HW8

Jinhong Du, 15338039

目录

1 Cancer

A case-control study was performed early in the NHS to assess the possible association between oral contraceptive (OC) use and ovarian cancer [50]. Forty seven ovarian cancer cases were identified at or before baseline (1976). For each case, 10 controls matched by year of birth and with intact ovaries at the time of the index woman's diagnosis were randomly chosen from questionnaire respondents free from ovarian cancer. The data in Table 13.56 were presented.

Duration OC use

Age at diagnosis	Never	< 3 years	3+ years	
Under 35	Case	9	2	0
	Control	55	42	12
35 – 44	Case	13	2	4
	Control	127	27	30

Age at diagnosis	Never	< 3 years	3+ years	
45+	Case	12	3	2
	Control	129	18	23

Table 13.56 Duration of OC use by age at diagnosis among women with ovarian cancer and controls

1.1 13.109 Use logistic regression methods to assess whether there is an association between ovarian cancer risk and duration of OC use while controlling for age. Provide a two-sided p-value. Assume that the average duration of use in the < 3 years group $= 1.5$ years and in the $3+$ years group $= 4$ years. Also, provide an estimate of the OR relating ovarian cancer risk per year of use of OCs and a 95% CI.

Let

$$x_{ij} = \begin{cases} 0 & \text{, if one never uses OC} \\ 1.5 & \text{, if duration OC use} < 3 \text{ years} \\ 4 & \text{, if duration OC use} \geq 3 \text{ years} \end{cases}$$

The conditional Logistic Regression model is given by

$$\ln \frac{p_{ij}}{1 - p_{ij}} = \alpha_i + \beta_1 x_{ij}$$

where α_i is the indicator variable for being in the ith matched set, which $= 1$ if a subject is in the ith matched set and $= 0$ otherwise.

To test

$$H_0 : \beta_1 = 0 \qquad H_1 : \beta_1 \neq 0$$

when successes and failures are more than 20 respectively, we have approximately

$$\frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \overset{H_0}{\sim} N(0, 1)$$

where $se(\hat{\beta}_1)$ is the first diagonal of the covariance matrix of coefficients $\hat{\Sigma} = I(\hat{\beta})^{-1} = -\mathbb{E}\left[\frac{\partial^2 l}{\partial \hat{\beta} \partial \hat{\beta}^\top}\right]$.

The estimate of the OR relating ovarian cancer risk per Δ years of use of OCs after controlling x_2 is given by $\hat{OR}_\Delta = e^{\hat{\beta}_1 \Delta}$ and its two-sided $100\% \times (1 - \alpha)$ CI for the trur OR is given by

$$(e^{[\hat{\beta}_1 - z_{1-\frac{\alpha}{2}} se(\hat{\beta}_1)]\Delta}, e^{[\hat{\beta}_1 + z_{1-\frac{\alpha}{2}} se(\hat{\beta}_1)]\Delta})$$

```
dataset <- matrix(c(9,2,0,
                    55,42,12,
                    13,2,4,
                    127,27,30,
                    12,3,2,
                    129,18,23), nrow = 6, ncol = 3, byrow = T)

library(survival)
y <- c(rep(1,sum(apply(dataset,1,sum)[c(1,3,5)])), rep(0,sum(apply(dataset,1,sum)[c(2,4,6)])))
x <- c(rep(0,sum(dataset[c(1,3,5),1])), rep(1.5,sum(dataset[c(1,3,5),2])),
        rep(4,sum(dataset[c(1,3,5),3])), rep(0,sum(dataset[c(2,4,6),1])),
        rep(1.5,sum(dataset[c(2,4,6),2])), rep(4,sum(dataset[c(2,4,6),3])))
x2 <- c(rep(c(0,1,2), times = dataset[c(1,3,5),1]),
        rep(c(0,1,2), times = dataset[c(1,3,5),2]),
        rep(c(0,1,2), times = dataset[c(1,3,5),3]),
        rep(c(0,1,2), times = dataset[c(2,4,6),1]),
        rep(c(0,1,2), times = dataset[c(2,4,6),2]),
        rep(c(0,1,2), times = dataset[c(2,4,6),3]))


model <- clogit(y ~ x + strata(x2))
summary(model)
```

```
## Call:
## coxph(formula = Surv(rep(1, 510L), y) ~ x + strata(x2), method = "exact")
##
##   n= 510, number of events= 47
##
##       coef exp(coef) se(coef)      z Pr(>|z|)
## x -0.05963   0.94212  0.11577 -0.515    0.607
##
##   exp(coef) exp(-coef) lower .95 upper .95
## x    0.9421      1.061    0.7509     1.182
##
## Rsquare= 0.001   (max possible= 0.445 )
## Likelihood ratio test= 0.28  on 1 df,   p=0.6
## Wald test            = 0.27  on 1 df,   p=0.6
## Score (logrank) test = 0.27  on 1 df,   p=0.6
```

The p-value is $0.61 > \alpha = 0.05$, so we cannot reject H_0, i.e., we cannot conclude that there is not an association between ovarian cancer risk and duration of OC use while controlling for age.

```
cat('The estimated OR is ',summary(model)$conf.int[1],
'\nThe 96% CI is (',summary(model)$conf.int[3],',',
summary(model)$conf.int[4],')\n')
```

```
## The estimated OR is  0.9421172
## The 96% CI is ( 0.750859 , 1.182093 )
```

1.2 13.110 Use logistic regression methods to assess whether there is an association between ever use of OCs and ovarian cancer risk, while controlling for age. Also, provide an estimate of the OR and a 95% CI about this estimate.

Let

$$x'_{ij} = \begin{cases} 0 & \text{, if one never uses OC} \\ 1 & \text{, if one used OC} \end{cases}$$

The conditional Logistic Regression model is given by

$$\ln \frac{p_{ij}}{1-p_{ij}} = \alpha_i + \beta'_1 x'_{ij}$$

The hypothesis is the same as the one in 13.109 and so does the estimated OR and its 95% CI.

```
x11 <- c(rep(0,sum(dataset[c(1,3,5),1])), rep(1,sum(dataset[c(1,3,5),c(2,3)])),
        rep(0,sum(dataset[c(2,4,6),1])), rep(1,sum(dataset[c(2,4,6),c(2,3)])))
model <- clogit(y ~ x11 + strata(x2))
summary(model)
```

```
## Call:
## coxph(formula = Surv(rep(1, 510L), y) ~ x11 + strata(x2), method = "exact")
##
##   n= 510, number of events= 47
##
##         coef exp(coef) se(coef)      z Pr(>|z|)
## x11 -0.2526    0.7768   0.3451 -0.732    0.464
##
##     exp(coef) exp(-coef) lower .95 upper .95
## x11    0.7768      1.287    0.3949     1.528
##
## Rsquare= 0.001   (max possible= 0.445 )
## Likelihood ratio test= 0.55  on 1 df,   p=0.5
## Wald test            = 0.54  on 1 df,   p=0.5
## Score (logrank) test = 0.54  on 1 df,   p=0.5
```

The p-value is $0.464 > \alpha = 0.05$, so we cannot reject H_0, i.e., we cannot conclude that there is an association between ever use of OCs and ovarian cancer risk while controlling for age.

```
cat('The estimated OR is ',summary(model)$conf.int[1],
'\nThe 96% CI is (',summary(model)$conf.int[3],',',
summary(model)$conf.int[4],')\n')
```

```
## The estimated OR is  0.77676
## The 96% CI is ( 0.39492 , 1.527793 )
```

2 Cardiovascular Disease

The Women's Health Study randomly assigned 39,876 initially healthy women ages 45 years or older to receive either 100 mg of aspirin on alternate days or placebo and monitored them for 10 years for a major cardiovascular event [52]. Table 13.58 shows the results stratified by age at randomization.

Age	Treatment group	CVD=yes	CVD=no
$45-54$	Aspirin	163	11,847
	Placebo	161	11,854
$55-64$	Aspirin	183	5693
	Placebo	186	5692
≥ 65	Aspirin	131	1917
	Placebo	175	1874

Table 13.58 Incidence of CVD by treatment group and age in the Women's Health Study

Use logistic regression methods to characterize the relationship between aspirin assignment and the odds of CVD, by doing the following.

2.1 13.113 Obtain the crude OR estimate, and provide a 95% CI for the crude OR.

Let

$$\ln \frac{p}{1-p} = \alpha + \beta_1 x$$

where $x = \begin{cases} 1 & \text{, Aspirin} \\ 0 & \text{, Placebo} \end{cases}$

```
dataset2 <- matrix(c(163,161,183,186,131,175,
                     11847,11854,5693,5692,1917,1874), 6,2)
a <- sum(dataset2[c(1,3,5),1])
c <- sum(dataset2[c(1,3,5),2])
b <-  sum(dataset2[c(2,4,6),1])
```

```
d <-  sum(dataset2[c(2,4,6),2])
y <- c(rep(1,a+b),rep(0,c+d))
x <- c(rep(1,a),rep(0,b),rep(1,c),rep(0,d))
dat2.glm <- glm(y ~ x, family = "binomial")
summary(dat2.glm)
```

```
##
## Call:
## glm(formula = y ~ x, family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.2303  -0.2303  -0.2201  -0.2201   2.7323
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.61639    0.04435 -81.537   <2e-16 ***
## x           -0.09205    0.06415  -1.435    0.151
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9338.9  on 39875  degrees of freedom
## Residual deviance: 9336.9  on 39874  degrees of freedom
## AIC: 9340.9
##
## Number of Fisher Scoring iterations: 6
```

```
crude_OR <- exp(dat2.glm$coefficients[2])
se_beta1 <- summary(dat2.glm)$coefficients[2,2]
cat('The crude OR estimate is ',crude_OR,
    '\nThe 95% CI is (',crude_OR*exp(-qnorm(0.975)*se_beta1),
    ',',crude_OR*exp(qnorm(0.975)*se_beta1),')')
```

```
## The crude OR estimate is  0.9120554
## The 95% CI is ( 0.8043008 , 1.034246 )
```

2.2 13.114 Test the null hypothesis of no association between aspirin assignment and CVD.

To test

$$H_0 : \beta_1 = 0 \qquad\qquad H_1 : \beta_1 \neq 0$$

Since the p-value$= 0.151 > \alpha = 0.05$, we cannot reject H_0, i.e., we cannot conclude that there is not an association between aspirin assignment and CVD.

2.3 13.115

Evaluate whether age confounds the CVD—aspirin relationship by using dummy variables for age categories; calculate the age-adjusted OR estimate and 95% CI.

Let

$$\ln \frac{p}{1-p} = \alpha + \beta_1 x + \beta_2 x_2$$

where

$$x = \begin{cases} 0 & \text{, Aspirin} \\ 1 & \text{, Placebo} \end{cases}$$

$$age_1 = \begin{cases} 1 & \text{, if one's age is between } 55-64 \\ 0 & \text{, otherwise} \end{cases}$$

$$age_2 = \begin{cases} 1 & \text{, if one's age is } \geq 65 \\ 0 & \text{, otherwise} \end{cases}$$

Here we choose one category $(45-54)$ to be the reference category and create two dummy variables to represent group membership in age groups 5564 and ≥ 45, respectively.

```r
age <- c(rep(c(0,1,2),times = dataset2[c(1,3,5),1]),
        rep(c(0,1,2),times = dataset2[c(2,4,6),1]),
        rep(c(0,1,2),times = dataset2[c(1,3,5),2]),
        rep(c(0,1,2),times = dataset2[c(2,4,6),2]))
dat2.glm2 <- glm(y ~ x + factor(age), family = "binomial")
summary(dat2.glm2)
```

```
##
## Call:
## glm(formula = y ~ x + factor(age), family = "binomial")
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4028  -0.2467  -0.1686  -0.1609   2.9507
```

```
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -4.24692    0.06387 -66.495   <2e-16 ***
## x            -0.09337    0.06461  -1.445    0.148
## factor(age)1  0.86331    0.07699  11.214   <2e-16 ***
## factor(age)2  1.77586    0.08162  21.759   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 9338.9  on 39875  degrees of freedom
## Residual deviance: 8888.8  on 39872  degrees of freedom
## AIC: 8896.8
##
## Number of Fisher Scoring iterations: 7
```

```
age_OR <- exp(dat2.glm2$coefficients[2])
se_beta2 <- summary(dat2.glm2)$coefficients[2,2]
cat('The age-adjusted OR estimate is ',age_OR,
    '\nThe 95% CI is (',age_OR*exp(-qnorm(0.975)*se_beta2),
    ',',age_OR*exp(qnorm(0.975)*se_beta2),')')
```

```
## The age-adjusted OR estimate is  0.9108603
## The 95% CI is ( 0.802525 , 1.03382 )
```

Since the age-adjusted OR $= 0.9108603$ lies in $(0.802525, 1.03382)$, the 95% CI of the crude OR, the age may not confound the CVD—aspirin relationship.

2.4 13.116 Evaluate whether age is an effect modifier of the relationship between aspirin and CVD.

For $i = 2, 3$, to test

$$H_0 : \beta_i = 0 \qquad\qquad H_1 : \beta_i \neq 0$$

The statistic is given be

$$\frac{\hat{\beta}_i}{se(\hat{\beta}_i)} \overset{H_0}{\sim} N(0,1)$$

approximately when there are more than 20 successes and failures respectively.

Since each p-value$< 2e - 16 < \alpha = 0.05$, we reject H_0, i.e, there is no difference of OR between people whose ages are between $45 - 55$ and whose ages are between $55 - 64$(or ≥ 65). It means that age is not an effect modifier of the relationship between aspirin and CVD.