STAT 226 Logistic Regression for Retrospective Studies

Yibi Huang

Example: Birdkeeping and Lung Cancer Data

A 1972–1981 health survey in The Hague, Netherlands, discovered an association between keeping pet birds and increased risk of lung cancer. To investigate bird-keeping as a risk factor, researchers conducted a **case–control** study of patients in 1985 at 4 hospitals in The Hague. They identified 49 cases of lung cancer among patients who were registered with a general practice, who were age 65 or younger, and who had resided in the city since 1965. They also selected 98 controls from a population of residents having the same general age structure [1].

```
birdkp = read.table(
  "http://www.stat.uchicago.edu/~yibi/s226/birdkeeping.txt",
  header=TRUE)
```

[1] Data from Chapter 20 of *The Statistical Sleuth*, 3ed, by Ramsey and Schafer

Birdkeeping and Lung Cancer Data Variables

- `LC`: Whether subject has lung cancer
- `FM`: Sex — `Male` or `Female`
- `AG`: Age, in years
- `SS`: Socioeconomic status — `High`or `Low`, determined by occupation of the household's principal wage earner
- `YR`: Years of smoking prior to diagnosis or examination
- `CD`: Average rate of smoking, in cigarettes per day
- `BK`: 2 levels: `Bird` or `NoBird`, indicating whether the subject kept caged birds at home for more than 6 consecutive months from 5 to 14 years before diagnosis (cases) or examination (controls)

Marginal OR between LR and BK

```
xtabs(~ BK + LC, data=birdkp)
        LC
BK         LungCancer NoCancer
  Bird             33       34
  NoBird           16       64
```

As the study is retrospective, the only prospective quantity that can be estimated is the odds ratio

$$OR = \frac{33 \times 64}{34 \times 16} \approx 3.88$$

- Odds of lung cancer among bird keepers were about 3.88 times as large as the odds among non-birdkeepers.
- However, several other variables need to be controlled, like, age, years of smoking, and so on.

OR between LC and BK Given Years of Smoking

```
birdkp$yr.smk = cut(birdkp$YR, breaks= seq(0,50,10),
                    include.lowest=TRUE, right=FALSE)
ftable(xtabs(~ BK + LC + yr.smk, data=birdkp),
       row.vars = c("yr.smk","BK"), col.vars=c("LC"))
               LC LungCancer NoCancer

yr.smk  BK
[0,10)  Bird               1        8
        NoBird             0       12
[10,20) Bird               2        2
        NoBird             1        9
[20,30) Bird               8       11
        NoBird             4       14
[30,40) Bird              11       10
        NoBird             6       14
[40,50] Bird              11        3
        NoBird             5       15
```

For bird keepers, odds of lung cancer remained higher comparing to non-keepers with similar years of smoking.

CMH test of the conditional independence of LC and BK given years of smoking:

```
options(digits=6)
mantelhaen.test(xtabs(~ BK + LC + yr.smk, data=birdkp), correct = F)

    Mantel-Haenszel chi-squared test without continuity correction

data:  xtabs(~BK + LC + yr.smk, data = birdkp)
Mantel-Haenszel X-squared = 14.01, df = 1, p-value = 0.000182
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 1.92775 9.01553
sample estimates:
common odds ratio
           4.1689
```

OR between LC and BK Given Age

```
birdkp$age = cut(birdkp$AG, breaks= seq(40,70,5),
                    include.lowest=TRUE, right=FALSE)
ftable(xtabs(~ BK + LC + age, data=birdkp),
       row.vars = c("age","BK"), col.vars=c("LC"))
```

```
               LC LungCancer NoCancer
age     BK
[40,45) Bird               2        1
        NoBird             1        4
[45,50) Bird               6        9
        NoBird             0        4
[50,55) Bird               3        4
        NoBird             2        5
[55,60) Bird               6        7
        NoBird             6       16
[60,65) Bird              12       11
        NoBird             6       25
[65,70] Bird               3        1
        NoBird             1        9
```

CMH test of the conditional independence of LC and BK given age:

```
mantelhaen.test(xtabs(~ BK + LC + age, data=birdkp), correct = F)


    Mantel-Haenszel chi-squared test without continuity correction

data:  xtabs(~BK + LC + age, data = birdkp)
Mantel-Haenszel X-squared = 14.15, df = 1, p-value = 0.000169
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 1.96925 9.17732
sample estimates:
common odds ratio
          4.25117
```

Advantage of Logistic Regression for Retrospective Studies

- CMH tests only work for $2 \times 2 \times K$ tables. To control for `YR` or `AG`, we need to turn them into grouping variables `yr.smk` and `age`, cannot use their numerical values.
- CMH tests can only control for one variable at a time
- Logistic regression can control for several variables at once, `YR` of smoking, `AG`, gender, social-economic status, all at once.
- Logistic regression can take the numerical values of a numerical control variable into account.

However...

Please note the logistic regression models the prospective probabilities

$$P(Y = 1 \mid X_1, X_2, \ldots, X_p) = \frac{\exp(\alpha + \beta_1 x_1 + \cdots + \beta_p x_p)}{1 + \exp(\alpha + \beta_1 x_1 + \cdots + \beta_p x_p)}.$$

Can we estimate the coefficients α and β_i's of the prospective probabilities using retrospective data?

- The intercept α **cannot** be estimated from retrospective data
- The coefficient β_i for X_i **can** be estimated from retrospective data if neither the probability that a case ($Y = 1$) is selected nor the probability that a control ($Y = 0$) is selected depend on X_i
- The prospective probability $P(Y = 1 \mid X_1, X_2, \ldots, X_p)$ cannot be estimated from retrospective data

Why Can Prospective β_i's Be Estimated Retrospectively?

For demonstration purpose, we use logistic regression with 2 predictors X_1 and X_2 as an example. Let

$$\begin{aligned}
\rho_{1,x_1,x_2} &= P(\text{selected} \mid Y = 1, X_1 = x_1, X_2 = x_2) \\
&= \text{the chance that a diseased case with } X_1 = x_1,\ X_2 = x_2 \\
&\quad\ \text{is included in the data} \\
\rho_{0,x_1,x_2} &= P(\text{selected} \mid Y = 0, X_1 = x_1, X_2 = x_2) \\
&= \text{the chance that a control case with } X_1 = x_1,\ X_2 = x_2 \\
&\quad\ \text{is included in the data}
\end{aligned}$$

As the disease is usually rare, to get enough diseased cases in the sample, usually in a retrospective study, the sampling rate among diseased cases ρ_{1,x_1,x_2} is much higher than the sampling rate among the control ρ_{0,x_1,x_2}.

Why Can Prospective β_i's Be Estimated Retrospectively?

Assume the correct model if the data are obtained **prospectively** is

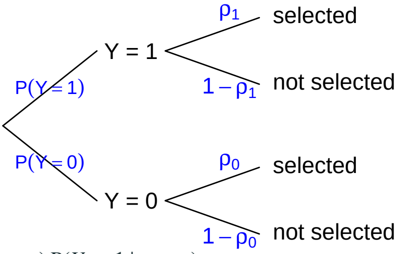
$$P(Y = 1 \mid x_1, x_2) = \frac{\exp(\alpha + \beta_1 x_1 + \beta_2 x_2)}{1 + \exp(\alpha + \beta_1 x_1 + \beta_2 x_2)}.$$

However, what would be the model if the data are obtained **retrospectively**? That is among the "selected", what's the probability that $Y = 1$ given $X_1 = x_1, X_2 = x_2$

$$P(Y = 1 \mid \text{selected},\ x_1, x_2) = ?$$

Why Can Prospective β_i's Be Estimated Retrospectively?

By Bayes' Theorem,



$$
\begin{aligned}
& P(Y = 1 \mid \text{selected}, x_1, x_2) \\
&= \frac{P(\text{selected}|Y = 1, x_1, x_2)P(Y = 1 \mid x_1, x_2)}{P(\text{selected}|Y = 0, x_1, x_2)P(Y = 0 \mid x_1, x_2) + P(\text{selected}|Y = 1, x_1, x_2)P(Y = 1 \mid x_1, x_2)} \\
&= \frac{\rho_{1,x_1,x_2} P(Y = 1 \mid x_1, x_2)}{\rho_{0,x_1,x_2} P(Y = 0 \mid x_1, x_2) + \rho_{1,x_1,x_2} P(Y = 1 \mid x_1, x_2)} \\
&= \frac{\rho_{1,x_1,x_2} P(Y = 1 \mid x_1, x_2)/P(Y = 0 \mid x_1, x_2)}{\rho_{0,x_1,x_2} + \rho_{1,x_1,x_2} P(Y = 1 \mid x_1, x_2)/P(Y = 0 \mid x_1, x_2)} \quad \left(\begin{array}{c}\text{divide both top and} \\ \text{bottom by } P(Y = 0 \mid x_1, x_2)\end{array}\right) \\
&= \frac{\rho_{1,x_1,x_2} \exp(\alpha + \beta_1 x_1 + \beta_2 x_2)}{\rho_{0,x_1,x_2} + \rho_{1,x_1,x_2} \exp(\alpha + \beta_1 x_1 + \beta_2 x_2)}
\end{aligned}
$$

where $P(Y = 1 \mid x_1, x_2)/P(Y = 0 \mid x_1, x_2) = \exp(\alpha + \beta_1 x_1 + \beta_2 x_2)$ because we assume the correct prospective model to be $P(Y = 1 \mid x_1, x_2) = \frac{\exp(\alpha+\beta_1 x_1+\beta_2 x_2)}{1+\exp(\alpha+\beta_1 x_1+\beta_2 x_2)}$.

Why Can Prospective β_i's Be Estimated Retrospectively?

If the sampling rates $\rho_{1,x_1,x_2} = \rho_1$ and $\rho_{0,x_1,x_2} = \rho_0$ do NOT depend on the predictors x_1 and x_2, then

$$
\begin{aligned}
P(Y=1\,|\,\text{selected}, x_1, x_2) &= \frac{\rho_{1,x_1,x_2}\exp(\alpha+\beta_1 x_1+\beta_2 x_2)}{\rho_{0,x_1,x_2}+\rho_{1,x_1,x_2}\exp(\alpha+\beta_1 x_1+\beta_2 x_2)} \\
&= \frac{\rho_1\exp(\alpha+\beta_1 x_1+\beta_2 x_2)}{\rho_0+\rho_1\exp(\alpha+\beta_1 x_1+\beta_2 x_2)} \\
&= \frac{(\rho_1/\rho_0)\exp(\alpha+\beta_1 x_1+\beta_2 x_2)}{1+(\rho_1/\rho_0)\exp(\alpha+\beta_1 x_1+\beta_2 x_2)} \\
&= \frac{\exp(\alpha'+\beta_1 x_1+\beta_2 x_2)}{1+\exp(\alpha'+\beta_1 x_1+\beta_2 x_2)} \quad \text{where } \alpha' = \alpha + \log(\rho_1/\rho_0).
\end{aligned}
$$

We can thus estimate the β_j's for a prospective model using retrospective since the retrospective data follow a logistic model $P(Y=1\,|\,\text{selected}, x_1, x_2)$ with identical β_j's for predictors as those for the prospective model $P(Y=1\,|\,x_1,x_2) = \frac{\exp(\alpha+\beta_1 x_1+\beta_2 x_2)}{1+\exp(\alpha+\beta_1 x_1+\beta_2 x_2)}$. Only the intercept is different.

Caution: Cannot Estimate Some β_i's Retrospectively if ...

If the sampling rates $\rho_{1,x_1,x_2} = \rho_{1,x_2}$ and $\rho_{0,x_1,x_2} = \rho_{0,x_2}$ depend x_2 but not x_1, then

$$
\begin{aligned}
P(Y = 1 \mid \text{selected}, x_1, x_2) &= \frac{\rho_{1,x_1,x_2} \exp(\alpha + \beta_1 x_1 + \beta_2 x_2)}{\rho_{0,x_1,x_2} + \rho_{1,x_1,x_2} \exp(\alpha + \beta_1 x_1 + \beta_2 x_2)} \\
&= \frac{\rho_{1,x_2} \exp(\alpha + \beta_1 x_1 + \beta_2 x_2)}{\rho_{0,x_2} + \rho_{1,x_2} \exp(\alpha + \beta_1 x_1 + \beta_2 x_2)} \\
&= \frac{(\rho_{1,x_2}/\rho_{0,x_2}) \exp(\alpha + \beta_1 x_1 + \beta_2 x_2)}{1 + (\rho_{1,x_2}/\rho_{0,x_2}) \exp(\alpha + \beta_1 x_1 + \beta_2 x_2)} \\
&= \frac{\exp(\beta_1 x_1 + c(x_2))}{1 + \exp(\beta_1 x_1 + c(x_2))}
\end{aligned}
$$

where $c(x_2) = \alpha + \log(\rho_{1,x_2}/\rho_{0,x_2}) + \beta_2 x_2$.

- The coefficient β_1 for X_1 in the retrospective model is identical to the β_1 in the prospective model
- The intercept, the coefficient β_2 (and how X_2 affects Y) in the retrospective model are different.

Logistic Model for Bird-Keeping & Lung Cancer Data

A logistic model for LC and BK controlling for both AG and YR:

```
fit1 = glm((LC == "LungCancer") ~ BK + AG + YR,
           family=binomial, data=birdkp)
summary(fit1)$coef
              Estimate Std. Error   z value   Pr(>|z|)
(Intercept)  0.3429642  1.5800186  0.217063 0.82815895
BKNoBird    -1.3765591  0.4007298 -3.435130 0.00059227
AG          -0.0460982  0.0342995 -1.343989 0.17895188
YR           0.0748529  0.0229553  3.260806 0.00111096
```

In this study, the controls were selected to have same general age distribution as the cancer cases. So the sampling rates ρ_1 and ρ_2 depend on AG.

- Can interpret β_i's for BK and YR prospectively.
- CANNOT interpret the intercept and the β for AG prospectively.

```
fit1$coef
(Intercept)    BKNoBird          AG          YR
  0.3429642  -1.3765591  -0.0460982   0.0748529
```

- The odds of lung cancer for bird keepers were $e^{1.376559} \approx 3.961248$ times the odds for non-keepers of the same age and same years of smoking.
- The odds of lung cancer become $e^{0.074853} \approx 1.077726$ times as large for every extra year of smoking, controlling for age and bird-keeping status

```
fit1$coef
(Intercept)    BKNoBird          AG          YR
  0.3429642  -1.3765591  -0.0460982   0.0748529
```

CANNOT interpret the coefficient -0.046098 for `AG` as "the odds of LC become $e^{-0.046098} \approx 0.954948$ times as large for every extra year in age, controlling for `BK` and `YR`"

- unreasonable negative coefficient -0.046098 for `AG`. The odds of disease would increase with age for most chronic diseases. We get this a negative (but insignificant) β because the 98 controls were selected to match the age distribution of the 49 cancer cases. Hence, we cannot infer the age effect on the odds of lung cancer from this study.

```
fit1$coef
(Intercept)    BKNoBird          AG          YR
  0.3429642  -1.3765591  -0.0460982   0.0748529
```

CANNOT estimate π = prob. of lung cancer retrospectively. E.g., cannot estimate π for 50-year-old bird keepers with 10 years of smoking as

$$\widehat{\pi} = \frac{\exp(0.342964 - 0.046098 \times 50 + 0.074853 \times 10)}{1 + \exp(0.342964 - 0.046098 \times 50 + 0.074853 \times 10)} = 0.229097$$

```
predict(fit1, data.frame(BK="Bird", AG=50, YR=10), type="response")
       1
0.229097
```