Two-Factor Log-Linear Model Example

May 14, 2020

0.1 Heights

0.1.1 Table 2 (Case 2)

This is a case of two response factors with a fixed total, n, of 205.

The probability density function is therefore:

$$f(\mathbf{y}|n) = \frac{f(\mathbf{y}, n)}{f(n)} = \frac{f(\mathbf{y})}{f(n)}$$

Which is a multinomial. We can model a multinomial via a Poisson distribution so long as we provide the fields which make up the fixed totals as part of the model.

```
In [2]: frequency <- c(18,20,12,
                       28,51,25,
                       14,28,9)

In [7]: husband <- factor(rep(c('T','M','S'),3))

In [20]: wife <- factor(c(rep('T',3),rep('M',3),rep('S',3)))

In [22]: married <- data.frame(frequency, husband, wife)

In [24]: rm(frequency, husband, wife)
         attach(married)
```

0.2 Fitting GLM

(NB: If you fit an interaction model here you will get the saturated model, and the deviance will be zero).

The null hypothesis here is of **independence**. That male height is independent of female height.

Under the null hypothesis, we don't have interaction terms (representing the dependence between males and females). A model with these terms **would be the saturated model**.

Therefore our deviance has meaning: if our independence model **requires more terms** with respect to the saturated model, then those additional terms **will be interaction terms**! *THAT'S CLEVER*!

The expected value of a Multinomial distributed random variable Y_{jk} is:

$$E[Y_{jk}] = n\theta_{jk},$$

where:

- n is the **fixed** total number of counts in the sample;
- θ_{jk} is the cell probability: the probability that a random observation will fall in the $j-k$th cell.

Our independence null hypothesis means that:

$$\theta_{jk} = \theta_{j.} \times \theta_{.k},$$

that is, the probability of landing in the $j-k$th cell is the probability of landing in the jth row multiplied by the probability of landing in the kth row.

This means that, **under the null hypothesis of independence**,

$$E[Y_{jk}] = n\theta_{jk} = n \times \theta_{j.} \times \theta_{.k}$$

Applying the **log link** function to this expected value yields the linear predictor, η_{jk}:

$$g(\mu_{jk}) = \ln(\mu_{jk}) = \eta_{jk}$$

$$\ln(E[Y_{jk}]) = \ln(\mu_{jk}) = \eta_{jk} = \ln(n \times \theta_{j.} \times \theta_{.k}) = \ln(n) + \ln(\theta_{j.}) + \ln(\theta_{.k}),$$

which can be reparameterised as:

$$\eta_{jk} = \mu + \alpha_j + \beta_k,$$

i.e. an interaction-less two-factor linear predictor.

The alternative hypothesis is that more parameters need to be added to the linear predictor, and thus that those parameters will represent **interaction** parameters, $(\alpha\beta)_{jk}$, where the full interaction linear predictor has the form:

$$\eta_{jk} = \mu + \alpha_j + \beta_k + (\alpha\beta)_{jk}$$

0.2.1 We're going to fit a Poisson distributed GLM to random variable that is actually Multinomial.

Important note: **For the Poisson distribution in generalised exponential form: the ϕ parameter = 1, therefore the scaled deviance is equal to the residual deviance in the R output below.**

```
In [29]: married.glm <- glm(frequency ~ husband +  wife, family = poisson(link='log'))

         summary(married.glm)
```

```
Call:
glm(formula = frequency ~ husband + wife, family = poisson(link = "log"))

Deviance Residuals:
      1        2        3        4        5        6        7        8
 0.8490  -0.8699   0.2304  -0.4482   0.1092   0.3404  -0.2424   0.6645
```

```
      9
-0.7508

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   3.9165     0.1218  32.152  < 2e-16 ***
husbandS     -0.7665     0.1784  -4.295 1.74e-05 ***
husbandT     -0.5008     0.1636  -3.061  0.00221 **
wifeS        -0.7126     0.1709  -4.168 3.07e-05 ***
wifeT        -0.7324     0.1721  -4.256 2.08e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 50.5890  on 8  degrees of freedom
Residual deviance:  2.9232  on 4  degrees of freedom
AIC: 56.57

Number of Fisher Scoring iterations: 4
```

```
In [30]: S <- 2.9232

         n <- 9

         # have fit 5 parameters, an intercept and two of the three levels for each of the two
         #ărecall the corner condition that R imposes to prevent overparameterisation in the l
         # alpha_1 = 0, beta_1 =  0
         p <- 5

         pchisq(S, n-p, lower.tail=FALSE)
```

0.570758868723671

0.2.2 Summary of the null hypothesis test

With a p-value of 0.57 associated with the scaled deviance of 2.9232, we have insufficient evidence to reject the null hypothesis that we have a sufficient number of parameters in our model.

We therefore do not need to add any additional parameters from the saturated model that are not already in the current model - **these additional parameters are interaction parameters**.

Therefore our analysis suggests that married couples find each other independently of their heights.

Note, the (scaled) deviance required to suggest that we'd need interaction terms, and thus that husbands and wives heights are not independent of each other would have been 9.49.

```
In [31]: qchisq(0.05, n-p, lower.tail = FALSE)
```

9.48772903678115

0.3 Cool extras

Because we included the factors which have the total count constraints on them, this is baked into the intercept.

If we sum the fitted values, we'll get the total count from the underlying data

```
In [33]: sum(fitted(married.glm))
```

205

```
In [ ]:
```