Log Linear Models in S-Plus

1 Example

The following data show are from a $2^4$ contingency table, with two levels of each of the following variables: 
$S =$ smoking status during pre-natal period, 1= non-smoker, 2=smoker; 
$D =$ one-year survival, 1=dead, 2=live at 1-year; 
$A =$ age of mother, 1=less than 30, 2=older than 30; 
$G =$ length of gestation, 1 = between 197-260 days, 2 = 260+ days.

\[
\begin{align*}
\text{freq} & \leftarrow c(50, 315, 24, 4012, 9, 40, 6, 459, 41, 147, 14, 1594, 4, 11, 1, 124) \\
\text{A} & \leftarrow c(\text{rep}(1, 8), \text{rep}(2, 8)) \\
\text{S} & \leftarrow c(\text{rep}(1, 4), \text{rep}(2, 4), \text{rep}(1, 4), \text{rep}(2, 4)) \\
\text{D} & \leftarrow \text{rep}(c(1, 2), 8) \\
\text{G} & \leftarrow \text{rep}(c(1, 1, 2, 2), 4) \\
\text{surv} & \leftarrow \text{data.frame}(\text{freq}, A=\text{as.factor(A)}, S=\text{as.factor(S)}, D=\text{as.factor(D)}, G=\text{as.factor(G)})
\end{align*}
\]

> surv
  freq A S D G
  1 50 1 1 1 1
  2 315 1 1 2 1
  3 24 1 1 1 2
  4 4012 1 1 2 2
  5 9 1 2 1 1
  6 40 1 2 2 1
  7 6 1 2 1 2
  8 459 1 2 2 2
  9 41 2 1 1 1
 10 147 2 1 2 1
 11 14 2 1 1 2
 12 1594 2 1 2 2
 13 4 2 2 1 1
 14 11 2 2 2 1
 15 1 2 2 1 2
 16 124 2 2 2 2

2 Log Linear Models for Contingency Tables

We are interested in how infant survival after one year and length of gestation depend on the other variables (smoking status, mother’s age). In this case we have 2 response or dependent variables ($D$ and $G$) and two explanatory variables ($S$ and $A$).

Models for the data can be fit using Poisson regression in S-Plus with \text{freq} (the counts) as the response, and all other variables coded as factors. For product-multinomial sampling, the Poisson model makes sense as long as we condition on the marginal totals for the variables that are fixed so that these variables correspond to explanatory variables. A sensible approach for models with 2 or more explanatory variables, is to also condition on the values of these factors, treating them as fixed (and marginal totals as fixed) even in cases when they are not. More on this later in the example...

2.1 Model Parameterization

There are several possible codings that lead to different parameter estimates, but equivalent deviances, fitted values, etc. The models are typically over-parameterized as in ANOVA models. The default coding
in S-Plus is using Helmert contrasts, contr.helmert. These are orthogonal contrasts, but may have little interpretability in log-linear models. The usual ANOVA contrasts are the restriction that the estimates sum to zero, contr.sum. The third, which is really not a contrast, is the treatment effects, contr.treatment, where the first coefficient is set to zero. Coefficients then have the interpretation as odds ratios. This is probably the most useful parameterization. To change the default, use

```r
options(contrasts=c("contr.treatment", "contr.poly"))
```

The first term specifies the type of contrasts used for factors.

### 2.2 Model Specification

The saturated model is specified by `D*G*S*A`:

```r
> glm(freq ~ D*G*S*A, family=poisson, dat=surv)
```

Call:
```
glm(formula = freq ~ D * G * S * A, family = poisson, data = surv)
```

Coefficients:

```
             (Intercept)          D           G          S          A          D:G          D:S
3.912023     1.840558 -0.7339692   -1.7147988 -0.1984509   3.278442    -0.3488948

0.3285041   -0.5636891 -0.3405456   -0.6124793   -0.4328058   0.1796423   0.08363515

             G:S:A          D:G:S:A
-0.6402837   0.7834025

Degrees of Freedom: 16 Total; 0 Residual
Residual Deviance: 6.545875e-13
```

Since we are treating S and A as explanatory variables, we are not particularly interested in whether they are dependent or not. To account for any (possible) dependence between, we will include all of their interactions in any model. So for example to explore the model of independence between D, G and the explanatory variables, we should allow for dependence between S and A, and fit the model `D + G + S*A`

```r
> glm(freq ~ D+ G + S*A, family=poisson, dat=surv)
```

Call:
```
glm(formula = freq ~ D + G + S * A, family = poisson, data = surv)
```

Coefficients:

```
             (Intercept)          D           G          S          A          S:A
2.15411     3.806208  2.312904    -2.147364   -0.8962698   -0.40431

Degrees of Freedom: 16 Total; 10 Residual
Residual Deviance: 360.1792
```

### 2.3 Model Selection

Model Selection may be based on the change in deviance, change in BIC, Bayes Factors, etc. A problem with BIC is deciding what is the correct value for N; is it the total number of counts or the total number of cells? From the change in deviance, it is clear that the model of independence does not fit. What other models provide reasonable fits? Try Forward or Backward Selection or all possible models, making sure that all interactions between S and A are included.
3 Drop in Deviance Tests

Consider models that include all A by S associations. To start, we will fit models with all main effects, then all two-way associations, all three-way associations and the four-way interaction or saturated model.

> main.glm = glm(freq ~ D + G + A*S, data=surv, family=poisson)
> main.glm
Call:
  glm(formula = freq ~ D + G + A * S, family = poisson, data = surv)

Coefficients:
(Intercept)         D         G         A         S         A:S
    2.15411  3.806208  2.312904 -0.8962698 -2.147364 -0.404311

Degrees of Freedom: 16 Total; 10 Residual
Residual Deviance: 360.1792

# model with all two-way interactions
> two.glm = glm(freq ~ D*G + D*A + D*S + G*A + G*S + A*S, data=surv, family=poisson)
> two.glm
Call:
  glm(formula = freq ~ D * G + D * A + D * S + G * A + G * S + A * S, family =
       poisson, data = surv)

Coefficients:
(Intercept)         D         G         A         S         D:G         D:A
    3.94097  1.811734 -0.7654219 -0.2914192 -1.698707  3.309402  -0.4638511
     D:S         G:A         G:S         A:S
    -0.4144251 -0.1671759 -0.04726655 -0.4119419

Degrees of Freedom: 16 Total; 5 Residual
Residual Deviance: 1.722536

# Model with all three-way interactions
> three.glm = glm(freq ~ D*G*A + D*G*S + D*A*S + G*A*S , data=surv, family=poisson)
> three.glm
Call:
  glm(formula = freq ~ D * G * A + D * G * S + D * A * S + G * A * S, family =
       poisson, data = surv)

Coefficients:
(Intercept)         D         G         A         S         D:G         D:A
    3.902543  1.851526 -0.7050239 -0.1775308 -1.65423 3.247882 -0.5893201
   -0.4152084 -0.4228238 0.1665819 -0.8099804 0.2594295 -0.2563785 0.339703
    G:A:S
    0.07929968

Degrees of Freedom: 16 Total; 1 Residual
Residual Deviance: 0.3593495
# saturated model
> four.glm   glm(freq ~ D*G*A*S, data=surv, family=poisson)

### 3.1 `anova()`

The `anova()` command can be used to perform the drop in deviance test for GLMs (likelihood ratio test). The `anova()` function takes a series of (nested) GLM models and computes the change in deviance between the models for testing whether the additional parameters are all equal to 0. The change in deviance has a Chi-squared distribution under the null hypothesis that degrees of freedom equal to the number of zero parameters under $H_0$, which is the change in residual degrees of freedom.

> anova(main.glm, two.glm, three.glm, four.glm, test="Chisq")

**Analysis of Deviance Table**

**Response: freq**

<table>
<thead>
<tr>
<th>Terms</th>
<th>Resid. Df</th>
<th>Resid. Dev</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>D + G + A * S</td>
<td>10 360.1792</td>
</tr>
<tr>
<td>2</td>
<td>D * G + D * A + D * S + G * A + G * S + A * S</td>
<td>5 1.7225</td>
</tr>
<tr>
<td>3</td>
<td>D * G * A + D * G * S + D * A * S + G * A * S</td>
<td>1 0.3593</td>
</tr>
<tr>
<td>4</td>
<td>D * G * A * S</td>
<td>0 0.0000</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test</th>
<th>Df</th>
<th>Deviance</th>
<th>Pr(Chi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>+D:G+D:A+D:S+G:A+G:S</td>
<td>5 358.4567</td>
<td>0.000000</td>
</tr>
<tr>
<td>3</td>
<td>+D:G:A+D:G:S+D:A:S+G:A:S</td>
<td>4 1.3632</td>
<td>0.8505675</td>
</tr>
<tr>
<td>4</td>
<td>+D:G:A:S</td>
<td>1 0.3593</td>
<td>0.5488677</td>
</tr>
</tbody>
</table>
From this we can concluded that we do not need to include the 4-way interaction or 3-way interactions. This means that the association between any two variables does not depend on the level of any of the other variables, and the two-factor interaction terms are log-odds ratios which measure the association. There do appear to be significant 2-way interactions beyond the A:S that is included in the independence model, so that we cannot conclude that survival and gestational length are independent given age and smoking status. Can we simply the model of two-way interaction? Look at Wald tests (Z-score based on estimate/standard error).

> summary(two(glm, cor=F))

Call: glm(formula = freq ~ D * G + D * A + D * S + G * A + G * S + A * S, family = poisson, data = surv)

Deviance Residuals:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.2056804</td>
<td>-0.002328818</td>
<td>0.01226882</td>
<td>0.02289173</td>
<td>-0.1361385</td>
<td>0.3093063</td>
<td>0.8366963</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.1551927</td>
<td>0.4056068</td>
<td>-0.08498143</td>
<td>-0.2952936</td>
<td>-0.00940606</td>
<td>-0.3131805</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>14</th>
<th>15</th>
<th>16</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.2500742</td>
<td>-0.6163622</td>
<td>0.2040947</td>
</tr>
</tbody>
</table>

Coefficients:

<table>
<thead>
<tr>
<th></th>
<th>Value</th>
<th>Std. Error</th>
<th>t value</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>3.94097028</td>
<td>0.12653287</td>
<td>31.1458226</td>
</tr>
<tr>
<td>D</td>
<td>1.81173357</td>
<td>0.13477441</td>
<td>13.4427121</td>
</tr>
<tr>
<td>G</td>
<td>-0.76542185</td>
<td>0.18366454</td>
<td>-4.1674995</td>
</tr>
<tr>
<td>A</td>
<td>-0.29141924</td>
<td>0.17034099</td>
<td>-1.7107992</td>
</tr>
<tr>
<td>S</td>
<td>1.69870684</td>
<td>0.24718385</td>
<td>6.8722404</td>
</tr>
<tr>
<td>D:G</td>
<td>3.30940172</td>
<td>0.18460710</td>
<td>17.9267303</td>
</tr>
<tr>
<td>D:A</td>
<td>-0.46385106</td>
<td>0.18004287</td>
<td>-2.5763367</td>
</tr>
<tr>
<td>D:S</td>
<td>-0.41442507</td>
<td>0.26168637</td>
<td>-1.5836708</td>
</tr>
<tr>
<td>G:A</td>
<td>-0.16717589</td>
<td>0.09613036</td>
<td>-1.7390542</td>
</tr>
<tr>
<td>G:S</td>
<td>-0.04726655</td>
<td>0.14907303</td>
<td>-0.3170698</td>
</tr>
<tr>
<td>A:S</td>
<td>-0.41194190</td>
<td>0.09951650</td>
<td>-4.1394334</td>
</tr>
</tbody>
</table>

(Dispersion Parameter for Poisson family taken to be 1 )

Null Deviance: 20311.07 on 15 degrees of freedom

Residual Deviance: 1.722536 on 5 degrees of freedom

Number of Fisher Scoring Iterations: 3

The Wald test implies that we can drop the G*S term (Z=-0.317, p-value=0.75), leading to the model [DG][DA][DS][GA][AS].
We can refit the model without the G*S term and continue with Wald tests and backward's elimination.

> two1.glm <- glm(freq ~ D*G + D*A + D*S + G*A + A*S, data=surv, family=poisson)

Call: glm(formula = freq ~ D * G + D * A + D * S + G * A + A * S, family = poisson, 
                 data = surv)

Deviance Residuals:
     Min          1Q  Median          3Q         Max
  -0.6526467  -0.2214703  -0.09177418  0.07485582  0.775331

Coefficients:
                     Value  Std. Error   t value
(Intercept)   3.9431194  0.12623457  31.236446
         D   1.8142233  0.13341728  13.496950
          G    -0.7723702  0.18236353  -4.235333
          A    -0.2919894  0.17032236  -1.714334
         S    -1.7131343  0.24305599   -7.048323
     D:G   3.3113469  0.18451525  17.946196
     D:A  -0.4648119  0.18002355  -2.581951
     D:S  -0.4437539  0.24467019  -1.813682
     G:A  -0.1655720  0.09599314  -1.724832
     A:S  -0.4113172  0.09949609  -4.134045

(Dispersion Parameter for Poisson family taken to be 1 )

Null Deviance: 20311.07 on 15 degrees of freedom
Residual Deviance: 1.822126 on 6 degrees of freedom

Number of Fisher Scoring Iterations: 3

For the G:A term, the Z-score is -1.72, while for the term D:S, the Z-score of -1.81, which are significant at α = 0.10, but not at α = 0.05. While not statistically significant, this may be practically or scientifically significant.

The coefficient for D:S is a log odds-ratio. For mothers who smoke (S=2), the odds of their infant dying after one year are $1.55 = \exp(.4475)$ times higher than the odds of dying for infants with mothers that did not smoke. This holds regardless of the age of the mother or gestational length. Verify from the fitted table:

> cbind(fitted(two1.glm), surv)

<table>
<thead>
<tr>
<th>fitted freq</th>
<th>A</th>
<th>S</th>
<th>D</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td>51.579245</td>
<td>50</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>316.506155</td>
<td>315</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>23.825327</td>
<td>24</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>4009.089272</td>
<td>4012</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>9.299727</td>
<td>9</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>36.614873</td>
<td>40</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>4.295701</td>
<td>6</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>463.789699</td>
<td>459</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>38.518166</td>
<td>41</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>148.493524</td>
<td>147</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>15.077262</td>
<td>14</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1593.911048</td>
<td>1594</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>4.602863</td>
<td>4</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>11.385448</td>
<td>11</td>
<td>2</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1.801710</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>122.209980</td>
<td>124</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>
odds-ratio $G = 1$, $A=1$ \(\frac{9.299727/36.61487}{51.57925/316.5062} = 1.5585\)

odds-ratio $G = 2$, $A=1$ \(\frac{4.295701/463.789699}{23.825327/4009.089272}\)

odds-ratio $G = 1$, $A=2$ \(\frac{4.602863/11.385448}{38.518166/148.493524}\)

odds-ratio $G = 2$, $A=2$ \(\frac{1.801710/122.209980}{15.077262/1593.911048}\)