                                                        LR chi2(3)      =      14.14
                                                        Prob > chi2     =     0.0027
Log likelihood = -313.23517                             Pseudo R2       =     0.0221

------------------------------------------------------------------------------
        case | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  currentpmh |   1.405057   .3758035     1.27   0.204     .8318163    2.373343
    ageblood |   1.008835   .0199317     0.45   0.656     .9705161    1.048666
ln_estradiol |   1.825742   .3105155     3.54   0.000     1.308188    2.548054
------------------------------------------------------------------------------
```

Analysis of NHS Blood Study data (cont.)

```
. logit

Logistic regression                                     Number of obs   =        510
                                                        LR chi2(3)      =      14.14
                                                        Prob > chi2     =     0.0027
Log likelihood = -313.23517                             Pseudo R2       =     0.0221

------------------------------------------------------------------------------
        case |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  currentpmh |    .340078   .2674649     1.27   0.204    -.1841437    .8642997
    ageblood |    .008796   .0197571     0.45   0.656    -.0299273    .0475193
ln_estradiol |   .6019865   .1700763     3.54   0.000      .268643    .9353299
       _cons |  -2.554535   1.308599    -1.95   0.051    -5.119343    .0102725
------------------------------------------------------------------------------
```

Analysis of NHS Blood Study data (cont.)

- If we compare slides 44 and 52, we see that the standard errors of the regression coefficients are smaller for the unconditional logistic regression compared with the conditional logistic regression, especially for the age at blood draw variable.

Analysis of NHS Blood Study data (cont.)

- This reflects the correlation induced by multiple subjects in the same matched set.

- The unconditional analysis assumes that the 510 sampling units are independent, which they probably are not.

- Another option is to include indicator variables for batch and use unconditional logistic regression. It is difficult to express results in terms of absolute levels of estradiol in this case.

Extensions of Logistic Regression

- 1. Correlated data
- (a) An assumption of ordinary logistic regression is that units of analysis are independent.
- (b) This assumption is violated if there are multiple siblings in a family or paired organ systems (e.g., eyes, ears) and the (eye,ear) is the unit of analysis, or multiple days of observation for a subject.

Extensions of Logistic Regression (cont.)

- (c) One could use conditional logistic regression to analyze the data, but effects are based on differences in exposure within a cluster.
- (d) Another approach is to use generalized estimating equation (GEE) methods to perform logistic regression but adjust se's for correlation (e.g., proc genmod of SAS or xtgee of Stata).

Dental Study - Design

- Longitudinal study of caries lesions on exposed roots of teeth
- 40 chronically ill subjects were followed for the development of root lesions over a 1 year period
- There were 11 males and 29 females in the dataset.
- 126 tooth surfaces were assessed among the 40 subjects.

Dental Study – Design (cont.)

- Outcome variable: incidence of caries lesions on a tooth surface over a 1 year period
- Independent variables: age, gender
- Issue: different tooth surfaces in the same subject are not independent
- Method of Analysis: Can use PROC GENMOD of SAS to perform GEE to analyze these data.

Dental Study - Model

- The following model was used:

$$\log it(p_{ij}) = \alpha + \beta_1 age + \beta_2 male,$$

$$Corr(p_{ij_1}, p_{ij_2}) = \rho,$$

where

p_{ij} = probability of developing caries lesions on the jth tooth surface for the ith subject,

ρ = correlation of presence of caries lesions on two surfaces from the same subject.

This is called an exchangeable correlation structure.

Dental Study - Results

- We have fit this model with PROC GENMOD of SAS with the following results:

Parameter	Beta	se	z	p-value	OR (95% CI)
Intercept	-3.0538	2.5527			
Age (10 yrs)	0.057	0.361	0.16	0.88	1.1 (0.5, 10.2)
Male gender	1.4569	0.7505	1.94	0.052	4.3 (1.0, 18.7)

- Exchangeable correlation = 0.117

Dental Study – Results (cont.)

- Thus, we see that age is not associated with incidence of caries lesions, while male gender is borderline significant (OR = 4.3, 95% CI = 1.0-18.7).

Comparison of approaches for handling correlated data

- The GEE approach to handling correlated data has the advantage that the odds ratios have a marginal interpretation, i.e., they quantify exposure-disease relationships over the entire population.
- Converesly, the conditional logistic regression approach provides odds ratios that have a conditional interpretation (i.e., within a matched set).

- Depending on the study design, one approach may have advantages over the other.

Extensions of Logistic Regression (cont.)

- 2. Nominal categorical data with more than 2 categories
- (a) For example, in a breast cancer setting we might have control subjects and two case groups of ER+ and ER- breast cancer.
- (b) we can use polytomous logistic regression (PLR) to estimate different odds ratios for ER+ breast cancer vs. control and ER- breast cancer vs. control
- (c) proc logistic of SAS with the generalized logit option or the mlogit command of Stata can implement these methods).

Extensions of Logistic Regression (cont.)

- 3. Ordered categorical data with more than 2 categories
- (a) In some datasets, there are more than 2 categories, but the categories can be ordered.

- (b) For example, we might perform an analysis among breast cancer cases and compare risk factors according to the size of the tumor (small = < 2 cm; medium = 2.0-3.9 cm; large = >= 4.0 cm)

Extensions of Logistic Regression (cont.)

(c) Ordinal logistic regression can be used to analyze the data where

OR = comparison of odds of increasing 1 category of the outcome given a 1 unit change in exposure (for example comparing a never HRT user vs. a current HRT user).

This OR is assumed to be the same comparing small to medium and medium to large tumors in the previous example.

Extensions of Logistic Regression (cont.)

- (d) Proc logistic of SAS or the ologit command of Stata can be used to analyze this type of data.

Summary

- 1. We have discussed methods of logistic regression in these 2 lectures.

- 2. Ordinary logistic regression can be used in datasets with independent observations and is a powerful method to relate a binary outcome variable to a combination of

Summary (cont.)

- (a) categorical risk factors with 2 or more categories
- (b) continuous risk factors,
- without stratifying the dataset by risk factors and losing power.
- 3. Both main effects of single risk factors and interaction effects between risk factors can be considered.

Summary (cont.)

- 4. Conditional logistic regression is an extension of ordinary logistic regression to the setting of matched sets, where effects are interpreted within a matched set rather than in the overall dataset. These methods inherently control for the correlation within matched sets.

- 5. If matching effects are very strong (e.g., twin pairs) the conditional logistic regression approach may be preferable.

Summary (cont.)

- 6. These methods were illustrated using Stata commands, but are also available in SAS using proc logistic(for ordinary logistic regression) and proc phreg (for conditional logistic regression).
- 6. Extensions of these methods exist for
- (a) correlated data
- (b) categorical outcome data with more than 2 categories