Log-linear Models for Contingency Tables

Statistics 149

Spring 2006

Log-linear Models for Two-way Contingency Tables

Example: Business Administration Majors and Gender

A study of the career plans of young men and women sent questionaires to all 722 members of the senior class in the College of Business Administration at the University of Illinois. One question asked which major within the business program the student had chosen.

Major	Women	Men
Accounting	68	56
Administration	91	40
Economics	5	6
Finance	61	59

Model for the data

- The joint PDF of (X, Y): $P[X = x_i, Y = y_i] = \pi_{ij}$
- Marginal PDF of X: $P[X = x_i] = \pi_{i+}$
- Marginal PDF of Y: $P[Y = Y_j] = \pi_{+j}$
- Expected cell counts: $\mu_{ij} = n\pi_{ij}$

  where $n = n_{++}$ is the total count.
- $N = rc$ is the effective sample size (number of observations).
- Poisson rate: π_{ij}
- Log-linear model on $\log \mu_{ij}$

Independence Model for Two-way Tables

If X and Y are independent, then

$$P[X = x_i, Y = y_i] = P[X = x_i] \times P[Y = y_i] = \pi_{i+}\pi_{+j}$$

and the expected count is

$$\mu_{ij} = n\pi_{ij} = n\pi_{i+}\pi_{+j}$$

This implies that the log-linear model satisfies

$$\begin{aligned} \log \mu_{ij} &= \log n + \log \pi_{i+} + \log \pi_{+j} \\ &= \lambda + \lambda_i^X + \lambda_j^Y \end{aligned}$$

The estimates for the marginal probabilities are

$$\hat{\pi}_{i+} = \frac{n_{i+}}{n} \qquad \hat{\pi}_{+j} = \frac{n_{+j}}{n}$$

The fitted values for this model are

$$\mu_{ij} = n\hat{\pi}_{i+}\hat{\pi}_{+j} = \frac{n_{i+}n_{+j}}{n}$$

This can seen by maximizing the log likelihood (under the constraint $\sum \pi_{i+} = \sum \pi_{+j} = \sum \pi_{ij} = 1$), which has the form

$$\begin{aligned} l(\boldsymbol{\pi}) &= \sum_{i=1}^{r}\sum_{j=1}^{c} \{y_{ij} \log n\pi_{ij} - n\pi_{ij}\} \\ &= \sum_{i=1}^{r}\sum_{j=1}^{c} \{y_{ij} \log \pi_{i+} + y_{ij} \log \pi_{+j} + y_{ij} \log n\} - n \end{aligned}$$

If using the

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y$$

parameterization you need deal with constraints on λ_i^X and λ_j^Y. These constraints are needed since $\sum \pi_{i+} = \sum \pi_{+j} = \sum \pi_{ij} = 1$ must hold. Not a problem in **R** as its approach to generating contrasts when dealing with categorical predictors is a valid approach.

One consequence of this that for the independence model, the number of parameters to be estimated is $r+c-1$. There is λ, $r-1$ λ_i^Xs and $c-1$ λ_j^Ys.

As in the ANOVA setting, there is no unique approach for dealing with constraints. Common choices in this setting are $\hat{\lambda}_1^X = \hat{\lambda}_1^Y = 0$ (**R** default) or $\hat{\lambda}_r^X = \hat{\lambda}_c^Y = 0$.

Any valid constraint must satisfy

$$\hat{\lambda}_r^X = \log\left(1 - e^{\hat{\lambda}_1^X} - \ldots - e^{\hat{\lambda}_{r-1}^X}\right)$$

(similarly for $\hat{\lambda}_c^Y$)

What is unique are the differences $\hat{\lambda}_i^X - \hat{\lambda}_j^X$ and $\hat{\lambda}_i^Y - \hat{\lambda}_j^Y$ (we need to deal with contrasts).

One reason this is important is to consider the odds

$$\frac{\pi_{ij}}{\pi_{ik}} = \frac{P[X = x_i, Y = y_j]}{P[X = x_i, Y = y_k]} = \frac{\mu_{ij}}{\mu_{ik}}$$

Under the independence assumption, the log odds satisfies

$$\begin{aligned}
\log \frac{\pi_{ij}}{\pi_{ik}} &= \log \pi_{ij} - \log \pi_{ik} \\
&= (\lambda + \lambda_i^X + \lambda_j^Y) - (\lambda + \lambda_i^X + \lambda_k^Y) \\
&= \lambda_j^Y - \lambda_k^Y
\end{aligned}$$

Note that this does not depend on the level of X, which should be the case when X and Y are independent. Also note that a similar relationship holds for $\log \frac{\pi_{ij}}{\pi_{kj}}$ (fix Y vary X)

Saturated Model for Two-way Tables

The saturated model for the two-way table can be written

$$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$$

The terms λ_{ij}^{XY} represent the interaction between X and Y.

In this case, the effect of Y depends on the level of X and similarly the effect of X depends on the level of Y.

One way to see this is to look at the odds ratio

$$\phi_{ik,jl} = \frac{\pi_{ij}\pi_{kl}}{\pi_{kj}\pi_{il}} = \frac{\mu_{ij}\mu_{kl}}{\mu_{kj}\mu_{il}}$$

The odds ratios are useful measures to describe dependency between categorical variables.

To compare accounting and administration for men and women

$$\hat{\phi} = \frac{68 \times 40}{56 \times 91} = 0.53$$

Major	Women	Men
Accounting	68	56
Administration	91	40
Economics	5	6
Finance	61	59

So the odds of women to men in accounting is about half of that in administration.

$\log \phi_{ik,jl}$ can be shown to have the form

$$\lambda_{ij}^{XY} - \lambda_{kj}^{XY} - \lambda_{il}^{XY} + \lambda_{kl}^{XY}$$

Under independence, all $\lambda_{ij}^{XY} = 0$ and the log odds ratio $= 0$ (odds ratio $=$ 1). Note that this goes the other way as well, if all the $\lambda_{ij}^{XY} = 0$, then X and Y must be independent.

So another way of thinking of the λ_{ij}^{XY} is that the describe the form of dependency between X and Y.

For the saturated model, there are rc different parameters to be fit: λ, $r-1$ λ_i^Xs, $c-1$ λ_j^Ys and $(r-1)(c-1)$ λ_{ij}^{XY}s. A consequence of this is that the fitted cell counts satisfy

$$\hat{\mu}_{ij} = n_{ij}$$

Deviances in Contingency Tables

This leads to the deviance of a contingency model having the form

$$X^2 = \sum_{\text{all cells}} 2y_{ij} \log \frac{y_{ij}}{\hat{\mu}_{ij}}$$

i.e. $\sum 2Obs \log \frac{Obs}{Exp}$

Note that this has the same form as the deviance for the binomial regression models discussed earlier. This makes sense that if we condition on n, the total count,

$$(n_{11}, \ldots, n_{rc})|n \sim Multinomial(n, \boldsymbol{\pi})$$

As before X^2 is approximately distributed χ^2_{df} where $df = rc - \#$ params.

For the independence model

$$df = rc - (r + c - 1) = (r - 1)(c - 1)$$

For the saturated model $df = 0$

As before X^2 can be used for Goodness of fit. For the business major example

```
> summary(business.ind)

Call:
glm(formula = n ~ major + gender, family = poisson(),
  data = business)

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)         4.28054    0.09959  42.981  < 2e-16 ***
majorAdministration  0.05492    0.12529   0.438  0.66117
majorEconomics      -2.42239    0.31460  -7.700 1.36e-14 ***
majorFinance        -0.03279    0.12805  -0.256  0.79790
genderMale          -0.33470    0.10323  -3.242  0.00119 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 168.473  on 7  degrees of freedom
Residual deviance:  11.017  on 3  degrees of freedom
AIC: 63.832

> pchisq(deviance(business.ind),df.residual(business.ind),
    lower.tail=F)
[1] 0.01163662
```

So there is some evidence of lack of fit. This is also supported by Pearson Chi-square test

$$X_p^2 = \sum_{\text{all cells}} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$$

which also has approximately a χ^2_{df} distribution.

```
> pearson.ind <- sum(resid(business.ind,type='pearson')^2)
> pearson.ind
[1] 10.82673
> pchisq(pearson.ind, df.residual(business.ind), lower.tail=F)
[1] 0.01270068
```

Note that for two-way tables, this test can be done in **R** by

```
> chisq.test(business.tab)

        Pearson's Chi-squared test

data:  business.tab X-squared = 10.8267, df = 3, p-value = 0.0127

Warning message: Chi-squared approximation may be incorrect in:
chisq.test(business.tab)
```

As discussed before, the χ^2_{df} approximation works better in both tests when μ_{ij} are big. In this case one of the fitted cell counts is just under 5.

To get a handle on where the problems are, lets look at the fits and residuals

	Observed		Expected	
Major	Female	Male	Female	Male
Accounting	68	56	72.28	51.72
Administration	91	40	76.36	54.64
Economics	5	6	6.41	4.59
Finance	61	59	69.95	50.05

	Deviance Residuals		Pearson Residuals	
Major	Female	Male	Female	Male
Accounting	-0.51	0.59	-0.50	0.60
Administration	1.63	-2.08	1.68	-1.98
Economics	-0.58	0.63	-0.56	0.66
Finance	-1.09	1.23	-1.07	1.26

It appears that the big difference occurs with administration. More women than men go into that major than would be expected under independence. The other majors tend to be closer to 50:50 women to men.

This can also be seen by looking at the $\hat{\phi}_{ik,jl}$. For example, when comparing accounting and finance

$$\hat{\phi} = \frac{68 \times 59}{56 \times 61} = 1.17$$

Another way of looking for problems in the fit is to examine the $\hat{\lambda}_{ij}^{XY}$. For the example

```
> summary(business.int)

Call:
glm(formula = n ~ major * gender, family = poisson(),
     data = business)
```

```
Deviance Residuals:
[1]  0  0  0  0  0  0  0  0

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                     4.2195     0.1213  34.795  < 2e-16 ***
majorAdministration             0.2914     0.1603   1.818   0.0691 .
majorEconomics                 -2.6101     0.4634  -5.633 1.77e-08 ***
majorFinance                   -0.1086     0.1764  -0.616   0.5379
genderMale                     -0.1942     0.1805  -1.076   0.2820
majorAdministration:genderMale -0.6278     0.2618  -2.398   0.0165 *
majorEconomics:genderMale       0.3765     0.6318   0.596   0.5513
majorFinance:genderMale         0.1608     0.2567   0.626   0.5310
---
(Dispersion parameter for poisson family taken to be 1)

    Null deviance:  1.6847e+02  on 7  degrees of freedom
Residual deviance: -8.8818e-16  on 0  degrees of freedom
AIC: 58.815
```

```
> anova(business.int, test="Chisq")
Analysis of Deviance Table

Model: poisson, link: log

Response: n

Terms added sequentially (first to last)


             Df Deviance Resid. Df Resid. Dev  P(>|Chi|)
NULL                             7    168.473
major         3  146.796         4     21.677  1.294e-31
gender        1   10.661         3     11.017      0.001
major:gender  3   11.017         0 -8.882e-16      0.012
```

How this gets exhibited can vary on the constraints placed on the λ_i^Xs, λ_j^Ys, and λ_{ij}^{XY}s.

For example, using the constraints

$$\sum_i \lambda_i^X = 0$$

$$\sum_j \lambda_j^Y = 0$$

$$\sum_i \lambda_{ij}^{XY} = 0 \quad \text{for each } j$$

$$\sum_j \lambda_{ij}^{XY} = 0 \quad \text{for each } i$$

leads to estimated parameters

```
> options(contrasts=c("contr.sum","contr.poly"))
> business.int2 <- glm(n ~ major * gender, family=poisson(),
   data=business)
```

```
> summary(business.int2)

Deviance Residuals:
[1]  0  0  0  0  0  0  0  0

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     3.50428    0.08556  40.955  < 2e-16 ***
major1          0.61815    0.10673   5.792 6.97e-09 ***
major2          0.59559    0.10872   5.478 4.30e-08 ***
major3         -1.80368    0.23055  -7.823 5.15e-15 ***
gender1         0.10839    0.08556   1.267  0.20522
major1:gender1 -0.01132    0.10673  -0.106  0.91557
major2:gender1  0.30260    0.10872   2.783  0.00538 **
major3:gender1 -0.19955    0.23055  -0.866  0.38674
---
(Dispersion parameter for poisson family taken to be 1)
```

```
    Null deviance:  1.6847e+02  on 7  degrees of freedom
Residual deviance: -9.1038e-15  on 0  degrees of freedom
AIC: 58.815

> anova(business.int2, test="Chisq")
Analysis of Deviance Table

Model: poisson, link: log

Response: n

Terms added sequentially (first to last)

             Df Deviance Resid. Df Resid. Dev  P(>|Chi|)
NULL                             7    168.473
major         3  146.796          4     21.677  1.294e-31
gender        1   10.661          3     11.017      0.001
major:gender  3   11.017          0 -9.104e-15      0.012
```

Note that we are not changing the model, just how it is described, so it is reasonable that different descriptions will lead to different descriptions of how the data differs from independence.

Notice that the deviances reported under the 2 parameterizations are the same, up to rounding differences.

Example: Belief in the afterlife

As part of the 1991 General Social Survey, conducted by the National Opinion Research Center asked participants about whether they believed in an afterlife. The data, broken down by gender are

Belief	Females	Males	Total
Yes	435	375	810
No	147	134	281
Total	582	509	1091

Lets examine whether belief and gender are associated.

```
> summary(afterlife.int)

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          4.99043    0.08248  60.506   <2e-16 ***
beliefYes            1.08491    0.09540  11.372   <2e-16 ***
genderMale          -0.09259    0.11944  -0.775    0.438
beliefYes:genderMale -0.05583    0.13868  -0.403    0.687

> anova(afterlife.int, test="Chisq")

             Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                                      3    272.685
belief        1  267.635          2      5.050 3.718e-60
gender        1    4.888          1      0.162     0.027
belief:gender 1    0.162          0  1.197e-13     0.687
```

In this case, there is little evidence for including the interaction and thus belief in an afterlife and gender appear to be independent.

Log-linear Model versus Logistic Regression

Lets suppose that X takes two levels, such as in the afterlife belief example. Then

$$\log \frac{\mu_{1j}}{\mu_{2j}} = \log \frac{\pi_j}{1-\pi_j}$$

is the log odds where $\pi_j = P[X = 1|Y = Y_j]$.

So $\log \phi_{12,jk}$ is the log odds ratio comparing Y at level j with Y at level k. Thus we are effectively modeling the same parameters in the saturated log-linear model as we are in logistic regression with Y as a categorical predictor.

It can be shown that inference in the two approaches leads to the same answer. To exhibit this, lets fit the afterlife belief data both ways

```
> summary(afterlife.int)
Call: glm(formula = n ~ belief * gender, family = poisson())

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          4.99043    0.08248  60.506   <2e-16 ***
beliefYes            1.08491    0.09540  11.372   <2e-16 ***
genderMale          -0.09259    0.11944  -0.775    0.438
beliefYes:genderMale -0.05583    0.13868  -0.403    0.687

> summary(afterlife.logit)
Call: glm(formula = agree ~ gen, family = binomial())

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.14074    0.21572   5.288 1.24e-07 ***
gen         -0.05583    0.13868  -0.403    0.687
```

```
> anova(afterlife.int, test="Chisq")
Analysis of Deviance Table

Model: poisson, link: log

              Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                              3    272.685
belief         1  267.635         2      5.050 3.718e-60
gender         1    4.888         1      0.162     0.027
belief:gender  1    0.162         0  1.197e-13     0.687

> anova(afterlife.logit, test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

     Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                     1    0.16200
gen   1  0.16200         0  8.415e-14   0.68733
```

Other Sampling Schemes

So far only tables generated by Poisson sampling schemes have been considered. However many other schemes lead to the same analysis. These come by fixing different marginal counts.

- Multinomial sampling: fix n.

  In the case we are still modeling the joint probabilities π_{ij}. However it is with 1 multinomial sample instead of rc Poisson samples.

- Product multinomial: fix row (n_{i+}) or column totals (n_{+j}).

  In this case of fixing the row totals, each row is considered as a multinomial sample we are modeling $P[Y = y_j|X = x_i]$. So there are r different independent multinomial samples.

  Instead of looking for independence, normally we would instead look for homogeneity of probabilities (i.e. $P[Y = y_j|X = x_i] = P[Y = y_j]$ for all x_i).

The justification of these statements can be made by showing that the likelihoods in these case have equivalent forms when dealing with the π_{ij}s. This relates to the fact that if $Y_1 \sim P(\mu_1)$ and $Y_2 \sim P(\mu_2)$ and Y_1 and Y_2 are independent

$$Y_1 | Y_1 + Y_2 = n \sim Bin(n, \pi)$$

where $\pi = \frac{\mu_1}{\mu_1 + \mu_2}$.