Conditional Logistic regression

Department of Statistics, University of South Carolina

Stat 705: Data Analysis II

Matching

The main objective of matching is to make the comparison groups same on everything except the variable of interest

First lets consider paired binary data (eg. data that comes from pre and post treatment, two eyes, twins)

Consider a hypothetical data of 595 subjects pre and post intervention and evaluated for their outcome

	outcome		
X	D=1	D=0	total
pre trt	166	429	595
post trt	276	319	595
Total	442	748	1190

Can we test change in outcome (H0: Pr(D=1/pre trt)=Pr(D=1/post trt)) using a χ^2 test based on this table? NO, because the test based on this table assumes the rows are INDEPENDENT samples, but we have the same people pre and post intervention.

Paired Match Data

As in paired t test, it is required to analyze the data as pairs as follows:

	post	
pre	D=0	D=1
D=0	n_{00}	n_{01}
D=1	n_{10}	n_{11}

Paired Match Data

- The concordant pairs (n_{00} and n_{11}) do not contribute any information about the effect of X or the intervention.
- So, we use the information in the discordant pairs (n_{01} and n_{10}) to measure treatment effect
- Under H0, we expect equal change from 0 to 1 and from 1 to 0, i.e $E(n_{10}) = E(n_{01})$. So, under the null,
  $n_{10}|(n_{01} + n_{10})$ is Binomial$(n_{01} + n_{10}, 1/2)$

$$Z = \frac{n_{10} - E(n_{10})}{\sqrt{(n_{01}+n_{10})1/2(1-1/2)}}. \sim N(0,1)$$

Z^2 is approximately distributed $\chi^2(1)$

- If the sample size is small, exact p-value based on binomial distribution can be obtained.

Paired Binary Data Analysis

The MLE for the odds ratio comparing pre and post trt groups is

$$\text{OR}=\frac{n_{01}}{n_{10}}$$

For the example data we considered above,

	post	
pre	D=0	D=1
D=0	251	178
D=1	68	98

Hypothesis: $H_0 : Pr(D = 1|pre\ trt) = Pr(D = 1|post\ trt))$

$$Z = \frac{n_{10}-E(n_{10})}{\sqrt{((n_{01}+n_{10})1/2(1-1/2))}} \sim N(0,1)$$
$$= \frac{178-(178+68)/2}{\sqrt{(178+68)/4}}=7.01 \text{ leads to } p\text{-value}<0.001$$

$$\text{OR}=178/68=2.62$$

Matching in case-control studies

- It can be "individually matched" where one case is matched to one or more controls or "group matched" where two or more cases are matched to one or more controls
- The most commonly used design is 1:1 matched
- once we match on certain factors, we are forfeiting estimating their effect
- So, they are nuisance parameters in the model

Conditional vs Unconditional Logistic Likelihood

The model for a matched data with $k = 1, ..., K$ strata is

$$logit[\pi_k(X)] = \alpha_k + \beta_1 X_1 + ... + \beta_p X_p$$

Where $\pi_k(X) = \Pr(D_{ik} = 1|X)$, α_k is log-odds in the kth stratum

- unless the number of subjects in each stratum is large, fitting these models using the unconditional ML does not work well
- in individually matched there is only one case in each stratum and hence we need some way of getting rid of the nuisance parameters
- Conditional likelihood - condition on a sufficient statistic for the nuisance parameter
- so the conditional likelihood for the k the stratum is obtained as the probability of the observed data conditional on the stratum total and the number of cases observed

Logistic regression for Matched data

Consider the simplest case, the 1:1 matched design with $k = 1, ..., K$ strata and p covariates

$$logit(\pi_k(X)) = \alpha_k + \beta' X$$

Where $\pi_k(X) = \Pr(D_{ik} = 1|X)$, α_k is log-odds in the kth stratum; X_{0k} be the data vector for the control and X_{1k} be the data vector for the case. $S_k = D_{0k} + D_{1k}$.

$$L_k(\beta) = \Pr(D_{1k} = 1, D_{0k} = 0|X_{1k}, X_{0k}, S_k = 1, n_k = 2)$$

Conditional Likelihood

$$
\begin{aligned}
L_k(\beta) &= \Pr(D_{1k}=1, D_{0k}=0 | X_{1k}, X_{0k}, S_k=1, n_k=2) \\
&= \frac{\Pr(D_{1k}=1|X_{1k})\Pr(D_{0k}=0|X_{0k})}{\Pr(D_{1k}=1|X_{1k})\Pr(D_{0k}=0|X_{0k}) + \Pr(D_{0k}=1|X_{0k})\Pr(D_{1k}=0|X_{1k})} \\
&= \frac{exp(\alpha_k + \beta' X_{1k})}{exp(\alpha_k + \beta' X_{1k}) + exp(\alpha_k + \beta' X_{0k})} \\
&= \frac{exp(\beta' X_{1k})}{exp(\beta' X_{1k}) + exp(\beta' X_{0k})} \\
L(\beta) &= \prod_{k=1}^{K} L_k
\end{aligned}
$$

This for binary univariate X results in the same OR reported above.

Matched Case-Control Study Of Esophageal Cancer

- The dataset contains 119 observations with 3 covariates.
- smoking
- rubber: whether the participant exposed to rubber industry
- alcohol: whether the participant drinks alcohol

R Code

```
> library(survival)
> library("epiDisplay")
>
> data(VC1to6)
> VC1to6[1:10,]
   matset case smoking rubber alcohol
1       1    1       1      0       0
2       1    0       1      0       0
3       2    1       1      0       1
4       2    0       1      1       0
5       3    1       1      1       0
6       3    0       1      1       0
7       4    1       1      0       0
8       4    0       1      1       1
9       4    0       0      1       1
10      5    1       0      0       1
> attach(VC1to6)
The following objects are masked from VC1to6 (pos = 3):

    alcohol, case, matset, rubber, smoking
```

R Code: Conditional Logistic Regression

```
> fitclogit<-clogit(case ~ smoking + rubber + alcohol + strata(matset), data=VC1to6)
> summary(fitclogit)
Call:
coxph(formula = Surv(rep(1, 119L), case) ~ smoking + rubber +
    alcohol + strata(matset), data = VC1to6, method = "exact")

  n= 119, number of events= 26

           coef exp(coef) se(coef)      z Pr(>|z|)
smoking  0.4398    1.5523   0.6462  0.681  0.49616
rubber  -0.4572    0.6331   0.6474 -0.706  0.48002
alcohol  1.6668    5.2951   0.5952  2.800  0.00511 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

        exp(coef) exp(-coef) lower .95 upper .95
smoking    1.5523     0.6442    0.4375     5.508
rubber     0.6331     1.5797    0.1780     2.251
alcohol    5.2951     0.1889    1.6490    17.003

Rsquare= 0.096   (max possible= 0.471 )
Likelihood ratio test= 12  on 3 df,   p=0.007
Wald test            = 9.18  on 3 df,   p=0.03
Score (logrank) test = 11.24  on 3 df,   p=0.01
```