Class 5: Chapter 8

EPH 705

Min Lu

Division of Biostatistics
University of Miami

Spring 2017

Overview

1. Object:
   - Extended Mantel-Haenszel statistics
   - Logistic regression
2. R Example
3. Exercise

Extended Mantel Haenszel statistics

R version of textbook

Table 6.2 summarizes the various types of extended Mantel-Haenszel statistics.

Extended Mantel-Haenszel Statistics

MH Statistic	Alternative Hypothesis	R output library("vcdExtra") CMHtest()	Degrees of Freedom	Scale Requirements	Nonparametric Equivalents
Q_{GMH}	general association	General Association	$(s-1)\times(r-1)$	none	
Q_{SMH}	mean score location shifts	Row Means Scores Differ	$(s-1)$	column variable ordinal	Kruskal-Wallis
Q_{CSMH}	linear association	Nonzero Correlation	1	row and column variable ordinal	Spearman correlation

Logistic regression

Assume that each i is associated either with a single Bernoulli trial or with n_i independent identically distributed trials, where the observation Y_i is the number of successes observed (the sum of the individual Bernoulli-distributed random variables), and hence follows a binomial distribution: $Y_i \ \sim \mathrm{Bin}(n_i, p_i)$, for $i = 1, \ldots, n$

$$\Pr(Y_i = y) = \binom{n_i}{y} p_i^y (1 - p_i)^{n_i - y}$$

Logistic regression for binomial outcome

$$\Pr(Y_i = y \mid \mathbf{X}_i) = \binom{n_i}{y} \left(\frac{1}{1 + e^{-\boldsymbol{\beta} \cdot \mathbf{X}_i}} \right)^y \left(1 - \frac{1}{1 + e^{-\boldsymbol{\beta} \cdot \mathbf{X}_i}} \right)^{n_i - y}.$$

Logistic regression

Outcome $Y_i \sim \mathrm{Bin}(n_i, p_i)$, for $i = 1, \ldots, n$

$$\Pr(Y_i = y) = \binom{n_i}{y} {p_i}^y (1 - p_i)^{n_i - y}$$

Logistic regression for binomial outcome

$$\Pr(Y_i = y \mid \mathbf{X}_i) = \binom{n_i}{y} \left(\frac{1}{1 + e^{-\boldsymbol{\beta} \cdot \mathbf{X}_i}}\right)^y \left(1 - \frac{1}{1 + e^{-\boldsymbol{\beta} \cdot \mathbf{X}_i}}\right)^{n_i - y}.$$

And the logistic function can now be written as: $F(x) = \dfrac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$.

For a continuous independent variable x the odds ratio can be defined as:

$$\mathrm{OR} = \frac{\mathrm{odds}(x+1)}{\mathrm{odds}(x)} = \frac{\left(\frac{F(x+1)}{1-F(x+1)}\right)}{\left(\frac{F(x)}{1-F(x)}\right)} = \frac{e^{\beta_0 + \beta_1 (x+1)}}{e^{\beta_0 + \beta_1 x}} = e^{\beta_1}$$

This exponential relationship provides an interpretation for β_1 : The odds multiply by e^{β_1} for every 1-unit increase in x.

BostonHousing Dataset

Data:Housing data (BostonHousing.csv) is 506 census tracts of Boston from the 1970 census. The dataframe contains the original data by Harrison and Rubinfeld (1979), with 506 observations on 14 variables, medv being the target variable:

variable name	discreption
crim	per capita crime rate by town
zn	proportion of residential land zoned for lots over 25,000 sq.ft
indus	proportion of non-retail business acres per town
chas	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
nox	nitric oxides concentration (parts per 10 million)
rm	average number of rooms per dwelling
age	proportion of owner-occupied units built prior to 1940
dis	weighted distances to five Boston employment centres
rad	index of accessibility to radial highways
tax	full-value property-tax rate per USD 10,000
ptratio	pupil-teacher ratio by town
b	$1000(B-0.63)^2$ where B is the proportion of blacks by town
lstat	percentage of lower status of the population
town	name of town
tract	census tract
lon	longitude of census tract
lat	latitude of census tract
cmedv	corrected median value of owner-occupied homes in USD 1000's

R code for Extended Mantel-Haenszel statistics

Import data and conduct contingency table analysis

```
# Give pretty variable names
BostonHousing <- read.csv("BostonHousing.csv")
BostonHousing$rm1 <- round(BostonHousing$rm/4)
BostonHousing$rad1 <- round(BostonHousing$rad/8)
x <- subset(BostonHousing, select = c(chas, rad1, rm1))
bost <- table(x)
names(dimnames(bost)) <- c("Charles River", "Highways accessibility", "Rooms number")
bost
```

```
## , , Rooms number = 1
##
##                Highways accessibility
## Charles River    0   1   3
##              0  63  46  53
##              1   5   4   2
##
## , , Rooms number = 2
##
##                Highways accessibility
## Charles River    0   1   3
##              0 118 120  71
##              1   6  12   6
```

R code for Extended Mantel-Haenszel statistics

Contingency Table analysis

```
# Give pretty variable names
mantelhaen.test(bost)
```

```
##
##  Cochran-Mantel-Haenszel test
##
## data:  bost
## Cochran-Mantel-Haenszel M^2 = 1.4793, df = 2, p-value = 0.4773
```

```
library("vcdExtra")
assocstats(bost)
```

```
## $`Rooms number:1`
##                     X^2 df P(> X^2)
## Likelihood Ratio 1.1202  2  0.57114
## Pearson          1.0237  2  0.59938
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.077
## Cramer's V        : 0.077
##
## $`Rooms number:2`
##                     X^2 df P(> X^2)
## Likelihood Ratio 1.8560  2  0.39535
## Pearson          1.7799  2  0.41068
##
## Phi-Coefficient   : NA
## Contingency Coeff.: 0.073
## Cramer's V        : 0.073
```

```
#CMHtest(bost)
#table12=rbind(CMHtest(bost)[[1]]$table,CMHtest(bost)[[2]]$table)
```

R code for publishable result

```r
# interpret overall association and interpret strata i
i <- 2
table2 <- CMHtest(bost)[[i]]$table
paste("In the Boston Housing census data, ",name[1]," and ",name[2]," are not significantly associated, Mantel-Haenszel Statistics X^2(",mantelhaen.test(bost)$parameter, ")=",round(mantelhaen.test(bost)$statistic,2),", p=",round(mantelhaen.test(bost)$p.value,3),". For the houses that have more than four rooms, ",name[1]," and ",name[2]," are not significantly associated, Mantel-Haenszel Statistics X^2(",table2[3,2],")=",round(table2[3,1],2),", p=",round(table2[3,3],3), "; Cramer's V is ",round(assocstats(bost)[[2]]$cramer,2),".",sep="")
```

R code for publishable result

- See output:

```
[1] "In the Boston Housing census data, Charles River and Highways accessibility are not significantly associated, Mantel-Haenszel Statistics X^2(2)=1.48, p=0.477. For the houses that have more than four rooms, Charles River and Highways accessibility are not significantly associated, Mantel-Haenszel Statistics X^2(2)=1.77, p=0.412; Cramer's V is 0.07."
```

Logistic Regression

```
object <- glm(chas ~ rm + crim + zn + tax, data = BostonHousing, binomial(link = "logit"))
summary(object)
```

```
##
## Call:
## glm(formula = chas ~ rm + crim + zn + tax, family = binomial(link = "logit"),
##     data = BostonHousing)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7625  -0.4108  -0.3663  -0.2934   2.4988
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -6.0519816  1.7636517  -3.432   0.0006 ***
## rm           0.5446217  0.2446163   2.226   0.0260 *
## crim        -0.0806952  0.0724800  -1.113   0.2656
## zn          -0.0165365  0.0101874  -1.623   0.1045
## tax          0.0008343  0.0017100   0.488   0.6256
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 254.50  on 505  degrees of freedom
## Residual deviance: 245.21  on 501  degrees of freedom
## AIC: 255.21
##
## Number of Fisher Scoring iterations: 7
```

```
# Display result using summary(object)
```

R code for logistic regression

Make publishable result

```
# Result looks messy
mytable <- summary(object)$coefficients
mytable
```

```
##                  Estimate  Std. Error    z value     Pr(>|z|)
## (Intercept) -6.0519815905 1.763651721 -3.4315061 0.0006002397
## rm           0.5446216692 0.244616274  2.2264327 0.0259852126
## crim        -0.0806951759 0.072480039 -1.1133434 0.2655609542
## zn          -0.0165365486 0.010187351 -1.6232432 0.1045374272
## tax          0.0008342876 0.001709957  0.4878997 0.6256208849
```

R code for logistic regression

Make publishable result

```
# Give pretty variable names
rownames(mytable) <- c("Intercept", "Room number", "Crime rate", "Residential land", "Property tax")
# Give pretty decimal
mytable[, 1:3] <- round(mytable[, 1:3], 2)
mytable[, 4] <- round(mytable[, 4], 3)
mytable
```

```
##                  Estimate Std. Error z value Pr(>|z|)
## Intercept           -6.05       1.76   -3.43    0.001
## Room number          0.54       0.24    2.23    0.026
## Crime rate          -0.08       0.07   -1.11    0.266
## Residential land    -0.02       0.01   -1.62    0.105
## Property tax         0.00       0.00    0.49    0.626
```

R code for logistic regression

Store publishable result

- Save as a table

```
write.csv(mytable,"BostonHousingResult.csv")
```

- Save in Latex

```
# have to intall package: install.packages("xtable")
library(xtable)
latex_table <- xtable(mytable)
```

UNIVERSITY OF MIAMI

R code for publishable result


```r
# interpret coefficient i
i <- 1
paste("The estimated OR for ", rownames(
 mytable)[i+1]," is ", round(exp(mytable
 [c(i+1),1]),2),", For each increase in
 1 unit of ", rownames(mytable)[i+1], "
 the estimated odds of outcome (tract
 bounds river) increases by roughly ",
 round(exp(mytable[i+1, 1]) - 1, 2),",
 when other variables hold the same.",
 sep="")

```

R code for publishable result

- See output:

```
[1] "The estimated OR for Room number is
  1.72, For each increase in 1 unit of
 Room number the estimated odds of
 outcome (tract bounds river) increases
 by roughly 0.72, when other variables
 hold the same."

```

In class exercise

Data:Housing data (BostonHousing.csv) is 506 census tracts of Boston from the 1970 census. The dataframe contains the original data by Harrison and Rubinfeld (1979), with 506 observations on 14 variables, medv being the target variable:

variable name	discreption
crim	per capita crime rate by town
zn	proportion of residential land zoned for lots over 25,000 sq.ft
indus	proportion of non-retail business acres per town
chas	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
nox	nitric oxides concentration (parts per 10 million)
rm	average number of rooms per dwelling
age	proportion of owner-occupied units built prior to 1940
dis	weighted distances to five Boston employment centres
rad	index of accessibility to radial highways
tax	full-value property-tax rate per USD 10,000
ptratio	pupil-teacher ratio by town
b	$1000(B - 0.63)^2$ where B is the proportion of blacks by town
lstat	percentage of lower status of the population
town	name of town
tract	census tract
lon	longitude of census tract
lat	latitude of census tract
cmedv	corrected median value of owner-occupied homes in USD 1000's

Task:Choose variables "rad", "dis", and "b" to predict "chas" in a linear regression model and check the result.

Take home exercise

Make a summary table of a linear model of your own. Choose your own outcome variable with 3-4 predictors. Save the result in an excel (.csv) file.

Class over