Matched Pair Analysis

1 Motivation and Set-up:

Matching provides an easy way to control for confounders.

- e.g. match on subjects that are of the same sex, age, etc.

Matching can be used in each of the study types previously discussed:

- Prospective cohort - match treatment A, treatment B subjects on *age* and *race*
- Case control study: match each case to a control of the same *age*

Matching preserves the generalizability of a study

Schemes

Various methods for matching:

- $1 : 1$
- $1 : m$
- $m : n$

The unit of analysis is the **matched set**

Analysis

Suppose we have data on m matched pairs

- In a given matched pair...
- Y_{i1} = response $(0, 1)$ from first subject in pair
- Y_{i2} = response $(0, 1)$ from second subject in pair

Typical data structure might look like this:

Pair	Placebo	Treatment
1	$1 \leftarrow Y_1 = 1 \vert x_1 =$ placebo	0
2	0	0
...	...	...
m	1	1

Which is then summarized as follows:

	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

Where the cells m_{10} and m_{01}, the discordant pairs, provide the most information regarding treatment effect.

McNemar's Test
$H_0 : P(Y_{i1} = 1) = P(Y_{i2} = 1)$

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

Regression Analysis of Matched Data

- View the matched sets as *strata*

  - e.g. matching on age groups: [0-14, 15-29,30-39,40-49] and diabetes type: [I, II, none] produces $K = 12$ strata (cartesian product of the two sets)

- If there are few strata, and many subjects in each, then stratum could be incorporated into the $\boldsymbol{x}_i$ via coding
- Technical issues arise when K is large

2 Matched Pairs Cohort Study

- Set- up:

  $k = 1, ..., K$ matched pairs

  observed outcome and covariates for each subject

  each pair consists of one *treated* and one *untreated* subject

  e.g. $x_{1k1} = 1, x_{2k1} = 0$

Data structure looks like:

$$k = 1$$
$$Y = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$
$$X = \begin{bmatrix} 1 & ... & x_{1q} \\ 0 & ... & x_{21} \end{bmatrix}$$

or alternatively

Pair	TX($X_{ik=1}$)	Placebo ($X_{2k=0}$)
1	1	1
2	0	0
3	1	0
$\vdots$	$\vdots$	$\vdots$
K	Y_{1K}	Y_{2K}

Model:

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

Conditional Logistic Regression

$L_k(\boldsymbol{\beta})$ = conditional likelihoood, stratum k

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

Derivation of Conditional Likelihood

$$\begin{aligned} P(Y_{1k} \neq Y_{2k}|x_{1k}, x_{2k}) &= P(Y_{1k} = 1, Y_{2k} = 0|x_{1k}, x_{2k}) + P(Y_{1k} = 0, Y_{2k} = 1|x_{1k}, x_{2k}) \\ &= P(Y_{1k} = 1|x_{1k})P(Y_{2k} = 0|x_{2k})P(Y_{1k} = 0|x_{1k})P(Y_{2k} = 1|x_{2k}) \\ &= \underbrace{\pi(x_{1k})(1-\pi(x_{2k}))}_{\gamma} + \underbrace{\pi(x_{2k})(1-\pi(x_{1k}))}_{\zeta} \end{aligned}$$

This implies that the conditional probabilities describing discordance can be described as follows

$$P(Y_{1k} = 1|\text{discordance}) = \frac{\gamma}{\gamma + \zeta}$$
$$P(Y_{2k} = 1|\text{discordance}) = \frac{\zeta}{\gamma + \zeta}$$

From the model...

$$\pi(\boldsymbol{x}_{ik}) = \frac{\exp\{\alpha_k + \boldsymbol{x}_{ik}^T \boldsymbol{\beta}\}}{1 + \exp\{\alpha_k + \boldsymbol{x}_{ik}^T \boldsymbol{\beta}\}}$$
$$\Rightarrow \gamma = \frac{\exp\{\alpha_k + \boldsymbol{x}_{1k}^T \boldsymbol{\beta}\}}{1 + \exp\{\alpha_k + \boldsymbol{x}_{1k}^T \boldsymbol{\beta}\}} \frac{1}{1 + \exp\{\alpha_k + \boldsymbol{x}_{2k}^T \boldsymbol{\beta}\}}$$
$$\zeta = \frac{\exp\{\alpha_k + \boldsymbol{x}_{2k}^T \boldsymbol{\beta}\}}{1 + \exp\{\alpha_k + \boldsymbol{x}_{2k}^T \boldsymbol{\beta}\}} \frac{1}{1 + \exp\{\alpha_k + \boldsymbol{x}_{1k}^T \boldsymbol{\beta}\}}$$
$$\Rightarrow \frac{\gamma}{\gamma + \zeta} = \frac{\exp\{\boldsymbol{x}_{1k} - \boldsymbol{x}_{2k})^T \boldsymbol{\beta}\}}{1 + \exp\{(\boldsymbol{x}_{1k} - \boldsymbol{x}_{2k})^T \boldsymbol{\beta}\}}$$
$$\Rightarrow \frac{\zeta}{\gamma + \zeta} = \frac{1}{1 + \exp\{(\boldsymbol{x}_{1k} - \boldsymbol{x}_{2k})^T \boldsymbol{\beta}\}}$$
$$\Rightarrow L_k(\boldsymbol{\beta}) = \left\{\frac{\gamma}{\gamma + \zeta}\right\}^{Y_{1k}(1-Y_{2k})} \left\{\frac{\zeta}{\gamma + \zeta}\right\}^{Y_{2k}(1-Y_{1k})}$$
$$\Rightarrow L_k(\boldsymbol{\beta}) = \left\{\frac{\exp\{(\boldsymbol{x}_{1k} - \boldsymbol{x}_{2k})^T \boldsymbol{\beta}\}}{1 + \exp\{(\boldsymbol{x}_{1k} - \boldsymbol{x}_{2k})^T \boldsymbol{\beta}\}}\right\}^{Y_{1k}(1-Y_{2k})} \times \left\{\frac{1}{1 + \exp\{(\boldsymbol{x}_{1k} - \boldsymbol{x}_{2k})^T \boldsymbol{\beta}\}}\right\}^{Y_{2k}(1-Y_{1k})}$$
$$L(\boldsymbol{\beta}) = \prod_{k=1}^{K} L_k(\boldsymbol{\beta})$$

- Equivalent to the typical logistic regression likelihood except:
    i. One record per matched pair

    ii. One record per matched pair

    iii. Response: $Y_k^* = Y_{1k}$

    iv. Covariate: $\boldsymbol{x}_k^* = \boldsymbol{x}_{1k} - \boldsymbol{x}_{1k}$

    v. No intercept term

    vi. Note that this model is fit using IRWLS with the full likelihood function as normal

Matched Pair Cohort: Example

Regardless of whether a cohort or a case-control study the three main steps of the regression are the same:

1. Split the data according to whether an individual was treated (cohort).
2. Rejoin this data together, and subtract the differing values from each other
3. Fit the logistic regression model without an intercept - interpret as usual

This is demonstrated with the images below: `to be added later (bugs in tex)`

3 Matched Case-Control Study

Mathced data set up:

- total of K strata: $k = 1, ..., K$
- 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
- let $\pi_{ik} = P(Y_{ik} = 1|\boldsymbol{x}_{ik})$

Data Structure:

Pair	$Y_{1k} = 1$	$Y_{2k} = 0$
1	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$
2	$\vdots$	$\vdots$
K	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$

Model:

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

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

Matched Pair Case-Control: Example

Regardless of whether a cohort or a case-control study the three main steps of the regression are the same:

1. Split the data according to whether an individual was a case (case-control)
2. Rejoin this data together, and subtract the differing values from each other
3. Fit the logistic regression model without an intercept - interpret as usual