2.3 Analysis of matched cohort and case-control studies

OR from unmatched/matched cohort

Assume in the population (for exposure X, confounder Z, outcome Y)

$$\text{P}[Y = 1] = \frac{e^{\alpha+\beta X+\gamma Z}}{1+e^{\alpha+\beta X+\gamma Z}}$$

UNMATCHED COHORT

Logistic regression model:

$$P(Y = 1|X) = \frac{e^{\alpha+\beta X++\gamma Z}}{1+e^{\alpha+\beta X++\gamma Z}}$$

"for a given X, Z"

MATCHED COHORT

Since matching only affects <u>independent</u> variables

For any regression model, we are free to choose "predictors" X, Z

So include matching factors in "usual" logistic model

OR from unmatched/matched case-control

Assume in the population $\quad P[Y = 1] = \frac{e^{\alpha+\beta X+\gamma Z}}{1+e^{\alpha+\beta X+\gamma Z}}$

UNMATCHED CASE-CONTROL

$\text{logit}(P[Y = 1 \mid X, sampled])$

$= \alpha^* + \beta X + \gamma Z$

Correct β, γ

different intercept due to

prevalence in sample ≠ population

$\alpha^* = \alpha + \log\frac{\pi_1}{\pi_0}$

π_1, π_0 sampling fractions of cases, controls

MATCHED (on Z strata)

Assuming common β in all strata

$\text{logit}(P[Y = 1 \mid X, Zs, sampled])$

$= \alpha_s^* + \beta X$

$\alpha_s^* = \alpha + \log\frac{\pi_{1s}}{\pi_{0s}}$

π_{1s}, π_{0s} sampling fractions of cases, controls <u>in stratum s</u>

By fitting "stratum effect" we recover the common β

Note difference between matched cohort and matched case-control

Matched cohort: can estimate the ”stratum effect” as this is just an independent variable, which we are free to choose.

Matched case-control: can adjust for stratum but cannot estimate the effect of stratum on outcome, as we have disturbed this (sampling depends on stratum and outcome!)

Terminology: stratum is “matched away”

A confounder whose effect is of interest should not be used for matching in case-control design

Regression model for matched pairs

For case-control data, denote by X_1 and X_0 the exposure level of case and control respectively

Logistic model for underlying probabilities:

$$P[Y = 1 | X_1] = \frac{e^{\alpha + \beta X1}}{1 + e^{\alpha + \beta X1}} \quad P[Y = 1 | X_0] = \frac{e^{\alpha + \beta X0}}{1 + e^{\alpha + \beta X0}}$$

For each pair, model the probability that event happens to individual with X_1, <u>conditional on one event in the pair</u>

Regression model for matched pairs

For case-control data, denote exposure of case and control as X_1, X_0,
Assume logistic model in population:

$$P[Y=1|X_1] = \frac{e^{\alpha+\beta X_1}}{1+e^{\alpha+\beta X_1}} \qquad P[Y=1|X_0] = \frac{e^{\alpha+\beta X_0}}{1+e^{\alpha+\beta X_0}}$$

$\mathrm{P}(X_1)$ $\qquad\qquad$ $\mathrm{P}(X_0)$

For each pair, model the probability that event happens to individual with X_1, <u>conditional on one event in the pair</u>

$$\text{Conditional probability: } = \frac{\mathrm{P}(x_1)[1-\mathrm{P}(x_0)]}{\mathrm{P}(x_1)[1-\mathrm{P}(x_0)]+[1-\mathrm{P}(x_1)]\mathrm{P}((x_0)}$$

$$= \frac{e^{\beta X_1}}{e^{\beta X_1}+e^{\beta X_0}}$$

Conditional logistic regression

Likelihood to be maximized $= \prod_i \frac{e^{\beta X_{1i}}}{e^{\beta X_{1i}+e^{\beta X_{0i}}}}$ (product over all pairs *"i"*)

$$= \prod \frac{e^{\beta(X_{1i}-X_{0i})}}{1+ e^{\beta(X_{1i}-X_{0i})}}$$

What do we notice about this function?

Where more than one control per case (more terms in denominator):

Likelihood for 1:3 matching $= \prod \frac{e^{\beta X_1}}{e^{\beta X_1+e^{\beta X_{01}+e^{\beta X_{02}+e^{\beta X_{03}}}}}}$

Matched (pairs) cohort study

Remember that OR is reversible (doesn't matter which variable is called "exposure" and which is called "outcome")

Can get adjusted OR from conditional logistic regression by swopping exposure and outcome labels!

For each pair, this models the probability that the case is exposed <u>conditional on one of pair exposed</u>

But we may prefer to have an adjusted RR, which can be obtained from other models (e.g. matched Poisson regression)

Summary of confounding control at analysis stage of matched design

For frequency matched data

stratum variable must be in the (unconditional) model:

- For matched cohort, stratum effect estimated
- For matched case-control, model adjusts for (but cannot estimate) the effects of matching factors

For individually matched data

conduct “conditional” analysis of the matched sets

Stratum not modelled (”matched away”)

Ignoring or breaking the matching

Lot of confusion regarding whether matching at the design stage can be ignored or broken at the analysis stage.

This is mostly due to unclear/inconsistent language.

We will discriminate between:

- **Ignoring** and **breaking** the matching,
- **Pooled**/marginal vs. **stratified** data
- **Conditional** vs. **unconditional** analysis
- **Adjusting** (or not) for matching variables

Ignoring vs. breaking the matching

Ignoring: proceed as if no matching had been used. This means matching varioable could be eliminated from the data set

most crude approach

Breaking: prior to analysis, matched sets are broken into individual records, but the matching factors may play a role in analysis

Three approaches to analysis
(from most crude to most correct)

1. Most crude

Ignore (completely!) the matching:

Combined data from all strata used to estimate crude OR or adjusted (for other confounders) OR.

2. Less strict, but recognises matching:

unconditional analysis, but adjust for the matching variable(s) in the model

3. Most strict/correct

Conditional analysis of matched sets thate were created, e.g. using Mantel-Haenszel, conditional logistic regression

Three approaches to analysis
(from most crude to most correct)

1. Most crude

Ignore (completely!) the matching:

Never appropriate for matched case-control studies, although balance can reduce the bias as seen earlier

Appropriate only under very specific conditions for matched cohort studies

Simple advice:

matching should be accommodated in some way in the analysis.

Three approaches to analysis
(from most crude to most correct)

2. Less strict, but recognises matching:

unconditional analysis, adjusted for matching

Situations where this can be useful for matched cohort data:

- recover loss of matched sets (due to… Quiz)
- undo bias from overmatching
- where matching is on a categorized continuous variable
  (continuous variable to be in the model)
- If matching was unnecessarily fine (e.g. 1 year age groups)

Three approaches to analysis
(from most crude to most correct)

2. Less strict, but recognises matching:

unconditional analysis, adjusted for matching

Prone to bias for matched case-control data

Magnitude of bias depends on

- exposure rate in controls,
- Strength of association (size of the true odds ratio)
- the size of the strata.

If many small strata bias can be serious, e.g. 1:1 (matched pairs)

unconditional OR = $\left[\frac{n_{10}}{n_{01}}\right]^2$ instead of correct $\left[\frac{n_{10}}{n_{01}}\right]$

Bias is less, but still present for larger sets

Three approaches to analysis
(from most crude to most correct)

2. Less strict, but recognises matching:

unconditional analysis, adjusted for matching

Prone to bias for matched case-control data

Advice: conditional analysis

If strong reasons for unconditional analysis (e.g. better precision):

- Only use if matched sets are large
- Check for bias by comparing estimates to conditional estimates

Quiz

If you have categorised a continuous variable to use it as a matching factor, but your model will contain the continuous variable, then the categorised version can be dropped from analysis of:

a) Matched case-control data
b) Matched cohort data
c) Both
d) Neither