Log Linear Analysis Walkthrough

Tom Dalton

09/02/2017

This walkthrough has been created in the process of following the tutorial here: https://ww2.coastal.edu/kingw/statistics/R-tutorials/loglin.html

This document records the steps taken to perform a log linear analysis in R. The first section makes use of a toy data set before complicating things in the second section and using data from a genalogical population generator and looking to assert the similarity of the generators input parameters to the resulting population.

Section 1 - Log linear analysis of Titanic data

Preliminary introductions...

First, lets read the data in:

```
data(Titanic)
dimnames(Titanic)
```

```
## $Class
## [1] "1st"  "2nd"  "3rd"  "Crew"
##
## $Sex
## [1] "Male"   "Female"
##
## $Age
## [1] "Child" "Adult"
##
## $Survived
## [1] "No"  "Yes"
```

```
margin.table(Titanic)
```

```
## [1] 2201
```

Log linear analysis will allow us to look for relationships among the varibles in the multiway contingency table that we have.

Log linear analysis assumes:

- All observations in the tables are independant *(observations **not** variables)*
- Cell by cell frequency are sufficiently high *(generally 5 or more)*

To start with the basics, lets look at the how surivial varies with gender. Here we have the data:

```r
margin.table(Titanic, c(2,4))
```

```
##         Survived
## Sex        No  Yes
##   Male   1364  367
##   Female  126  344
```

Lets consider the odds for survival for each gender:

```r
male_survival_odds = male_survivors / male_deaths
female_survival_odds = female_survivors / female_deaths
survival_odds_ratio = female_survival_odds / male_survival_odds
```

```
## 10.14697
```

We can see from this that the a women was much more likely to survive.

The likelihood ratio of a female vs. a male surviving is:

```r
male_likelihood = male_survivors / total_males
female_likelihood = female_survivors / total_females
likelihood_ratio = female_likelihood / male_likelihood
```

```
## 3.452165
```

This being the proportion of females who survived divided by the proportion of males who survived.

Thus females were almost 3.5 times more likely to survive.

This is the langugae of log linear analysis. In this case applying a Pearson Chi-square test would reveal a highly significant interaction between these two factors *(sex and survival)*.

The underpinning of a log linear analysis is based on the **likelihood ratio chi square**. The advantage of this is that the likelihood ratio chi squares are additive, meaning we can the chi squares derived from simpler models can be added together to produce the chi squares derived from more complex models.

We can perform a likelihood ratio chi square test for a our two way table above *(although this would not be the normal way to analyse a two way table)*.

```r
sex.survived = margin.table(Titanic, c(2,4))   # create the contingency table
likelihood.test(sex.survived)
```

```
##    LR-chisq          df     p-value Pears-chisq
##    434.4688      1.0000      0.0000    456.8742
```

Here we see a p value of 0.000, thus we can confirm our ealier assertion that women are more likely to survive, thus rejecting the null hypothersis. However is we define a data table to be:

```
##    Died Survived
## gA 1364      367
## gB 1364      366
```

We can see that here where the there is virtually no difference between the outcomes of the two groups then the null hypothesis can be accepted and we can assert that the interaction between these two factors is not significant:

```
likelihood.test(table)
```

```
##    LR-chisq          df     p-value Pears-chisq
##      0.0011      1.0000      0.9738      0.0011
```

The Log Linear Analysis

```
library("MASS")
```

Firstly we can conduct the same analysis we did before on a two way table simply by doing:

```
loglm( ~ Sex + Survived, data = sex.survived)
```

```
## Call:
## loglm(formula = ~Sex + Survived, data = sex.survived)
##
## Statistics:
##                       X^2 df P(> X^2)
## Likelihood Ratio 434.4688  1        0
## Pearson          456.8742  1        0
```

We've passed in the contingency table and specified the variables to be looked at. We can check back and see that this has achieved the same results as out likelihood test.

Now lets include all the variables in the model.

```
loglm( ~ Class + Sex + Age + Survived, data = Titanic)
```

```
## Call:
## loglm(formula = ~Class + Sex + Age + Survived, data = Titanic)
##
## Statistics:
##                       X^2 df P(> X^2)
## Likelihood Ratio 1243.663 25        0
## Pearson          1637.445 25        0
```

This is a four-way chi-squared test of independance, the p value allows us to reject the null hypothesis thus indicating somewhere in our data there are factors that are interacting with one another to produce the observed cell frequencies.

Here we should introduce the term **saturated model**: A model is saturated when it includes all effects for each factor thus including all possible interactions between them. As such a model would explain the cell frequencies perfectly, it would have chi squared statistic of zero on zero degrees of freedom. This being the case it would have no explanatory power at all. Here is an exmaple of such a model:

```
loglm( ~ Class * Sex * Age * Survived, data = Titanic)
```

```
## Call:
## loglm(formula = ~Class * Sex * Age * Survived, data = Titanic)
##
## Statistics:
##                   X^2 df P(> X^2)
## Likelihood Ratio   0  0        1
## Pearson          NaN  0        1
```

We can remove particular interactions from our model by doing this to remove the fourway interaction:

```
loglm( ~ Class * Sex * Age * Survived - Class:Sex:Age:Survived, data = Titanic)
```

```
## Call:
## loglm(formula = ~Class * Sex * Age * Survived - Class:Sex:Age:Survived,
##     data = Titanic)
##
## Statistics:
##                           X^2 df  P(> X^2)
## Likelihood Ratio 0.0002728865  3 0.9999988
## Pearson                   NaN  3       NaN
```

Therefore this shows us that the four way interaction in the model was expendable.

The next model however with the simple factors plus the interaction between Age and Survived does predict that expected frequencies are significantly difference from the observed frequencies.

```
loglm( ~ Class + Sex + Age + Survived + Age:Survived, data = Titanic)
```

```
## Call:
## loglm(formula = ~Class + Sex + Age + Survived + Age:Survived,
##     data = Titanic)
##
## Statistics:
##                       X^2 df P(> X^2)
## Likelihood Ratio 1224.103 24        0
## Pearson          1596.846 24        0
```

Testing a Specific Hypothesis

Lets say we set out with thr hypothersis that gender was related to survival on the Titanic. The two way chi-square test we have done supports this. But by the same method sex is strongly related to class and age, which both in turn seem to be strongly related to survived. Thus we need to understand class and ages part in the relationship between sex and survived.

Log linear analysis will allow us to tease apart these effects.

If we remove all interaction terms that involve both sex and survived and the model still fits the obseved frequencies adequately, then we can conclude that gender and survival were unrelated. So:

```
sat.model = loglm(~ Class * Sex * Age * Survived, data=Titanic)
model2 = update(sat.model, ~.-(Class:Sex:Age:Survived + Sex:Age:Survived + Class:Sex:Survived + Sex:Sur
model2
```

```
## Call:
## loglm(formula = ~Class + Sex + Age + Survived + Class:Sex + Class:Age +
##     Sex:Age + Class:Survived + Age:Survived + Class:Sex:Age +
##     Class:Age:Survived, data = Titanic)
##
## Statistics:
##                       X^2 df P(> X^2)
## Likelihood Ratio 436.2715  8        0
## Pearson               NaN  8      NaN
```

As we can see this leave us with a model where the expected frequencies differ significantly from the obsevered frequencies. This we are able to assert that there is an interaction between the two variables.

Lets try another model:

```
model3 = update(sat.model, ~.-(Class:Sex:Age:Survived + Sex:Age:Survived + Class:Sex:Survived))
model3
```

```
## Call:
## loglm(formula = ~Class + Sex + Age + Survived + Class:Sex + Class:Age +
##     Sex:Age + Class:Survived + Sex:Survived + Age:Survived +
##     Class:Sex:Age + Class:Age:Survived, data = Titanic)
##
## Statistics:
##                       X^2 df     P(> X^2)
## Likelihood Ratio 76.90406  7 5.884182e-14
## Pearson               NaN  7          NaN
```

This model also has to be rejected.

Is this getting tedious - fear not, now we'll tell you the quick way…

The step() function

Here's how we automate:

```
step(sat.model, direction="backward")
```

```
## Start:  AIC=64
## ~Class * Sex * Age * Survived
##
##                          Df AIC
## - Class:Sex:Age:Survived  3  58
## <none>                       64
##
```

```
## Step:  AIC=58
## ~Class + Sex + Age + Survived + Class:Sex + Class:Age + Sex:Age +
##     Class:Survived + Sex:Survived + Age:Survived + Class:Sex:Age +
##     Class:Sex:Survived + Class:Age:Survived + Sex:Age:Survived
##
##                      Df     AIC
## - Sex:Age:Survived    1  57.685
## <none>                   58.000
## - Class:Sex:Age       3  61.783
## - Class:Age:Survived  3  89.263
## - Class:Sex:Survived  3 117.013
##
## Step:  AIC=57.69
## ~Class + Sex + Age + Survived + Class:Sex + Class:Age + Sex:Age +
##     Class:Survived + Sex:Survived + Age:Survived + Class:Sex:Age +
##     Class:Sex:Survived + Class:Age:Survived
##
##                      Df     AIC
## <none>                   57.685
## - Class:Sex:Age       3  71.953
## - Class:Age:Survived  3  95.899
## - Class:Sex:Survived  3 126.904
```

```
## Call:
## loglm(formula = ~Class + Sex + Age + Survived + Class:Sex + Class:Age +
##     Sex:Age + Class:Survived + Sex:Survived + Age:Survived +
##     Class:Sex:Age + Class:Sex:Survived + Class:Age:Survived,
##     data = Titanic, evaluate = FALSE)
##
## Statistics:
##                       X^2 df  P(> X^2)
## Likelihood Ratio 1.685479  4 0.7933536
## Pearson               NaN  4       NaN
```

The result here shows us that the most parsimonious model as indicated by the AIC (Akaike's Information Criterion). R has identifies this model by removing the interactions bewteen the four way interaction and the Sex:Age:Survived interaction.

This it appears the relationship between Sex and Survived is contioned on class…

This shows us that the relationship between Sex and Survived is conditioned on class. We can view the tables in this arrangment:

```
margin.table(Titanic, c(2,4,1))
```

```
## , , Class = 1st
##
##         Survived
## Sex       No Yes
##   Male   118  62
##   Female   4 141
##
## , , Class = 2nd
##
```

```
##         Survived
## Sex       No Yes
##   Male   154  25
##   Female  13  93
##
## , , Class = 3rd
##
##         Survived
## Sex       No Yes
##   Male   422  88
##   Female 106  90
##
## , , Class = Crew
##
##         Survived
## Sex       No Yes
##   Male   670 192
##   Female   3  20
```

The odds ratios for these tables being:

```
### odds ratio 1st class
(141/4) / (62/118)
```

```
## [1] 67.08871
```

```
### odds ratio 2nd class
(93/13) / (25/154)
```

```
## [1] 44.06769
```

```
### odds ratio 3rd class
(90/106) / (88/422)
```

```
## [1] 4.071612
```

```
### odds ratio crew
(20/3) / (192/670)
```

```
## [1] 23.26389
```

In all classes the odds of a female surviving were better than the odds of a male suriviving, although this varies significantly. **? Thus we can say that class has a meaningful effect on survival ?**

Getting more information from the model...

Lets store the model in a data object:

```r
loglm(formula = ~Class + Sex + Age + Survived + Class:Sex + Class:Age +
    Sex:Age + Class:Survived + Sex:Survived + Age:Survived +
    Class:Sex:Age + Class:Sex:Survived + Class:Age:Survived,
    data = Titanic, evaluate = FALSE) -> step.model
```

We can view the model's expected frequencies:

```r
fitted(step.model)
```

```
## Re-fitting to get fitted values

## , , Age = Child, Survived = No
##
##       Sex
## Class      Male   Female
##   1st   0.00000  0.00000
##   2nd   0.00000  0.00000
##   3rd  37.43281 14.56719
##   Crew  0.00000  0.00000
##
## , , Age = Adult, Survived = No
##
##       Sex
## Class      Male  Female
##   1st  118.0000  4.0000
##   2nd  154.0000 13.0000
##   3rd  384.5672 91.4328
##   Crew 670.0000  3.0000
##
## , , Age = Child, Survived = Yes
##
##       Sex
## Class      Male   Female
##   1st   5.00000  1.00000
##   2nd  10.98493 13.01507
##   3rd  10.56718 16.43282
##   Crew  0.00000  0.00000
##
## , , Age = Adult, Survived = Yes
##
##       Sex
## Class       Male    Female
##   1st   57.00000 140.00000
##   2nd   14.02291  79.97709
##   3rd   77.43281  73.56719
##   Crew 192.00000  20.00000
```

_We should note that our EFs do contain zeros and so there is a danger of our model being inaccurate on account of this.

We can view the model's standardised residuals:

```r
resid(step.model)
```

```
## Re-fitting to get frequencies and fitted values

## , , Age = Child, Survived = No
##
##       Sex
## Class           Male        Female
##   1st   0.000000e+00  0.000000e+00
##   2nd   0.000000e+00  0.000000e+00
##   3rd  -4.020602e-01  6.208006e-01
##   Crew  0.000000e+00  0.000000e+00
##
## , , Age = Adult, Survived = No
##
##       Sex
## Class           Male        Female
##   1st   0.000000e+00  0.000000e+00
##   2nd   0.000000e+00  0.000000e+00
##   3rd   1.239264e-01 -2.555637e-01
##   Crew  0.000000e+00  0.000000e+00
##
## , , Age = Child, Survived = Yes
##
##       Sex
## Class           Male        Female
##   1st   2.142148e-08 -4.552313e-08
##   2nd   4.546268e-03 -4.178434e-03
##   3rd   7.221252e-01 -6.159455e-01
##   Crew  0.000000e+00  0.000000e+00
##
## , , Age = Adult, Survived = Yes
##
##       Sex
## Class           Male        Female
##   1st  -5.572219e-08  3.881541e-08
##   2nd  -6.118921e-03  2.561368e-03
##   3rd  -2.779358e-01  2.820973e-01
##   Crew  0.000000e+00  0.000000e+00
```

Using the glm() function for Log Linear Modelling

To use a glm we need data frame rather than a contingency table. We can create this by doing:

```r
ti = as.data.frame(Titanic)
ti
```

```
##    Class    Sex   Age Survived Freq
## 1    1st   Male Child       No    0
## 2    2nd   Male Child       No    0
## 3    3rd   Male Child       No   35
## 4   Crew   Male Child       No    0
```

```
## 5    1st Female Child       No    0
## 6    2nd Female Child       No    0
## 7    3rd Female Child       No   17
## 8   Crew Female Child       No    0
## 9    1st   Male Adult       No  118
## 10   2nd   Male Adult       No  154
## 11   3rd   Male Adult       No  387
## 12  Crew   Male Adult       No  670
## 13   1st Female Adult       No    4
## 14   2nd Female Adult       No   13
## 15   3rd Female Adult       No   89
## 16  Crew Female Adult       No    3
## 17   1st   Male Child      Yes    5
## 18   2nd   Male Child      Yes   11
## 19   3rd   Male Child      Yes   13
## 20  Crew   Male Child      Yes    0
## 21   1st Female Child      Yes    1
## 22   2nd Female Child      Yes   13
## 23   3rd Female Child      Yes   14
## 24  Crew Female Child      Yes    0
## 25   1st   Male Adult      Yes   57
## 26   2nd   Male Adult      Yes   14
## 27   3rd   Male Adult      Yes   75
## 28  Crew   Male Adult      Yes  192
## 29   1st Female Adult      Yes  140
## 30   2nd Female Adult      Yes   80
## 31   3rd Female Adult      Yes   76
## 32  Crew Female Adult      Yes   20
```

Isn't this form of the data so much more sensible!!!

With this data we can perform the Log Linear Analysis (here on the saturated model). Then we can use an extractor function (in this case ANOVA) to identify the importance of the different interactions

```
glm.model = glm(Freq ~ Class * Age * Sex * Survived, data = ti, family = poisson)
anova(glm.model, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: Freq
##
## Terms added sequentially (first to last)
##
##
##                         Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                       31     4953.1
## Class                    3   475.81        28     4477.3 < 2.2e-16 ***
## Age                      1  2183.56        27     2293.8 < 2.2e-16 ***
## Sex                      1   768.32        26     1525.4 < 2.2e-16 ***
## Survived                 1   281.78        25     1243.7 < 2.2e-16 ***
## Class:Age                3   148.33        22     1095.3 < 2.2e-16 ***
## Class:Sex                3   412.60        19      682.7 < 2.2e-16 ***
```

```
## Age:Sex                1     6.09        18      676.6   0.01363 *
## Class:Survived         3   180.90        15      495.7 < 2.2e-16 ***
## Age:Survived           1    25.58        14      470.2 4.237e-07 ***
## Sex:Survived           1   353.58        13      116.6 < 2.2e-16 ***
## Class:Age:Sex          3     4.02        10      112.6   0.25916
## Class:Age:Survived     3    35.66         7       76.9 8.825e-08 ***
## Class:Sex:Survived     3    75.22         4        1.7 3.253e-16 ***
## Age:Sex:Survived       1     1.69         3        0.0   0.19421
## Class:Age:Sex:Survived  3     0.00         0        0.0   1.00000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the output we can see that the interactions that we want to investigate for possible elimination are: Class:Age:Sex, Age:Sex:Survived and Class:Age:Sex:Survived. This is due to the p values on these rows indivating that the addition of these factors to the model gave no significant benifit to ‘reducing deviance’ between the Efs and the observed frequencies.

```
anova(update(glm.model, .~.-(Class:Age:Sex:Survived + Age:Sex:Survived + Class:Age:Sex)), test = "Chisq
```

```
## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: Freq
##
## Terms added sequentially (first to last)
##
##
##                    Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                                 31     4953.1
## Class               3   475.81        28     4477.3 < 2.2e-16 ***
## Age                 1  2183.56        27     2293.8 < 2.2e-16 ***
## Sex                 1   768.32        26     1525.4 < 2.2e-16 ***
## Survived            1   281.78        25     1243.7 < 2.2e-16 ***
## Class:Age           3   148.33        22     1095.3 < 2.2e-16 ***
## Class:Sex           3   412.60        19      682.7 < 2.2e-16 ***
## Age:Sex             1     6.09        18      676.6   0.01363 *
## Class:Survived      3   180.90        15      495.7 < 2.2e-16 ***
## Age:Survived        1    25.58        14      470.2 4.237e-07 ***
## Sex:Survived        1   353.58        13      116.6 < 2.2e-16 ***
## Class:Age:Survived  3    29.21        10       87.4 2.024e-06 ***
## Class:Sex:Survived  3    65.43         7       22.0 4.066e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From this we can calculate the p value by taking the final Redidual Deviation and the Degrees of freedom:

```
1- pchisq(22, df=7)
```

```
## [1] 0.002540414
```

Based on the p value we can see due to it’s small size that it is not a good model (note how the step() function before retained the Class:Age:Sex interaction). We need to put one of these interactions back in - to get a model that is acceptable.

Viewing the results of a log linear analysis as a mosaic plot can also be helpful:

```
mosaicplot(Titanic, shade = T)
```