BIOS 6110
Applied Categorical Data Analysis

Instructor: Yuan Huang, Ph.D.

Department of Biostatistics

Fall 2017

Part VIII

Models for Matched Pairs

Overview

- Previously, we have worked on the data where the subjects are independently collected.
- This section introduces methods for comparing categorical responses for two samples that have a natural pairing between each subject in one sample and a subject in the other sample.
- Because each observation in one sample pairs with an observation in the other sample, the responses in the two samples are matched pairs.
- Because of the matching, the samples are statistically dependent.
- In this case, methods that treat the two sets of observations as independent samples are inappropriate

We will focus on 1-1 Match with binary response in this chapter.

Objectives

1. Data display
   - Population-Average table
   - Subject-Specific tables
2. Compare two dependent proportions
   - McNemar test for testing marginal homogeneity
   - Estimate and CI of the proportion difference
   - Mantel-Haenszel odds ratio
3. Logistic regression for matched pairs
   - Marginal models
   - Conditional Logistic regression for matched-pairs
   - *Why unconditional logistic model is biased for matched-pairs

Two motivating data examples:

- 2000 General Social Survey (used in objective 2)
- Diabetes-MI Matched Pairs Case-Control Study (used in objective 3)

Motivating Example: 2000 General Social Survey

In the 2000 General Social Survey, 1144 subjects were asked whether, to help the environment, they would be willing to (1) pay higher taxes or (2) accept a cut in living standards.

	Cut Living Standards		
Pay Higher Taxes	Yes	No	Total
Yes	227	132	359
No	107	678	785
Total	334	810	1144

How can we compare the probabilities of a “yes” outcome for the two environmental questions?

Comparing Dependent Proportions

Generally, we have

Pay Higher Taxes	Cut Living Standards		
	Yes	No	Total
Yes	n_{11}	n_{12}	n_{1+}
No	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

Define sample estimates

$$\begin{aligned} p_{ij} &= n_{ij}/n, \quad i = 1,2, j = 1,2 \\ p_{i+} &= p_{i1} + p_{i2}, \quad i = 1,2 \\ p_{+j} &= p_{1j} + p_{2j}, \quad j = 1,2 \end{aligned}$$

The population value corresponding to p_{ij} is denoted by π_{ij}. When $\pi_{+1} = \pi_{1+}$ and $\pi_{+2} = \pi_{2+}$, there is *marginal homogeneity*. From this condition, we have

$$\pi_{1+} - \pi_{+1} \quad = \quad (\pi_{11} + \pi_{12}) - (\pi_{11} + \pi_{21}) = \pi_{12} - \pi_{21}$$

That is, marginal homogeneity is equivalent to $\pi_{12} = \pi_{21}$.

Testing Marginal Homogeneity: McNemar Test

For matched pairs data with a binary response, a test of marginal homogeneity has null hypothesis

$$H_0 : \pi_{12} = \pi_{21}, \quad \text{or equivalently} \quad H_0 : \pi_{+1} = \pi_{1+}$$

As off-diagonal counts are $n_{12} + n_{21}$, under H_0, both n_{12} and n_{21} are distributed as Binomial($n_{12} + n_{21}$, 0.5).

Construct test based on n_{12}. When $n_{12} + n_{21}$ is large, the binomial can be approximately by normal distribution with mean $(n_{12} + n_{21}) \times 0.5$ and variance $(n_{12} + n_{21}) \times (0.5) \times (1 - 0.5)$. Therefore, the z statistics can be constructed accordingly.

$$z = \frac{n_{12} - (n_{12} + n_{21}) \times 0.5}{\sqrt{(n_{12} + n_{21}) \times 0.5 \times 0.5}} = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}}$$

The $z \sim$ N(0,1), or z^2 is approximately χ^2 with df $= 1$ under H_0.

The chi-squared test (z^2) for a comparison of two dependent proportions is called McNemar test.

Example: 2000 General Social Survey (cont.)

The McNemar's test statistic is

$$\frac{(132-107)^2}{132+107} = 2.615$$

with *p*-value

```
> 1-pchisq(2.615,1)
[1] 0.1058575
```

There is no sufficient evidence that the probability of approval was greater for higher taxes than for a lower standard of living

Estimating Differences of Proportions

The point estimate of $\pi_{1+} - \pi_{+1}$ is $p_{1+} - p_{+1}$

The estimated standard error (SE) of $p_{1+} - p_{+1}$ is

$$
\begin{aligned}
& \left\{ \frac{p_{1+}(1-p_{1+}) + p_{+1}(1-p_{+1}) - 2(p_{11}p_{22} - p_{12}p_{21})}{n} \right\}^{1/2} \\
= \quad & n^{-1}\sqrt{(n_{12}+n_{21}) - n^{-1}(n_{12}-n_{21})^2}
\end{aligned}
$$

95% Confidence interval: $(p_{1+} - p_{+1}) \pm 1.96 \times SE$

- Matched-pairs data usually show a positive association, which implies a odds ratio > 1. A sample odds ratio exceeding 1.0 corresponds to $p_{11}p_{22} > p_{12}p_{21}$, a negative contribution from the third term.
- Thus, an advantage of using dependent samples, rather than independent samples, is a smaller variance for the estimated difference in proportions.

Example: 2000 General Social Survey (cont.)

The point estimate is

$$359/1144 - 334/1144 = 0.314 - 0.292 = 0.022$$

The estimated SE is

$$1144^{-1}\sqrt{(132 + 107) - 1144^{-1}(132 - 107)^2} = 0.0135$$

A 95% CI is

$$0.022 \pm 1.96(0.0135) = (-0.004, 0.048)$$

Population-Average Table vs Subject-Specific Tables

In the 2000 Social Survey Example, the previous table cross-classifies in a single table the two response for **all subjects**. This is called the **population-average table**.

We can consider **each matched pair as a cluster**, and present 2×2 partial table for each pair. In this case, each of the 2×2 partial table shows 2 observations from the matched pairs. Because there are 1144 subject, we will get $2 \times 2 \times 1144$ three-way table. This is called the **subject-specific tables**.

E.g., each partial table can be displayed as

	Response	
	Yes	No
Pay Higher Taxes	XXX	XXX
Cut Living Standards	XXX	XXX

Example: 2000 General Social Survey (cont.)

(a)	Response	
	Yes	No
Pay Higher Taxes	1	0
Cut Living Standards	1	0

(b)	Response	
	Yes	No
Pay Higher Taxes	1	0
Cut Living Standards	0	1

(c)	Response	
	Yes	No
Pay Higher Taxes	0	1
Cut Living Standards	1	0

(d)	Response	
	Yes	No
Pay Higher Taxes	0	1
Cut Living Standards	0	1

- There are 227 such partial table (a), which correspond to n_{11}.
- There are 132 such partial table (b), which correspond to n_{12}.
- There are 107 such partial table (c), which correspond to n_{21}.
- There are 678 such partial table (d), which correspond to n_{22}.

Mantel-Haenszel Odds Ratio - Based on Subject-Specific Tables

Previously in XYZ three-way table, where X is the binary treatment, Y is binary response, and Z is the centers ($k = 1, \ldots, K$).

	Diseased	Non-diseased	Totals
Exposed	n_{11k}	n_{12k}	n_{1+k}
Unexposed	n_{21k}	n_{22k}	n_{2+k}
Totals	n_{+1k}	n_{+2k}	n_{++k}

The Mantel-Haenszel method

- assumes that there is a true odds ratio which is consistent across k
- provides a pooled estimate of the common odds ratio. In essence, it is a weighted average of the odds ratios from the individual strata

The Mantel-Haenszel estimate of the odds ratio is

$$\hat{\theta}_{MH} = \frac{\sum_{k=1}^{K} n_{11k} n_{22k} / n_{++k}}{\sum_{k=1}^{K} n_{21k} n_{12k} / n_{++k}}$$

Note that, in our current matched pair case, the K is number of pairs.

Example: 2000 General Social Survey (cont.)

(a)	Response	
	Yes	No
Pay Higher Taxes	1	0
Cut Living Standards	1	0

(b)	Response	
	Yes	No
Pay Higher Taxes	1	0
Cut Living Standards	0	1

(c)	Response	
	Yes	No
Pay Higher Taxes	0	1
Cut Living Standards	1	0

(d)	Response	
	Yes	No
Pay Higher Taxes	0	1
Cut Living Standards	0	1

- Pairs as (a) → contribute 0 to denominator and 0 to numerator
- Pairs as (b) → contribute 0 to denominator and 1 to numerator
- Pairs as (c) → contribute 1 to denominator and 0 to numerator
- Pairs as (d) → contribute 0 to denominator and 0 to numerator
- $$\hat{\theta}_{MH} = \frac{\sum_{k=1}^{K} n_{11k} n_{22k} / n_{++k}}{\sum_{k=1}^{K} n_{21k} n_{12k} / n_{++k}} = \frac{n_{12}}{n_{21}} = \frac{132}{107} = 1.234$$

Marginal Models for Marginal Proportions

A marginal table can be obtained by adding partial tables across stratas.

Issue	Response		
	Yes	No	Total
Pay Higher Taxes	359	785	1144
Cut Living Standards	334	810	1144

A marginal model aims to study the marginal distributions of response for the two observations. The third dimension, i.e., pair stratas, is omitted.

$$\frac{359/785}{334/810} = 1.11.$$

The population odds of willingness to pay higher taxes are estimated to be 11% higher than the population odds of willingness to accept cuts in living standards.

Marginal Models for Marginal Proportions (cont.)

- We already know that this analysis can also be obtained by using the following logistic regression

$$\text{logit}[\Pr(Y = \text{``Yes''})] = \alpha + \beta x$$

  where x is an indicator for question 1.

- Chapter 9 will discuss marginal models in details.

Objectives

1. Data display
   - Population-Average table
   - Subject-Specific tables
2. Compare two dependent proportions
   - McNemar test for testing marginal homogeneity
   - Estimate and CI of the proportion difference
   - Mantel-Haenszel odds ratio
3. Logistic regression for matched pairs
   - Marginal models
   - Conditional Logistic regression for matched-pairs
   - *Why unconditional logistic model is biased for matched-pairs

Two motivating data examples:

- 2000 General Social Survey (used in objective 2)
- Diabetes-MI Matched Pairs Case-Control Study (used in objective 3)

Motivating Example: Diabetes-MI Match Pairs Case-Control Study

A study of acute myocardial infarction (MI) among Navajo Indians matched 144 victims of MI (MI Cases) according to age and gender with 144 individuals free of MI (MI Controls). Subjects were then asked whether they had ever been diagnosed as having diabetes ($x = 0$, no; $x = 1$, yes)

	MI Cases		
MI Controls	Diabetes	No Diabetes	Total
Diabetes	9	16	25
No diabetes	37	82	119
Total	46	98	144

- Note, this is not only a case-control study, but case-control with individual match.

We can obtain marginal table.

	Disease	
Exposure	MI	Non-MI
Diabetics	9	16
No Diabetics	37	82

As introduced in last lecture, there are the four types of partial tables, including 9 Table (a), 16 Table (b), 37 Table (c), and 82 Table (d).

(a)	Disease	
Exposure	MI	Non-MI
Diabetics	1	1
No Diabetics	0	0

(b)	Disease	
Exposure	MI	Non-MI
Diabetics	0	1
No Diabetics	1	0

(c)	Disease	
Exposure	MI	Non-MI
Diabetics	1	0
No Diabetics	0	1

(d)	Disease	
Exposure	MI	Non-MI
Diabetics	0	0
No Diabetics	1	1

$$\hat{\theta}_{MH} = \frac{\sum_{k=1}^{K} n_{11k} n_{22k} / n_{++k}}{\sum_{k=1}^{K} n_{21k} n_{12k} / n_{++k}} = \frac{37}{16}$$

Conditional logistic regression (Introduction)

To handle stratum/pairs, one possibility is to estimate the following model

$$\text{logit}(\Pr(MI = 1)) = \alpha_i + \beta \; Diabetes,$$

where

$$MI = \begin{cases} 1 & \text{MI} \\ 0 & \text{Non-MI} \end{cases} \qquad Diabetes = \begin{cases} 1 & \text{Diabetes} \\ 0 & \text{No Diabetes} \end{cases}$$

The odds that a subject with diabetes is an MI case equal e^{β} times the odds of a subject without diabetes is an MI case.

However, it can have a lot of nuisance parameters attributing to α_is.

To get a better estimation of β, we want to use conditional logistic regression. By doing conditioning on a discordant pair, the conditional likelihood does not involves parameters α_is, it depends only on β.

Conditional logistic regression (Details)

For the i-th stratum/pair, based on $\text{logit}(\Pr(MI = 1)) = \alpha_i + \beta\ Diabetes$, we can fill in probabilities of the following 2×2 partial table.

	Disease	
	MI	Non-MI
Diabetics	$\frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}$	$1 - \frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}} = \frac{1}{1+e^{\alpha_i+\beta}}$
Non-Diabetics	$\frac{e^{\alpha_i}}{1+e^{\alpha_i}}$	$1 - \frac{e^{\alpha_i}}{1+e^{\alpha_i}} = \frac{1}{1+e^{\alpha_i}}$

Table (b): MI with diabetes AND Non-MI without diabetes

(b)	Disease	
Exposure	MI	Non-MI
Diabetics	0	1
No Diabetics	1	0

$$\frac{e^{\alpha_i}}{1+e^{\alpha_i}} \times \frac{1}{1+e^{\alpha_i+\beta}}$$

Table (c): MI without diabetes AND Non-MI with diabetes

(c)	Disease	
Exposure	MI	Non-MI
Diabetics	1	0
No Diabetics	0	1

$$\frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}} \times \frac{1}{1+e^{\alpha_i}}$$

Conditional logistic regression (Details, Cont.)

Instead of modeling probability of observing a partial table, conditional logistic regression modeling the conditional probability of observing a partial table given a discordant pair.

Therefore, for conditional logistic regression, only discordant stratum/pair can contribute terms to the likelihood.
− This is similar to McNemar Test.

A discordant pair means either Table (b) or Table (c)

$$\frac{e^{\alpha_i}}{1+e^{\alpha_i}} \times \frac{1}{1+e^{\alpha_i+\beta}} + \frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}} \times \frac{1}{1+e^{\alpha_i}}$$

$$P(\text{Table (b)}|\text{Discordant pair}) = \frac{\frac{e^{\alpha_i}}{1+e^{\alpha_i}} \times \frac{1}{1+e^{\alpha_i+\beta}}}{\frac{e^{\alpha_i}}{1+e^{\alpha_i}} \times \frac{1}{1+e^{\alpha_i+\beta}} + \frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}} \times \frac{1}{1+e^{\alpha_i}}}$$

$$P(\text{Table (c)}|\text{Discordant pair}) = \frac{\frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}} \times \frac{1}{1+e^{\alpha_i}}}{\frac{e^{\alpha_i}}{1+e^{\alpha_i}} \times \frac{1}{1+e^{\alpha_i+\beta}} + \frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}} \times \frac{1}{1+e^{\alpha_i}}}$$

Conditional logistic regression (Details, Cont.)

$$
\begin{aligned}
P(\text{Table (c)}|\text{Discordant pair}) &= \frac{\frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}} \times \frac{1}{1+e^{\alpha_i}}}{\frac{e^{\alpha_i}}{1+e^{\alpha_i}} \times \frac{1}{1+e^{\alpha_i+\beta}} + \frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}} \times \frac{1}{1+e^{\alpha_i}}} \\
&= \frac{e^{\alpha_i+\beta}}{e^{\alpha_i+\beta}+e^{\alpha_i}} = \frac{e^{\beta}}{e^{\beta}+1} \\
P(\text{Table (b)}|\text{Discordant pair}) &= 1 - \frac{e^{\beta}}{e^{\beta}+1} = \frac{1}{e^{\beta}+1}
\end{aligned}
$$

Conditional likelihood depends only on β.

Because there are 16 Table (b) and 37 Table (c), so the conditional likelihood is

$$
\left(\frac{1}{e^{\beta}+1}\right)^{16}\left(\frac{e^{\beta}}{e^{\beta}+1}\right)^{37},
$$

which is maximized at $e^{\beta} = \frac{37}{16}$. the same as $\hat{\theta}_{MH}$.

* An Ordinary Logistic Regression Approach to 1:1 Matching

For 1-1 match, let

$$y_i^* = \begin{cases} 1 & \text{when} \quad y_{i1} = 1, y_{i2} = 0 \\ 0 & \text{when} \quad y_{i1} = 0, y_{i2} = 1 \end{cases}$$

Also let

$$\begin{aligned} x_{1i}^* &= x_{1i1} - x_{1i2} \\ & \dots \\ x_{ki}^* &= x_{ki1} - x_{ki2} \end{aligned}$$

Fit the ordinary logistic regression to the y^* with predictors $\{x_1^*, \dots, x_k^*\}$, forcing the intercept parameter as 0. The resulting estimates are the same as the conditional logistic regression.

* Why Ordinary/Unconditional Logistic Regression is Inappropriate

Recall that for the *i*-th stratum/pair, we have

	Disease	
	MI	Non-MI
Diabetics	$\frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}$	$1-\frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}=\frac{1}{1+e^{\alpha_i+\beta}}$
Non-Diabetics	$\frac{e^{\alpha_i}}{1+e^{\alpha_i}}$	$1-\frac{e^{\alpha_i}}{1+e^{\alpha_i}}=\frac{1}{1+e^{\alpha_i}}$

Without conditioning, all tables (a) - (d) contribute to the likelihood.

There are 9 Table (a): MI with diabetes AND Non-MI with diabetes

(a)	Disease	
Exposure	MI	Non-MI
Diabetics	1	1
No Diabetics	0	0

$$\frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}} \times \frac{1}{1+e^{\alpha_i+\beta}}$$

Regardless of β, the α_i is maximized at $-\beta$. So for estimating β, it will contribute to the likelihood the following term

$$l_{(a)} = \left(\frac{e^{-\beta+\beta}}{1+e^{-\beta+\beta}} \times \frac{1}{1+e^{-\beta+\beta}} \right)^9 = (1/4)^9$$

	Disease	
	MI	Non-MI
Diabetics	$\frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}$	$1-\frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}=\frac{1}{1+e^{\alpha_i+\beta}}$
Non-Diabetics	$\frac{e^{\alpha_i}}{1+e^{\alpha_i}}$	$1-\frac{e^{\alpha_i}}{1+e^{\alpha_i}}=\frac{1}{1+e^{\alpha_i}}$

There are 16 Table (b): MI without diabetes AND Non-MI with diabetes

(b)	Disease	
Exposure	MI	Non-MI
Diabetics	0	1
No Diabetics	1	0

$$\frac{e^{\alpha_i}}{1+e^{\alpha_i}} \times \frac{1}{1+e^{\alpha_i+\beta}}$$

Regardless of β, the α_i is maximized at $-\beta/2$. So for estimating β, it will contribute to the likelihood the following term

$$l_{(d)} = \left(\frac{e^{-\beta/2}}{1+e^{-\beta/2}} \times \frac{1}{1+e^{-\beta/2+\beta}}\right)^{16} = \left(\frac{1}{1+e^{\beta/2}}\right)^{2\times 16}$$

	Disease	
	MI	Non-MI
Diabetics	$\dfrac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}$	$1-\dfrac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}=\dfrac{1}{1+e^{\alpha_i+\beta}}$
Non-Diabetics	$\dfrac{e^{\alpha_i}}{1+e^{\alpha_i}}$	$1-\dfrac{e^{\alpha_i}}{1+e^{\alpha_i}}=\dfrac{1}{1+e^{\alpha_i}}$

There are 37 Table (c): MI with diabetes AND Non-MI without diabetes

(c)	Disease	
Exposure	MI	Non-MI
Diabetics	1	0
No Diabetics	0	1

$$\frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}} \times \frac{1}{1+e^{\alpha_i}}$$

Regardless of β, the α_i is maximized at $-\beta/2$. So for estimating β, it will contribute to the likelihood the following term

$$l_{(c)} = \left(\frac{e^{-\beta/2+\beta}}{1+e^{-\beta/2+\beta}} \times \frac{1}{1+e^{-\beta/2}}\right)^{37} = \left(\frac{e^{\beta/2}}{1+e^{\beta/2}}\right)^{2\times 37}$$

	Disease	
	MI	Non-MI
Diabetics	$\frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}$	$1-\frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}}=\frac{1}{1+e^{\alpha_i+\beta}}$
Non-Diabetics	$\frac{e^{\alpha_i}}{1+e^{\alpha_i}}$	$1-\frac{e^{\alpha_i}}{1+e^{\alpha_i}}=\frac{1}{1+e^{\alpha_i}}$

There are 82 Table (d): MI without diabetes AND Non-MI without diabetes

(d)	Disease	
Exposure	MI	Non-MI
Diabetics	0	0
No Diabetics	1	1

$$\frac{e^{\alpha_i}}{1+e^{\alpha_i}} \times \frac{1}{1+e^{\alpha_i}}$$

Regardless of β, the α_i is maximized at 0. So for estimating β, it will contribute to the likelihood the following term

$$l_{(d)} = \left(\frac{e^0}{1+e^0} \times \frac{1}{1+e^0}\right)^{82} = (1/4)^{82}$$

In order to maximize the likelihood function

$$l = l_{(a)} l_{(b)} l_{(c)} l_{(d)}$$

it is suffice to maximize $l_{(b)} l_{(c)}$, or equivalently, to maximize

$$\begin{aligned} \log(l_{(b)} l_{(c)}) &= 2 \times 37 \times \log\left(\frac{e^{\beta/2}}{1+e^{\beta/2}}\right) + 2 \times 16 \times \log\left(\frac{1}{1+e^{\beta/2}}\right) \\ &= -2 \times 37 \times \log(1+e^{-\beta/2}) - 2 \times 16 \times \log(1+e^{\beta/2}) \end{aligned}$$

The partial derivative with respect to β is

$$\frac{37}{1+e^{\beta/2}} - \frac{16 \times e^{\beta/2}}{1+e^{\beta/2}}$$

Setting it to 0, we have

$$\exp(\beta/2) = 37/16$$

Or

$$\exp(\beta) = (37/16)^2 = (OR_{MH})^2$$

Therefore, the unconditional logistic regression is not appropriate.