Matching and Conditional Likelihood

David M. Rocke

May 2, 2017

Matched Pairs

- Suppose we are studying MI = myocardial infarction and want to examine the effect of smoking on risk of MI.

- We have 100 cases, and we match each case with a control also in the hospital who has not had an MI and is matched on age, race, sex, and hospital status.

- If we tried to use ordinary logistic regression, we would have to use 99 strata variables and one exposure variable with 200 cases. This would not end well.

Four Possible Outcomes

	SMK = 0	SMK = 1
No MI	1	0
MI	1	0

	SMK = 0	SMK = 1
No MI	1	0
MI	0	1

	SMK = 0	SMK = 1
No MI	0	1
MI	1	0

	SMK = 0	SMK = 1
No MI	0	1
MI	0	1

There is always one observation per row, but either 0/2 or 1/1 per column.

Upper left and lower right are indifferent to SMK $\rightarrow$ MI.

Upper right tends to show that smoking is associated with MI.

Lower left tends to show that not smoking is associated with MI.

McNemar's Test

	SMK = 0	SMK = 1
No MI	1	0
MI	1	0

	SMK = 0	SMK = 1
No MI	1	0
MI	0	1

	SMK = 0	SMK = 1
No MI	0	1
MI	1	0

	SMK = 0	SMK = 1
No MI	0	1
MI	0	1

Let X_{ij} represent the number of pairs out of 100 in each of the four sub-tables. We have that X_{11} and X_{22} are uninformative. If smoking is a hazard, then we expect that $X_{21} > X_{12}$ and the reverse if smoking is protective. The statistic

$$\frac{(X_{21} - X_{12})^2}{X_{21} + X_{12}}$$

Has a χ_1^2 distribution asymptotically.

McNemar's Test

	SMK = 0	SMK = 1		SMK = 0	SMK = 1
No MI	1	0	No MI	1	0
MI	1	0	MI	0	1

	SMK = 0	SMK = 1		SMK = 0	SMK = 1
No MI	0	1	No MI	0	1
MI	1	0	MI	0	1

$$\frac{(X_{21} - X_{12})^2}{X_{21} + X_{12}}$$

Has a χ_1^2 distribution asymptotically. This is called McNemar's test and it is numerically identical to the Cochran-Mantel-Haenszel test. It is conditional since we fix the margins on each table of outcomes, one per matched pair.

More General Conditional Logistic Regression

- Conditional logistic regression is often used when the data are divided into many strata, which often happens when we have a matched design.
- The book's MI data set has 39 MI patients, each matched on age, race, sex, and hospital status by two contol patients.
- The primary exposure of interest is SMK = current smoking status (0/1).
- We also have systolic blood pressure, SBP in mm mercury and ECG abnormality (0/1).

mi data set

```
> summary(mi)
    MATCH          PERSON           MI
 Min.   : 1    Min.   : 1     Min.   :0.0000
 1st Qu.:10    1st Qu.: 30    1st Qu.:0.0000
 Median :20    Median : 59    Median :0.0000
 Mean   :20    Mean   : 59    Mean   :0.3333
 3rd Qu.:30    3rd Qu.: 88    3rd Qu.:1.0000
 Max.   :39    Max.   :117    Max.   :1.0000
     SMK           SBP             ECG
 Min.   :0.0000   Min.   :120.0   Min.   :0.0000
 1st Qu.:0.0000   1st Qu.:120.0   1st Qu.:0.0000
 Median :0.0000   Median :140.0   Median :0.0000
 Mean   :0.2821   Mean   :136.4   Mean   :0.2051
 3rd Qu.:1.0000   3rd Qu.:140.0   3rd Qu.:0.0000
 Max.   :1.0000   Max.   :160.0   Max.   :1.0000
```

Conditional Logistic Regression in R

```
library(survival)
> summary(clogit(MI~SMK+SBP+ECG+strata(MATCH),data=mi))
Call:
coxph(formula = Surv(rep(1, 117L), MI) ~ SMK + SBP + ECG + strata(MATCH),
    data = mi, method = "exact")

  n= 117, number of events= 39

        coef exp(coef) se(coef)     z Pr(>|z|)
SMK 0.72906   2.07313  0.56126 1.299  0.19395
SBP 0.04564   1.04670  0.01525 2.994  0.00276 **
ECG 1.59926   4.94938  0.85341 1.874  0.06094 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Conditional Logistic Regression in R

```
library(survival)
> summary(clogit(MI~SMK+SBP+ECG+strata(MATCH),data=mi))
Call:
coxph(formula = Surv(rep(1, 117L), MI) ~ SMK + SBP + ECG + strata(MATCH),
    data = mi, method = "exact")

  n= 117, number of events= 39

    exp(coef) exp(-coef) lower .95 upper .95
SMK     2.073     0.4824    0.6901     6.228
SBP     1.047     0.9554    1.0159     1.078
ECG     4.949     0.2020    0.9292    26.362

Rsquare= 0.173   (max possible= 0.519 )
Likelihood ratio test= 22.2  on 3 df,   p=5.925e-05
Wald test            = 13.68  on 3 df,   p=0.003382
Score (logrank) test = 19.68  on 3 df,   p=0.0001979
```

```
> summary(glm(MI~SMK+SBP+ECG+strata(MATCH),binomial,data=mi))

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)
(Intercept)            -1.251e+01  3.704e+00  -3.378 0.000731 ***
SMK                     1.218e+00  7.175e-01   1.697 0.089607 .
SBP                     7.330e-02  1.997e-02   3.671 0.000242 ***
ECG                     2.784e+00  1.140e+00   2.442 0.014607 *
strata(MATCH)MATCH=2   -4.062e-14  3.054e+00   0.000 1.000000
strata(MATCH)MATCH=3    1.325e+00  2.632e+00   0.503 0.614678
...........
strata(MATCH)MATCH=38  -1.150e+00  2.700e+00  -0.426 0.670136
strata(MATCH)MATCH=39   1.804e+00  2.681e+00   0.673 0.500901
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 148.94  on 116  degrees of freedom
Residual deviance: 113.75  on 75  degrees of freedom
AIC: 197.75

Number of Fisher Scoring iterations: 5
```