536Final

Xinyu Gao

December 3, 2019

Problem 1

Question 1

Data

```r
data <- c(1105, 4624, 411111, 157342, 14, 497, 483, 1008)
data.array <- array(data, c(2,2,2))
data.array
```

```
## , , 1
##
##      [,1]   [,2]
## [1,] 1105 411111
## [2,] 4624 157342
##
## , , 2
##
##      [,1] [,2]
## [1,]   14  483
## [2,]  497 1008
```

The most complex model: The Saturated Log-linear Model

$$logm_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3k} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)} + u_{123(ijk)}$$

Interpretations:

We set the variable “Safety Equipment in Use” be X_1, “Whether Ejected” be X_2, and “Injury” be X_3.

1. u represents the mean of logarithms of the expected counts;
2. $u_{1(i)}$ represents the deviation from the grand mean u assocaited with category i of X_1;
3. $u_{2(j)}$ represents the deviation from the grand mean u assocaited with category j of X_2;
4. $u_{3(k)}$ represents the deviation from the grand mean u assocaited with category k of X_3;
5. $u_{12(ij)}$ represents thedeviation from $u + u_{1(i)} + u_{2(j)}$ assocaited with the interaction between category i of X_1 and category j of X_2;
6. $u_{13(ik)}$ represents thedeviation from $u + u_{1(i)} + u_{3(k)}$ assocaited with the interaction between category i of X_1 and category k of X_3;
7. $u_{23(jk)}$ represents thedeviation from $u + u_{2(j)} + u_{3(k)}$ assocaited with the interaction between category j of X_2 and category k of X_3;
8. $u_{123(ijk)}$ represents thedeviation from $u + u_{1(i)} + u_{2(j)} + u_{123(ijk)}$ assocaited with the interaction between category i of X_1, category j of X_2 and category k of X_3.

Then the R commands:

```
saturated.loglin <- loglin(data.array, margin = list(c(1,2,3)))
```

```
## 2 iterations: deviation 0
```

Complete Independence Model

We set to zero the first and second order interaction terms

$$u_{12(ij)} = u_{13(ik)} = u_{23(jk)} = u_{123(ijk)} = 0$$

in the saturated model, which can be expressed as

$$logm_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3(k)}$$

We assume each variable is independent of the other two variables, X_1 independent (X_2, X_3), X_2 independent (X_1, X_3), X_3 independent (X_1, X_2).//

Then R commands:

```
indep.login = loglin(data.array,margin = list(1,2,3),param = TRUE,fit = TRUE)
```

```
## 2 iterations: deviation 5.820766e-11
```

```
indep.login$param
```

```
## $`(Intercept)`
## [1] 7.366401
##
## $`1`
## [1]  0.4630584 -0.4630584
##
## $`2`
## [1] -2.257279  2.257279
##
## $`3`
## [1]  2.8294 -2.8294
```

```
paste("The number of degree of freedom is ", indep.login$df)
```

```
## [1] "The number of degree of freedom is  4"
```

```
paste("The value of the likelihood ratio statistic is ", indep.login$lrt)
```

```
## [1] "The value of the likelihood ratio statistic is  11444.3753501372"
```

```
paste("The value of X2 statistic is", indep.login$pearson)
```

```
## [1] "The value of X2 statistic is 47427.6315484939"
```

Then the p-value for testing the null hypothesis

$$H_0 : u_{12(ij)} = u_{13(ik)} = u_{23(jk)} = u_{123(ijk)} = 0$$

based on G^2 is given by $P(\chi_{1}2^2 \geq 1078.8)$ and is obtained with the R call

```
1 - pchisq(indep.login$lrt, indep.login$df)
```

```
## [1] 0
```

The p-value for testing H_0 based on X^2 is obtained with the call

```
1 - pchisq(indep.login$pearson, indep.login$df)
```

```
## [1] 0
```

From both the test above, we can reject H_0 and conclude that the complete independence model does not fit well in this data.

Models with One variable Independent of the Other Two

In thi section, we will try three models: 1. [1][23], i.e. The model of independece of X_1 and (X_2, X_3), and the corresponding H_0 is

$$H_0 : u_{12(ij)} = u_{13(ik)} = u_{123(ijk)} = 0$$

2. [2][13], i.e. The model of independece of X_2 and (X_1, X_3), and the corresponding H_0 is

$$H_0 : u_{12(ij)} = u_{23(jk)} = u_{123(ijk)} = 0$$

3. [3][12], i.e. The model of independece of X_3 and (X_1, X_2), and the corresponding H_0 is

$$H_0 : u_{13(ik)} = u_{23(jk)} = u_{123(ijk)} = 0$$

Model1: [1][23] contains an interaction term between X_2 and X_3, but no interaction terms between X_1 and X_2 or between X_1 and X_3, which can be expressed as

$$logm_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3k} + u_{23(jk)}$$

We fit model1 by R code

```
X1indepX2X3 = loglin(data.array, margin = list(1, c(2,3)), fit = TRUE, param = TRUE)
```

```
## 2 iterations: deviation 5.820766e-11
```

```
P_val <- function(model){
  p1 = 1 - pchisq(model$lrt, model$df)
  p2 = 1 - pchisq(model$pearson, model$df)
  paste("The p-value based on G2 is", p1, ", The p-value based on X2 is", p2)
}
P_val(X1indepX2X3)
```

```
## [1] "The p-value based on G2 is 0 , The p-value based on X2 is 0"
```

we reject H_0, which indicates that model [1][23] does not fit the data.

Model2: [2][13] contains an interaction term between X_1 and X_3, but no interaction terms between X_2 and X_1 or between X_2 and X_3, which can be expressed as

$$logm_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3k} + u_{13(ik)}$$

We fit model2 by R code

```
X2indepX1X3 = loglin(data.array, margin = list(2, c(1,3)), fit = TRUE, param = TRUE)
```

```
## 2 iterations: deviation 0
```

```
P_val(X2indepX1X3)
```

```
## [1] "The p-value based on G2 is 0 , The p-value based on X2 is 0"
```

we reject H_0, which indicates that model [2][13] does not fit data.

Model3: [3][12] contains an interaction term between X_1 and X_2, but no interaction terms between X_3 and X_1 or between X_3 and X_2, which can be expressed as

$$logm_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3k} + u_{12(ij)}$$

We fit model3 by R code

```
X3indepX1X2 = loglin(data.array, margin = list(3, c(1,2)), fit = TRUE, param = TRUE)
```

```
## 2 iterations: deviation 0
```

```
P_val(X3indepX1X2)
```

```
## [1] "The p-value based on G2 is 0 , The p-value based on X2 is 0"
```

we reject H_0, which indicates that model [3][12] does not fit data.

Model of Conditional Independece

In this section, we will try three models: model1: [12][13] the model of conditional independece of X_2 and X_3 given X_1, and H_0 is

$$H_0 : u_{23(jk)} = u_{123(ijk)} = 0$$

model2: [12][23] the model of conditional independece of X_1 and X_3 given X_2, and H_0 is

$$H_0 : u_{13(ik)} = u_{123(ijk)} = 0$$

model3: [13][23] the model of conditional independece of X_1 and X_2 given X_3, and H_0 is

$$H_0 : u_{12(ij)} = u_{123(ijk)} = 0$$

Model1: [12][13] contains an interaction term between X_1 and X_2, an interaction between X_1 and X_3, but no interaction terms between X_2 and X_3, which can be expressed as

$$logm_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3k} + u_{12(ij)} + u_{13(ik)}$$

We fit model1 by R code

```
X2indepX3givenX1 = loglin(data.array, margin = list(c(1,2), c(1,3)), fit = TRUE, param = TRUE)
```

```
## 2 iterations: deviation 0
```

```
P_val(X2indepX3givenX1)
```

```
## [1] "The p-value based on G2 is 0 , The p-value based on X2 is 0"
```

Both p-values are less than 0.05, hence we will reject H_0, which indicates that model [12][13] does not fit data.

Model2: [12][23] contains an interaction term between X_1 and X_2, an interaction between X_2 and X_3, but no interaction terms between X_1 and X_3, which can be expressed as

$$logm_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3k} + u_{12(ij)} + u_{23(jk)}$$

We fit model1 by R code

```
X1indepX3givenX2 = loglin(data.array, margin = list(c(1,2), c(2,3)), fit = TRUE, param = TRUE)
```

```
## 2 iterations: deviation 0
```

```
P_val(X1indepX3givenX2)
```

```
## [1] "The p-value based on G2 is 0 , The p-value based on X2 is 0"
```

Both p-values are less than 0.05, hence we will reject H_0, which indicates that model [12][23] does not fit data.

Model3: [13][23] contains an interaction term between X_1 and X_3, an interaction between X_2 and X_3, but no interaction terms between X_1 and X_2, which can be expressed as

$$logm_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3k} + u_{13(ik)} + u_{23(jk)}$$

We fit model1 by R code

```
X1indepX2givenX3 = loglin(data.array, margin = list(c(1,3), c(2,3)), fit = TRUE, param = TRUE)
```

```
## 2 iterations: deviation 0
```

```
P_val(X1indepX2givenX3)
```

```
## [1] "The p-value based on G2 is 0 , The p-value based on X2 is 0"
```

Both p-values are less than 0.05, hence we will reject H_0, which indicates that model [13][23] does not fit data.

The Model of No Second Order Interaction

This model [12][13][23] is obtained from the saturated log-linear model by setting the second order interaction terms to zero:

$$u_{123(ijk)} = 0$$

It contains an interaction term between X_1 and X_2, between X_1 and X_3, and between X_2 and X_3

$$logm_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3k} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)}$$

In this model, the H_0 is

$$u_{123(ijk)} = 0$$

R codes:

```
no2ind.loglin = loglin(data.array, margin = list(c(1,2), c(1,3), c(2,3)), fit=TRUE, param = TRUE)
```

```
## 5 iterations: deviation 0.08833306
```

```
P_val(no2ind.loglin)
```

```
## [1] "The p-value based on G2 is 0.0911456482999311 , The p-value based on X2 is 0.109892596865728"
```

We will not reject H_0 based on p-values, which indicates that no second order interaction model fits data well, and a log-linear model that is representative for the associations among “Safety Equipment in Use”, “Whether Ejected” and “Injury” is the model [12][13][23],

$$logm_{ijk} = u + u_{1(i)} + u_{2(j)} + u_{3k} + u_{12(ij)} + u_{13(ik)} + u_{23(jk)}$$

Question 2

Since we have chosen the model[12][13][23], we find the parameters of this model are

```
no2ind.loglin$param
```

```
## $`(Intercept)`
## [1] 7.831966
##
## $`1`
## [1] -0.5489897  0.5489897
##
## $`2`
## [1] -1.663277  1.663277
##
## $`3`
## [1]  2.251693 -2.251693
##
## $`1.2`
##            [,1]       [,2]
## [1,] -0.5999082  0.5999082
## [2,]  0.5999082 -0.5999082
##
## $`1.3`
##            [,1]       [,2]
## [1,]  0.4293324 -0.4293324
## [2,] -0.4293324  0.4293324
##
## $`2.3`
##            [,1]       [,2]
## [1,] -0.6994481  0.6994481
## [2,]  0.6994481 -0.6994481
```

Again We set the variable "Safety Equipment in Use" be X_1, "Whether Ejected" be X_2, and "Injury" be X_3. Hence the logistic regression here is

$$log\frac{P(X_3 = 2|X_1, X_2)}{P(X_3 = 1|X_1, X_2)}$$

where $X_3 = 2$ represents "Fatal", while $X_3 = 1$ represents "Non-fatal".

We then assume "$X_1 = 1$" represents "Seat belt" and $X_1 = 2$ represents "None" in variable "Safety Equipment in Use"; $X_2 = 1$ presents "Yes" and $X_2 = 2$ represents "No" in variable "Whether Ejected". Then the regression is expressed using the u-terms of the log-linear model [12][13][23] as follows:

$$log\frac{P(X_3 = k_1|X_1 = i, X_2 = j)}{P(X_3 = k_2|X_1 = i, X_2 = j)} = (u_{3(k_1)} - u_{3(k_2)}) + (u_{13(ik_1)} - u_{13(ik_2)}) + (u_{23(jk_1)} - u_{23(jk_2)})$$

We first calcualte the odds of "Fatal"($X_3 = 2$) vs. "Non-fatal"($X_3 = 1$) for Seat belt in use ($X_1 = 1$) and Ejected($X_2 = 1$), that is

$$\frac{P(X_3 = 2|X_1 = 1, X_2 = 1)}{P(X_3 = 1|X_1 = 1, X_2 = 1)} = exp((\hat{u}_{3(2)} - \hat{u}_{3(1)}) + (\hat{u}_{13(12)} - \hat{u}_{13(11)}) + (\hat{u}_{23(12)} - \hat{u}_{23(11)}))$$

The estimate of the odds is

```
exp((-2.251693 - ( 2.251693)) + ( -0.4293324 - 0.4293324) + ( 0.6994481 - (- 0.6994481))  )
```

```
## [1] 0.01900307
```

We then calcualte the odds of “Fatal”($X_3 = 2$) vs. “Non-fatal”($X_3 = 1$) for “No Safety Equipment in Use ($X_1 = 2$) and Ejected($X_2 = 1$), that is

$$\frac{P(X_3 = 2|X_1 = 2, X_2 = 1)}{P(X_3 = 1|X_1 = 2, X_2 = 1)} = exp((\hat{u}_{3(2)} - \hat{u}_{3(1)}) + (\hat{u}_{13(22)} - \hat{u}_{13(21)}) + (\hat{u}_{23(12)} - \hat{u}_{23(11)}))$$

The estimate of the odds is

```
exp((-2.251693 - (+ 2.251693)) + (0.4293324 - (-0.4293324)) + (0.6994481 - (-0.6994481))  )
```

```
## [1] 0.1058402
```

We then calcualte the odds of “Fatal”($X_3 = 2$) vs. “Non-fatal”($X_3 = 1$) for Seat belt in use ($X_1 = 1$) and not Ejected($X_2 = 2$), that is

$$\frac{P(X_3 = 2|X_1 = 1, X_2 = 2)}{P(X_3 = 1|X_1 = 1, X_2 = 2)} = exp((\hat{u}_{3(2)} - \hat{u}_{3(1)}) + (\hat{u}_{13(12)} - \hat{u}_{13(11)}) + (\hat{u}_{23(22)} - \hat{u}_{23(21)}))$$

The estimate of the odds is

```
exp(( -2.251693 - (2.251693)) + (-0.4293324 - 0.4293324) + (- 0.6994481 -0.6994481 )  )
```

```
## [1] 0.001158132
```

We finally calcualte the odds of “Fatal”($X_3 = 2$) vs. “Non-fatal”($X_3 = 1$) for Seat belt not in use ($X_1 = 2$) and not Ejected($X_2 = 2$), that is

$$\frac{P(X_3 = 2|X_1 = 2, X_2 = 2)}{P(X_3 = 1|X_1 = 2, X_2 = 2)} = exp((\hat{u}_{3(2)} - \hat{u}_{3(1)}) + (\hat{u}_{13(22)} - \hat{u}_{13(21)}) + (\hat{u}_{23(22)} - \hat{u}_{23(21)}))$$

The estimate of the odds is

```
exp((-2.251693 - (2.251693)) + (0.4293324 - (-0.4293324)) + ( -0.6994481 - 0.6994481)  )
```

```
## [1] 0.006450372
```

Based on this regression, the chance of having a fatal injury is much smaller than that of having a non-fatal injury, whatever the condition is. We also can find that in automobile accident, people with seat belt in use and not ejected are the most likely to be “Non-fatal” than to “Fatal”, while people with no safety equipment in use and ejected are the most likely to be “Fatal” than to “Non-fatal”. This analysis tells us that we should at least use seat belt as a tool of safety equipment when driving.

Question 3

```
data.array
```

```
## , , 1
##
##      [,1]   [,2]
## [1,] 1105 411111
## [2,] 4624 157342
##
## , , 2
##
##      [,1] [,2]
## [1,]   14  483
## [2,]  497 1008
```

model1

```
mydata = matrix(c(rep(c(1,1,1),1105),rep(c(2,1,1),4624),rep(c(1,2,1),411111),rep(c(2,2,1),157342),rep(c
```

In question 3, We denote the response "Injury" as Y, and $Y = 1$ represents "Non-fatal", $Y = 2$ represents "Fatal".

We denote "Safety Equipment in Use", "Whether Ejected" as X_1, X_2, respectively. We will logistic regression models

$$log\frac{P(Y = 1|X)}{P(Y = 2|X)} = X\beta$$

where X can be any of $\{1\}$, $\{1, X_1\}$, $\{1, X_2\}$, $\{1, X_1, X_2\}$,hence 4 models in total.//

model1:

$$log\frac{P(Y = 1|X)}{P(Y = 2|X)} = \beta_0$$

```
mylogit = glm(factor(mydata[,3])~1, family=binomial(link=logit))
```

model2:

$$log\frac{P(Y = 1|X)}{P(Y = 2|X)} = \beta_0 + \beta_1 * X_1$$

```
mylogit_X1 = glm(factor(mydata[,3])~factor(mydata[,1]), family=binomial(link=logit))
```

model3:

$$log\frac{P(Y = 1|X)}{P(Y = 2|X)} = \beta_0 + \beta_2 * X_2$$

```
mylogit_X2= glm(factor(mydata[,3])~factor(mydata[,2]), family=binomial(link=logit))
```

model4:

$$log\frac{P(Y = 1|X)}{P(Y = 2|X)} = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2$$

```
mylogit_all= glm(factor(mydata[,3])~factor(mydata[,1])+factor(mydata[,2]),family=binomial(link=logit))
```

```
# AIC
print("------AIC-----")
```

```
## [1] "------AIC-----"
```

```
AIC(mylogit)
```

```
## [1] 26670.81
```

```
AIC(mylogit_X1)
```

```
## [1] 24785.5
```

```
AIC(mylogit_X2)
```

```
## [1] 24249.72
```

```
AIC(mylogit_all)
```

```
## [1] 23109.94
```

```
print("------BIC-----")
```

```
## [1] "------BIC-----"
```

```r
#BIC
BIC(mylogit)
```

```
## [1] 26682.07
```

```r
BIC(mylogit_X1)
```

```
## [1] 24808.02
```

```r
BIC(mylogit_X2)
```

```
## [1] 24272.25
```

```r
BIC(mylogit_all)
```

```
## [1] 23143.73
```

Based on both AIC and BIC, we prefer to choose the model with both X_1 and X_2 as the explantory variable„ i.e. the model

$$log\frac{P(Y=1|X)}{P(Y=2|X)} = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2$$

fits the data best.

```r
summary(mylogit_all)
```

```
##
## Call:
## glm(formula = factor(mydata[, 3]) ~ factor(mydata[, 1]) + factor(mydata[,
##     2]), family = binomial(link = logit))
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.4486  -0.1134  -0.0481  -0.0481   3.6775
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -3.96315    0.06944  -57.07   <2e-16 ***
## factor(mydata[, 1])2  1.71732    0.05401   31.79   <2e-16 ***
## factor(mydata[, 2])2 -2.79779    0.05526  -50.63   <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: 26669  on 576183  degrees of freedom
## Residual deviance: 23104  on 576181  degrees of freedom
## AIC: 23110
##
## Number of Fisher Scoring iterations: 9
```

Interpretation: the log odds of non-fatal v.s fatal is -3.96 when X1, X2 is 1. when X1 changing from 1 to 2, the log odds of non-fatal injury V.S. fatal injury changed by 1.72 holding X2 constant;when X2 changing from 1 to 2, the log odds of non-fatal injury V.S. fatal injury changed by -2.80 holding X1 constant.

Question 4

Summary of my findings.

Based on Question 1, we can find that the model without second order interaction term fits the data best.

Based on Question 2, we can find that in automobile accident, people with seat belt in use and not ejected are the most likely to be "Non-fatal" than to "Fatal" (i.e. the safest way), while people with no safety equipment in use and ejected are the most likely to be "Fatal" than to "Non-fatal" (i.e. the most dangerous way). This analysis tells us that we should at least use seat belt as a toll of safety equipment when driving.

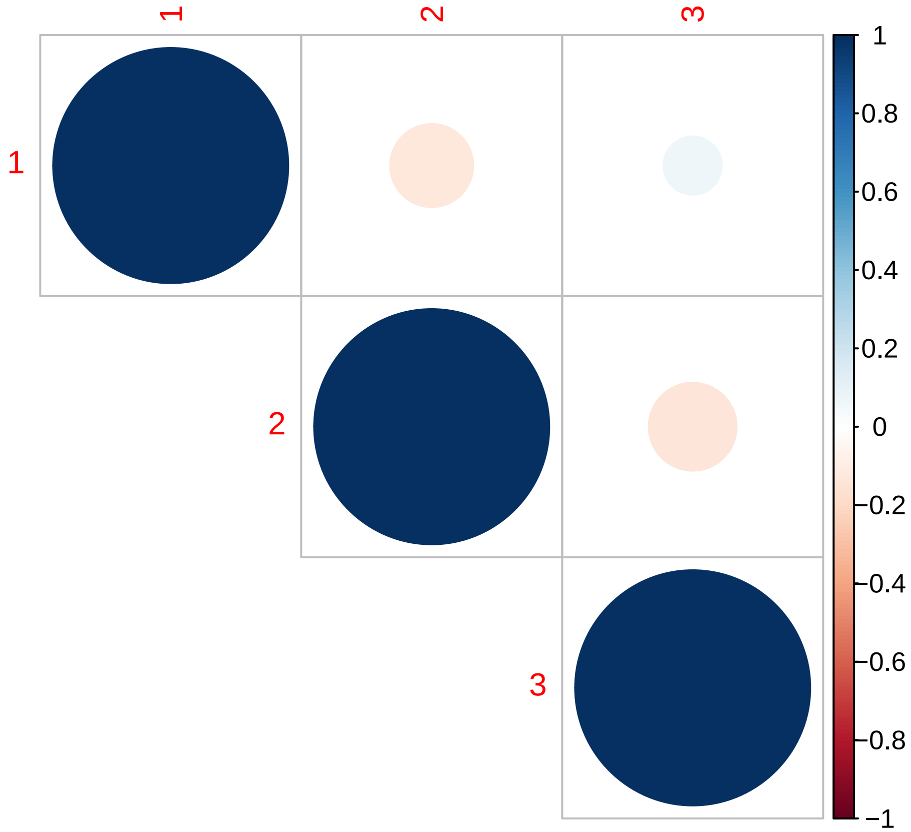
Based on Question 3, the factor "Whether Ejected" and "Safety in use" seem to be the determining the seriousness of the injuries sustained after a car accident.

We then explore the relationship between these factors.

```
library(corrplot)
```

```
## corrplot 0.84 loaded
```

```
library(RColorBrewer)
M <-cor(mydata)
corrplot(M, type="upper")
```



From the correlation matrix, we can find that there exists a positive relationship between "Safety Equipment in Use" and "Whether Ejected", and between "Injury" and "Whether Ejected"; there exists a negative relationship between "Injury" and "Safety Equipment in Use".

Bayesian Testing

```
bayes.test = function(var1,var2)
{
```

```r
#two-way table
obstable = table(var1,var2);
#first one-way marginal
rowtable = table(var1);
#second one-way marginal
columntable = table(var2);
#calculate the log-marginal likelihood under the saturated log-linear model
alpha = 0.25;
logmargSaturated = lgamma(4*alpha)-4*lgamma(alpha);
logmargSaturated = logmargSaturated +sum(lgamma(as.vector(obstable+alpha)));
logmargSaturated = logmargSaturated -lgamma(sum(as.vector(obstable+alpha)));
#calculate the log-marginal likelihood under the log-linear model of independence
logmargIndep = 2*lgamma(4*alpha)-4*lgamma(2*alpha);
logmargIndep = logmargIndep + sum(lgamma(as.vector(rowtable+2*alpha)));
logmargIndep = logmargIndep+sum(lgamma(as.vector(columntable+2*alpha)));
logmargIndep = logmargIndep - lgamma(sum(as.vector(rowtable+2*alpha)));
logmargIndep = logmargIndep -lgamma(sum(as.vector(columntable+2*alpha)));
return(2*(logmargSaturated-logmargIndep));
}
```

```r
bayes.test(mydata[,1], mydata[,3])
```

```
## [1] 1877.463
```

```r
bayes.test(mydata[,2], mydata[,3])
```

```
## [1] 2414.74
```

Based on Bayesian Testing, we can find that there exist strong relationship between "Injury" and other two explantory variables.

Problem 2

We will perform different methods for model selection.

```r
library(gRbase)
library(gRim)
library(graph)
```

```
## Loading required package: BiocGenerics

## Loading required package: parallel

##
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:parallel':
##
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, parApply, parCapply, parLapply,
##     parLapplyLB, parRapply, parSapply, parSapplyLB

## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs

## The following objects are masked from 'package:base':
```

```
##
##     anyDuplicated, append, as.data.frame, basename, cbind, colnames,
##     dirname, do.call, duplicated, eval, evalq, Filter, Find, get, grep,
##     grepl, intersect, is.unsorted, lapply, Map, mapply, match, mget,
##     order, paste, pmax, pmax.int, pmin, pmin.int, Position, rank,
##     rbind, Reduce, rownames, sapply, setdiff, sort, table, tapply,
##     union, unique, unsplit, which, which.max, which.min
```

```
library(Rgraphviz)
```

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

```
# data set
data(reinis)
str(reinis)
```

```
##  'table' num [1:2, 1:2, 1:2, 1:2, 1:2, 1:2] 44 40 112 67 129 145 12 23 35 12 ...
##  - attr(*, "dimnames")=List of 6
##   ..$ smoke  : chr [1:2] "y" "n"
##   ..$ mental : chr [1:2] "y" "n"
##   ..$ phys   : chr [1:2] "y" "n"
##   ..$ systol : chr [1:2] "y" "n"
##   ..$ protein: chr [1:2] "y" "n"
##   ..$ family : chr [1:2] "y" "n"
```

We display the saturated model:

```
m<-dmod(~.^.,data=reinis)
formula(m)
```

```
## ~smoke * mental * phys * systol * protein * family
```

We denote these 6 variables as A, B, C, D, E, F:

A indicates whether or not the worker "smokes",

B corresponds to "strenuous mental work"

C corresponds to "strenuous physical work"

D corresponds to "systolic blood pressure"

E corresponds to "ratio of β and α lipoproteins"

F represents "family anamnesis of coronary heart disease".

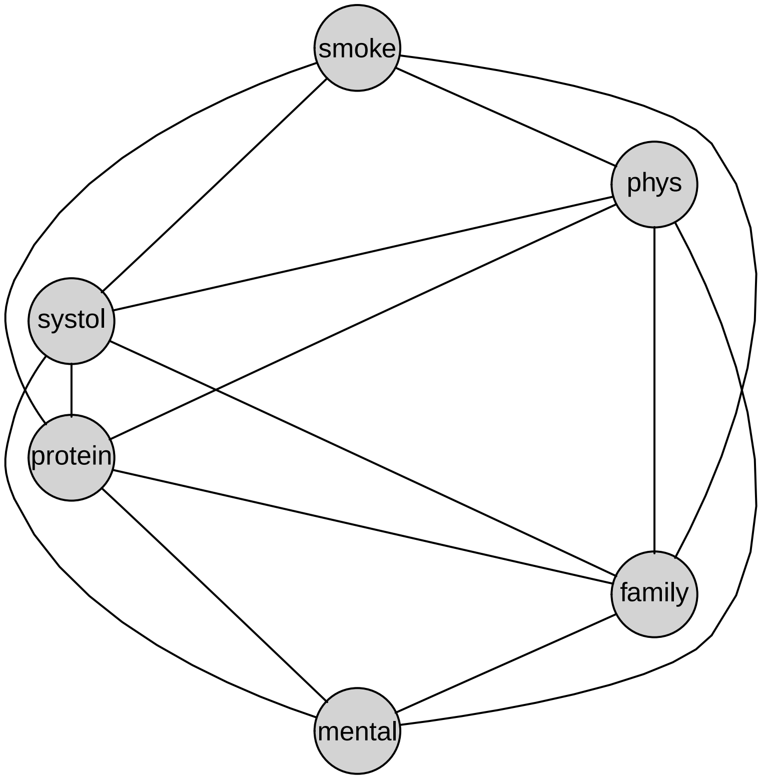
Model selection based on the paper"A fast procedure for model search in multidimensional contingency tables"

The first step consists of removing one edge at a time from the saturated model, i.e. fitting the candidate models C_0,

$$C_0 = \{(AB)^{-1}, (AC)^{-1}, ..., (EF)^{-1}\}$$

there are 6 choose 2, 15 models in total, and each of them can be derived by deleting one edge at a time from the saturated model. We will take the model $(AB)^{-1}$ as an example.

```
m_AB<-update(m,list(dedge=~smoke:mental))
plot(m_AB)
```



The Figure of model $(AB)^{-1}$ can verify the deleting edge is AB.

Then we need to list all of the candidates and compare these 15 models with the saturated model by likelihood ratio test.

```
# list all of the candidate models
combinations <- combn(c("smoke", "mental", "phys", "systol","protein", "family"), 2)
combinations
```

```
##      [,1]     [,2]    [,3]     [,4]      [,5]     [,6]     [,7]     [,8]
## [1,] "smoke"  "smoke" "smoke"  "smoke"   "smoke"  "mental" "mental" "mental"
## [2,] "mental" "phys"  "systol" "protein" "family" "phys"   "systol" "protein"
##      [,9]     [,10]    [,11]     [,12]    [,13]     [,14]    [,15]
## [1,] "mental" "phys"   "phys"    "phys"   "systol"  "systol" "protein"
## [2,] "family" "systol" "protein" "family" "protein" "family" "family"
```

```
# deleting one edge
# A.
m_AB <- update(m,list(dedge=~smoke:mental))
m_AC <- update(m,list(dedge=~smoke:phys))
m_AD <- update(m,list(dedge=~smoke:systol))
m_AE <- update(m,list(dedge=~smoke:protein))
m_AF <- update(m,list(dedge=~smoke:family))

# B.
m_BC<-update(m,list(dedge=~phys:mental))
m_BD<-update(m,list(dedge=~systol:mental))
m_BE<-update(m,list(dedge=~mental:protein))
m_BF<-update(m,list(dedge=~mental:family))

# C.
```

```
m_CD<-update(m,list(dedge=~systol:phys))
m_CE<-update(m,list(dedge=~phys:protein))
m_CF<-update(m,list(dedge=~phys:family))

# D.
m_DE<-update(m,list(dedge=~systol:protein))
m_DF<-update(m,list(dedge=~systol:family))

# E.
m_EF<-update(m,list(dedge=~protein:family))
```

We perform G2 and X2 test and calculate p-values based on models.

```
P_val <- function(model){
  #model is the model deleting one edge
  p1 = 1 - pchisq(model$fitinfo$dev, model$fitinfo$dimension[4] )
  p2 = 1 - pchisq(model$fitinfo$pearson, model$fitinfo$dimension[4])
  cat("The p-value based on G2 is", p1, ", The p-value based on X2 is", p2, "\n")
  return (list(p1=p1, p2=p2))
}

candidates <- list(m_AB,m_AC,m_AD,m_AE,m_AF,m_BC,m_BD,m_BE,m_BF,m_CD,m_CE,m_CF,m_DE,m_DF,m_EF)
s <- sapply(candidates ,P_val)
```

```
## The p-value based on G2 is 0.1233634 , The p-value based on X2 is 0.1704837
## The p-value based on G2 is 0.0002990934 , The p-value based on X2 is 0.0004500081
## The p-value based on G2 is 0.0258639 , The p-value based on X2 is 0.03659285
## The p-value based on G2 is 0.0007723243 , The p-value based on X2 is 0.001177167
## The p-value based on G2 is 0.1670744 , The p-value based on X2 is 0.2335899
## The p-value based on G2 is 0 , The p-value based on X2 is 0
## The p-value based on G2 is 0.728312 , The p-value based on X2 is 0.8103852
## The p-value based on G2 is 0.3711028 , The p-value based on X2 is 0.4451269
## The p-value based on G2 is 0.1195338 , The p-value based on X2 is 0.1104507
## The p-value based on G2 is 0.5387137 , The p-value based on X2 is 0.6413533
## The p-value based on G2 is 0.2883525 , The p-value based on X2 is 0.3529375
## The p-value based on G2 is 0.1383237 , The p-value based on X2 is 0.1884649
## The p-value based on G2 is 0.01322325 , The p-value based on X2 is 0.01688678
## The p-value based on G2 is 0.3040622 , The p-value based on X2 is 0.3730389
## The p-value based on G2 is 0.3057192 , The p-value based on X2 is 0.3345246
```

```
# at 5% level
combinations[, which(s[1,] < 0.05)]
```

```
##      [,1]    [,2]     [,3]      [,4]     [,5]
## [1,] "smoke" "smoke"  "smoke"   "mental" "systol"
## [2,] "phys"  "systol" "protein" "phys"   "protein"
```

```
combinations[, which(s[2,] < 0.05)]
```

```
##      [,1]    [,2]     [,3]      [,4]     [,5]
## [1,] "smoke" "smoke"  "smoke"   "mental" "systol"
## [2,] "phys"  "systol" "protein" "phys"   "protein"
```

From the results of both tests, we will reject these five models: m_AC, m_AD, m_AE, m_BC, m_DE, i.e. {(AC)-, (AD)-, (AE)-, (BC)-,(DE)-}, and accept models: m_AB, m_AF, m_BD, m_BE, m_BF, m_CD, m_CF, m_DF, m_DF.

The second step is fit the model (AC,AD,AE,BC,DE)+,

```
m_second <-  dmod(~smoke*protein*systol+smoke*phys+mental*phys+family,data=reinis)
p <- P_val(m_second)
```

```
## The p-value based on G2 is 0.002607725 , The p-value based on X2 is 0.003897921
```

At 5% level, this model is rejected, so the rejected models now are {(AC)-, (AD)-, (AE)-, (BC)-,(DE)-, (AC,AD,AE,BC,DE)+}.

The third step is to fit the models that contain the edges AC, AD, AE, BC, and DE, plus one edge from {AB, AF, BD, BE, BF, CD, CF, DF, DF}, upward stepping from the model fitted in the second step.

```
m_second_AB <- update(m_second, list(aedge=~smoke:mental))
m_second_AF <- update(m_second, list(aedge=~smoke:family))
m_second_BD<-update(m_second,list(aedge=~mental:systol))
m_second_BE<-update(m_second,list(aedge=~mental:protein))
m_second_BF<-update(m_second,list(aedge=~mental:family))
m_second_CD<-update(m_second,list(aedge=~phys:systol))
m_second_CE<-update(m_second,list(aedge=~phys:protein))
m_second_CF<-update(m_second,list(aedge=~phys:family))
m_second_DF<-update(m_second,list(aedge=~systol:family))
m_second_EF<-update(m_second,list(aedge=~protein:family))
```

```
m_list <- list(m_second_AB, m_second_AF, m_second_BD, m_second_BE, m_second_BF, m_second_CD,m_second_CE
s <- sapply(m_list, P_val)
```

```
## The p-value based on G2 is 0.005523942 , The p-value based on X2 is 0.005509597
## The p-value based on G2 is 0.00248871 , The p-value based on X2 is 0.003368513
## The p-value based on G2 is 0.002057461 , The p-value based on X2 is 0.003186275
## The p-value based on G2 is 0.1023395 , The p-value based on X2 is 0.1229055
## The p-value based on G2 is 0.005523628 , The p-value based on X2 is 0.008379627
## The p-value based on G2 is 0.001733692 , The p-value based on X2 is 0.002810451
## The p-value based on G2 is 0.09938997 , The p-value based on X2 is 0.1349491
## The p-value based on G2 is 0.002033665 , The p-value based on X2 is 0.002937662
## The p-value based on G2 is 0.00251954 , The p-value based on X2 is 0.004334444
## The p-value based on G2 is 0.003812355 , The p-value based on X2 is 0.005798924
```

```
which(s[1,]>0.05)
```

```
## [1] 4 7
```

```
which(s[2,]>0.05)
```

```
## [1] 4 7
```

based on both tests, only two models are not rejected: m_second_BE and m_second_CE, namely (AC,AD,AE,BC,DE,BE)+, (AC,AD,AE,BC,DE,CE)+.

Thus we obtain A = {(AC,AD,AE,BC,DE,BE)+, (AC,AD,AE,BC,DE,CE)+} since all other accepted models include one of these.

The fourth step consists of examination if $D_r(A)/R$.

$$D_r(A)/R = (BE, CE)^-$$

```
# (BE, CE)-
m_BECE <- update(m, list(dedge=~mental:protein+phys:protein))
p <- P_val(m_BECE)
```

```
## The p-value based on G2 is 0.01114833 , The p-value based on X2 is 0.01621571
```

P-values of both tests are less than 0.05, hence this model is rejected and we can interpret this as meaning that either BE or CE must be in any acceptable model.

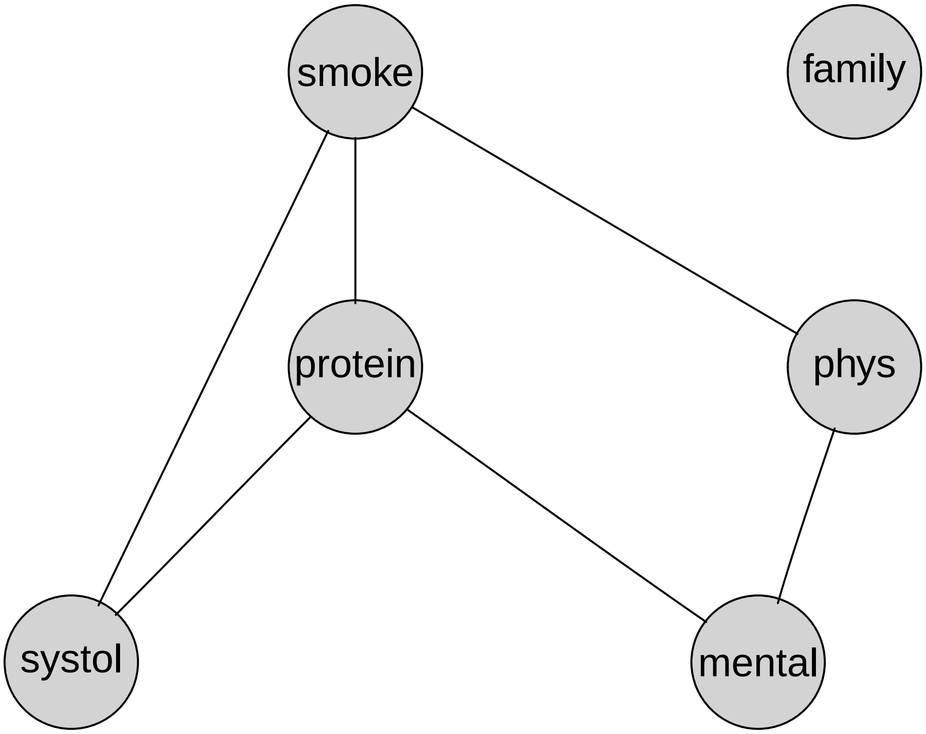
We infer that the two models accepted at step three constitude a complete accepted set, and the procedure stops. In the generating set notation, these models are [AC, ADE, BC, BE, F] and [ACE, ADE, BC, F]. Explan: in model (AC,AD,AE,BC,DE,BE)+, AD, DE, AE -> ADE, in model (AC,AD,AE,BC,DE,CE)+, AC, CE, AE -> ACE, AD, AE, DE -> ADE

Hence the final two models:

```
# [AC, ADE, BC, BE, F]
formula(m_second_BE)
```

```
## ~smoke * protein * systol + smoke * phys + mental * phys + family +
##     mental * protein
```

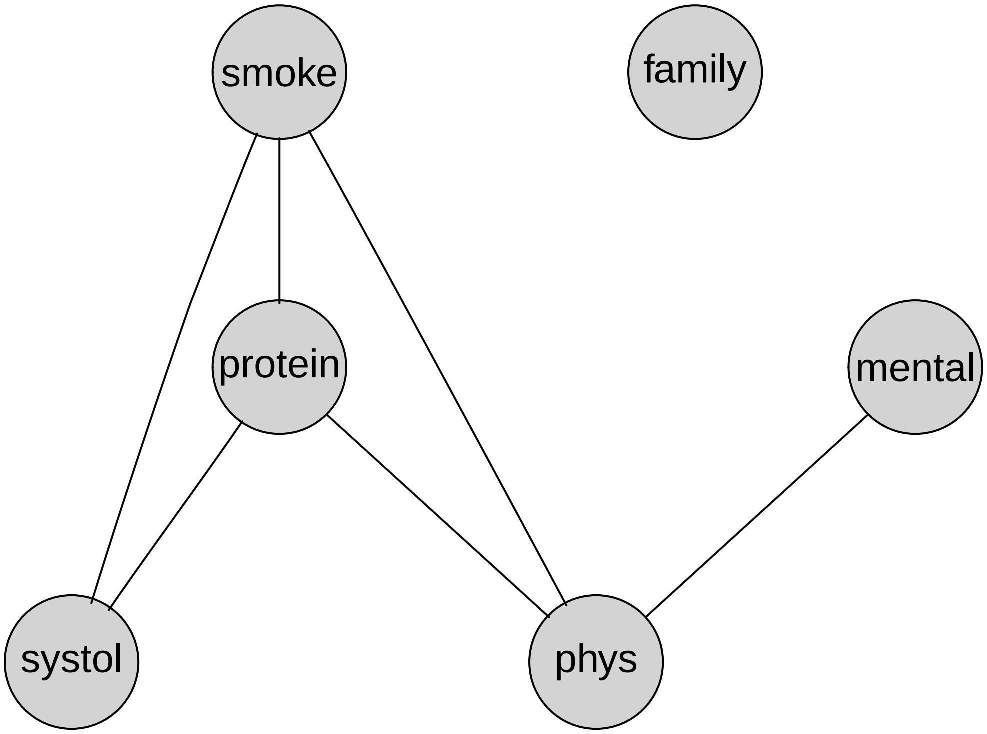
```
plot(m_second_BE)
```



```
# [ACE, ADE, BC, F]
formula(m_second_CE)
```

```
## ~smoke * protein * systol + mental * phys + family + phys * protein *
##     smoke
```

```
plot(m_second_CE)
```



Interpretations: The most striking feature of both models is the independence of the family anamnesis, F. A dependence on D, E. The two models differ in thepresence or absence of the edges BE and CE.

Model selection based on AIC and BIC

Another way to select model is based on AIC and BIC. AIC minimizes the negative of a penalized likelihood

$$AIC(k) = -2log(L) + kdim(M)$$

where $dim(M)$ is the number of independent parameters in model M.

BIC also penalized the likelihood, but in a more severe way

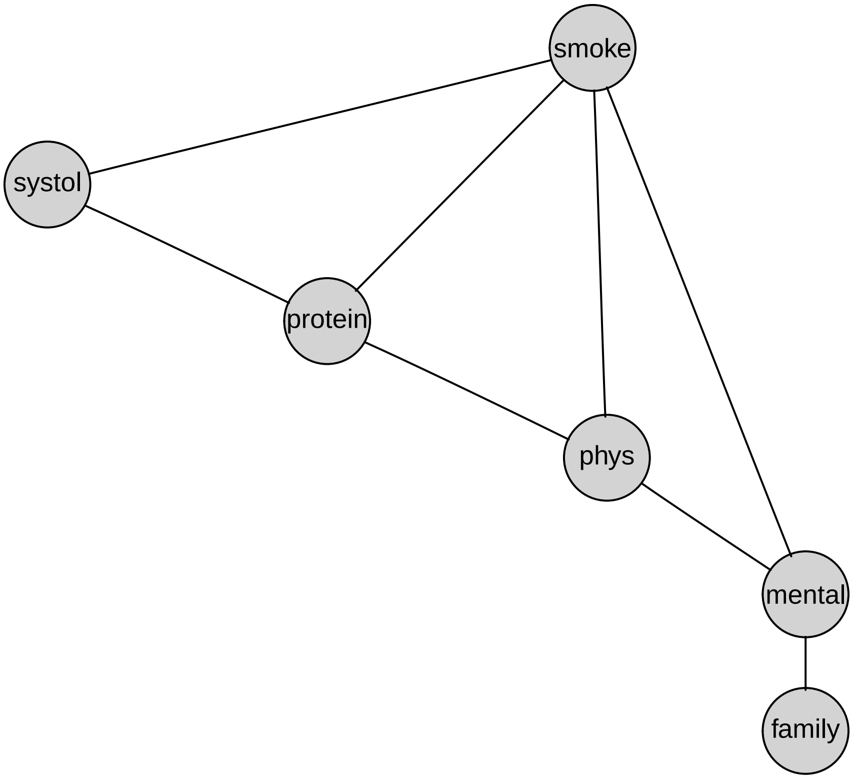
$$BIC(k) = -2log(L) + log(n)dim(M)$$

where n is the number of observations. The results based on AIC and BIC are as follows

```
# AIC
aic <-stepwise(m)
formula(aic)
```

```
## ~smoke * systol * protein + smoke * phys * protein + smoke *
##     mental * phys + mental * family
```

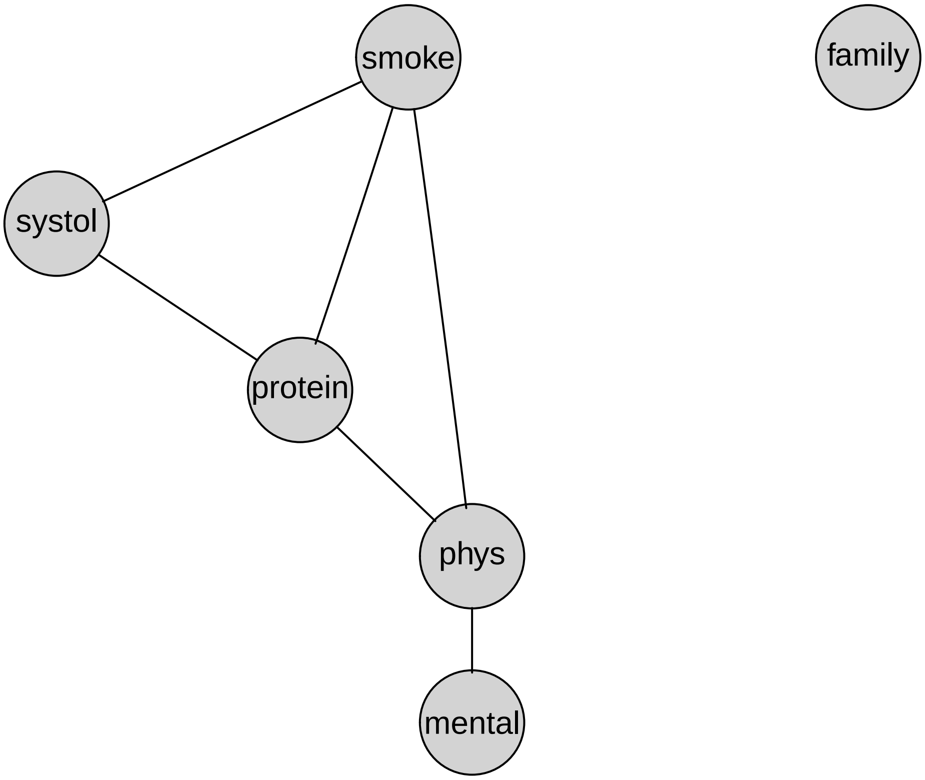
```
plot(aic)
```



```
# BIC
bic <- stepwise(m, k =log(sum(reinis)))
formula(bic)
```

```
## ~smoke * systol * protein + smoke * phys * protein + mental *
##     phys + family
```

```
plot(bic)
```



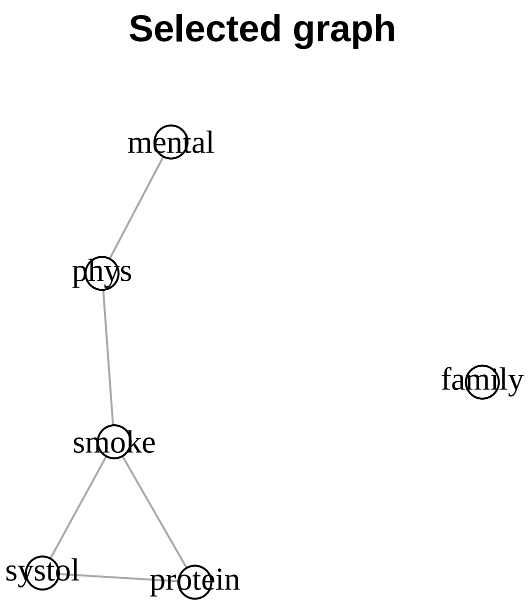
The differnece between the model selected by AIC and by BIC is (1). the independence or dependence of F; (2) the presence or absence of the dependence of mental and smoke.

Model selection based on BDMCMC algorithm

```
library(BDgraph)
data("reinis")
sample <- bdgraph.mpl(data=reinis, method="dgm-binary",  iter = 10000, burnin = 6000)
```

```
## 10000 iteration is started.
 Iteration  1000
##  Iteration  2000
##  Iteration  3000
##  Iteration  4000
##  Iteration  5000
##  Iteration  6000
##  Iteration  7000
##  Iteration  8000
##  Iteration  9000
##  Iteration  10000
```

```
summary(sample)
```

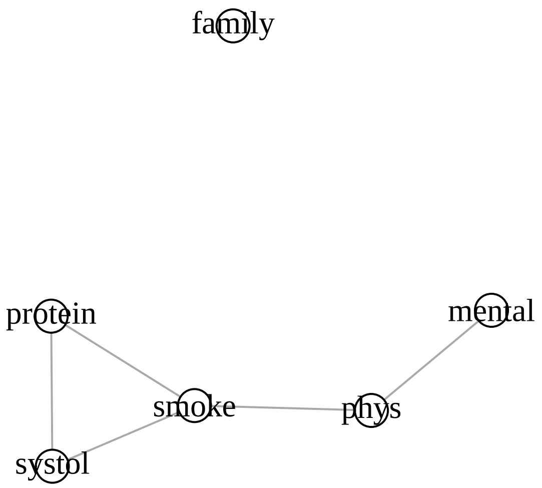


Graph with edge posterior probability > 0.5

```
## $selected_g
##      smoke mental phys systol protein family
## [1,]     0      0    1      1       1      0
## [2,]     0      0    1      0       0      0
## [3,]     0      0    0      0       0      0
## [4,]     0      0    0      0       1      0
## [5,]     0      0    0      0       0      0
## [6,]     0      0    0      0       0      0
##
## $p_links
##      smoke mental phys systol protein family
## [1,]     0      0    1   0.76    1.00   0.00
## [2,]     0      0    1   0.00    0.11   0.06
## [3,]     0      0    0   0.00    0.01   0.00
## [4,]     0      0    0   0.00    0.99   0.00
## [5,]     0      0    0   0.00    0.00   0.00
## [6,]     0      0    0   0.00    0.00   0.00
```

```r
select(sample, cut=0.5,vis = TRUE)
```

Graph with links posterior probabilities > 0.5



```
##      smoke mental phys systol protein family
## [1,]     0      0    1      1       1      0
## [2,]     0      0    1      0       0      0
## [3,]     0      0    0      0       0      0
## [4,]     0      0    0      0       1      0
## [5,]     0      0    0      0       0      0
## [6,]     0      0    0      0       0      0
```

Besides the model selected by AIC, the other models shows the independence of variable F, i.e. “family anamnesis of coronary heart disease. Hence we infer that variables A, B, C, D, E are not directly determinant of variable F and we will not include any of them into logistic regression model as variables.