BIOSTAT 651
Notes #11:
Conditional Logistic Regression

- Lecture Topics:
  - Matching
  - Analysis of matched pairs
  - Conditional logistic regression
  - Matched case-control studies

Control of confounding

- *Restriction* is a frequently employed method of eliminating confounding in biomedical studies
  - Study on exercise and heart disease
    - Age and gender are confounding factors
    - Restricted the study to men aged 40-65
- Sample restriction has its pros and cons
  - Adv: successfully eliminates confounding
  - Disadv:
    - Can't evaluate the effects of factors that have been restricted for
    - Reduces generality of study's findings

Matching

- Alternative to restriction: *Matching*
  - i.e., match subjects who are very similar with respect to the confounder of concern
- For example, common to match on age, sex, diagnosis
- Matching can be used in each of the studies we've previously described
  - prospective cohort study:

    e.g., match treatment A and treatment B subjects by gender and race
  - case-control studies:

    e.g., match each case to a control of the same age

Matching (continued)

- Matching eliminates the need to adjust for confounders upon which matching is based
  - not able to estimate the effect of matched covariates
- Unlike restriction, matching need not reduce generality of study
- Analyses of matched data often require methods distinct from those of unmatched study

Matching Schemes

- Various methods are available for matching subjects:
  - $1:1$ matching
  - $1:m$ matching
  - $m:n$ matching
- Depending on the nature of the analysis, matched sets of unequal size may be permitted
- The unit of analysis is typically the *matched set*, as opposed to the subject

Analysis of Matched Pairs

- Consider a study consisting of matched pairs
  - each pair consists of two responses:

    Y_{i1} = response (0,1) from subject 1

    Y_{i2} = response (0,1) from subject 2

    $i = 1, \ldots, m$
- Unit of analysis: pair
- Ex. Vaccine study
  - 500 matched pairs. Each pair is matched on gender and age.
  - For example, Pair 1 might be two men, both age 23. Pair 2 might be two women, both age 22.

Pair	Placebo (Y_{i1})	Treatment (Y_{i2})
1	1	0
2	0	0
...	...	...
500	1	1

Analysis of Matched Pairs

- Summarize the observed data by the following table:

	Y_{i2}=0	Y_{i2}=1	total
Y_{i1}=0	m_{00}	m_{01}	m_{0+}
Y_{i1}=1	m_{10}	m_{11}	m_{1+}
total	m_{+0}	m_{+1}	m

McNemar's Test

- Set $\pi_1 = P(Y_{i1} = 1)$ and $\pi_2 = P(Y_{i2} = 1)$
- McNemar's Test
  - $H_0 : \pi_1 = \pi_2$
  - uses only off-diagonal elements
  - test statistic:

$$X_M^2 = \frac{(m_{10} - m_{01})^2}{(m_{10} + m_{01})} \quad \sim \quad \chi_1^2$$

Example: McNemar's Test

- Example: Vaccine study (Page 6)

  Y_{i1} = Placebo

  Y_{i2} = Treatment

- Responses summarized in the following table:

	Y_{i2}=0	Y_{i2}=1	total
Y_{i1}=0	384	18	402
Y_{i1}=1	91	7	98
total	475	25	500

- McNemar's Test:

McNemar's Test: SAS Code

- We can carry out McNemar's Test using PROC FREQ:

```
data vaccine;
 input placebo treatment count;
 datalines;
  0   0  384
  0   1  18
  1   0  91
  1   1  7
  ;
 run;
```

```
proc freq data=vaccine;
 tables placebo*treatment / agree cmh;
 weight count;
run;
```

Regression Analysis of Matched Data

Regression Analysis: Matched Data

- Matched sets are often viewed as *strata*
  - e.g., matching on state (MI, OH, WI, NC) produces $K = 4$ strata
  - e.g., matching by age group (0-14, 15-29, 30-39, 40-49) and diabetes type (I, II, none) produces $K = 12$ strata
- If there are few strata and many subjects in each, then stratum could be incorporated into the $\mathbf{x}_i$ vector
- However, technical issues arise when K is large

Matched Pairs Cohort Study

- The matched-pairs cohort study provides an interesting application of conditional likelihood
- Set-up is as follows:
  - cohort study
  - data consist of matched pairs: $k = 1, \ldots, K$
  - observed data for each subject: $(Y_{ik}, \mathbf{x}_{ik})$

    covariate: $\mathbf{x}_{ik} = (x_{ik1}, x_{ik2}, \ldots, x_{ikq})^T$

    parameter of interest: $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)^T$
  - each pair consists of one *treated* and one *untreated* subject

    $x_{1k1} = 1$, $x_{2k1} = 0$
- Model:

$$\log\left\{\frac{\pi_{ik}}{1-\pi_{ik}}\right\} = \alpha_k + \mathbf{x}_{ik}^T\boldsymbol{\beta}$$

Estimation Issues: Stratified Logistic Regression

- Issues in estimating $(\alpha_1, \alpha_2, \ldots\ldots, \alpha_K, \boldsymbol{\beta}^T)$:

Conditional Logistic Regression

- Alternative to standard logistic regression (applicable to matched data):
    - *conditional* logistic regression
    - uses conditional likelihood
- Set $L_k(\boldsymbol{\beta})$ = conditional likelihood, stratum k

  $\propto$ probability of observed data in stratum k *given some characteristic of stratum*
- Conditional Likelihood:

$$L(\boldsymbol{\beta}) = \prod_{k=1}^{K} L_k(\boldsymbol{\beta})$$

Matched Pairs: Conditional Likelihood

- Recall (from McNemar's Test): principle that concordant matched pairs provide little information on $\boldsymbol{\beta}$
  - therefore, we form the likelihood by conditioning on discordance
- In particular,

$$
\begin{aligned}
L(\boldsymbol{\beta}) &= \prod_{k=1}^{K} L_k(\boldsymbol{\beta}) \\
L_k(\boldsymbol{\beta}) &\propto P(\text{observed data, stratum } k) \\
&\propto P(\text{observed data, stratum } k | \text{discordance})
\end{aligned}
$$

Matched Pairs: Conditional Likelihood (continued)

- Probability of discordant pair:

$$
\begin{aligned}
& P(Y_{1k} = 1|\mathbf{x}_{1k})P(Y_{2k} = 0|\mathbf{x}_{2k}) \\
& \quad + P(Y_{1k} = 0|\mathbf{x}_{1k})P(Y_{2k} = 1|\mathbf{x}_{2k}) \\
= \quad & \pi(\mathbf{x}_{1k})\{1 - \pi(\mathbf{x}_{2k})\} \qquad (1) \\
& \quad + \pi(\mathbf{x}_{2k})\{1 - \pi(\mathbf{x}_{1k})\} \qquad (2)
\end{aligned}
$$

- Conditional probabilities:

$$P(Y_{1k} = 1|\text{discordance}) \quad =$$

$$P(Y_{2k} = 1|\text{discordance}) \quad =$$

Forming Conditional Likelihood: MPC

- Recall that under the assumed model,

$$\pi(\mathbf{x}_{ik}) = \frac{e^{\alpha_k + \mathbf{x}_{ik}^T \boldsymbol{\beta}}}{1 + e^{\alpha_k + \mathbf{x}_{ik}^T \boldsymbol{\beta}}}$$

- Therefore, we have

$$(1) = \frac{e^{\alpha_k + \mathbf{x}_{1k}^T \boldsymbol{\beta}}}{1 + e^{\alpha_k + \mathbf{x}_{1k}^T \boldsymbol{\beta}}} \frac{1}{1 + e^{\alpha_k + \mathbf{x}_{2k}^T \boldsymbol{\beta}}}$$

$$(2) = \frac{1}{1 + e^{\alpha_k + \mathbf{x}_{1k}^T \boldsymbol{\beta}}} \frac{e^{\alpha_k + \mathbf{x}_{2k}^T \boldsymbol{\beta}}}{1 + e^{\alpha_k + \mathbf{x}_{2k}^T \boldsymbol{\beta}}}$$

- We then obtain

$$\frac{(1)}{(1) + (2)} = \frac{e^{(\mathbf{x}_{1k} - \mathbf{x}_{2k})^T \boldsymbol{\beta}}}{1 + e^{(\mathbf{x}_{1k} - \mathbf{x}_{2k})^T \boldsymbol{\beta}}}$$

$$\frac{(2)}{(1) + (2)} = \frac{1}{1 + e^{(\mathbf{x}_{1k} - \mathbf{x}_{2k})^T \boldsymbol{\beta}}}$$

Matched Pair: Conditional Likelihood

- Finally, the conditional likelihood is then given by:

$$
\begin{aligned}
L_k(\boldsymbol{\beta}) \quad &= \quad \left\{ \frac{e^{(\mathbf{x}_{1k} - \mathbf{x}_{2k})^T \boldsymbol{\beta}}}{1 + e^{(\mathbf{x}_{1k} - \mathbf{x}_{2k})^T \boldsymbol{\beta}}} \right\}^{Y_{1k}(1 - Y_{2k})} \\
&\quad \times \left\{ \frac{1}{1 + e^{(\mathbf{x}_{1k} - \mathbf{x}_{2k})^T \boldsymbol{\beta}}} \right\}^{Y_{2k}(1 - Y_{1k})}
\end{aligned}
$$

- Equal to the typical logistic regression likelihood, except:
  - using only discordant pairs
  - one record per matched pair
  - response: $Y_k^* = Y_{1k}$
  - covariate: $\mathbf{x}_k^* = \mathbf{x}_{1k} - \mathbf{x}_{2k}$
  - no intercept term

Matched Case-Control Studies

Matched Case-Control Study

- Matched-data set-up:
  - case-control study
  - total of K strata: $k = 1, \ldots, K$
  - n_{1k} cases and n_{0k} controls in stratum k
  - set $n_k = n_{0k} + n_{1k}$
  - K can be quite large, with n_k generally small
  - set $\pi_{ik} = P(Y_{ik} = 1 | \mathbf{x}_{ik})$
- Model:

$$\log\left\{\frac{\pi_{ik}}{1-\pi_{ik}}\right\} = \alpha_k + \mathbf{x}_{ik}^T\boldsymbol{\beta}$$

$\boldsymbol{\beta} =$

$\mathbf{x}_{ik}^T =$

Conditional Likelihood: Case-Control Study

- Set $L_k(\boldsymbol{\beta})$ = conditional likelihood, stratum k

  $\propto$ probability of observed data in stratum k *given the total number of cases*

- Note: among n_k subjects, number of case assignments:

$$\binom{n_k}{n_{1k}} \equiv c_k$$

- Matched pair: $c_k = 2$

Conditional Likelihood: Matched pair

- Probability of observed data (stratum k):

$$\prod_{i=1}^{n_k} P(\mathbf{x}_{ik}|Y_{ik}=1)^{Y_{ik}} P(\mathbf{x}_{ik}|Y_{ik}=0)^{1-Y_{ik}}$$

- *Conditional* probability of observed data (stratum k), *given* the total number of cases (stratum k):

$$\frac{\prod_{i=1}^{n_k} P(\mathbf{x}_{ik}|Y_{ik}=1)^{Y_{ik}} P(\mathbf{x}_{ik}|Y_{ik}=0)^{1-Y_{ik}}}{\sum_{j=1}^{c_k}\prod_{i=1}^{n_k} P(\mathbf{x}_{ik}|Y_{i(j)k}=1)^{Y_{i(j)k}} P(\mathbf{x}_{ik}|Y_{i(j)k}=0)^{1-Y_{i(j)k}}}$$

- Re-write key probability:

$$\begin{aligned} P(\mathbf{x}_{ik}|Y_{ik}=1) &= \frac{P(Y_{ik}=1|\mathbf{x}_{ik})P(\mathbf{x}_{ik})}{P(Y_{ik}=1)} \\ &= \frac{\pi(\mathbf{x}_{ik})P(\mathbf{x}_{ik})}{P(Y_{ik}=1)} \end{aligned}$$

Constructing Conditional Likelihood (continued)

- We can then write:

$$L_k(\boldsymbol{\beta}) \quad \propto \quad \frac{\prod_{i=1}^{n_k} \pi(\mathbf{x}_{ik})^{Y_{ik}}\{1-\pi(\mathbf{x}_{ik})\}^{1-Y_{ik}}}{\sum_{j=1}^{c_k}\prod_{i=1}^{n_k} \pi(\mathbf{x}_{ik})^{Y_{i(j)k}}\{1-\pi(\mathbf{x}_{ik})\}^{1-Y_{i(j)k}}}$$

- Now, we recall that

$$\pi(\mathbf{x}_{ik}) \quad = \quad \frac{e^{\alpha_k+\mathbf{x}_{ik}^T\boldsymbol{\beta}}}{1+e^{\alpha_k+\mathbf{x}_{ik}^T\boldsymbol{\beta}}}$$

such that the denominators in the above RHS cancel out

Constructing Conditional Likelihood (cont'd)

- We are then left with

$$L_k(\boldsymbol{\beta}) \quad \propto \quad \frac{\prod_{i=1}^{n_k} e^{(\alpha_k + \mathbf{x}_{ik}^T \boldsymbol{\beta}) Y_{ik}}}{\sum_{j=1}^{c_k} \prod_{i=1}^{n_k} e^{(\alpha_k + \mathbf{x}_{ik}^T \boldsymbol{\beta}) Y_{i(j)k}}}$$

- Finally, we arrive at our conditional likelihood:

$$\begin{aligned} L_k(\boldsymbol{\beta}) &= \frac{\prod_{i=1}^{n_k} e^{\mathbf{x}_{ik}^T \boldsymbol{\beta} Y_{ik}}}{\sum_{j=1}^{c_k} \prod_{i=1}^{n_k} e^{\mathbf{x}_{ik}^T \boldsymbol{\beta} Y_{i(j)k}}} \\ L(\boldsymbol{\beta}) &= \prod_{k=1}^{K} L_k(\boldsymbol{\beta}) \end{aligned}$$

- It has been shown that $L(\boldsymbol{\beta})$ possesses the key properties of a typical likelihood function ...

Conditional Likelihood: Case-Control v. Cohort

- Q: How does this version of $L_k(\boldsymbol{\beta})$ relate to that used for matched cohort studies?
  - recall: for MPC data:

$$L_k(\boldsymbol{\beta}) = \left\{ \frac{e^{(\mathbf{x}_{1k}-\mathbf{x}_{2k})^T\boldsymbol{\beta}}}{1+e^{(\mathbf{x}_{1k}-\mathbf{x}_{2k})^T\boldsymbol{\beta}}} \right\}^{Y_{1k}(1-Y_{2k})} \times \left\{ \frac{1}{1+e^{(\mathbf{x}_{1k}-\mathbf{x}_{2k})^T\boldsymbol{\beta}}} \right\}^{Y_{2k}(1-Y_{1k})}$$

  - and, for matched case-control data (1:1 matching with $Y_{1k}=1$ and $Y_{2k}=0$)

$$L_k(\boldsymbol{\beta}) = \frac{e^{\mathbf{x}_{1k}^T\boldsymbol{\beta}}}{e^{\mathbf{x}_{1k}^T\boldsymbol{\beta}}+e^{\mathbf{x}_{2k}^T\boldsymbol{\beta}}} = \left\{ \frac{e^{(\mathbf{x}_{1k}-\mathbf{x}_{2k})^T\boldsymbol{\beta}}}{1+e^{(\mathbf{x}_{1k}-\mathbf{x}_{2k})^T\boldsymbol{\beta}}} \right\}^{Y_{1k}(1-Y_{2k})}$$