Notes on Log-Linear Models

Robin Evans
evans@stats.ox.ac.uk

Michaelmas 2020

This version: August 27, 2020

These notes are a supplement to the Graphical Models notes, designed to help bridge the gap between the material in Part A Statistics and the analyses of contingency tables. This material is not examinable (except insofar as it may appear in the lecture notes or problem sheets), but is to help explain why we use the Poisson GLM when fitting graphical models on contingency tables.

As ever, if you find any mistakes or omissions, I'd be very grateful to be informed.

1 Introduction

Suppose we have a collection of i.i.d. pairs $(X^{(i)}, Y^{(i)})$ for $i = 1, \dots, n$, where $X^{(i)}$ and $Y^{(i)}$ are categorical variables with d_1 and d_2 levels respectively. (Here we choose notation to be consistent with the Graphical Models lecture notes.) Since there are only $d_1 d_2$ distinct possible values, it makes sense to summarize our data with a **two-way contingency table**, in which we have counts

$$n_{xy} = \sum_{i=1}^{n} \mathbb{1}\{X^{(i)} = x, Y^{(i)} = y\}.$$

An example is shown below:

Death Penalty?	Defendant's Race	
	White	Black
Yes	53	15
No	430	176

This table classifies 674 individuals sentenced to death in Florida during the 1970s. Note that $\sum_{x,y} n_{xy} = n = 674$, because each observation is represented in exactly one cell.

We can model this as a multinomal distribution via

$$\boldsymbol{n} = (n_{11}, n_{21}, \dots, n_{d_1 d_2}) \sim \text{Multinom}(n, \boldsymbol{\pi}),$$

where $\boldsymbol{\pi} = (\pi_{11}, \dots, \pi_{d_1 d_2})$ is a vector of unknown probabilities. This has likelihood

$$L(\boldsymbol{\pi}; \boldsymbol{n}) = \binom{n}{n_{11} \dots n_{d_1 d_2}} \prod_{x=1}^{d_1} \prod_{y=1}^{d_2} \pi_{xy}^{n_{xy}} \qquad \sum_{xy} \pi_{xy} = 1, \pi_{xy} \geq 0$$

and log-likelihood

$$l(\boldsymbol{\pi}; \boldsymbol{n}) = \sum_{x=1}^{d_1} \sum_{y=1}^{d_2} n_{xy} \log \pi_{xy}.$$

To maximize the likelihood with respect to $\boldsymbol{\pi}$, we can just add a Lagrange multiplier (to ensure that the probabilities will sum to 1) and differentiate to obtain that the MLEs are

$$\hat{\pi}_{xy} = \frac{n_{xy}}{n}.$$

You will also have seen that, if X and Y are assumed to be independent, (i.e. the X-category an observation is in gives no information about which Y-category it is in, and vice versa), then (using the definition of ordinary independence) this corresponds to $\pi_{xy} = \alpha_x \beta_y$ for some vectors $\boldsymbol{\alpha}, \boldsymbol{\beta}$;

$$\alpha_x = \sum_y \pi_{xy}, \qquad \beta_y = \sum_x \pi_{xy}.$$

By replacing π_{xy} with $\alpha_x \beta_y$ in the log-likelihood above and maximizing with respect to these quantities, we obtain that the MLE under the assumption of independence is

$$\hat{\alpha}_x = \frac{\sum_y n_{xy}}{n}, \qquad \hat{\beta}_y = \frac{\sum_x n_{xy}}{n}. \tag{1}$$

This gives $\hat{\pi}_{xy} = \frac{n_{x+} \cdot n_{+y}}{n^2}$, where n_{x+} is the sum of the xth row of the contingency table $\boldsymbol{n}$, and similarly for n_{+y}.

1.1 Fitting an Independence Model

Using the formula (1) above applied to the death penalty table, we obtain

```
> dat <- read.table("deathpen.txt", header=TRUE)
> nxyz <- table(dat[,1:3])
> nxyz[] <- rev(dat[,4])
> nxyz  # our data

, , Victim = Black

        Defendant
DeathPen Black White
     No    139    16
     Yes     4     0

, , Victim = White

        Defendant
DeathPen Black White
     No     37   414
     Yes    11    53
```

```
> n <- sum(nxyz) # total observations
> nxy <- margin.table(nxyz, 1:2)
> nxy  # two-way table we're interested in

        Defendant
DeathPen Black White
     No    176   430
     Yes    15    53

> (alpha_hat <- margin.table(nxy, 1)/n)

DeathPen
 No Yes
0.9 0.1

> (beta_hat <- margin.table(nxy, 2)/n)

Defendant
Black White
 0.28  0.72

> alpha_hat %*% t(beta_hat)

        Defendant
DeathPen Black White
     No  0.255 0.644
     Yes 0.029 0.072
```

Is this a good fit? We can check with a likelihood ratio test, comparing the fit of the independence model with the fit of the saturated model (i.e. the model with no restriction on the probability distribution). This gives

$$\begin{aligned}\Lambda &= 2\left\{l(\hat{\boldsymbol{\pi}};\boldsymbol{n}) - l(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\beta}};\boldsymbol{n})\right\}\\ &= 2\sum_{x,y} n_{xy}\log\frac{n_{xy}n}{n_{x+}n_{+y}}\end{aligned}$$

(the first equation is just the definition of the likelihood ratio, the second you should derive for yourself). Since the expected number of observations in each cell is $e_{xy} \equiv \mathbb{E}n_{xy} = n\pi_{xy}$, this is sometimes written as

$$\Lambda = 2\sum_{x,y} n_{xy}\log\frac{n_{xy}}{e_{xy}},$$

where $e_{xy} = n_{x+}n_{+y}/n$ (in the case of the independence model). We can also use a Pearson's χ^2-test, in which the test statistic is

$$\mathrm{X}^2 = \sum_{x,y}\frac{(n_{xy}-e_{xy})^2}{e_{xy}}.$$

This is approximately the same as Λ, as can be seen by a Taylor expansion; see Sections 1.5.4 and 1.5.5 of Agresti (2002).

In our case, the values are indeed quite similar.

```
> exy <- n*alpha_hat %*% t(beta_hat)
> (Lambda <- 2*sum(nxy*log(nxy/exy)))

[1] 1.536

> (X2 <- sum((nxy-exy)^2/exy))

[1] 1.469
```

To test goodness of fit, we compare the likelihood ratio statistic to a χ^2-distribution with 1 degree of freedom, since this is the number of additional parameters in the larger model (in general this is given by $(d_1 - 1)(d_2 - 1)$, as shown in Part A Stats).

Exercise. (See also Worksheet 0.) Suppose we have a three-way contingency table (n_{xyz}) with $d_1 d_2 d_3$ levels. Show that the MLE for $\boldsymbol{\pi}$ under the **conditional independence** model $X \perp\!\!\!\perp Y \mid Z$ is given by

$$\pi_{xyz} = \frac{n_{x+z} \cdot n_{+yz}}{n^2}.$$

[Hint: this model is the same as setting $\pi_{xyz} = \gamma_z \alpha_{x|z} \beta_{x|z}$ for suitably defined $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}$.]

2 Log-Linear Models

An alternative way to fit the model above is to use a Poisson GLM:

```
> glm1 <- glm(Freq ~ DeathPen + Defendant, data=as.data.frame(nxy),
+             family=poisson)
> summary(glm1)


Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     5.14592    0.07350   70.01   <2e-16 ***
DeathPenYes    -2.18737    0.12789  -17.10   <2e-16 ***
DefendantWhite  0.92774    0.08548   10.85   <2e-16 ***

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 625.8487  on 3  degrees of freedom
Residual deviance:   1.5362  on 1  degrees of freedom
```

Notice that the **residual deviance** is exactly the value of the likelihood ratio calculated above. Why does this work? Well, suppose that $n_{xy} \sim \text{Poisson}(\mu_{xy})$ independently; then

(easy exercise) we have

$$n = \sum n_{xy} \sim \text{Poisson}(\mu)$$

where $\mu = \sum_{xy} \mu_{xy}$, and

$$\boldsymbol{n} \mid n \sim \text{Multinom}(n, \boldsymbol{\pi}),$$

where $\pi_{xy} = \mu_{xy}/\mu$; in other words, if we *fix* the total count, then the Poisson model reduces to the multinomial model we used earlier.

It follows that the two models are interchangeable, in the sense that maximizing the log-likelihood of one model is the same as maximizing that of the other, except that in the Poisson model we have the additional parameter μ whose MLE is the total count $\hat{\mu} = n$.

2.1 The Independence Model

Note that, if $\pi_{xy} = \alpha_x \beta_y$ then (assuming the probabilities are all positive) we have

$$\begin{aligned} \log \pi_{xy} &= \log \alpha_x + \log \beta_y \\ \log \mu_{xy} &= \log \mu + \log \alpha_x + \log \beta_y. \end{aligned} \tag{2}$$

This parameterization is not identifiable unless we restrict $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in some way, since otherwise we could multiply α by m and divide μ by m, and then

$$\begin{aligned} \log \mu_{xy} &= \log \frac{\mu}{m} + \log m\alpha_x + \log \beta_y \\ &= \log \mu + \log \alpha_x + \log \beta_y, \end{aligned}$$

leaving the model unchanged. [Of course in the multinomial setting we usually constrain $\sum_x \alpha_x = 1$, but this is not so convenient for the Poisson model as it a restriction on the exponents of the parameters.]

We generally write (2) as

$$\log \mu_{xy} = \lambda_\emptyset + \lambda_X(x) + \lambda_Y(y),$$

and specify that $\lambda_X(1) = \lambda_Y(1) = 0$ to give the identifiability. In the GLM output above, we see that:

$$\hat{\lambda}_\emptyset = 5.146, \qquad \hat{\lambda}_X(2) = -2.187, \qquad \hat{\lambda}_Y(2) = 0.928.$$

This fits MLEs for the true expected counts as

$$\begin{aligned} e^{5.146} &= 171.73 & e^{5.146-2.187} &= 19.27 \\ e^{5.146+0.928} &= 434.27 & e^{5.146-2.187+0.928} &= 48.73; \end{aligned}$$

when normalised these give the same probabilities as we calculated in Section 1 (i.e. 0.255, 0.029, 0.644, 0.072).

3 General Graphical Models

In general, if we want to fit a graphical model with cliques $C_1, \ldots, C_k$, we just need to specify that we want those sets as interactions in a hierarchical model. We can try this with the simulated data from lectures used to illustrate the IPF algorithm. Those were:

```
> set.seed(124)
> dat <- c(rmultinom(1, size=96, prob=rexp(16)))
> dat

 [1]  9  9  0  8  6  4  4  3 22  0  2  6  5  3 10  5
```

A bit of formatting is needed to have this as a data frame (don't worry about exactly what this is doing):

```
> dim(dat) <- rep(2,4)
> dimnames(dat) <- list(x1=0:1, x2=0:1, x3=0:1, x4=0:1)
> dat <- as.data.frame(as.table(dat))
```

Now we can apply the `glm()` function as before:

```
> glm2 <- glm(Freq ~ x1*x2 + x2*x3 + x3*x4 + x1*x4, data=dat,
+             family=poisson)
> glm2$fitted.values   # compare to the IPF output in the slides

     1      2      3      4      5      6      7      8      9     10     11
10.067  7.409  2.294  6.230  3.874  2.851  2.765  7.510 18.698  3.825  4.260
    12     13     14     15     16
 3.216  9.360  1.915  6.681  5.044

> glm2$deviance  # compared to 7 degrees of freedom, model is not a good fit!

[1] 26.91
```

There are 8 parameters in this model (count the non-empty complete subsets of vertices), compared with 15 in the saturate model. The deviance is therefore large (26.907 on $15 - 8 = 7$ degrees of freedom), and the model is not a good fit.