Log-linear models

The ASTA team

Contents

1 Log-linear models

1.1 Model specification

- We shall consider the data set `living`, which is a Danish survey on standard of living. The variables are
  - `B` (Housing, Bolig): bad/acceptable/good
  - `H` (Health, Helbred): bad/good
  - `I` (Isolated, Isoleret): yes/no
  - `A` (Anxiety, Angst): yes/no
- `N`: The number of respondents for each combination of the 4 factors above

- We want to order the factors such that the reference is "negative".

```
living <- read.delim("https://asta.math.aau.dk/datasets?file=living.txt", stringsAsFactors = TRUE)
```

```
living$B <- relevel(living$B,"Bad")
living$I <- relevel(living$I,"Yes")
living$A <- relevel(living$A,"Yes")
```

1.2 Example

- At first we study interaction between housing and health. So we aggregate data and only look at the association between `B` and `H` without controlling for `I` and `A`:

```
BH <- aggregate(N ~ B + H, FUN = sum, data = living)
BH
```

```
##            B    H    N
## 1        Bad  Bad  211
## 2 Acceptable  Bad  327
## 3       Good  Bad 1734
## 4        Bad Good  145
## 5 Acceptable Good  211
## 6       Good Good 1855
```

1.3 Model specification

- Like last time we can write down a model for (the logarithm of) the expected frequencies f_e by using dummy variables.
- We let z_{b1}, z_{b2} and z_{h1} denote the different levels of `B` and `H` (the reference level has all dummy variables equal to 0):

$$\log(f_e) = \alpha + \beta_{b1} z_{b1} + \beta_{b2} z_{b2} + \beta_{h1} z_{h1} + \beta_{b1h1} z_{b1} z_{h1} + \beta_{b2h1} z_{b2} z_{h1}.$$

- Note that this time we have included an interaction term, which in this case implies, that we do not assume independence between `B` and `H` in the model.
- This model contains all possible terms and there are as many parameters(6) as there are cells(6) in the table. This is called the **saturated** model.

1.4 Example

- We fit the model using `glm`:

```
model <- glm(N ~ B * H, family = poisson, data = BH)
```

- The parameter estimates (of the contrasts, i.e. differences to the reference level (`B: Bad, H: Bad`) are

```
summary(model)
```

```
##
## Call:
## glm(formula = N ~ B * H, family = poisson, data = BH)
##
## Deviance Residuals:
## [1]  0  0  0  0  0  0
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        5.35186    0.06884  77.740  < 2e-16 ***
## BAcceptable        0.43810    0.08830   4.961 7.00e-07 ***
## BGood              2.10633    0.07291  28.889  < 2e-16 ***
## HGood             -0.37512    0.10787  -3.478 0.000506 ***
## BAcceptable:HGood -0.06298    0.13940  -0.452 0.651438
## BGood:HGood        0.44258    0.11292   3.919 8.88e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance:  4.2103e+03  on 5  degrees of freedom
## Residual deviance: -2.9532e-14  on 0  degrees of freedom
## AIC: 59.485
##
## Number of Fisher Scoring iterations: 2
```

- The combination `Good:Good` increases the response, i.e. an over representation of people with good housing conditions and good health.

1.5 Model form

- Log-linear models quickly become cumbersome to write down.
- For log-linear models the structure of the model is often more interesting than the value of the parameters.
- Therefore we typically just use the model form

$$\mathrm{B} + \mathrm{H} + \mathrm{B} : \mathrm{H}$$

  where $\mathrm{B} + \mathrm{H}$ are the main effects and $\mathrm{B} : \mathrm{H}$ means that we have interaction between B and H.
- We shall stick to the **hierarchical principle**, which means that if $\mathrm{B} : \mathrm{H}$ is included then we must include $\mathrm{B} + \mathrm{H}$. For that reason we use the shorthand notation

$$\mathrm{B} * \mathrm{H} = \mathrm{B} + \mathrm{H} + \mathrm{B} : \mathrm{H}$$

1.6 Model comparison

- If we suggest a simpler model than the saturated model it will always provide a poorer fit to the given data.
- We need to find a model which is as simple (few parameters) as possible but which at the same time fits the data well.
- Last time we used the χ^2 statistic to judge whether the model $\mathrm{B} + \mathrm{H}$ (independence between B and H) was good enough compared to the saturated model $\mathrm{B} * \mathrm{H}$. This is only possible for models with two variables.

- More generally we shall consider the **Deviance of a model**, which is - approximately - given by

$$\sum \frac{(\texttt{observed} - \texttt{expected})^2}{\texttt{expected}}$$

1.7 Deviance

- We look at the output from the function `drop1`:

```
drop1(model, test = "Chisq")
```

```
## Single term deletions
##
## Model:
## N ~ B * H
##        Df Deviance    AIC    LRT  Pr(>Chi)
## <none>       0.000 59.485
## B:H     2   40.766 96.251 40.766 1.405e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- The deviance for the saturated model ($\langle$none$\rangle$) is zero as observed=expected.
- Deviance is increased by 40.8, when we remove the interaction B:H.
- Removing B:H corresponds to the null hypothesis $H_0 : \beta_{b1h1} = \beta_{b2h1} = 0$.
- The deviance is compared with a $\chi^2(Df = 2)$ distribution to determine whether it is significant.
- Since the p-value is 1.4×10^{-9} the interaction is significant and cannot be left out.

2 Higher order interaction

2.1 Higher order interaction

- We shall consider models with interaction between more that two variables.
- E.g. there may be a combined effect of adding water, fertilizer and light to a plant at the same time, i.e. the effect of adding water and fertilizer depends on whether the light is on.
- We call this a **3-way interaction**.
- We shall still respect the hierarchical principle:
  - If we include a 3-way interaction, then we must include all main effects and 2-way interactions of the 3 variables.
- Again we use short hand notation

$$\mathrm{B} * \mathrm{H} * \mathrm{A} = \mathrm{B} + \mathrm{H} + \mathrm{A} + \mathrm{B} : \mathrm{H} + \mathrm{B} : \mathrm{A} + \mathrm{H} : \mathrm{A} + \mathrm{B} : \mathrm{H} : \mathrm{A}$$

- Similar considerations hold for 4-way interactions like $\mathrm{B} * \mathrm{H} * \mathrm{A} * \mathrm{I}$, etc.

2.2 Bigger example

- We fit the saturated model to the full dataset and look at the estimated parameters (only the last half of the values are printed to save space):

```r
satmodel <- glm(N ~ B * H * A * I, family = poisson, data = living)
summary(satmodel)
```

```
##
## Call:
## glm(formula = N ~ B * H * A * I, family = poisson, data = living)
##
## Deviance Residuals:
##  [1]  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
## [24]  0
##
## Coefficients:
##                             Estimate Std. Error z value Pr(>|z|)
## (Intercept)                2.079e+00  3.536e-01   5.882 4.06e-09 ***
## BAcceptable                5.596e-01  4.432e-01   1.263 0.206710
## BGood                      1.417e+00  3.941e-01   3.596 0.000323 ***
## HGood                     -2.438e+01  4.225e+04  -0.001 0.999540
## ANo                        4.855e-01  4.494e-01   1.080 0.279943
## INo                        1.749e+00  3.831e-01   4.566 4.96e-06 ***
## BAcceptable:HGood          2.284e+01  4.225e+04   0.001 0.999569
## BGood:HGood                2.268e+01  4.225e+04   0.001 0.999572
## BAcceptable:ANo            5.349e-02  5.613e-01   0.095 0.924076
## BGood:ANo                 -3.077e-02  5.015e-01  -0.061 0.951069
## HGood:ANo                  2.343e+01  4.225e+04   0.001 0.999558
## BAcceptable:INo            6.192e-03  4.801e-01   0.013 0.989710
## BGood:INo                  5.473e-01  4.244e-01   1.290 0.197159
## HGood:INo                  2.405e+01  4.225e+04   0.001 0.999546
## ANo:INo                    6.557e-01  4.802e-01   1.365 0.172142
## BAcceptable:HGood:ANo     -2.345e+01  4.225e+04  -0.001 0.999557
## BGood:HGood:ANo           -2.254e+01  4.225e+04  -0.001 0.999574
## BAcceptable:HGood:INo     -2.303e+01  4.225e+04  -0.001 0.999565
## BGood:HGood:INo           -2.267e+01  4.225e+04  -0.001 0.999572
## BAcceptable:ANo:INo       -2.516e-01  6.007e-01  -0.419 0.675370
## BGood:ANo:INo              2.827e-01  5.329e-01   0.531 0.595707
## HGood:ANo:INo             -2.339e+01  4.225e+04  -0.001 0.999558
## BAcceptable:HGood:ANo:INo  2.365e+01  4.225e+04   0.001 0.999553
## BGood:HGood:ANo:INo        2.301e+01  4.225e+04   0.001 0.999565
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 1.0965e+04  on 23  degrees of freedom
## Residual deviance: 4.1199e-10  on  0  degrees of freedom
## AIC: 179.1
##
## Number of Fisher Scoring iterations: 20
```

- We see that the four-way interaction and all three-way interactions look like they could be left out.
- However, the p-values are related to removing one and leaving everything else in the model so we have to remove terms of the model one at a time and check against the new model.
- This can be done by successive use of `drop1` and `update` as explained in the following.

```
drop1(satmodel, test = "Chi")
```

```
## Single term deletions
##
## Model:
## N ~ B * H * A * I
##         Df Deviance    AIC    LRT Pr(>Chi)
## <none>       0.0000 179.10
## B:H:A:I  2   3.5112 178.61 3.5112   0.1728
```

- The output of `drop1` reveals that the four-way interaction is insignificant and we remove it and save the updated model like this:

```
reducedmodel <- update(satmodel, .~.-B:H:A:I)
```

- We use `drop1` again:

```
drop1(reducedmodel, test = "Chi")
```

```
## Single term deletions
##
## Model:
## N ~ B + H + A + I + B:H + B:A + H:A + B:I + H:I + A:I + B:H:A +
##     B:H:I + B:A:I + H:A:I
##        Df Deviance    AIC    LRT Pr(>Chi)
## <none>      3.5112 178.61
## B:H:A   2   7.1419 178.24 3.6307   0.1628
## B:H:I   2   3.5316 174.63 0.0205   0.9898
## B:A:I   2   5.5332 176.63 2.0220   0.3639
## H:A:I   1   4.5857 177.68 1.0745   0.2999
```

- We see `B:H:I` could be removed and use `update` again:

```
reducedmodel <- update(reducedmodel, .~.-B:H:I)
```

- We can continue this way as long as we like.
- If instead we have a specific simple model in mind, we can define this model and test it against the saturated model with `anova` as shown in the following.

2.3 Test of simple model against the saturated model

- We fit the simpler model to the full dataset:

```
simplemodel <- glm(N ~ B*H + B*A + B*I + H*A + H*I + I*A, family = poisson, data = living)
```

- We compare the two models using `anova`:

```r
anova(simplemodel, satmodel, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: N ~ B * H + B * A + B * I + H * A + H * I + I * A
## Model 2: N ~ B * H * A * I
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1         9     10.697
## 2         0      0.000  9   10.697   0.2971
```

- The deviance between the two models is 10.697 with $df = 9$, which has a p-value of 29.7%. So we prefer the simpler model without four- and 3-way interactions.

2.4 Model reduction

- Let us check whether we can make further model reductions, where we again use the function `drop1`.

```r
drop1(simplemodel, test = "Chisq")
```

```
## Single term deletions
##
## Model:
## N ~ B * H + B * A + B * I + H * A + H * I + I * A
##        Df Deviance    AIC    LRT  Pr(>Chi)
## <none>      10.697 171.79
## B:H     2   38.170 195.27 27.474 1.082e-06 ***
## B:A     2   37.912 195.01 27.215 1.231e-06 ***
## B:I     2   35.245 192.34 24.548 4.671e-06 ***
## H:A     1   41.947 201.04 31.250 2.268e-08 ***
## H:I     1   56.048 215.14 45.352 1.646e-11 ***
## A:I     1   26.449 185.54 15.753 7.218e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Conclusion: All of the pairwise interaction terms are significant.

2.5 Final model and parameter estimates

- In conclusion our final model is `simplemodel`.

```r
summary(simplemodel)
```

```
##
## Call:
## glm(formula = N ~ B * H + B * A + B * I + H * A + H * I + I *
##     A, family = poisson, data = living)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
```

```
## -1.72955  -0.48421  -0.00248   0.31663   1.23187
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        2.17042    0.22962   9.452  < 2e-16 ***
## BAcceptable        0.65323    0.26323   2.482 0.013081 *
## BGood              1.13785    0.23365   4.870 1.12e-06 ***
## HGood             -1.76785    0.21028  -8.407  < 2e-16 ***
## ANo                0.35067    0.19249   1.822 0.068488 .
## INo                1.74803    0.23116   7.562 3.97e-14 ***
## BAcceptable:HGood -0.04121    0.14119  -0.292 0.770370
## BGood:HGood        0.37773    0.11441   3.302 0.000961 ***
## BAcceptable:ANo   -0.12759    0.15834  -0.806 0.420355
## BGood:ANo          0.39413    0.13265   2.971 0.002966 **
## BAcceptable:INo   -0.14021    0.25873  -0.542 0.587878
## BGood:INo          0.72038    0.22799   3.160 0.001579 **
## HGood:ANo          0.44076    0.07938   5.552 2.82e-08 ***
## HGood:INo          1.12352    0.17982   6.248 4.16e-10 ***
## ANo:INo            0.66875    0.16277   4.109 3.98e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 10964.662  on 23  degrees of freedom
## Residual deviance:    10.697  on  9  degrees of freedom
## AIC: 171.79
##
## Number of Fisher Scoring iterations: 4
```

- We see that all significant interactions are positive - as expected(why?).
- `(Intercept)` is log of the expected number, when all factors are "bad". So the expected number is $\exp(2.17) = 8.76$, whereas the observed number is 8.

2.6 Predicted values

- What is the expected number of people without anxiety (A), with acceptable housing (B), good health (H) that are isolated (I)?

```
summary(simplemodel)
```

```
##
## Call:
## glm(formula = N ~ B * H + B * A + B * I + H * A + H * I + I *
##     A, family = poisson, data = living)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.72955  -0.48421  -0.00248   0.31663   1.23187
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)        2.17042    0.22962   9.452  < 2e-16 ***
## BAcceptable        0.65323    0.26323   2.482 0.013081 *
## BGood              1.13785    0.23365   4.870 1.12e-06 ***
## HGood             -1.76785    0.21028  -8.407  < 2e-16 ***
## ANo                0.35067    0.19249   1.822 0.068488 .
## INo                1.74803    0.23116   7.562 3.97e-14 ***
## BAcceptable:HGood -0.04121    0.14119  -0.292 0.770370
## BGood:HGood        0.37773    0.11441   3.302 0.000961 ***
## BAcceptable:ANo   -0.12759    0.15834  -0.806 0.420355
## BGood:ANo          0.39413    0.13265   2.971 0.002966 **
## BAcceptable:INo   -0.14021    0.25873  -0.542 0.587878
## BGood:INo          0.72038    0.22799   3.160 0.001579 **
## HGood:ANo          0.44076    0.07938   5.552 2.82e-08 ***
## HGood:INo          1.12352    0.17982   6.248 4.16e-10 ***
## ANo:INo            0.66875    0.16277   4.109 3.98e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 10964.662  on 23  degrees of freedom
## Residual deviance:    10.697  on  9  degrees of freedom
## AIC: 171.79
##
## Number of Fisher Scoring iterations: 4
```

- Logarithm of the expected:

$$2.17042 + 0.65323 - 1.76785 + 0.35067 - 0.04121 - 0.12759 + 0.44076 = 1.67843$$

- so the expected number is $\exp(1.67843) = 5.36$, whereas the observed number is 5.

- We can of course make **R** calculate all the table values expected by the model and add them to the data:

```
living$expected <- fitted(simplemodel)
living[1:12,]
```

```
##             B    H   I   A   N   expected
## 1         Bad Good Yes  No   5   3.300261
## 2         Bad Good  No  No 107 113.784094
## 3         Bad Good Yes Yes   0   1.495663
## 4         Bad Good  No Yes  33  26.419981
## 5         Bad  Bad Yes  No  13  12.442145
## 6         Bad  Bad  No  No 144 139.473499
## 7         Bad  Bad Yes Yes   8   8.761930
## 8         Bad  Bad  No Yes  46  50.322426
## 9  Acceptable Good Yes  No   5   5.357140
## 10 Acceptable Good  No  No 155 160.536219
## 11 Acceptable Good Yes Yes   3   2.758237
## 12 Acceptable Good  No Yes  48  42.348403
```

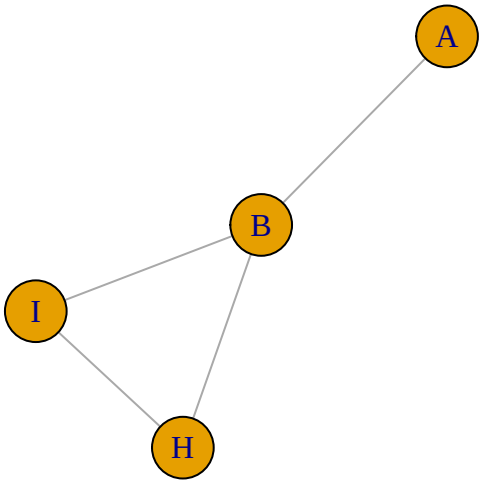
3 Graphical representation

3.1 Graphical representation

- We make a graphical representation by
  - drawing a circle for each variable.
  - connecting variables which enter the same model term.
- Example: Assume the model is

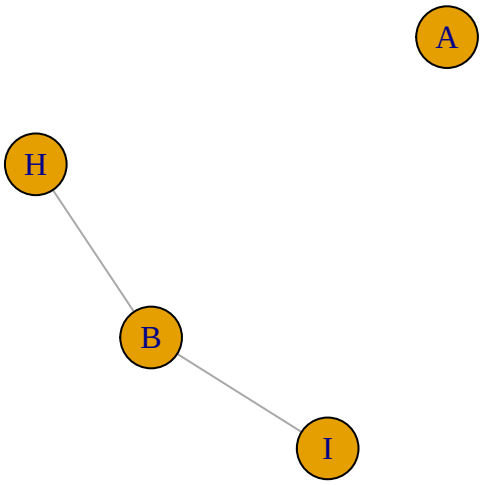
$$A * B + B * H * I$$

- Then the graphical representation is

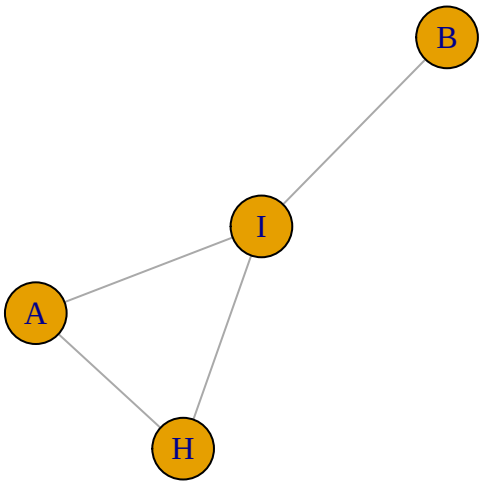


3.2 Interpretation of graphical model

- **Independence**: If A enters in the model formula, but A doesn't enter in any other terms (e.g. $A * B$, $A * H$, etc.), then A is independent of the other variables.
- E.g. $A + B * H + B * I$

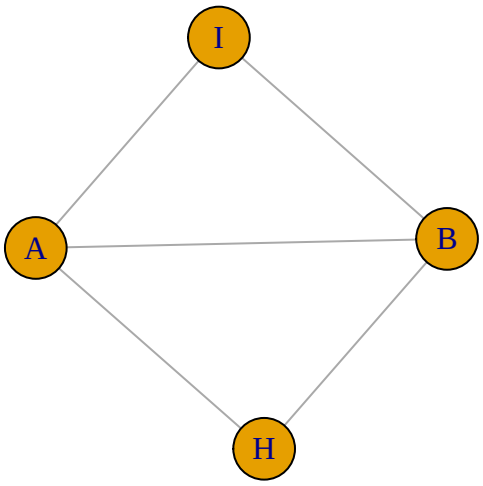


- **Explained association**. If B and H are "connected" via other terms, but don't enter the same term, then the association is explained by other variables. I.e. the model cannot include e.g. $B * H$, $B * H * A$ or $A * B * H * I$.
- E.g. $B * I + A * H * I$

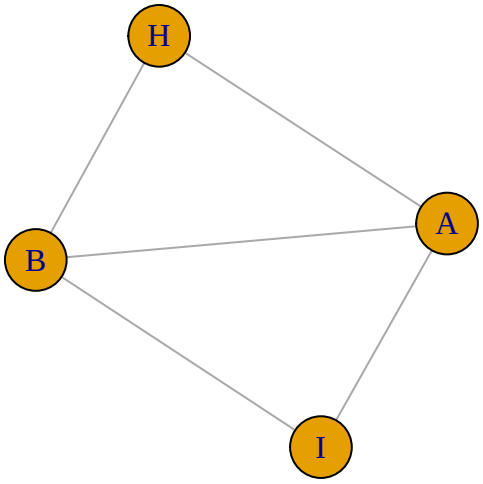


3.3 Interpretation of graphical model

- **Homogeneous association**: If $A * H$ enters the model, but $A * H$ doesn't enter more complicated terms, then the association between A and H is homogeneous.
- I.e. the model cannot contain $A * H * I$, $A * B * H$ or $A * B * H * I$.
- E.g. $A * H + A * I * B + B * H$



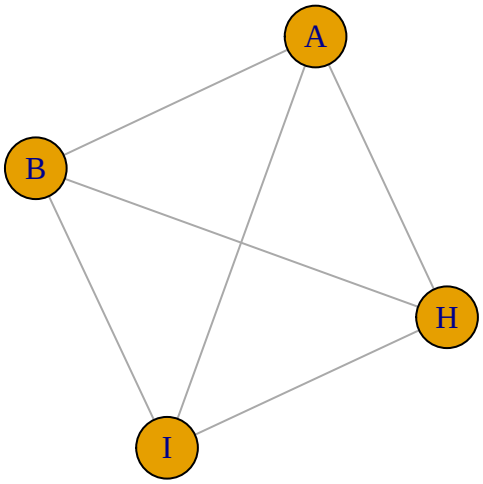
- **Heterogeneous association**: If $A * H$ enter the model as a part of a more complicated term, then the association between A and H is heterogeneous.
- I.e. the model *must* contain $A * H * I$, $A * B * H$ or $A * B * H * I$.
- E.g. $A * B * H + A * I * B$



3.4 Final model of the example

- In the example the final model was:

$$B * I + H * I + I * A + B * H + B * A + H * A$$



- We can directly see from the graph, that:
  - we don't have any independent variables since all variables are connected
  - we do not have an explained association.
- From the formula we have homogeneous association between all pairs of variables.