536HW6

Coco_Luo

2022-11-29

```
geno = c(1167, 1509, 763, 234, 107, 14,
         377, 16, 225, 2, 29, 0,
         186, 179, 130, 19, 11, 2)
geno.array = array(geno, c(2, 3, 3))
geno.array
```

```
## , , 1
##
##      [,1] [,2] [,3]
## [1,] 1167  763  107
## [2,] 1509  234   14
##
## , , 2
##
##      [,1] [,2] [,3]
## [1,]  377  225   29
## [2,]   16    2    0
##
## , , 3
##
##      [,1] [,2] [,3]
## [1,]  186  130   11
## [2,]  179   19    2
```

1 Saturated Log-linear Model

```
geno1 = c(1167, 1509, 763, 234, 107, 14,
         377, 16, 225, 2, 29, 0.00001,
         186, 179, 130, 19, 11, 2)
geno.array1 = array(geno1, c(2, 3, 3))
#geno.array1
```

```
saturated.loglin1 = loglin(geno.array1,margin=list(c(1,2,3)),fit = T,param = T)
```

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

```
#saturated.loglin1$param
```

I first fit the saturated log-linear model. Since the saturated log-linear model considers that all three factors (disease status, genotype at SNP1 nad SNP2) are related in any possible way. The saturated log-linear model implies:

- knowing a one's genotype at loci SNP 1 and SNP 2 tells something about disease status

- knowing a one's disease status and genotype at SNP 1 tells something about their genotype at SNP 2
- knowing a one's disease status and genotype at SNP 2 tells something about their genotype at SNP 1.

For example, the regression associated with X_1 is:

$$log\frac{P(X_1 = i_1|X_2 = j, X_3 = k)}{P(X_1 = i_2|X_2 = j, X_3 = k)} = log\frac{P(X_1 = i_1, X_2 = j, X_3 = k)}{P(X_1 = i_2, X_2 = j, X_3 = k)}$$
$$= (u_{1(i_1)} - u_{1(i_2)}) + (u_{12(i_1j)} - u_{12(i_2j)}) + (u_{13(i_1k)} - u_{13(i_2k)}) + (u_{123(i_1jk)} - u_{123(i_2jk)})$$

This model does not have any interpretation in terms of independence or conditional independence relationships among the three variables. It fits the data well and provides little information about what is going on in our dataset.

2 Complete Independence Model

The Complete Independence Model considers that all three factors (disease status, genotype at SNP1 nad SNP2) are independent to each other. That is knowing the disease status tells nothing about genotypes, knowing the disease status and genotype at locus SNP 1 tells nothing about their genotype at locus SNP 2 and knowing genotype at loci SNP 1 and SNP 2 tells nothing about their disease status. We used the R function to obtain the p value $P(\chi^2_{12} \geq 1078.846$ for our null hypothesis: $H_0 : \mu_{12(ij)} = \mu_{23(jk)} = \mu_{13(ik)} = \mu_{123(ijk)} = 0$ based on G^2 and χ^2, our p values are both 0. Thus, we reject the null and conclude that the complete independence model does not fit our data well.

```
ind.logln = loglin(geno.array,margin=list(1,2,3), fit=TRUE,param=TRUE)
```

```
## 2 iterations: deviation 9.094947e-13
```

```
1-pchisq(ind.logln$lrt,ind.logln$df)
```

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

```
1-pchisq(ind.logln$pearson,ind.logln$df)
```

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

3 Models of conditional independence

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

```
## 2 iterations: deviation 2.273737e-13
```

```
1 - pchisq(X2indepX3givenX1.loglin$lrt, X2indepX3givenX1.loglin$df)
```

```
## [1] 0.552064
```

```
1 - pchisq(X2indepX3givenX1.loglin$pearson, X2indepX3givenX1.loglin$df)
```

```
## [1] 0.6075692
```

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

```
## 2 iterations: deviation 2.273737e-13
```

```
1 - pchisq(X1indepX3givenX2.loglin$lrt, X1indepX3givenX2.loglin$df)
```

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

```
1 - pchisq(X1indepX3givenX2.loglin$pearson, X1indepX3givenX2.loglin$df)
```

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

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

```
## 2 iterations: deviation 2.273737e-13
```

```
1 - pchisq(X1indepX2givenX3.loglin$lrt, X1indepX2givenX3.loglin$df)
```

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

```
1 - pchisq(X1indepX2givenX3.loglin$pearson, X1indepX2givenX3.loglin$df)
```

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

We have three models of conditional independece.

1. The model of conditional independence of X_1 and X_2 given X_3 indicates that one's disease status is independent of their genotype at SNP 1 given their genotype at SNP 2. In other words, if we already know a person's genotype at SNP 2, knowing their disease status does not provide additional information about their genotype at SNP 1. Moreover, if we know a person's genotype at SNP2, knowing their genotype at SNP1 does not provide any additional information about their disease status.
2. The model of conditional independence of X_1 and X_3 given X_2 indicates that a person's disease status is independent of their genotype at SNP 2 given their genotype at SNP 1. If we already know a person's genotype at SNP 1, knowing their genotype at SNP 2 does not provide any extra information about their disease status. If we already know a person's genotype at SNP 1, knowing their disease status does not provide any additional information about their genotype at SNP 2.
3. The model of conditional independence of X_2 and X_3 given X_1 indicates that one's genotype at SNP 1 and SNP 2 are independent given their disease status. If we already know a person's disease status, knowing their genotype at SNP 1 does not provide any additional information about their genotype at SNP 2. And if we already know a person's disease status, knowing their genotype at SNP 2 does not provide any additional information about their genotype at SNP 1.

We fit the log-linear model [12][13] with the number of degrees of freedom = 8 and the p value based on G^2 related to [12][13] is about 0.55 thus the model fits the data well.

```
x2indx3givx1.loglin = loglin(geno.array,margin=list(c(1,2),c(1,3)), fit=TRUE,param=TRUE)
```

```
## 2 iterations: deviation 2.273737e-13
```

```
x2indx3givx1.loglin$df
```

```
## [1] 8
```

```
1-pchisq(X2indepX3givenX1.loglin$lrt,X2indepX3givenX1.loglin$df)
```

```
## [1] 0.552064
```

We fit the log-linear model [12][23] with the number of degrees of freedom = 8 and the p value based on G^2 related to [12][23] is about 0 thus the model does not fit the data well.

```
x1indx3givx2.loglin = loglin(geno.array,margin=list(c(1,2),c(2,3)), fit=TRUE,param=TRUE)
```

```
## 2 iterations: deviation 2.273737e-13
```

```
1-pchisq(x1indx3givx2.loglin$lrt,x1indx3givx2.loglin$df)
```

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

We fit the log-linear model [13][23] with the number of degrees of freedom = 8 and the p value based on G^2 related to [13][23] is about 0 thus the model does not fit the data well.

```
x1indx2givx3.loglin = loglin(geno.array,margin=list(c(1,3),c(2,3)), fit=TRUE,param=TRUE)
```

```
## 2 iterations: deviation 2.273737e-13
```

```
1-pchisq(x1indx2givx3.loglin$lrt,x1indx2givx3.loglin$df)
```

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

From the above three models, we can see that the only conditional independence model that fits the data is [12][13]. This model has interaction terms between SNP1, SNP2 and disease status.

4 One Variable Independent of the Other Two

We have three different models: the first one [1][23] is a model of independence of X_1 and $\{X_2, X_3\}$, in our case, the is the model of independence of disease from SNP1 and SNP 2. We cannot have any more information about genotype SNP1 and SNP2 even we know one's disease status. The model also implies that if we know something about genotype at SNP1, we have information about their genotype at SNP2. If we know one's genotype at SNP 2, we have information about their genotype at SNP 1. We fit the model below with df = 8 and the p value is about 0 for both G^2 and χ^2 statistics. Our p value indicates that [1][23] does not fit he data well.

```
x1indx2x3.loglin = loglin(geno.array,margin=list(1,c(2,3)), fit=TRUE,param=TRUE)
```

```
## 2 iterations: deviation 4.547474e-13
```

```
x1indx2x3.loglin$df
```

```
## [1] 8
```

```
1-pchisq(x1indx2x3.loglin$lrt,x1indx2x3.loglin$df)
```

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

```
1-pchisq(x1indx2x3.loglin$pearson,x1indx2x3.loglin$df)
```

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

The second model [2][13] is a model of independence of X_2 and $\{X_1, X_3\}$, in our case, this is the model of independence of genotype at SNP1 from disease and SNP2. We cannot have any information about disease status and genotype at SNP2 even we know one's genotype at SNP1. The model also implies that if we know something about one's disease status, we have information about their genotype at SNP2. If we know one's genotype at SNP 2, we have information about their disease status. We fit the model below with df = 8 and the p value is about 0 for both G^2 and χ^2 statistics. Our p value indicates that [2][13] does not fit he data well.

```
x2indx1x3.loglin = loglin(geno.array,margin=list(2,c(1,3)), fit=TRUE,param=TRUE)
```

```
## 2 iterations: deviation 4.547474e-13
```

```
1-pchisq(x2indx1x3.loglin$lrt,x2indx1x3.loglin$df)
```

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

```
1-pchisq(x2indx1x3.loglin$pearson,x2indx1x3.loglin$df)
```

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

The third model [3][12] is a model of independence of X_3 and $\{X_1, X_2\}$, in our case, this is the model of independence of genotype at SNP2 from disease and genotype at SNP1. We cannot have any information about disease status and genotype at SNP1 even we know one's genotype at SNP2. The model also implies that if we know something about one's disease status, we have information about their genotype at SNP1. If we know one's genotype at SNP 1, we have information about their disease status. We fit the model below with df = 8 and the p value is about 0 for both G^2 and χ^2 statistics. Our p value indicates that [3][12] does not fit he data well.

```
x3indx1x2.loglin = loglin(geno.array,margin=list(3,c(1,2)), fit=TRUE,param=TRUE)
```

```
## 2 iterations: deviation 4.547474e-13
```

```
1-pchisq(x3indx1x2.loglin$lrt,x3indx1x2.loglin$df)
```

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

```
1-pchisq(x3indx1x2.loglin$pearson,x3indx1x2.loglin$df)
```

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

5 No second order interaction [12][13][23]

We fit the log-linear model [12][13][23] below where the p value related to G^2 is about 0.462. Since the conditional independence model also fits the data well and it is a sub-model of it, we do expect that the no second order interaction model fits the data well. Moreover, the no second order interaction model does not have any interpretation in terms of independence or conditional independence relationships among X1, X2 and X3.

```
no2log = loglin(geno.array,margin=list(c(1,2),c(1,3),c(2,3)), fit=TRUE,param=TRUE)
```

```
## 4 iterations: deviation 0.05994974
```

```
1-pchisq(no2log$lrt,no2log$df)
```

```
## [1] 0.4619473
```

6 Model Selection

From the above analysis, we found that [12][13] and [12][13][23]. To decide on a model, we want to test H0: $u_{23(ij)} = 0$ using a likelihood ratio test:

```
1-pchisq(X2indepX3givenX1.loglin$lrt-no2log$lrt, X2indepX3givenX1.loglin$df-no2log$df)
```

```
## [1] 0.5166232
```

The difference in the values of G^2 follows an asymptotic Chi-squared distribution with a number of degrees of freedom equal to the difference in the number of degrees of freedom between the two log-linear models. We fail to reject H0. We should abandon [12][13][23] and choose [12][13], which is the most suitable log linear model for our data.