Lecture 15 (Part 1): Logistic Regression & Common Odds Ratios

Dipankar Bandyopadhyay, Ph.D.

BMTRY 711: Analysis of Categorical Data Spring 2011

Division of Biostatistics and Epidemiology

Medical University of South Carolina

TABLES IN 3 DIMENSIONS–Using Logistic Regression

- Previously, we examined higher order contingency tables
- We were concerned with partial tables and conditional associations
- In most problems, we consider one variable the outcome, and all others as covariates.
- In the example we will study, BIRTH OUTCOME will be considered as the outcome of interest, and CARE and CLINIC as predictors or covariates.
- We are mainly interested in estimating a common partial odds ratio between two variables (OUTCOME VS CARE), conditional or controlling for a third variable (CLINIC).

Study Data

CLINIC=1 OUTCOME

CARE	died	lived	Total
less	3	176	179
more	4	293	297
Total	7	469	476

CLINIC=2 OUTCOME

CARE	died	lived	Total
less	17	197	214
more	2	23	25
Total	19	220	239

Interpretation

- With the tables constructed as presented, we are interested in the ODDS of a poor birth outcome (fetal death) as a function of care
- For Clinic 1: OR = 1.2. Accordingly, the odds of a poor delivery (death) are 1.24 times higher in mothers that receive less prenatal care than those mothers that receive “more” (regular checkups, fetal heart monitoring, kick counts, gestational diabetes screening etc).
- For Clinic 2: OR = 1.0 (no association)
- We will explore various methods to estimate the “common” odds ratio for this data

- Suppose $W = CLINIC$, $X = CARE$, and $Y = OUTCOME$.
  Let

$$Y_{jk\ell} = \text{number of subjects with } W = j, X = k, Y = \ell,$$

  and $m_{jk\ell} = E(Y_{jk\ell})$.

- We are going to explore the use of logistic regression to calculate the conditional associations while thinking of BIRTH OUTCOME as the outcome, and CARE and the CLINIC as covariates.
- Suppose

$$Y = \begin{cases} 1 \text{ if died} \\ 0 \text{ if lived} \end{cases}.$$

- We are interested in modelling

$$P[Y = 1 | W = j, X = k] = p_{jk}$$

- Now, in the notation of the $(2 \times 2 \times 2)$ table, the CARE by CLINIC margins $n_{jk} = y_{jk+}$ are fixed (either by design, or conditioning). In particular, each row for the two clinic (2×2) are fixed.
- Also,

$$Y_{jk1} = \#\text{ died when CLINIC=}j\text{ and CARE=}k$$

- For ease of notation, we drop the last subscript 1, to give

$$Y_{jk} \sim Bin(n_{jk}, p_{jk}) \qquad j, k = 1, 2,$$

  which are 4 independent binomials.
- In general, the likelihood is the product of 4 independent binomials (the 4 CARE by CLINIC combinations):

$$\prod_{j=1}^{2} \prod_{k=1}^{2} \left(\begin{array}{c} n_{jk} \\ y_{jk} \end{array} \right) p_{jk}^{y_{jk}} (1 - p_{jk})^{n_{jk} - y_{jk}}$$

- You then use maximum likelihood to estimate the parameters of the model with SAS or STATA.

- The logistic regression model is

$$\text{logit}\{P[Y=1|W=w, X=x]\} = \beta_0 + \alpha w + \beta x,$$

where

$$w = \begin{cases} 1 \text{ if CLINIC=1} \\ 0 \text{ if CLINIC=2} \end{cases},$$

and

$$x = \begin{cases} 1 \text{ if CARE = less} \\ 0 \text{ if CARE = more} \end{cases},$$

- Think of α as a nuisance parameter
- We are primarily interested in β, the log-odds ratio of a death given less care

In other words, plugging in the four possible values of (W, X)

- 1. For CLINIC = 1, CARE = LESS: $(W = 1, X = 1)$

$$\text{logit}\{P[Y = 1|w = 1, x = 1]\} = \beta_0 + \alpha + \beta$$

- 2. For CLINIC = 1, CARE = MORE: $(W = 1, X = 0)$

$$\text{logit}\{P[Y = 1|w = 1, x = 0]\} = \beta_0 + \alpha$$

- 3. For CLINIC = 2, CARE = LESS: $(W = 0, X = 1)$

$$\text{logit}\{P[Y = 1|w = 0, x = 1]\} = \beta_0 + \beta$$

- 4. For CLINIC = 2, CARE = MORE: $(W = 0, X = 0)$

$$\text{logit}\{P[Y = 1|w = 0, x = 0]\} = \beta_0$$

- In this model, the log-odds ratio between X and Y, controlling for $W = w$ is

$$\log\left(\frac{P[Y=1|w,x=1]}{1-P[Y=1|w,x=1]}\right) - \log\left(\frac{P[Y=1|w,x=0}{1-P[Y=1|w,x=0}\right) =$$

$$\text{logit}\{P[Y=1|w,x=1\} - \text{logit}\{P[Y=1|w,x=0\} =$$

$$[\beta_0 + \alpha w + \beta(1)] - [\beta_0 + \alpha w + \beta(0)] = \beta$$

- This logistic model says that there is a common odds ratio between CARE (X) and OUTCOME (Y) controlling for CLINIC (W), which equals

$$\exp(\beta) = OR_w^{XY.W},$$

- Also, you can show that this model says there is a common odds ratio between CLINIC (W) and OUTCOME (Y) controlling for CARE (X), which equals

$$\exp(\alpha) = (OR_k^{WY.X}).$$

- $X = k$ where k indexes the site

SAS Proc Logistic

```
data one;
input clinic care out count;
  clinic = 2 - clinic; /* To code the regression model with */
  care = 2 - care;     /* appropriate dummy codes           */
  out = 2 - out;
cards;
1 1 1   3 /* out: 2 - 1 = 1 => success */
1 1 2 176 /* out: 2 - 2 = 0 => failure */
1 2 1   4
1 2 2 293
2 1 1  17
2 1 2 197
2 2 1   2
2 2 2  23
;

proc logistic descending data=one;
 model out = clinic care ;
 freq count;
run;
```

Selected output

```
The LOGISTIC Procedure

              Model Information

Data Set                          WORK.ONE
Response Variable                 out
Number of Response Levels         2
Frequency Variable                count
Model                             binary logit
Optimization Technique            Fisher's scoring **** Note Fisher's Scoring


Number of Observations Read           8
Number of Observations Used           8
Sum of Frequencies Read             715 **** Should be your N
Sum of Frequencies Used             715
```

Response Profile

Ordered Value	out	Total Frequency
1	1	26
2	0	689

Probability modeled is out=1. **** Always check this ****

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.
**** Our iterative process converged

```
/* SELECTED OUTPUT */

            Analysis of Maximum Likelihood Estimates

                                 Standard          Wald
Parameter    DF    Estimate       Error    Chi-Square    Pr > ChiSq

Intercept     1     -2.5485      0.5606       20.6644        <.0001
clinic        1     -1.6991      0.5307       10.2520        0.0014
care          1      0.1104      0.5610        0.0387        0.8440
```

The parameters in the logistic regression model correspond to the (conditional) association between the response (Y) and the particular covariate (X) or (W).

Here, the level of care appears to be less important. However, where you receive the care maybe of interest to patients.

Interpretation

- The logistic regression estimate of the 'common odds ratio' between CARE (X) and OUTCOME (Y) controlling for CLINIC (W) is

$$\exp(\hat{\boldsymbol{\beta}}) = \exp(.1104) = 1.1167,$$

with a 95% confidence interval,

`[.3719, 3.3533],`

which contains 1. Thus, babies who have 'LESS' care have 1.1 times the odds of dying than babies who have 'MORE' care; however, this association is not statistically significant at $\alpha = 0.05$.

- A test for conditional independence of CARE (X) and OUTCOME (Y) given CLINIC (W),

$$H_0 : \beta = 0,$$

can be performed using the likelihood ratio, the WALD statistic, and the SCORE.

Estimating a Common OR from J, (2×2) tables

Randomized trial to see if Salk Vaccine is effective in preventing paralysis

Age	Salk Vaccine	Paralysis: No	Paralysis: Yes
0-4	Yes	20	14
	No	10	24
5-9	Yes	15	12
	No	3	15
10-14	Yes	3	2
	No	3	2
15-19	Yes	12	3
	No	7	5
20+	Yes	1	0
	No	3	2

Question: When controlled for the effects of age, is Salk vaccine effective at reducing the rate of paralysis from polio?

- For now, in this dataset, you assume, or have prior information that there is a common odds ratio among the J tables.
- You want to estimate a common odds ratio among the J tables, and you also want to see if, controlling or stratifying on age (W), paralysis (Y) and taking vaccine (X) are independent.
- It is easier to directly think of this table in terms of a logistic regression model.
- Again, we are interested in modelling

$$P[Y = 1|W = j, X = k] = p_{jk}$$

where

$$Y = \begin{cases} 1 \text{ if not paralyzed} \\ 0 \text{ if paralyzed} \end{cases} .$$

- Now, the AGE by VACCINE margins $n_{jk} = y_{jk+}$ (rows in the table) are considered fixed.
- Also,

$$\begin{aligned} Y_{jk1} \quad &= \quad \# \text{ not paralyzed in the} \\ & \qquad j^{th} \text{ age group and } k^{th} \text{ vaccine group} \end{aligned}$$

- For ease of notation, we again drop the last subscript, to give

$$Y_{jk} \sim Bin(n_{jk}, p_{jk}) \qquad j = 1, , , , J, k = 1, 2,$$

which are independent binomials.

- In general, the likelihood is the product of $J \times 2$ independent binomials (the J strata by covariate VACCINE combinations):

$$\prod_{j=1}^{J} \prod_{k=1}^{2} \left(\begin{array}{c} n_{jk} \\ y_{jk} \end{array} \right) p_{jk}^{y_{jk}} (1 - p_{jk})^{n_{jk} - y_{jk}}$$

- The logistic regression model is

$$\text{logit}\{P[Y = 1 | W = j, X = x]\} = \beta_0 + \alpha_j + \beta x$$

$$= \beta_0 + \alpha_1 w_1 + \alpha_2 w_2 + ... + \alpha_{J-1} w_{J-1} + \beta x$$

where we constrain $\alpha_J = 0$, and

$$Y = \begin{cases} \text{1 if not paralyzed} \\ \text{0 if paralyzed} \end{cases} . \quad x = \begin{cases} \text{1 if VACCINE=YES} \\ \text{0 if VACCINE=NO} \end{cases} , \quad w_j = \begin{cases} \text{1 if } W = j \\ \text{0 if } W \neq j \end{cases}$$

- In this model, the log-odds ratio between VACCINE (X) and PARALYSIS (Y), given AGE$= W = j$, is

$$\log\left(\frac{P[Y=1|W=j,x=1]}{1-P[Y=1|W=j,x=1]}\right) - \log\left(\frac{P[Y=1|W=j,x=0}{1-P[Y=1|W=j,x=0}\right) =$$

$$\text{logit}\{P[Y=1|W=j,x=1\}-$$
$$\text{logit}\{P[Y=1|W=j,x=0\} =$$

$$[\beta_0 + \alpha_j + \beta(1)] - [\beta_0 + \alpha_j + \beta(0)] = \beta$$

- Then, in this model, the conditional odds ratio between VACCINE (X) and PARALYSIS (Y), given AGE (W) is

$$\exp(\beta) = OR_j^{XY.W}$$

- The logistic regression estimate of the 'common odds ratio' between X and Y given W is

$$\exp(\hat{\boldsymbol{\beta}})$$

- A test for conditional independence

$$H_0 : \beta = 0$$

can be performed using the likelihood ratio, the WALD statistic, and the SCORE.

Data Analysis of PARALYSIS DATA

- The logistic regression estimate of the 'common odds ratio' between VACCINE (X) and PARALYSIS (Y) controlling for AGE (W) is

$$\exp(\hat{\boldsymbol{\beta}}) = \exp(1.2830) = 3.607,$$

with a 95% confidence interval,

`[1.791,7.266]`

which does not contain 1. Thus, individuals who take the vaccine have 3.6 times the odds of not getting POLIO than individuals who do not take the vaccine.

- A test for conditional independence $H_0 : \beta = 0$ using the WALD statistic rejects the null,

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square
VAC	1	1.2830	0.3573	12.8949	0.0003

- Thus, even controlling for AGE (W), VACCINE (X) and PARALYSIS (Y) do not appear to be independent.

Using SAS

```
data one;
/* vac = 1 = yes, 0 = no */
/* y = # not paralyzed,  n = sample size */

input age vac y n;
cards;
1  1  20 34
1  0  10 34
2  1  15 27
2  0   3 18
3  1   3  5
3  0   3  5
4  1  12 15
4  0   7 12
5  1   1  1
5  0   3  5
;
proc logistic data=one;
 class age;
 model y/n  = vac age ;
run;
```

```
/* SELECTED OUTPUT */

             Analysis of Maximum Likelihood Estimates

                                    Standard          Wald
Parameter        DF    Estimate       Error    Chi-Square    Pr > ChiSq

Intercept         1     -0.3122      0.2965        1.1082        0.2925
vac               1      1.2830      0.3573       12.8948        0.0003
age       1       1     -0.5907      0.3236        3.3317        0.0680
age       2       1     -0.9106      0.3634        6.2790        0.0122
age       3       1      0.1185      0.5822        0.0415        0.8387
age       4       1      0.5461      0.4240        1.6590        0.1977

               Odds Ratio Estimates

                   Point          95% Wald
Effect          Estimate      Confidence Limits

vac                3.607       1.791       7.266 ***** What is of interest
age 1 vs 5         0.240       0.039       1.489
age 2 vs 5         0.174       0.027       1.143
age 3 vs 5         0.488       0.055       4.360
age 4 vs 5         0.748       0.107       5.223
```

Other Test Statistics

The Cochran, Mantel-Haenzel Test

- Mantel-Haenszel think of the data as arising from J, (2×2) tables:

Stratum j (of W)

Variable (X) \ Variable (Y)	1	2	
1	Y_{j11}	Y_{j12}	Y_{j1+}
2	Y_{j21}	Y_{j22}	Y_{j2+}
	Y_{j+1}	Y_{j+2}	Y_{j++}

- For each (2×2) table, Mantel-Haenzel proposed conditioning on both margins (i.e., assuming both are fixed).
- We discussed that this is valid for a single (2×2) table regardless of what the design was, and it also generalizes to J, (2×2) tables. Thus, the following test will be valid for any design, including both prospective and case-control studies.
- Since Mantel and Haenszel condition on both margins, we only need to consider one random variable for each table, say Y_{j11},

- Under
  H_0: no association between Y and X given $W = j$,
- Conditional on both margins of the j^{th} table, the data follow a (central) hypergeometric distribution, with
- 1. Usual hypergeometric mean

$$E_j = E(Y_{j11}|y_{jk+}y_{j+\ell}) = \frac{y_{j1+}y_{j+1}}{y_{j++}}$$

- 2. and usual hypergeometric variance

$$\begin{aligned} V_j &= Var(Y_{j11}|y_{jk+}y_{j+\ell}) \\ &= \frac{y_{j1+}y_{j2+}y_{j+1}y_{j+2}}{y_{j++}^2(y_{j++}-1)} \end{aligned}$$

- Under the null of no association, with y_{j++} large,

$$Y_{j11} \sim N(E_j, V_j)$$

- Then, pooling over the J strata, since the sum of normals is normal, under the null

$$\sum_{j=1}^{J} Y_{j11} \sim N(\sum_{j=1}^{J} E_j, \sum_{j=1}^{J} V_j),$$

or, equivalently

$$Z = \frac{\sum_{j=1}^{J}[Y_{j11} - E_j]}{\sqrt{\sum_{j=1}^{J} V_j}} \sim N(0,1)$$

- The Mantel-Haenszel test statistic for

$$H_0 : \text{no association between } Y \text{ and } X \text{ given } W,$$

Or, equivalently,

$$H_0 : \beta = 0$$

in the logistic regression model,

$$\text{logit}\{P[Y = 1|W = j, X = x]\} = \beta_0 + \alpha_j + \beta x$$

is

$$Z^2 = \frac{\left(\sum_{j=1}^{J}[O_j - E_j]\right)^2}{\sum_{j=1}^{J} V_j} \sim \chi^2_1$$

where

$$O_j = Y_{j11},$$

$$E_j = E(Y_{j11}|y_{jk+}y_{j+\ell}) = \frac{y_{j1+}y_{j+1}}{y_{j++}},$$

and

$$\begin{aligned} V_j &= Var(Y_{j11}|y_{jk+}y_{j+\ell}) \\ &= \frac{y_{j1+}y_{j2+}y_{j+1}y_{j+2}}{y^2_{j++}(y_{j++}-1)}. \end{aligned}$$

Notes About the Mantel-Haenszel test statistic

- Cochran's name was added to the test, because he proposed what amounts to the logistic regression score test for

$$H_0 : \beta = 0$$

in the model

$$\text{logit}\{P[Y = 1|W = j, X = x]\} = \beta_0 + \alpha_j + \beta x$$

and this score test is approximately identical to the Mantel-Haenszel test.

- Mantel-Haenszel derived their test conditioning on both margins of each (2×2) table.
- Cochran, and the logistic regression score test treats one margin fixed and one margin random; in this test, O_j and E_j are the same as the Mantel-Haenszel test, but Cochran used

$$V_j = Var(Y_{j11}) = \frac{y_{j1+}y_{j2+}y_{j+1}y_{j+2}}{y_{j++}^3}$$

as opposed to Mantel-Haenzel's

$$V_j = Var(Y_{j11}) = \frac{y_{j1+}y_{j2+}y_{j+1}y_{j+2}}{y_{j++}^2(y_{j++} - 1)}$$

for large strata (y_{j++} large), they are almost identical.

Cochran Mantel-Haenzel Using SAS PROC FREQ

```
data two;
 input age placebo   para count; cards; 1  0  0 20 1  0  1
14 1  1  0 10 1  1  1 24 2  0  0 15 <<more data>> ;

proc freq  data=two;
 table age*placebo*para /relrisk CMH NOROW NOCOL NOPERCENT;
 /* put in W*X*Y when controlling for W */
 weight count;
run;
```

A brief aside

- Tired of seeing the “fffffffffff” in your SAS output?
- Use this SAS statement

```
OPTIONS FORMCHAR="|----|+|---+=|-/\<>*";
```

- This reverts the formatting back to the classic (i.e., mainframe) SAS platform friendly font (as opposed to the true type font with the f’s)

```
Table 1 of placebo by para
Controlling for age=1

placebo     para

Frequency|       0|       1|  Total
---------+--------+--------+
       0 |     20 |     14 |     34
---------+--------+--------+
       1 |     10 |     24 |     34
---------+--------+--------+
Total          30       38       68



Case-Control (Odds Ratio)      3.4286       95% CI (1.2546, 9.3695)
(as presented: The odds of no paralysis are
```

Table 1 of placebo by para Controlling for age=2

placebo para

Frequency	0	1	Total
0	15	12	27
1	3	15	18
Total	18	27	45

Case-Control (Odds Ratio) 6.2500 95% CI (1.4609, 26.7392)

Table 3 of placebo by para
Controlling for age=3

placebo para

Frequency	0	1	Total
0	3	2	5
1	3	2	5
Total	6	4	10

Case-Control (Odds Ratio) 1.0000 95% CI (0.0796, 12.5573)

Table 4 of placebo by para
Controlling for age=4

placebo para

Frequency	0	1	Total
0	12	3	15
1	7	5	12
Total	19	8	27

Case-Control (Odds Ratio) 2.8571 95% CI (0.5177, 15.7674)

Table 5 of placebo by para
Controlling for age=5

placebo para

Frequency	0	1	Total
0	1	0	1
1	3	2	5
Total	4	2	6

OR not calculated by SAS due to the zero cell, the empirical OR = 2.14

SUMMARY STATISTICS FOR VAC BY PARA
CONTROLLING FOR AGE
Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic	Alternative Hypothesis	DF	Value	Prob
1	Nonzero Correlation	1	13.047	0.0003
2	Row Mean Scores Differ	1	13.047	0.0003
3	General Association	1	13.047	0.0003

Estimates of the Common Relative Risk (Row1/Row2)

Type of Study	Method	Value	95% Confidence Bounds	
Case-Control	Mantel-Haenszel	3.591	1.795	7.187
(Odds Ratio)	Logit *	3.416	1.696	6.882

* denotes that the logit estimators use a correction of 0.5 in every cell of those tables that contain a zero.

Example - Age, Vaccine, Paralysis Data

- The Cochran-Mantel Haenzel Statistic was

$$Z^2 = 13.047, \qquad df = 1, \qquad 0.000$$

- Thus, Vaccine and Paralysis are not conditionally independent given age group.
- Recall, the WALD test for conditional independence in the logistic regression model,

$$H_0 : \beta = 0$$

was similar,

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square
VAC	1	1.2830	0.3573	12.8949	0.0003

- The Mantel-Haenszel and the Wald Stat are very similar.

Exact $p-$value for Mantel-Haenzsel Test

- Suppose the cell counts are small in many of the (2×2) tables; for example, tables 4 and 5 have small cell counts in the previous example.
- With small cell counts, the asymptotic approximations we discussed may not be valid
- Actually, the Mantel-Hanzel Statistic is usually approximately normal (chi-square) as long as one of the two following things hold:
  1. If the number of strata, J, is small, then y_{j++} should be large.
  2. If the strata sample sizes (y_{j++}) are small, then the number of strata J should be large.

- One can see this by looking at the statistic

$$Z = \frac{\sum_{j=1}^{J}[Y_{j11} - E_j]}{\sqrt{\sum_{j=1}^{J} V_j}} \sim N(0,1)$$

- 1. If each Y_{j++} is large, then each Y_{j11} will be approximately normal (via central limit theorem), and the sum of normals is normal, so Z will be normal.
- 2. If each Y_{j++} is small, then Y_{j11} will not be approximately normal; however, if J is large, then the sum

$$\sum_{j=1}^{J} Y_{j11}$$

  will be the sum of a lot of random variables, and we can apply the central limit theorem to it, so Z will be normal.
- However, if both J is small and y_{j++} is small, then the normal approximation may not be valid, and can use an 'exact' test.

- Under the null of conditional independence

$$\text{H}_0\text{:}OR_j^{XY.W} = 1,$$

for each j, the data (Y_{j11}) in the j^{th} table follow the central hypergeometric distribution,

$$P[Y_{j11} = y_{j11} | OR_j^{XY.W} = 1] \quad = \quad \frac{\left(\begin{array}{c} y_{j+1} \\ y_{j11} \end{array} \right) \left(\begin{array}{c} y_{j+2} \\ y_{j21} \end{array} \right)}{\left(\begin{array}{c} y_{j++} \\ y_{j1+} \end{array} \right)}$$

- The distribution of the data under the null is the product over these tables

$$\prod_{j=1}^{J} P[Y_{j11} = y_{j11} | OR_j^{XY.W} = 1] \quad = \quad \prod_{j=1}^{J} \frac{\left(\begin{array}{c} y_{j+1} \\ y_{j11} \end{array} \right) \left(\begin{array}{c} y_{j+2} \\ y_{j21} \end{array} \right)}{\left(\begin{array}{c} y_{j++} \\ y_{j1+} \end{array} \right)}$$

- This null distribution can be used to construct the exact $p-$value for the Mantel-Haenszel Statistic

- Let T be Mantel-Haensel statistic.
- Then, an exact, $p-$value for testing the null

$$\text{H}_0\text{:}OR_j^{XY.W} = 1,$$

for each j, is given by

$$p\text{-value} = P[T \geq t_{\text{observed}} | \text{H}_0\text{:cond. ind}]$$

where

$$P[T = t | \text{H}_0\text{:cond. ind}]$$

is obtained from the above product of (central) hypergeometric distributions.

- In particular, given the fixed margins of all J, (2×2) tables, we could write out all possible tables with margins fixed. For each possible set of J (2×2) tables, we could write out the Mantel-Hanesel statistic T, and the corresponding probability from the product hypergeometric.
- To get the $p-$value, we then sum all of the probabilities corresponding to the T's greater than or equal to the observed Mantel-Hanesel statistic T_{obs}.

Example - Age, Vaccine, Paralysis Data

Age	Salk Vaccine	Paralysis: No	Paralysis: Yes
0-4	Yes	20	14
	No	10	24
5-9	Yes	15	12
	No	3	15
10-14	Yes	3	2
	No	3	2
15-19	Yes	12	3
	No	7	5
20+	Yes	1	0
	No	3	2

Large Sample $p-$value is .0003

```
proc freq;
 table age*vac*para /cmh ;
  weight count;
run;

/* SELECTED OUTPUT */

Summary Statistics for vac by para
Controlling for age

  Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic    Alternative Hypothesis    DF       Value      Prob
------------------------------------------------------------------
    1        Nonzero Correlation        1     13.0466    0.0003
    2        Row Mean Scores Differ     1     13.0466    0.0003
    3        General Association        1     13.0466    0.0003
```

Exact Statistic using PROC FREQ

```
proc freq  data=two;
 table age*placebo*para /relrisk CMH NOROW NOCOL NOPERCENT;
 /* put in W*X*Y when controlling for W */
 weight count;
 exact comor;
run;
```

Note: This is exactly the same as before with the exception that “exact comor” has been added (comor = common odds ratio)

Selected Results

```
Summary Statistics for placebo by para
Controlling for age

         Common Odds Ratio
-------------------------------------
Mantel-Haenszel Estimate      3.5912

Asymptotic Conf Limits
95% Lower Conf Limit          1.7811
95% Upper Conf Limit          7.2406

Exact Conf Limits
95% Lower Conf Limit          1.6667
95% Upper Conf Limit          7.4704
```

```
Exact Test of H0: Common Odds Ratio = 1

Cell (1,1) Sum (S)     51.0000
Mean of S under H0     40.0222

One-sided Pr >=  S   2.381E-04
Point    Pr  =  S   1.754E-04

Two-sided P-values
2 * One-sided        4.763E-04 <-- Note quite correct
Sum <= Point         4.770E-04     for same reason as before
Pr >= |S - Mean|     4.770E-04
```

- The exact $p-$value can be also obtained using SAS Proc Logistic
- Recall that the Mantel-Haenszel test statistic is the logistic regression score test for

$$H_0 : \beta = \log(OR_j^{XY.W}) = 0$$

in the model

$$\text{logit}\{P[Y = 1|W = j, X = x]\} = \beta_0 + \alpha_j + \beta x$$

- For the Age, Vaccine, Paralysis Data, we want to test that the odds ratio between Vaccine (X) and Paralysis (Y) is conditionally independent given AGE (W), i.e., we are testing

$$H_0 : \beta = \log(OR_j^{Vac,Par.Age}) = 0$$

in the model

$$\text{logit}\{P[Par = 1|Age = j, Vac = x]\} = \beta_0 + \alpha_j + \beta x$$

Results from SAS Proc Logistic

The Exact $p-$value is .0005.

```
proc logistic descending data=one;
 class age;
 model para = vac age  ; /* model y = x w */
  exact vac ;           /* exact x */
  freq  count;
run;

/* SELECTED OUTPUT */

Exact Conditional Analysis

            Conditional Exact Tests

                                    --- p-Value ---
Effect   Test          Statistic    Exact      Mid

vac      Score          13.0466    0.0005   0.0004 /* Exact MH */
         Probability   0.000175    0.0005   0.0004
```

Mantel-Haenszel Estimator of Common Odds Ratio

- Mantel and Haenszel also proposed an estimator of the common odds ratio
- For table $W = j$, the observed odds ratio is

$$\widehat{OR}_j^{XY.W} = \frac{y_{j11}y_{j22}}{y_{j21}y_{j12}}$$

- If there is a common OR across tables, we could estimate the common OR with a 'weighted estimator':

$$\widehat{OR}_{MH} = \frac{\sum_{j=1}^{J} w_j \widehat{OR}_j^{XY.W}}{\sum_{j=1}^{J} w_j},$$

for some 'weights' w_j. (Actually, any weight will give you an asymptotically unbiased estimate).

- Mantel-Haenszel chose weights

$$w_j = \frac{y_{j21} y_{j22}}{y_{j++}}$$

when $OR_j^{XY \cdot W} = 1$, giving

$$\widehat{OR}_{MH} = \frac{\sum_{j=1}^{J} y_{j11} y_{j22} / y_{j++}}{\sum_{j=1}^{J} y_{j21} y_{j12} / y_{j++}}$$

- A good (consistent) estimate the variance of $\log[\widehat{OR}_{MH}]$ is (Robbins, et. al, 1985), based on a Taylor series expansion,

$$\widehat{Var}[\log \widehat{OR}_{MH}] =$$

$$\frac{\sum_{j=1}^{J} P_j R_j}{2[\sum_{j=1}^{J} R_j]^2} + \frac{\sum_{j=1}^{J} P_j S_j + Q_j R_j}{2[\sum_{j=1}^{J} R_j][\sum_{j=1}^{J} S_j]} + \frac{\sum_{j=1}^{J} Q_j S_j}{2[\sum_{j=1}^{J} S_j]^2},$$

where

$$P_j = (Y_{j11} + Y_{j22})/Y_{j++}$$

$$Q_j = (Y_{j12} + Y_{j21})/Y_{j++}$$

$$R_j = \frac{Y_{j11} Y_{j22}}{Y_{j++}}$$

$$S_j = \frac{Y_{j12} Y_{j21}}{Y_{j++}}$$

which is given in SAS.

Notes about M-H estimate

- 1. This estimate is easy to calculate (non-iterative), although its variance estimate is a little more complicated.
- 2. Asymptotically normal and unbiased with large strata (strata sample size y_{j++} large).
- 3. When each y_{j++} is large, the Mantel-Haensel estimate is not as efficient as the MLE, but close to MLE for logistic regression, which is iterative. When each y_{j++} is small, the MLE from logistic model could have a lot of bias.
- 4. Just like the Mantel-Haenzel statistic, unlike the logistic MLE, this estimator actually works well when the strata sample sizes are small (y_{j++} small), as long as the number of strata J is fairly large. (When doing large sample approximations, something must be getting large, either y_{j++} or J, or both).

Example - Age, Vaccine, Paralysis Data

- We showed earlier that the logistic regression estimate of the 'common odds ratio' between VACCINE (X) and PARALYSIS (Y) controlling for AGE (W) is

$$\exp(\hat{\boldsymbol{\beta}}) = \exp(1.2830) = 3.607,$$

  with a 95% confidence interval,

  `[1.791,7.266]`

  which does not contain 1. Thus, individuals who take the vaccine have 3.6 times the odds of not getting POLIO than individuals who do not take the vaccine.

- The Mantel-Haenzel Estimator of the common Odds Ratio is

$$\widehat{OR}^{MH} = 3.591$$

  with a 95% confidence interval of

  `[1.781,7.241]`

- Thus, individuals who take the vaccine have about 3.6 times the odds of not getting POLIO than individuals who do not take the vaccine.

Confounding in Logistic Regression

- Here, we are interested in using logistic regression to see if W confounds the relationship between X and Y.
- For simplicity, suppose we have 3 dichotomous variables
- In the logistic regression model,

$$w = \begin{cases} 1 \\ 0 \end{cases} . \qquad x = \begin{cases} 1 \\ 0 \end{cases} . \qquad Y = \begin{cases} 1 \\ 0 \end{cases} .$$

- The logistic regression model of interest is

$$\text{logit}\{P[Y = 1|w, x]\} = \beta_0 + \alpha w + \beta x.$$

The conditional odds ratio between Y and X given W is

$$\exp(\beta) = OR^{XY.W}.$$

- The marginal odds ratio between Y and X can be obtained from logistic regression model

$$\text{logit}\{P[Y = 1|x_]\} = \beta_0^* + \beta^* x,$$

  and is

$$\exp(\beta^*) = OR^{XY}.$$

- If there is no confounding, then

$$\beta = \beta^*$$

- Basically, you can fit both models, and, if

$$\hat{\boldsymbol{\beta}} \approx \hat{\boldsymbol{\beta}}^*,$$

  then you see that there is no confounding.

More Formal Check of Confounding of W

- However, to be more formal about checking for confounding, one would check to see if
- 1. W and Y are conditionally independent given X,
  or
- 2. W and X are conditionally independent given Y.
- To check these two conditions, you could fit a logistic model in which you make W the response, and X and Y covariates;

$$\text{logit}\{P[W = 1|x, Y]\} = \alpha_0 + \tau x + \alpha y,$$

- In this model, α is the conditional log-odds ratio between W and Y given X, and is identical to α in the logistic model with Y as the response and W and x as the covariates,

$$\alpha = \log[OR^{WY.X}]$$

- Also, τ is the conditional log-odds ratio between W and X given Y

$$\tau = \log[OR^{WX.Y}].$$

- Thus, if there is no confounding, the test for one of these two conditional OR's equalling 0 would not be rejected, i.e., you would either not reject $\alpha = 0$, or you would not reject $\tau = 0$.

Alternative Procedure

- However, if it was up to me, if you really want to see if there is confounding, I would just fit the two models:

$$\text{logit}\{P[Y=1|w,x]\} = \beta_0 + \alpha w + \beta x$$

and

$$\text{logit}\{P[Y=1|x]\} = \beta_0^* + \beta^* x,$$

and see if

$$\hat{\boldsymbol{\beta}} \approx \hat{\boldsymbol{\beta}}^*$$

- Rule of thumb in Epidemiology is that

$$\left| \frac{\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}^*}{\hat{\boldsymbol{\beta}}^*} \right| \leq 20\%?$$

- If there were many other covariates in the model, this is probably what you would do.

If $J > 2$, then you would fit the two models

$$\text{logit}\{P[Y = 1|W = j, X = x]\} = \beta_0 + \alpha_j + \beta x$$

and

$$\text{logit}\{P[Y = 1|X = x]\} = \beta_0^* + \beta^* x,$$

and see if

$$\hat{\boldsymbol{\beta}} \approx \hat{\boldsymbol{\beta}}^*$$

Notes about Models

- In journal papers, the analysis with the model

$$\text{logit}\{P[Y=1|x]\} = \beta_0^* + \beta^* x,$$

  is often called **univariate** (or unadjusted) analysis (the univariate covariate with the response)

- The analysis with the model

$$\text{logit}\{P[Y=1|w,x]\} = \beta_0 + \alpha w + \beta x$$

  is often called a **multivariate** analysis (more than one covariate with the response).

- Strictly speaking,

$$\text{logit}\{P[Y=1|w,x]\} = \beta_0 + \alpha w + \beta x$$

  is a **multiple** logistic regression analysis.

- In general, you state the results from a multiple regression as **adjusted** ORs.

Efficiency Issues

- Suppose you fit the two models, and there is no confounding,
- Then, in the models

$$\text{logit}\{P[Y=1|w,x]\} = \beta_0 + \alpha w + \beta x$$

and

$$\text{logit}\{P[Y=1|x]\} = \beta_0^* + \beta^* x,$$

we have

$$\beta = \beta^*$$

- Suppose, even though there is no confounding, W is an important predictor of Y, and should be in the model.
- Even though $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}^*$ are both asymptotically unbiased (since they are both estimating the same β), you can show that

$$Var(\hat{\boldsymbol{\beta}}) \quad \leq \quad Var(\hat{\boldsymbol{\beta}}^*)$$

$$\text{FULLER} \quad \leq \quad \text{REDUCED}$$

Quasi-proof

- Heuristically, this is true because W is explaining some of the variability in Y that is not explained by X alone,
- and thus, since more variability is being explained, the variance of the estimates from the fuller model (with W) will be smaller.

Suppose $\alpha = 0$.

- Now, suppose that, in real life, you have overspecified the model, i.e., $\alpha = 0$, so that W and Y are conditionally independent given X, i.e., the true model is

$$\text{logit}\{P[Y = 1|w, x]\} = \beta_0^* + \beta x$$

- However, suppose you estimate (β_0, α, β) in the model

$$\text{logit}\{P[Y = 1|w, x]\} = \beta_0 + \alpha w + \beta x$$

  you are estimating β from an 'overspecified' model in which we are (unnecessarily) estimating α, which is 0.

- In this case, $\hat{\boldsymbol{\beta}}$ from the overspecified model will still be asymptotically unbiased, however estimating a parameter α that is 0 actually adds more error to the model, and

$$\begin{array}{ccc} Var(\hat{\boldsymbol{\beta}}) & \geq & Var(\hat{\boldsymbol{\beta}}^*) \\ \text{FULLER} & \geq & \text{REDUCED} \end{array}$$