Log-linear Models with R Part 1

2-D tables

```
> # Playing with how to do it in R -- loglin command
> # H0: (Prisoner's race)(Victim's race)
> # help(loglin)
> racetable1 = rbind(c(151,9),
+                    c(63,103))
> test1 = chisq.test(racetable1,correct=F); test1

	Pearson's Chi-squared test

data:  racetable1
X-squared = 115.0083, df = 1, p-value < 2.2e-16

> try1 = loglin(racetable1,margin=list(1,2)); try1
2 iterations: deviation 0
$lrt
[1] 129.7977

$pearson
[1] 115.0083

$df
[1] 1

$margin
$margin[[1]]
[1] 1

$margin[[2]]
[1] 2
```

```
> # Look at estimated expected frequencies and parameter
> # estimates under H0
> try2 = loglin(racetable1,margin=list(1,2),fit=T,param=T); try2
2 iterations: deviation 0
$lrt
[1] 129.7977

$pearson
[1] 115.0083

$df
[1] 1

$margin
$margin[[1]]
[1] 1

$margin[[2]]
[1] 2

$fit
         [,1]     [,2]
[1,] 105.0307 54.96933
[2,] 108.9693 57.03067

$param
$param$`(Intercept)`
[1] 4.348921

$param$`1`
[1] -0.01840699  0.01840699

$param$`2`
[1]  0.3237386 -0.3237386
> log(sum(try2$fit[1,])/sum(try2$fit[2,]))/2
[1] -0.01840699

> # try2$fit are the usual expected frequencies
> sum(racetable1); sum(try2$fit) # Both = n
[1] 326
[1] 326
> # LR test by "hand"
> G2 = 2 * sum(racetable1 * log(racetable1/try2$fit)) ; G2
[1] 129.7977
```

```
> # Try a saturated model. Recall last command:
> # try2 = loglin(racetable1,margin=list(1,2),fit=T,param=T)
> try3 = loglin(racetable1,margin=list(c(1,2)),fit=T,param=T)
2 iterations: deviation 0
> try3
$lrt
[1] 0

$pearson
[1] 0

$df
[1] 0

$margin
$margin[[1]]
[1] 1 2

$fit
     [,1] [,2]
[1,]  151    9
[2,]   63  103

$param
$param$`(Intercept)`
[1] 3.998092
```

λ

```
$param$`1`
[1] -0.3908398  0.3908398
```

λ_1^X λ_2^X

```
$param$`2`
[1]  0.5821152 -0.5821152
```

λ_1^Y λ_2^Y

```
$param$`1.2`
           [,1]       [,2]
[1,]  0.8279124 -0.8279124
[2,] -0.8279124  0.8279124
```

λ_{ij}^{XY}

Berkeley Admissions Data

```
> UCBAdmissions # From package MASS: Admit x Gender x Dept
, , Dept = A

          Gender
Admit      Male Female
  Admitted  512     89
  Rejected  313     19

, , Dept = B

          Gender
Admit      Male Female
  Admitted  353     17
  Rejected  207      8

, , Dept = C

          Gender
Admit      Male Female
  Admitted  120    202
  Rejected  205    391

, , Dept = D

          Gender
Admit      Male Female
  Admitted  138    131
  Rejected  279    244

, , Dept = E

          Gender
Admit      Male Female
  Admitted   53     94
  Rejected  138    299

, , Dept = F

          Gender
Admit      Male Female
  Admitted   22     24
  Rejected  351    317
```

```
> GxA = margin.table(UCBAdmissions,c(2,1)); GxA
        Admit
Gender   Admitted Rejected
  Male       1198     1493
  Female      557     1278
>
> prop.table(GxA,1) # Proportion of dimension 1 (rows)
        Admit
Gender    Admitted  Rejected
  Male   0.4451877 0.5548123
  Female 0.3035422 0.6964578

> G2 = loglin(GxA,margin=list(1,2))$lrt ; G2   # LR Test of independence
2 iterations: deviation 0
[1] 93.44941
>
> 1-pchisq(G2,1)
[1] 0

> chisq.test(GxA,correct=F)

	Pearson's Chi-squared test

data:  GxA
X-squared = 92.2053, df = 1, p-value < 2.2e-16


> DxA =  margin.table(UCBAdmissions,c(3,1))
> prop.table(DxA,1)
    Admit
Dept   Admitted   Rejected
   A 0.64415863 0.35584137
   B 0.63247863 0.36752137
   C 0.35076253 0.64923747
   D 0.33964646 0.66035354
   E 0.25171233 0.74828767
   F 0.06442577 0.93557423
```

```
> # Exploratoy model fitting strategy (common)
> # Find a model that fits almost as well as the saturated model.
> # Saturated is full, candidate model is reduced.
> # Can also compare candidate models if they are nested.
> mod1 = loglin(UCBAdmissions,margin=list(1,2,3)) # Complete independence
2 iterations: deviation 4.547474e-13
> mod1
$lrt
[1] 2097.671

$pearson
[1] 2000.328

$df
[1] 16

$margin
$margin[[1]]
[1] "Admit"

$margin[[2]]
[1] "Gender"

$margin[[3]]
[1] "Dept"


> length(UCBAdmissions)
[1] 24
> 23-5-1-1 # df
[1] 16
> # mod2 will have (Admit,Dept) and (Dept,Gender), but not (Admit,Gender)
> # Conditional independence, and innocence
> mod2 = loglin( UCBAdmissions, margin=list(c(1,3),c(2,3)) ) ; mod2
2 iterations: deviation 2.842171e-14
$lrt
[1] 21.73551

$pearson
[1] 19.93841

$df
[1] 6
```

```
$margin
$margin[[1]]
[1] "Admit" "Dept"

$margin[[2]]
[1] "Gender" "Dept"

> # Does it fit?
> 1-pchisq(mod2$lrt,mod2$df)
[1] 0.001351993

> # No. Is it an improvement on mod1?
> # mod1 is full, mod2 is reduced
> G2 = mod1$lrt-mod2$lrt; DF = mod1$df-mod2$df
> G2; DF; 1-pchisq(G2,DF)
[1] 2075.936
[1] 10
[1] 0
> # So it's a big improvement but still not good enough. Add (Gender,Admit)
> mod3 = loglin( UCBAdmissions, margin=list(c(1,2),c(1,3),c(2,3)) ) ; mod3
9 iterations: deviation 0.04920393
$lrt
[1] 20.20428

$pearson
[1] 18.82376

$df
[1] 5

$margin
$margin[[1]]
[1] "Admit"  "Gender"

$margin[[2]]
[1] "Admit" "Dept"

$margin[[3]]
[1] "Gender" "Dept"
```

```
> # Does Model 3 fit?
> 1-pchisq(mod3$lrt,mod3$df)
[1] 0.001144077
>
> # No, and the p-value is similar to mod2. Is it an improvement?
> G2 = mod2$lrt-mod3$lrt; DF = mod2$df-mod3$df
> G2; DF; 1-pchisq(G2,DF)
[1] 1.531229
[1] 1
[1] 0.2159281
>
> # This is where the university administration can relax, but it's a bit
> # technical. They look at separate tables by department, and then they
> # are really happy.
>
> # One statistician is not satisfied. Does the relationship of Gender
> # to Admission DEPEND on the Department? How about a formal test?
>
> # The second statistician says "We've already tested this, fool."
>
> #  The first statistician says "Relationship DEPENDS means the model has
> #  a 3-way interaction. Models are hierarchical, so the
> #  (Gender,Admit) term is in the model too, but now it does not mean
> #  guilt. When higher-order interactions are present, the lower-order
> #  interactions involving those terms become difficult to interpret. "
>
> # The second statistician says "Do you think I don't know this?"
>
> # The first statistician says "We'll compare the fit of a model with
> #  the 3-way interaction to mod3, which has all the 2-ways. We could
> #  also compare  it to mod2, which is our best model so far.
>
```

```
> mod4 = loglin( UCBAdmissions, margin=list(c(1,2,3)) ) ; mod4
2 iterations: deviation 5.684342e-14
$lrt
[1] 0

$pearson
[1] 0

$df
[1] 0

$margin
$margin[[1]]
[1] "Admit"  "Gender" "Dept"


>
> # The second statistician says "Haha, I told you so! The model with the
> # highest order interaction is always saturated, so the test of fit for
> # mod3 is also the test of mod3 vs. mod4."
>
> # The first statistician says "The Dean is taking me out to lunch, but
for some reason you are not invited."
```