Computer practical 2: Topics in Statistics III/IV, Term 1

Georgios Karagiannis

Aim

- To perform Fisher exact test
- To Mantel-Haenszel test
- To analyse contigency tables with Log-linear models
  - Fit the model
  - Perform inference
  - Specify the Identifiability constraint (Contrasts)
  - Perform model comparison (Nested, or Non-nested models)

Load the libraries

```
library(MASS) # for contigency tables and log-linear models
library(vcd) # to visualise categorical data
```

```
## Loading required package: grid
```

Fisher's Exact test

Description

To perform FIsher's exact test we can use the command `fisher.test {stats}`, as,

- `fisher.test(x)`

where `x` is either a two-dimensional contingency table in matrix form, or a factor object.

For the output of the command, please chck help `?fisher.test`

Dataset

[Cancer dataset] The table below shows the results of a retrospective study comparing radiation therapy with surgery in treating cancer of the larynx. The response indicates whether the cancer was controlled for at least two years following treatment.

```
# load the data
## I will do this for you

Cancer.frame<-data.frame(count=c(21,2,15,3),
                         expand.grid(
                         Cancer=factor(c("Controled","Not-controled"),
                                       levels=c("Controled","Not-controled")),
                         Therapy=factor(c("Surgery","Radiation"),
```

```
                                    levels=c("Surgery","Radiation"))
                     )
  )

## print the obs.frame
Cancer.frame
```

```
##   count        Cancer   Therapy
## 1    21     Controled   Surgery
## 2     2 Not-controled   Surgery
## 3    15     Controled Radiation
## 4     3 Not-controled Radiation
```

Question

Create a matrix `Cancer.xtabs` from `Cancer.frame` by using `xtabs()`. Print the result on the screen.

Answer

Question

Perform Fisher's Exact test

- Recall the type of the command's argument, and tranform it suitably by using `xtabs()`

Answer

Question

Report and interpret the P-value for Fisher's exact test whose alternative hypothesis is that the odds ratio is larger than 1.

$$H_0 : \theta = 1 \quad \text{versus} \quad H_1 : \theta > 1$$

Answer

Write your discussion here . . . .
.
.
.
.
.
.
.

Question

Explain how the P-values are calculated for this particular test. (you can check your notes: Handout: Contigency tables)

Answer

Write your discussion here . . . .
.
.
.
.
.
.
.

Chi-squared Test

Description

To perform Chi-squared Test on a contigency tables, we can use the command `chisq.test()` from the package `stats`, as,

- `chisq.test(x,...)`

where `x` a numeric vector or matrix holding the data

For the output of the command, see in the help page `?chisq.test`

Dataset

Use the [Cancer dataset]

Question

- Perform the Chi-squared Test in order to test if Cancer control and Therapy are independent.
- What is the conclusion?

Answer

Write your discussion here . . . .
.
.

Mantel–Haenszel test

Description

We wish to perform Mantel-Haenszel chi-squared test of the null that two nominal variables are conditionally independent in each stratum, assuming that there is no three-way interaction.

We can use the command `mantelhaen.test(x,...)`, as,

- mantelhaen.test(x,alternative="two.sided")

where `x` is a 3-dimensional contingency table in array form where each dimension is at least 2 and the last dimension corresponds to the strata.

For more options check help `?mantelhaen.test`

Dataset

[Marijuana data-set] Consider the following table where refers to a 1992 survey by the Wright State University School of Medicine and the United Health Services in Dayton, Ohio. The survey asked 2276 students in their final year of high school in a nonurban area near Dayton, Ohio whether they had ever used alcohol, cigarettes, or marijuana. Denote the variables in this $2 \times 2 \times 2$ table by A for alcohol use, C for cigarette use, and M for marijuana use.

```
# I will do this for you

## load the data

marijuana.frame<-data.frame(count=c(911,538,44,456,3,43,2,279),
                      expand.grid(
```

```
  marijuana=factor(c("Yes","No"),levels=c("No","Yes")),
  cigarette=factor(c("Yes","No"),levels=c("No","Yes")),
  alcohol=factor(c("Yes","No"),levels=c("No","Yes")))
  )

## print the obs.frame

marijuana.frame
```

```
##   count marijuana cigarette alcohol
## 1   911       Yes       Yes     Yes
## 2   538        No       Yes     Yes
## 3    44       Yes        No     Yes
## 4   456        No        No     Yes
## 5     3       Yes       Yes      No
## 6    43        No       Yes      No
## 7     2       Yes        No      No
## 8   279        No        No      No
```

Question

Test the hypothesis that the use of Marijuana and the use of cigarete are independent accross the levels of the use of alcohol, by using Mantel–Haenszel test.

Report the result of your inference.

If you wish, you can play with the alternative hypothesis (see the help page)

Answer

Write your discussion here . . . .
.
.

Log-linear models

Data

Consider the [Marijuana data-set] used previously

```
marijuana.frame
```

```
##   count marijuana cigarette alcohol
## 1   911       Yes       Yes     Yes
## 2   538        No       Yes     Yes
## 3    44       Yes        No     Yes
## 4   456        No        No     Yes
## 5     3       Yes       Yes      No
## 6    43        No       Yes      No
## 7     2       Yes        No      No
## 8   279        No        No      No
```

Fit the Log-linear model

Fitting a loglinear model can be done using Iterative Proportional Fitting (`loglm` of the package `MASS`) or Newton Raphson (`glm` with poisson family, of the package `stats`). The former uses `loglm` from `MASS`. `loglm`

accepts as input table output, crosstabs output (`xtabs` in R), or a formula using variables from a data frame.

Specify a model

Below is an example for fiting the Log-linear model [ACM]. Notice that [ACM] it can be writen in different ways. I set the arguments `fit` and `param` to `T` so that I can get fitted values and parameter estimates, as we see later.

```
fitAC.AM.CM_ <- loglm(count~1+alcohol+cigarette+marijuana
                      +alcohol:cigarette
                      +cigarette:marijuana
                      +alcohol:marijuana
                      ,data=marijuana.frame,
                      param=T,fit=T) # ACM

fitAC.AM.CM <- loglm(count ~ alcohol*cigarette
                 +alcohol*marijuana
                 +cigarette*marijuana,
                 data=marijuana.frame,
                 param=T,fit=T) # ACM

fitAC.AM.CM_
```

```
## Call:
## loglm(formula = count ~ 1 + alcohol + cigarette + marijuana +
##     alcohol:cigarette + cigarette:marijuana + alcohol:marijuana,
##     data = marijuana.frame, param = T, fit = T)
##
## Statistics:
##                       X^2 df  P(> X^2)
## Likelihood Ratio 0.3739859  1 0.5408396
## Pearson          0.4011039  1 0.5265197
```

```
fitAC.AM.CM
```

```
## Call:
## loglm(formula = count ~ alcohol * cigarette + alcohol * marijuana +
##     cigarette * marijuana, data = marijuana.frame, param = T,
##     fit = T)
##
## Statistics:
##                       X^2 df  P(> X^2)
## Likelihood Ratio 0.3739859  1 0.5408396
## Pearson          0.4010998  1 0.5265218
```

Question

Notice the details in the two equivalent executions.

Regarding the `formula` above, what is the relation between terms (1, alcohol, cigarette, alcohol:cigarette) and (alcohol*cigarette) ?

Answer

Write your discussion here . . . .

.

.

Update a model

Based on a given model, we can build, new ones. E.g., the [ACM] and the [AC,M],

```
fitACM<-update(fitAC.AM.CM,
               .~. + alcohol:cigarette:marijuana) # ACM

fitAC.M<-update(fitAC.AM.CM,
               .~. - alcohol:marijuana - cigarette:marijuana) # AC, M
```

Question

Fit models [AM, CM], and [A, C, M] and print the output.

Answer

Question

Get the estimates of the model parameters (the lambdas) by applying `...$param` to the output object of the `loglm` or `update`.

Get the fitted values by applying `...$fitted` the command `fitted()` to the output object of the `loglm` or `update`.

Answer

Inference

Likelihood ratio chi-squared test statistics are output using the `summary()` function for `loglm`. The functin provides:

- The formula of the model as `Formula:`
- The design matrix X in the vectorized form of the log-linear model $\log(\mu) = X\beta$ as `attr(,"factors")`
- The Pearson and Likelihood ratio Goodness-of-fit test

Question

Apply `summary()` function on the output object of the function `loglm` for the homogeneous association model object.

Does the data-set support the homogeneous association model at sig. level 5% ?

Answer

Write your discussion here . . . .
.
.

Stepwise model selection

Two popular procedures can be used in order to perform Variable selection in linear models. By saying variable here, we mean the factors of the log-linear model, or (equivalently) the classifier variables of the associated contigency table. They are the follwoing:

- Forward selection.
  - Forward selection adds terms sequentially until further additions do not improve the fit. E.g., [A,C,M]->[AC,M]->[AC,AM]->. . . At each stage it selects the term giving the greatest improvement in fit. The selection criterion can be the minimum P-value for testing the term in the model. Then the procedure may stop when adding more terms may result in insignifical p-values.

- Backward elimination. Backward elimination begins with a complex model and sequentially removes terms until we reach a simple model such as reducing it further can will not fit the data well. At each stage, it selects the term for which its removal has the least damaging effect on the model e.g., largest P-value. It stops at the model such that removal of any of its term lead to rejection of the Goodness-of-Fit test.

There is not a commonly accepted answer to the question which procedure is better. Both are okay and in use... What is your personal opinion? Discuss it with your fellow students; alternatively ask the instructor.

Model comparison among a set of nested models

Comment

Comparison of nested models can be done using the anova method. It gives the likelihood ratio tests comparing hierarchical loglinear models given in the list of arguments. For example,

```
anova(fitAC.M, fitAC.AM.CM)
```

```
## LR tests for hierarchical log-linear models
##
## Model 1:
##  count ~ alcohol + cigarette + marijuana + alcohol:cigarette
## Model 2:
##  count ~ alcohol * cigarette + alcohol * marijuana + cigarette * marijuana
##
##              Deviance df  Delta(Dev) Delta(df) P(> Delta(Dev)
## Model 1   843.8266437  3
## Model 2     0.3739859  1 843.4526577         2        0.00000
## Saturated   0.0000000  0   0.3739859         1        0.54084
```

The column `Deviance` is the Devience... (i.e., the Likelihood ratio statistic of the model vs the saturated)

Regarding the likelihood ratio test of

- [AC, M] vs. [AC,AM,CM] , the statistic is equal to 843.4526577, the degrees of freedom are equal to 2 and the p-value equal to 0.00000
- [AC,AM,CM] vs. [ACM], the statistic is equal to 0.3739859, the degrees of freedom are equal to 1 and the p-value equal to 0.54084

Question

Compare models [ACM], [AC,AM,CM], [AM, CM], [AC, M], [A,C,M]. What is your conclusion?

Answer

Write your discussion here ... .
.
.

Model comparison among a set of non-nested models

Select a subset of predictor (classifier) variables from a larger set (e.g., stepwise selection) is a controversial topic.

You can perform stepwise selection (forward, backward) using the `stepAIC( )` function from the `MASS` package. `stepAIC( )` performs stepwise model selection by exact AIC.

Comment

Model selection based on either Forward selection, or Backword elimination procedures can be perforemed by using AIC as a criterion.

The command to perofrm this procedure automatically is `stepAIC()`.

- Forward selection, starting from fitA.C.M: `stepAIC(fitA.C.M, direction = "forward")$anova`
- Backward elimination, starting from fitACM: `stepAIC(fitACM, direction ="backward")$anova`

The `...$anova` is because we are interested in the `...$anova` return value only.

Question

Apply Forward selection, and Backward elimination by using AIC.

- What is your conclusion about the preferable model by using each method?
- Do the two procedures produce the same result?

Answer

Write your discussion here . . . .
.
.

Practice at home

Data

Use the Marijuana data, and consider additional classifier variables Gender (G), and race (R).

The dataset is given below.

```
marijuana_new.frame<-data.frame(count=c(405,13,1,1,268,218,17,117,453,28,1,1,228,201,17,
                                        133,23,2,0, 0,23, 19,1,12,30,1,1,0,19,18,8,17),
                                expand.grid(cigarette=c("Yes","No"),
                                            alcohol=c("Yes","No"),marijuana=c("Yes","No"),
                                            sex=c("female","male"),
                                      race=c("white","other"))
                                )
marijuana_new.frame
```

```
##    count cigarette alcohol marijuana    sex  race
## 1    405       Yes     Yes       Yes female white
## 2     13        No     Yes       Yes female white
## 3      1       Yes      No       Yes female white
## 4      1        No      No       Yes female white
## 5    268       Yes     Yes        No female white
## 6    218        No     Yes        No female white
## 7     17       Yes      No        No female white
## 8    117        No      No        No female white
## 9    453       Yes     Yes       Yes   male white
## 10    28        No     Yes       Yes   male white
## 11     1       Yes      No       Yes   male white
## 12     1        No      No       Yes   male white
## 13   228       Yes     Yes        No   male white
## 14   201        No     Yes        No   male white
## 15    17       Yes      No        No   male white
## 16   133        No      No        No   male white
## 17    23       Yes     Yes       Yes female other
## 18     2        No     Yes       Yes female other
## 19     0       Yes      No       Yes female other
```

```
## 20     0        No      No       Yes female other
## 21    23       Yes     Yes        No female other
## 22    19        No     Yes        No female other
## 23     1       Yes      No        No female other
## 24    12        No      No        No female other
## 25    30       Yes     Yes       Yes   male other
## 26     1        No     Yes       Yes   male other
## 27     1       Yes      No       Yes   male other
## 28     0        No      No       Yes   male other
## 29    19       Yes     Yes        No   male other
## 30    18        No     Yes        No   male other
## 31     8       Yes      No        No   male other
## 32    17        No      No        No   male other
```

Question

- Perform a model selection (by using the procedure of your preference) in order to find the best model that prepresents the dependences of the variables.
- For the selected model, compute the fitted values, and the estimations of the parameters.
- Discuss the conclusions of your inference.

*** Answer ***

Write your discussion here . . . .

.

.

Save me

Generate the document as a Notebook, PDF, Word, or HTML by choosing the relevant option (from the pop-up menu next to the Preview button). Then save your Markdown code by choosing the relevant option (from the task bar menu).

Save the *.Rmd script, so that you can edit it later.