Lecture 14: Logistic regression for case-control data and conditional logistic regression

Reading:

Regression Analysis, Fall 2015
Institute of Statistics, National Chiao Tung University

December 15, 2015

Brief outline

1. Risk and odds ratio in case-control study
2. Logistic regression of case-control data
3. When there are many strata with few observations in each
4. Conditional logistic regression

Risk and odds ratio in case-control study

- So far all derivations and analyses with logistic regression have assumed cohort sampling so that both the risk and the odds ratio can be estimated.
- However, we know that risks cannot be estimated from case-control studies without additional, external information.
- Under case-control studies, logistic regression can be used to estimate odds ratios, but not risks.

Logistic regression for case-control data

- Under the cohort study, we can write logistic regression for an outcome D as

$$\Pr(D = 1|x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}, \tag{1}$$

  which means, the probability of disease given the value of a predictor x.

- In a case-control study, we have selected cases and controls from the population. We observe and have to model $\Pr(D = 1|x, S = 1)$, where $S = 1$ indicates selection into the study.
  In other words, the probability of being a case in the sample, depends both on the risk factor and on being in the study.

- Utilizing conditional probability and Bayes formula, we obtain

$$\Pr(D = 1|x, S = 1) = \frac{\exp(\ln(\frac{p_1}{p_0}) + \beta_0 + \beta_1 x)}{1 + \exp(\ln(\frac{p_1}{p_0}) + \beta_0 + \beta_1 x)}, \quad (2)$$

  $p_1 = \Pr(S = 1|D = 1)$: the probability that a case in the population is selected into the study, and
  $p_0 = \Pr(S = 1|D = 0)$: the probability that a control in the population is selected into the study.
  Notice that this is just the logistic formula (1) we started with, except that the intercept is different.

- What we have just shown is that modeling a case-control study by logistic regression just as if it were a cohort study yields the correct odds ratio, but unless we know the sampling probabilities, we can't calculate the disease risk.

- Example: IVH and antenatal steroids (BETA) use
  1. Determine which neonates had IVH from discharge records.
  2. Do a case-control study including all neonates with IVH, but applying a sampling probability of 40% to those without IVH.
  3. Fit a logistic regression for IVH on BETA, pretending the data are from a cohort study.

- 4. We get the result

$$\ln[\frac{\Pr(\text{IVH}=1|\text{BETA},S=1)}{\Pr(\text{IVH}=0|\text{BETA},S=1)}] = -8.24 + 2.17 \times \text{BETA}$$

The odds ratio

$$\begin{aligned} & \frac{\text{odds}(\text{IVH}=1|\text{BETA}=1,S=1)}{\text{odds}(\text{IVH}=1|\text{BETA}=0,S=1)} \\ = & \frac{\text{odds}(\text{IVH}=1|\text{BETA}=1)}{\text{odds}(\text{IVH}=1|\text{BETA}=0)} \\ = & \exp(2.17) = 8.72, \end{aligned}$$

the "true" intercept $= -8.24 - \ln(1/0.4) = -9.16$, and

$$\begin{aligned} \Pr(\text{IVH}=1|\text{BETA}=1) &= \frac{\exp(-9.16+2.17\times 1)}{1+\exp(-9.16+2.17\times 1)} \\ &= 0.00092 \end{aligned}$$

Many strata with few observations in each

- When grouping continuous variables into categories,
  1. there is always a concern that the covariate should be grouped into small enough strata to be sure confounding has been properly adjusted for.
  2. But many strata with few observations in each may lead to problems with likelihood inference.
- When there is matching in the study. For example, each subject with a risk factor may be matched to a neighbor without the risk factor.
  1. This procedure is designed to simultaneously control for many confounders, some of which may not be explicitly measurable.
  2. In a matched study, the number of strata may increase almost as rapidly as the number of subjects, definitely invalidating likelihood inference.

Modeling many strata in logistic regression

- Formulate a logistic regression with indicator variables for all strata

$$\Pr(Y_i = 1|x_i) = \frac{\exp(\beta_0 + \sum_{k=2}^{K} \beta_{0k} STRA_{ik} + \beta_1 x_i)}{1 + \exp(\beta_0 + \sum_{k=2}^{K} \beta_{0k} STRA_{ik} + \beta_1 x_i)},$$

where

$$STRA_{ik} = \begin{cases} 1 & \text{if } Y_i \in \text{stratum } k \\ 0 & \text{if } Y_i \notin \text{stratum } k \end{cases}$$

- Above formula can also be written

$$\Pr(Y_i = 1|x_i) = \frac{\exp(\alpha_k + \beta_1 x_i)}{1 + \exp(\alpha_k + \beta_1 x_i)}$$

if $Y_i \in$ stratum k

- If K is very large, we have problems with likelihood inference in regression coefficients.

Conditional logistic regression

- The procedure that has been devised for the situation with many small strata is conditional logistic regression.
  1. Conditional logistic regression is a cross between Fisher's exact test and logistic regression.
  2. By conditioning on the total number of events in each stratum, it allows one to avoid explicitly fitting stratum indicator.
  3. The likelihood is constructed from the conditional probability that, out of all possible combinations, the events occur in the people actually being observed to have them.

- Except for the fact that there are no stratum indicators, and no intercept, everything works the same for conditional logistic regression as it did for (unconditional) logistic regression. Coefficient interpretation is the same, likelihood inference is the same, and it works for both cohort and case-control studies.
- When using conditional logistic regression, we can not obtain the risk estimate because the intercept and stratum indicators are not estimable.
- Conditional logistic regression can be fit by tricking the software program for a Cox model (e.g., function `clogit` in R package `survival`) or by the EXACT logistic regression (e.g., function `clogistic` in R package `Epi`). Practical experience to date indicates that the latter option is practical only for small sample sizes.

- Eample: neonatal mortality (DEATH) and birth year (YR89, YR90), birth weight (BW).
  1. (Unconditional) logistic regression

$$\ln\left(\frac{\Pr(\text{DEATH})}{\Pr(\text{ALIVE})}\right) = -1.59 - 0.61(\text{YR89}) - 0.37(\text{YR90}) - 0.004(\text{BW}) + 0.000004(\text{BW})^2$$

  2. We may still be worried that the quadratic fit for birth weight may be inadequate for complete control of confounding.

- 3. To control for confounding more convincingly, we group birth weights into 30g intervals, and fit conditional logistic regression:
  The results are:
  coefficient of YR89 = -0.65
  coefficient of YR90 = -0.40
- 4. We see that results are quite similar from unconditional and conditional logistic regression. We conclude that either method can be used to adjust for confounding BW.