Statistics 305/605: Introduction to Biostatistical Methods for Health Sciences

Chapter 20, part 3: Multiple Logistic Regression

Jinko Graham

2017-11-13

Heart Data For Multiple Logistic Regression

- To study the association between atherosclerotic heart disease (AHD) and maximum heart rate (MaxHR), investigators randomly sampled 303 patients presenting with chest pain at a large hospital.
- They recorded information on the patients'
  - age in years,
  - sex (1=male, 0=female), and
  - MaxHR, the maximum heart rate in beats per minute
  - AHD diagnosis (1=Yes, 0=No) based on a coronary angiogram.

```
uu<-url("http://people.stat.sfu.ca/~jgraham/Teaching/S305_18/Data/hrt.csv")
heart <- read.csv(uu)
head(heart)
```

```
##   X Age Sex MaxHR AHD
## 1 1  63   1   150   0
## 2 2  67   1   108   1
## 3 3  67   1   129   1
## 4 4  37   1   187   0
## 5 5  41   0   172   0
## 6 6  56   1   178   0
```

Multiple Logistic Regression

- Multiple logistic regression allows us to
  - investigate statistical interaction between explanatory variables
  - adjust for potential confounders
- Example: Could `sex` or `age` modify the relationship between `MaxHR` and the odds of `AHD`?
- If not, does either variable confound the relationship between `MaxHR` and the odds of `AHD`?

Multiple Logistic Regression Model

- We may model the log-odds of AHD as a function of q explanatory variables $X_1, X_2, \ldots, X_q$; i.e.,

$$\log\left[\frac{p}{1-p}\right] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_q X_q,$$

  - where:
    - log is the natural logarithm and
    - p is the probability of AHD given X_1,. . . ,X_q.

- Letting $LO = \alpha + \beta_1 X_+ \beta_2 X_2 + \ldots \beta_q X_q$, we have the logistic function:

$$p = \frac{e^{LO}}{1 + e^{LO}}$$

Statistical Interaction

- Interaction terms such as X_1X_2 appear as products of main effect terms such as X_1 and X_2:
  - $\log\left[\frac{p}{1-p}\right] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$
- In general, we allow the slopes for X_1 to depend on the value of the modifying variable X_2.
- Let's focus on the two values $X_2 = x_2$ and $X_2 = x_2 + 1$ that are one unit apart:
  - Line for $X_2 = x_2$ baseline group:
    - $\log\left[\frac{p}{1-p}\right] = \alpha + \beta_1 X_1 + \beta_2 x_2 + \beta_3 X_1 x_2$
    - intercept is $\alpha + \beta_2 x_2$ and
    - slope for X_1 is $\beta_1 + \beta_3 x_2$
  - Line for $X_2 = x_2 + 1$ group:
    - $\log\left[\frac{p}{1-p}\right] = \alpha + \beta_1 X_1 + \beta_2 (x_2 + 1) + \beta_3 X_1 (x_2 + 1)$
    - intercept is $\alpha + \beta_2 (x_2 + 1)$ and
    - slope for X_1 is $\beta_1 + \beta_3 (x_2 + 1)$
  - Difference between the slopes for X_1 in the two groups is $\beta_1 + \beta_3 (x_2 + 1) - (\beta_1 + \beta_3 x_2) = \beta_3$.

- The interaction coefficient β_3 is the difference between the slopes for X_1 in two groups that are defined by a one-unit increase in X_2.
  - If X_2 is a binary variable such as `sex`, a one-unit increase is from $X_2 = 0$ to $X_2 = 1$; i.e., from female to male.
- If $\beta_3 = 0$ then the slopes are the same.
  - Therefore, X_2 does **not** modify the effect of X_1 on Y.
- To assess statistical interaction between X_1 and X_2, test the hypotheses $H_0 : \beta_3 = 0$ *vs.* $H_a : \beta_3 \neq 0$.

Example: Interaction between MaxHR and Age

- Fitting ht model with MaxHR-by-Age interaction gives the following table of coefficients:

```
hfit2 <- glm(AHD ~ MaxHR+Age+MaxHR:Age,data=heart,family=binomial)
summary(hfit2)$coefficients
```

```
##                 Estimate   Std. Error   z value     Pr(>|z|)
## (Intercept) 26.414122478 7.0633107300  3.739623 1.842961e-04
## MaxHR       -0.184174006 0.0467014100 -3.943650 8.025083e-05
## Age         -0.363512892 0.1204382548 -3.018251 2.542382e-03
## MaxHR:Age    0.002562051 0.0008041757  3.185934 1.442873e-03
```

- Estimated $\hat{\beta}_3 = .00256$ difference between slopes for MaxHR in 2 groups defined by a one-unit change in Age.
- E.G. One group of patients of median Age 56 years and another of patients of Age 57 years.
  - In patients of Age 57 years, the slope for MaxHR is estimated to be 0.00256 more than in patients of Age 56 years.

Effect of MaxHR on AHD in patients of different ages

- The linear predictor or log-odds is

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2.$$

- In patients aged 57 years, simplifies to

$$\alpha + \beta_1 X_1 + \beta_2 57 + \beta_3 X_1 57 = \alpha + \beta_2 57 + (\beta_1 + \beta_3 57) X_1$$

- Slope for maxHR (X_1) in patients aged 57 years is $\beta_1 + \beta_3 57$; slope in patients aged 56 years is $\beta_1 + \beta_3 56$.
- We estimate that the effect of MaxHR on the log-odds of AHD is:
  - $\hat{\beta}_1 + \hat{\beta}_3 57 = -0.18417 + 0.00256 \times 57 = -0.03825$ in patients aged **57** years, and
  - $\hat{\beta}_1 + \hat{\beta}_3 56 = -0.18417 + 0.00256 \times 56 = -0.04081$ in patients aged **56** years
- i.e., an increase of one beat-per-minute in MaxHR is associated with estimated decreases of
  - 0.038 in the log-odds of AHD in patients aged 57 years, and
  - 0.041 in the log-odds of AHD in patients aged 56 years.

Does Age modify the effect of MaxHR on AHD?

- To address this question, test for statistical interaction between MaxHR and Age at the 5% level.

```
summary(hfit2)$coefficients
```

```
##                 Estimate   Std. Error   z value     Pr(>|z|)
## (Intercept) 26.414122478 7.0633107300  3.739623 1.842961e-04
## MaxHR       -0.184174006 0.0467014100 -3.943650 8.025083e-05
## Age         -0.363512892 0.1204382548 -3.018251 2.542382e-03
## MaxHR:Age    0.002562051 0.0008041757  3.185934 1.442873e-03
```

- Compare the p-value for the interaction term MaxHR:Age to the level 0.05.
  - Since the p-value of 0.001 is less than the level, we reject the null hypothesis of no interaction.
- Conclude that Age **does** modify the effect of MaxHR on AHD.

Confounding

Sex as a potential confounding variable

- Age has been declared to modify the effect of MaxHR on AHD and so we can't consider it as a potential confounding variable
- However, the binary variable Sex is not declared as an effect modifier at level 5% (results not shown).
  - We may thus consider Sex as a potential confounding variable.
- We fit a model with MaxHR and Sex main effects, and a model with a main effect only for MaxHR:

```
coefficients(glm(AHD~MaxHR+Sex,data=heart,family=binomial))
```

```
## (Intercept)       MaxHR         Sex
##  5.60185719 -0.04508903  1.40621009
```

```
coefficients(glm(AHD~MaxHR,data=heart,family=binomial))
```

```
## (Intercept)       MaxHR
##  6.32494975 -0.04341112
```

- We find that the estimated MaxHR effect changes by only

$$\frac{|-0.0451-(-0.0434)|}{|-0.0451|}\times 100\% = 3.7\%.$$

- Notice that we use the estimate from the larger model in the denominator.
- As the change is less than 10%, we follow convention and declare that Sex is not a confounder.
- A logistic regression model with a main effect for `MaxHR` is therefore sufficient and is summarized by:

```
hfitFinal <- glm(AHD~MaxHR,data=heart,family=binomial)
summary(hfitFinal)$coefficients
```

```
##                Estimate  Std. Error   z value     Pr(>|z|)
## (Intercept)  6.32494975 0.984366768  6.425400 1.315236e-10
## MaxHR       -0.04341112 0.006510412 -6.667954 2.593944e-11
```

Model checking and residual diagnostics

- There are measures of goodness-of-fit and residual diagnostics for logistic regression.
  - However, these are difficult to interpret and beyond the scope of the course.
- See Stat 475 **Applied Discrete Data Analysis** if interested.
  - Stat 305 is a pre-requisite for Stat 475.