STAT 226 Lecture 26

Section 8.2 Logistic Regression For Matched Pairs

Yibi Huang

Estimation of Odds Ratio for Matched-Pair Data

- Population-Avaraged Models (a.k.a. Marginal Models)
- Subject-Specific Models (a.k.a. Conditional Models)

Population-Avaraged Models

Suppose a matched-pair is selected at random from the population. Let (Y_1, Y_2) denote the two responses from the selected pair, where

$$Y_i = \begin{cases} 1 & \text{for category 1 (success)} \\ 2 & \text{for category 2} \end{cases}$$

In population probabilities:

	$Y_2 = 1$	$Y_2 = 2$	Total
$Y_1 = 1$	π_{11}	π_{12}	π_{1+}
$Y_1 = 2$	π_{21}	π_{22}	π_{2+}
Total	π_{+1}	π_{+2}	1

Then

$$P(Y_1 = 1) = \pi_{1+}, \quad P(Y_1 = 0) = \pi_{2+}$$

$$P(Y_2 = 1) = \pi_{+1}, \quad P(Y_2 = 0) = \pi_{+2}$$

Population-Avaraged Models (a.k.a. Marginal Models)

Suppose

$$\text{logit}[\mathrm{P}(Y_1 = 1)] = \alpha + \beta, \quad \text{i.e.,} \quad \frac{\mathrm{P}(Y_1 = 1)}{\mathrm{P}(Y_1 = 0)} = \frac{\pi_{1+}}{\pi_{2+}} = e^{\alpha+\beta}$$

$$\text{logit}[\mathrm{P}(Y_2 = 1)] = \alpha, \qquad \text{i.e.,} \quad \frac{\mathrm{P}(Y_2 = 1)}{\mathrm{P}(Y_2 = 0)} = \frac{\pi_{+1}}{\pi_{+2}} = e^{\alpha}$$

Consequently, $e^{\beta} = \dfrac{\mathrm{P}(Y_1 = 1)/\mathrm{P}(Y_1 = 0)}{\mathrm{P}(Y_2 = 1)/\mathrm{P}(Y_2 = 0)} = \dfrac{\pi_{1+}/\pi_{2+}}{\pi_{+1}/\pi_{+2}}$,

which means, at <u>population level</u>, the odds of success for response 1 are e^{β} times the odds of success for response 2. This OR is called the *marginal OR*.

The MLE of the marginal OR = $e^{\widehat{\beta}}$ is

$$\frac{\widehat{\pi}_{1+}/\widehat{\pi}_{2+}}{\widehat{\pi}_{+1}/\widehat{\pi}_{+2}} = \frac{n_{1+}/n_{2+}}{n_{+1}/n_{+2}}$$

Example (Matched-Pair Case-Control Study of MI & Diabetes)

A study of acute myocardial infarction (MI) among Navajo Indians matched 144 victims of MI according to age and gender with 144 people free of heart disease and recored whether they had ever been diagnosed diabetes.

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

Estmiated marginal OR is

$$\frac{n_{1+}/n_{2+}}{n_{+1}/n_{+2}} = \frac{46/98}{25/119} \approx 2.234.$$

Example (Matched-Pair Case-Control Study of MI & Diabetes)

A study of acute myocardial infarction (MI) among Navajo Indians matched 144 victims of MI according to age and gender with 144 people free of heart disease and recored whether they had ever been diagnosed diabetes.

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

Estmiated marginal OR is

$$\frac{n_{1+}/n_{2+}}{n_{+1}/n_{+2}} = \frac{46/98}{25/119} \approx 2.234.$$

Two interpretations:

- The population odds of diabetes for MI cases are estimated to be 2.234 times the population odds of diabetes for MI controls.
- The population odds of MI for those w diabetes were estimated to be 2.234 times the population odds of MI for those without diabetes

SE for log(Marginal OR)

- The estimated marginal OR = $\dfrac{n_{1+}/n_{2+}}{n_{+1}/n_{+2}}$ has a skewed sampling distribution. Its normal approximation is NOT good.
- Sampling distribution of for $\log\left(\dfrac{n_{1+}/n_{2+}}{n_{+1}/n_{+2}}\right)$ is closer to normal with a large sample variance

$$\frac{1}{n}\left(\frac{1}{\pi_{1+}}+\frac{1}{\pi_{2+}}+\frac{1}{\pi_{+1}}+\frac{1}{\pi_{+2}}-2\frac{\pi_{11}\pi_{22}-\pi_{12}\pi_{21}}{\pi_{1+}\pi_{2+}\pi_{+1}\pi_{+2}}\right),$$

estimated by

$$\frac{1}{n}\left(\frac{n}{n_{1+}}+\frac{n}{n_{2+}}+\frac{n}{n_{+1}}+\frac{n}{n_{+2}}-2\frac{n^2(n_{11}n_{22}-n_{12}n_{21})}{n_{1+}n_{2+}n_{+1}n_{+2}}\right).$$

- The large sample SE is

$$\mathrm{SE}=\sqrt{\frac{1}{n_{1+}}+\frac{1}{n_{2+}}+\frac{1}{n_{+1}}+\frac{1}{n_{+2}}-\frac{2n(n_{11}n_{22}-n_{12}n_{21})}{n_{1+}n_{2+}n_{+1}n_{+2}}}$$

If we ignore pairing and rewrite the table as the 2-way table for MI (Case, Control) and Diabete (Yes, No),

	MI Controls		
MI Cases	diabetes	no diabetes	total
diabetes	n_{11}	n_{12}	n_{1+}
no diabetes	n_{21}	n_{22}	n_{2+}
total	n_{+1}	n_{+1}	n

$\Rightarrow$

	Diabetes		
MI	Yes	No	Total
case	n_{1+}	n_{2+}	n
control	n_{+1}	n_{+2}	n

the marginal OR would be in the usual "cross-product" form:

$$\frac{n_{1+}/n_{2+}}{n_{+1}/n_{+2}} = \frac{n_{1+}n_{+2}}{n_{2+}n_{+1}}.$$

Large sample SE of log(marginal OR) for matched-pair data

$$\text{SE} = \sqrt{\frac{1}{n_{1+}} + \frac{1}{n_{2+}} + \frac{1}{n_{+1}} + \frac{1}{n_{+2}} - \frac{2n(n_{11}n_{22} - n_{12}n_{21})}{n_{1+}n_{2+}n_{+1}n_{+2}}}$$

is usually less than the SE for $\log$(OR) for two-sample data

$$\text{SE} = \sqrt{\frac{1}{n_{1+}} + \frac{1}{n_{2+}} + \frac{1}{n_{+1}} + \frac{1}{n_{+2}}}$$

CI for Marginal OR

CI for $\log$(marginal OR):

$$(L, U) = \log\left(\frac{n_{1+}/n_{2+}}{n_{+1}/n_{+2}}\right) \pm z_{\alpha/2}\mathrm{SE}$$

where the SE is given on the previous page

CI for marginal OR is (e^L, e^U).

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

$$\log(\text{marginal OR}) = \log\left(\frac{46 \times 119}{98 \times 25}\right) \approx 0.8039.$$

with

$$\text{SE} = \sqrt{\frac{1}{n_{1+}} + \frac{1}{n_{2+}} + \frac{1}{n_{+1}} + \frac{1}{n_{+2}} - \frac{2n(n_{11}n_{22} - n_{12}n_{21})}{n_{1+}n_{2+}n_{+1}n_{+2}}}$$

$$= \sqrt{\frac{1}{46} + \frac{1}{98} + \frac{1}{25} + \frac{1}{119} - \frac{2 \times 144(9 \times 82 - 37 \times 16)}{46 \times 98 \times 25 \times 119}} \approx 0.2779.$$

95% CI for log(marginal OR):

$0.8039 \pm 1.96 \times 0.2779 \approx (0.2592, 1.3486)$

95% CI for marginal OR: $(e^{0.2592}, e^{1.3486}) \approx (1.296, 3.852)$.

If we ignore pairing,

	MI Controls		
MI Cases	diabetes	no diabetes	total
diabetes	9	37	46
no diabetes	16	82	98
total	25	119	144

$\Rightarrow$

	Diabetes		
MI	Yes	No	Total
case	46	98	144
control	25	119	144

- $\log(\text{OR}) = \log\left(\frac{46 \times 119}{98 \times 25}\right) \approx 0.8039$ (same as paired data)
- SE = $\sqrt{\frac{1}{46} + \frac{1}{98} + \frac{1}{25} + \frac{1}{119}} \approx 0.2835$ is bigger than the SE for paired data
- 95% CI for log(OR): $0.8039 \pm 1.96 \times 0.2835 \approx (0.2482, 1.3596)$
- 95% CI for OR: $(e^{0.2592}, e^{1.3486}) \approx (1.282, 3.894)$
  wider than the CI $(1.282, 3.894)$ for marginal OR of paired data

Subject Specific Models

The population-avaraged model do not reflect the correlation within a pair. Let (Y_{1i}, Y_{2i}) denote the two responses from the ith pair

$$\text{logit}[\mathrm{P}(Y_{1i} = 1)] = \alpha_i + \beta, \quad \text{logit}[\mathrm{P}(Y_{2i} = 1)] = \alpha_i$$

i.e., $\mathrm{P}(Y_{1i} = 1) = \dfrac{e^{\alpha_i + \beta}}{1 + e^{\alpha_i + \beta}}, \quad \mathrm{P}(Y_{2i} = 1) = \dfrac{e^{\alpha_i}}{1 + e^{\alpha_i}}$.

The subject-specific model allows dependence within a pair by including a "subject effect" α_i.

- If $\alpha_i > 0$ is large, both Y_{1i} and Y_{2i} are likely to be 1
- If $\alpha_i < 0$ and is large in magnitude, both Y_{1i} and Y_{2i} are likely to be 0

For each subject, the odds of success for response 1 are e^{β} times the odds of success for response 2. e^{β} is called the *conditional odds ratio*.

Population-Averaged v.s. Subject Specific Models

Suppose the population contains N pairs. Based on the subject specific model, the responses of the ith pair (Y_{1i}, Y_{2i}) have the distribution

$$\mathrm{P}(Y_{1i} = 1) = \frac{e^{\alpha_i+\beta}}{1 + e^{\alpha_i+\beta}}, \quad \mathrm{P}(Y_{2i} = 1) = \frac{e^{\alpha_i}}{1 + e^{\alpha_i}}.$$

If a pair (Y_1, Y_2) is selected at random from the population,

$$\begin{aligned}
\pi_{1+} &= \mathrm{P}(Y_1 = 1) = \frac{1}{N}\sum\nolimits_{i=1}^{N} \frac{e^{\alpha_i+\beta}}{1 + e^{\alpha_i+\beta}}, \\
\pi_{2+} &= \mathrm{P}(Y_1 = 0) = \frac{1}{N}\sum\nolimits_{i=1}^{N} \frac{1}{1 + e^{\alpha_i+\beta}}, \\
\pi_{+1} &= \mathrm{P}(Y_2 = 1) = \frac{1}{N}\sum\nolimits_{i=1}^{N} \frac{e^{\alpha_i}}{1 + e^{\alpha_i}} \\
\pi_{+2} &= \mathrm{P}(Y_2 = 0) = \frac{1}{N}\sum\nolimits_{i=1}^{N} \frac{1}{1 + e^{\alpha_i}}.
\end{aligned}$$

Population-Averaged v.s. Subject Specific Models

The odds ratio in the population averaged model is

$$
\begin{aligned}
\frac{\pi_{1+}/\pi_{2+}}{\pi_{+1}/\pi_{+2}} &= \frac{\mathrm{P}(Y_1 = 1)/\mathrm{P}(Y_1 = 0)}{\mathrm{P}(Y_2 = 1)/\mathrm{P}(Y_2 = 0)} = \frac{\mathrm{P}(Y_1 = 1)\mathrm{P}(Y_2 = 0)}{\mathrm{P}(Y_2 = 1)\mathrm{P}(Y_1 = 0)} \\
&= \frac{\sum_{i=1}^{N} \frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}} \sum_{i=1}^{N} \frac{1}{1+e^{\alpha_i}}}{\sum_{i=1}^{N} \frac{e^{\alpha_i}}{1+e^{\alpha_i}} \sum_{i=1}^{N} \frac{1}{1+e^{\alpha_i+\beta}}} \\
&\neq e^{\beta} \quad \text{in general, unless } \alpha_i = \alpha \text{ for all } i
\end{aligned}
$$

So, the β in the subject specific model is different from the β in the population averaged model.

Estimate of β in the Subject Specific Model

- Ordinary ML do not work well for the subject-specific model for having as many subject parameters $\{\alpha_i\}$ as the # of pairs.
- A remedy: for pairs with $Y_{1i} + Y_{2i} = 1$, can show in the next slide that
$$\mathrm{P}(Y_{1i} = 1 | Y_{1i} + Y_{2i} = 1) = \frac{e^\beta}{1 + e^\beta},$$
i.e., the conditional distribution of Y_{1i} given $Y_{1i} + Y_{2i} = 1$ is free of α_i.
- In matched-paired data, there are $n^* = n_{12} + n_{21}$ independent pairs with $Y_{1i} + Y_{2i} = 1$. Given $n^* = n_{12} + n_{21}$, the conditional distribution of n_{12} is
$$n_{12} \sim \text{Binomial}(n^*, \frac{e^\beta}{1 + e^\beta})$$
based on which one can obtain a *conditional likelihood* for β that is free of α_i's and the maximal conditional likelihood estimator for e^β is $\widehat{e^\beta} = n_{12}/n_{21}$, or $\widehat{\beta} = \log(n_{12}/n_{21})$.

$$
\begin{aligned}
\mathrm{P}(Y_{1i} = 1 | Y_{1i} + Y_{2i} = 1) &= \frac{\mathrm{P}(Y_{1i} = 1, Y_{1i} + Y_{2i} = 1)}{\mathrm{P}(Y_{1i} + Y_{2i} = 1)} \\
&= \frac{\mathrm{P}(Y_{1i} = 1, Y_{2i} = 0)}{\mathrm{P}(Y_{1i} = 1, Y_{2i} = 0) + \mathrm{P}(Y_{1i} = 0, Y_{2i} = 1)} \\
&= \frac{\frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}} \frac{1}{1+e^{\alpha_i}}}{\frac{e^{\alpha_i+\beta}}{1+e^{\alpha_i+\beta}} \frac{1}{1+e^{\alpha_i}} + \frac{1}{1+e^{\alpha_i+\beta}} \frac{e^{\alpha_i}}{1+e^{\alpha_i}}} \\
&= \frac{e^{\alpha_i+\beta}}{e^{\alpha_i+\beta} + e^{\alpha_i}} = \frac{e^{\beta}}{1 + e^{\beta}}
\end{aligned}
$$

Population-Averaged v.s. Subject Specific (Section 8.2.3)

We can rewrite the data below as a 3-way $2 \times 2 \times 144$ table of the 3 variables

X = MI (Cases, Controls)

Y = Diabetes (Yes, No)

Z = Pair ID (1 to 144)

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

where the XY partial table for a pair is one of the following 4

9 pairs:

	Diabetes	
MI	Yes	No
case	1	0
control	1	0

37 pairs:

	Diabetes	
MI	Yes	No
case	1	0
control	0	1

16 pairs:

	Diabetes	
MI	Yes	No
case	0	1
control	1	0

82 pairs:

	Diabetes	
MI	Yes	No
case	0	1
control	0	1

and the XY marginal table is

	Diabetes		
MI	Yes	No	Total
case	46	98	144
control	25	119	144

.

For the subject specific model

$$\text{logit}[\mathrm{P}(Y_{1i} = 1)] = \alpha_i + \beta, \quad \text{logit}[\mathrm{P}(Y_{2i} = 1)] = \alpha_i$$

The conditional OR e^{β} for the subject specific model is the conditional OR of (X, Y) given Z = pairing

The marginal OR for the population average model is the marginal OR of (X, Y) ignoring Z = pairing

McNemar's Test is CMH Test for Matched-Pair Data (1)

While we rewrite matched-pair data below as a 3-way table of the 3 variables

X = MI (Cases, Controls)

Y = Diabetes (Yes, No)

Z = Pair ID (1 to 144)

	MI Controls		
MI Cases	diabetes	no diabetes	Total
diabetes	n_{11}	n_{12}	n_{1+}
no diabetes	n_{21}	n_{22}	n_{2+}
Total	n_{+1}	n_{+2}	n

where each (X, Y) partial table for a pair is one of the following 4

	Diabetes	
MI	Yes	No
case	1	0
control	1	0

n_{11} *pairs*

	Diabetes	
MI	Yes	No
case	1	0
control	0	1

n_{12} *pairs*

	Diabetes	
MI	Yes	No
case	0	1
control	1	0

n_{21} *pairs*

	Diabetes	
MI	Yes	No
case	0	1
control	0	1

n_{22} *pairs*

We can test the conditional independence of (X, Y) given Z by apply CMH test on the 3-way table of XYZ.

McNemar's Test is CMH Test for Matched-Pair Data (2)

Recall the CMH statistic is

$$\text{CMH} = \frac{\sum_k (n_{11k} - \mathrm{E}[n_{11k}])}{\sqrt{\sum_k \mathrm{Var}(n_{11k})}}, \text{ where } \mathrm{E}[n_{11k}] = \frac{R_{1k}C_{1k}}{T_k}, \ \mathrm{Var}(n_{11k}) = \frac{R_{1k}R_{2k}C_{1k}C_{2k}}{T_k^2(T_k-1)}$$

if the XY partial table for $Z = k$ is

	$Y=1$	$Y=2$	total
$X=1$	n_{11k}	n_{12k}	R_{1k}
$X=2$	n_{21k}	n_{22k}	R_{2k}
total	C_{1k}	C_{2k}	T_k

n_{11} *pairs*

	Diabetes	
MI	Yes	No
case	1	0
control	1	0

$\mathrm{E}[n_{11}] = 1$, $\mathrm{Var}(n_{11}) = 0$

n_{12} *pairs*

	Diabetes	
MI	Yes	No
case	1	0
control	0	1

$\mathrm{E}[n_{11}] = 1/2$, $\mathrm{Var}(n_{11}) = 1/4$

n_{21} *pairs*

	Diabetes	
MI	Yes	No
case	0	1
control	1	0

$\mathrm{E}[n_{11}] = 1/2$, $\mathrm{Var}(n_{11}) = 1/4$

n_{22} *pairs*

	Diabetes	
MI	Yes	No
case	0	1
control	0	1

$\mathrm{E}[n_{11}] = 0$, $\mathrm{Var}(n_{11}) = 0$

$$\text{CMH} = \frac{n_{11}(0) + n_{12}(1-1/2) + n_{21}(0-1/2) + n_{22}(0)}{\sqrt{n_{11}(0) + n_{12}(1/4) + n_{21}(1/4) + n_{22}(0)}} = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}} = \text{McNemar's}$$

Estimate for the Conditional OR is Mantel-Haenszel's Estimate

Recall Mantel-Haenszel's estimate of the common odds ratio of several tables is

$$\widehat{\theta}_{MH} = \frac{\sum_k n_{11k} n_{22k}/T_k}{\sum_k n_{12k} n_{21k}/T_k}$$

if kth partial table is

	$Y = 1$	$Y = 2$	total
$X = 1$	n_{11k}	n_{12k}	R_{1k}
$X = 2$	n_{21k}	n_{22k}	R_{2k}
total	C_{1k}	C_{2k}	T_k

n_{11} *pairs*

	Diabetes	
MI	Yes	No
case	1	0
control	1	0

n_{12} *pairs*

	Diabetes	
MI	Yes	No
case	1	0
control	0	1

n_{21} *pairs*

	Diabetes	
MI	Yes	No
case	0	1
control	1	0

n_{22} *pairs*

	Diabetes	
MI	Yes	No
case	0	1
control	0	1

$$\widehat{\theta}_{MH} = \frac{n_{11}(1 \cdot 0/2) + n_{12}(1 \cdot 1/2) + n_{21}(0 \cdot 0/2) + n_{22}(0 \cdot 1/2)}{n_{11}(0 \cdot 1/2) + n_{12}(0 \cdot 0/2) + n_{21}(1 \cdot 1/2) + n_{22}(1 \cdot 0/2)} = \frac{n_{12}}{n_{21}}.$$

which is exactly the estimate for the conditional OR of the subject-specific model.

CI for Conditional OR

- The estimated conditional OR $= n_{12}/n_{21}$ has a skewed sampling distribution. Its normal approximation is NOT good.
- Sampling distribution of for $\log n_{12}n_{21}$ is closer to normal with the large sample SE

$$\text{SE} = \sqrt{\frac{1}{n_{12}} + \frac{1}{n_{21}}}$$

- CI for $\log(\text{conditional OR})$:

$$(L, U) = \log(n_{12}/n_{21}) \pm z_{\alpha/2}\text{SE}$$

- CI for conditional OR: (e^L, e^U).

Example

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

- $\log(\text{conditional OR}) = \log(37/16) \approx 0.8383.$
- $\text{SE} = \sqrt{\frac{1}{n_{12}} + \frac{1}{n_{21}}} = \sqrt{\frac{1}{37} + \frac{1}{16}} \approx 0.2992$
- 95% CI for log(conditional OR):
  $0.8383 \pm 1.96 \times 0.2992 \approx (0.25191.4247)$
- 95% CI for conditional OR: $(e^{0.2519}, e^{1.4247}) \approx (1.286, 4.157)$
- Interpretation: For a subject w/ diabetes, his/her odds of MI are 1.286 to 4.157 times the odds for someone w/o diabetes, with 95% confidence.

Which Model to Use? Population-Averaged or Subject Specific?

Both are useful, depending on the application

- If interested in mechanism on individuals, use subject specific model.
- If the goal is to compare the relative frequency of occurrence of some outcome for different groups in a population (e.g., in surveys or epidemiological studies), use population-averaged model