Note Set #2 for Thursday March 5, 1998

More on The Sampling Model and the Log-Linear Model

Consider data which might come in a table that looks like,

$x_{1,1,1}$	$x_{1,1,2}$	$x_{1,1,3}$		$x_{2,1,1}$	$x_{2,1,2}$	$x_{2,1,3}$
$x_{1,2,1}$	$x_{1,2,2}$	$x_{1,2,3}$		$x_{2,2,1}$	$x_{2,2,2}$	$x_{2,2,3}$

where i indexes the faces of a cube, j indexes the rows and k indexes the columns.

The first question to ask is *what sort of sampling schemes could have generated data like these*? and consequently, *what sort of (conditional) independence statements can one make*?

There are many ways that these data could have been sampled/generated. I'll list some of these here.

1. Multinomial sampling. Assume that the total number of observations in the table $x_{+,+,+}$ is fixed and we just cross-classify these into $\mathrm{I} \times \mathrm{J} \times \mathrm{K}$ categories. (This implies 3 response variables)
2. Poisson Each cell in the table is an independent Poisson observation, so the grand total is not fixed. (This too implies 3 response variables)
3. Product-Multinomial. Lets assume that the category i is an explanatory variable, categories j and k are responses. That is we know $x_{i,+,+}$ for $i = 1, 2$. Thus the two faces of the table are multinomials. We are interested in comparisons between the two populations, such as how does the interaction change from one population to the other.
4. Product-Multinomial. Both i and j are explanatory variables, so we know $x_{i,j,+}$. That is we sample $\mathrm{I} \times \mathrm{J}$ multinomials, each with K cells.

Each of these cases corresponds to a different probability distribution. Each of the following looks different, but they are all very similar, modulo some conditioning arguments. The probability distributions are:

1.
$$\mathrm{P}(x_{i,j,k}) = \binom{x_{+,+,+}}{x_{1,1,1},\cdots,x_{I,J,K}} \prod_{i,j,k} ({p_{i,j,k}}^{x_{i,j,k}})$$
with the constraint that $\sum_{i,j,k} p_{i,j,k} = 1$. That is, we have one big multinomial with probabilities, $p_{i,j,k}$.

2.
$$\mathrm{P}(x_{i,j,k}) = \prod_{i,j,k} \frac{e^{-m_{i,j,k}} {m_{i,j,k}}^{x_{i,j,k}}}{(x_{i,j,k}!)}$$
That is we have I $\times$ J $\times$ K independent Poissons each with mean $m_{i,j,k}$.

3.
$$\mathrm{P}(x_{i,j,k}) = \prod_{i} \left[\binom{x_{i,+,+}}{x_{i,1,1},\cdots,x_{i,J,K}} \prod_{j,k} ({p_{i,j,k}}^{x_{i,j,k}}) \right]$$
where $\sum_{j,k} p_{i,j,k} = 1$ for each i. That is, we have lots of independent multinomials.

4.
$$\mathrm{P}(x_{i,j,k}) = \prod_{i,j} \left[\binom{x_{i,j,+}}{x_{i,j,1},\cdots,x_{i,j,K}} \prod_{k} ({p_{i,j,k}}^{x_{i,j,k}}) \right]$$
where $\sum_{k} p_{i,j,k} = 1$ for each combination of i and j. That is, we have lots and lots of independent multinomials.

It is not an accident that each of these likelihoods can be written in the form
$$k(x_{i,j,k}) \times \prod_{i,j,k} e^{-m_{i,j,k}} {m_{i,j,k}}^{x_{i,j,k}}$$
with some appropriate constraints on the $m_{i,j,k}$.

Thus each of the situations is essentially the same *except* the likelihoods require that the SAMPLING CONSTRAINTS be recognized. That is for the last case where we have I $\times$ J independent multinomials we must make sure that we get
$$\sum_{i} \hat{p}_{i,j,k} = 1, \quad \forall i, j$$

Models

Let us consider the loglinear model

$$\log(m_{i,j}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(i,j)} + u_{13(i,k)} + u_{23(j,k)} + u_{123(i,j,k)}$$

the model that include the (so-called) (i) main effects, (ii) two-factor interactions, (iii) three-factor interaction. The linear space spanned by this model is (quite complicated but) written out below.

$$u \text{ spanned by } \left\{ \begin{array}{|c|c|c|}\hline 1&1&1\\\hline 1&1&1\\\hline\end{array}, \begin{array}{|c|c|c|}\hline 1&1&1\\\hline 1&1&1\\\hline\end{array} \right\} \text{ depends on no index}$$

$$u_{1(i)} \text{ spanned by } \left\{ \begin{array}{|c|c|c|}\hline 1&1&1\\\hline 1&1&1\\\hline\end{array}, \begin{array}{|c|c|c|}\hline 0&0&0\\\hline 0&0&0\\\hline\end{array} \right\}, \left\{ \begin{array}{|c|c|c|}\hline 0&0&0\\\hline 0&0&0\\\hline\end{array}, \begin{array}{|c|c|c|}\hline 1&1&1\\\hline 1&1&1\\\hline\end{array} \right\}, \text{depends only on } i \text{ index}$$

$$u_{2(j)} \text{ spanned by } \left\{ \begin{array}{|c|c|c|}\hline 1&1&1\\\hline 0&0&0\\\hline\end{array}, \begin{array}{|c|c|c|}\hline 1&1&1\\\hline 0&0&0\\\hline\end{array} \right\}, \left\{ \begin{array}{|c|c|c|}\hline 0&0&0\\\hline 1&1&1\\\hline\end{array}, \begin{array}{|c|c|c|}\hline 0&0&0\\\hline 1&1&1\\\hline\end{array} \right\}, \text{depends only on } j \text{ index}$$

$$u_{12(i,j)} \text{ spanned by } \begin{array}{c} \left\{ \begin{array}{|c|c|c|}\hline 1&1&1\\\hline 0&0&0\\\hline\end{array}, \begin{array}{|c|c|c|}\hline 0&0&0\\\hline 0&0&0\\\hline\end{array} \right\}, \left\{ \begin{array}{|c|c|c|}\hline 0&0&0\\\hline 1&1&1\\\hline\end{array}, \begin{array}{|c|c|c|}\hline 0&0&0\\\hline 0&0&0\\\hline\end{array} \right\}, \\ \left\{ \begin{array}{|c|c|c|}\hline 0&0&0\\\hline 0&0&0\\\hline\end{array}, \begin{array}{|c|c|c|}\hline 1&1&1\\\hline 0&0&0\\\hline\end{array} \right\}, \left\{ \begin{array}{|c|c|c|}\hline 0&0&0\\\hline 0&0&0\\\hline\end{array}, \begin{array}{|c|c|c|}\hline 0&0&0\\\hline 1&1&1\\\hline\end{array} \right\} \end{array}, \text{depends only on } i,j \text{ index}$$

and so on for the other indices.

Obviously there need to be some constraints, of the form $\sum_i u_{1(i)} = 0$, and $\sum_j u_{2(j)} = 0$, and $\sum_i u_{12(i,j)} = \sum_j u_{12(i,j)} = 0$. $\sum_{i,j} u_{12(i,j)} = 0$.

These models are (from experience) useful ways of modelling data. They are also (from theory) useful ways of describing independence and conditional independence relationships.

FACT Let $\mathcal{M}$ be the model we are considering (defined by parameters, or whatever), then the sufficient statistics (i.e., minimal summary of the data) are $\mathrm{P}_{\mathcal{M}}(x)$ and the likelihood equations are:

1. $\mathrm{P}_{\mathcal{M}}(x) = \mathrm{P}_{\mathcal{M}}(\hat{m})$
2. $\log \hat{m} \in \mathcal{M}$

EXAMPLE Consider the independence model

$$\log(m_{i,j,k}) = u + u_{1(i)} + u_{2(j)} + u_{3(k)}$$

then the sufficient statistics are $x_{i,+,+}$, $x_{+,j,+}$, and $x_{+,+,k}$, and the likelihood equations are

$$\begin{aligned} x_{i,+,+} &= \hat{m}_{i,+,+} \\ x_{+,j,+} &= \hat{m}_{+,j,+} \\ x_{+,+,k} &= \hat{m}_{+,+,k} \\ \text{and} & \\ \log \hat{m}_{i,j,k} &= u + u_{1(i)} + u_{2(j)} + u_{3(k)} \end{aligned}$$

The solution to these equations is *easily* (once you know the answer) seen to be

$$\hat{m}_{i,j,k} = \frac{x_{i,+,+}x_{+,j,+}x_{+,+,k}}{x_{+,+,+}x_{+,+,+}}$$

The indices on the u-term parameterization give the form of the sufficient statistics. (That is, with some experience one can just read these things off).

This model was OK if we had three response variables, and could be interpreted as the independence model for all three factors. What if (say) variable 1 was an explanatory variable—that is the margin corresponding to that variable was fixed by the sampling scheme? If that were the case, then the first condition of the likelihood equations, $x_{i,+,+} = \hat{m}_{i,+,+}$ is not a MODEL constraint but is rather a SAMPLING constraint. In other words, we can **only** consider models which include the sampling constraint or in other words, include the u_1 term. We can **not** talk about a model that is like

$$\log(m_{i,j,k}) = u + u_{2(j)} + u_{3(k)}$$

because the fitted values need not (and probably will not) satisfy the sampling constraint.

What about the product-multinomial case where both i and j are explanatory. This implies that all valid models must include at least the $u_{1(i)} + u_{2(j)} + u_{12(i,j)}$ terms. An example of an acceptable model is

$$\log(m_{i,j,k}) = u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(i,j)} + u_{13(i,k)}$$

with corresponding likelihood equations of

$$\begin{aligned} x_{i,j,+} &= \hat{m}_{i,j,+} \\ x_{i,+,k} &= \hat{m}_{i,+,k} \\ \text{and} & \\ \log(m_{i,j,k}) &= u_{1(i)} + u_{2(j)} + u_{3(k)} + u_{12(i,j)} + u_{13(i,k)} \end{aligned}$$

which implies MLEs of

$$\hat{m}_{i,j,k} = \frac{x_{i,j,+}x_{i,+,k}}{x_{i,+,+}}$$

Models need not have closed-form MLEs—we were just lucky here.

A Categorical Data Example

To illustrate these ideas, we will do the "Red Dye Number 2" example from Fienberg (1987, pp. 52–53), in Splus. You might try to do the same thing in SAS (using PROC CATMOD, PROC GENMOD and/or PROC INSIGHT). Here is the problem and data.

The following table contains one version of the results of an experiment in which the food additive Red Dye No. 2 was fed to four different groups of rats at varying dosages. Some of the rats died during the experiment (i.e., between 0 and 131 weeks). The remainder were sacrificed at the end of 131 weeks. All rats were examined for tumors. Try to understand the data, and in particular what are the explanatory and response variables, and find some good models to describe the data.

	Age at Death	0–131 weeks		Terminal sacrifice	
	Dosage Group	Low	High	Low	High
Cancer	Present	4	7	0	7
	Absent	26	16	14	14
	Totals	30	23	14	21

This is a 2×2 table, with 3 categorizing variables. It seems reasonable to regard the dosage level as an explanatory variable with age at death and presence or absence of chance as both response variables. One could argue that the age at death variable might be an explanatory variable, but the number of observations in this group is not fixed (in advance) so for most purposes it makes sense to regard it as a response variable.

The distinction of response or explanatory variable does *not* change how we read in or manipulate the data, but does change the class of models we might consider.

In the following, I choose to enter the data the fancy way, using the `fac.design` command. Had I been lazy, I would have just generated the design by hand.

```
402 > reddye <- fac.design(c(2,2,2),c("Dose","Age","Cancer"))
402 > reddye$Resp <- c(4,7,0,7,26,16,14,14)
402 > reddye
   Dose  Age Cancer Resp
1 Dose1 Age1  Cncr1    4
2 Dose2 Age1  Cncr1    7
3 Dose1 Age2  Cncr1    0
4 Dose2 Age2  Cncr1    7
5 Dose1 Age1  Cncr2   26
6 Dose2 Age1  Cncr2   16
7 Dose1 Age2  Cncr2   14
8 Dose2 Age2  Cncr2   14
```

Now let's try to explore some models for these data. The command we need to use it `glm` and the appropriate family is `Poisson`.

The simplest model we can fit includes just the `dose` term. If we do not include `dose` then we risk fitting models which have fitted numbers of animals at the various dose levels different from the oberserved numbers—and that is not a good idea.

Thus to fit the simplest model (in this case as dose is explanatory),

```
402 > anova(mymod <- glm( Resp ~ Dose, family=poisson, data=reddye))
Analysis of Deviance Table
```

```
Poisson model

Response: Resp

Terms added sequentially (first to last)
     Df      Deviance Resid. Df Resid. Dev
NULL                          7   49.47734
Dose  1 -4.214286e-09         6   49.47734
```

and the corresponding Pearson X2 is

```
402 > Pearson(mymod)
Degrees of Freedom: 8 Total; 6 Residual
Residual Pearson X^2 42.72691
```

There is little to be said here. The dosage group is not a predictor of the raw response. We shouldn't be surprised by this.

The next step is to fit various models to the data. Here goes with an unstructured look at things. We'll see more structured approaches shortly.

```
402 > anova(mymod <- glm( Resp ~ Dose + Age, family=poisson, data=reddye))
Analysis of Deviance Table

Poisson model

Response: Resp

Terms added sequentially (first to last)
     Df Deviance Resid. Df Resid. Dev
NULL                     7   49.47734
Dose  1 0.000000         6   49.47734
 Age  1 3.707932         5   45.76941

402 > anova(mymod <- glm( Resp ~ Dose*Age, family=poisson, data=reddye))
Analysis of Deviance Table

Poisson model
```

```
Response: Resp

Terms added sequentially (first to last)
         Df Deviance Resid. Df Resid. Dev
    NULL                     7   49.47734
    Dose  1 0.000000         6   49.47734
     Age  1 3.707932         5   45.76941
Dose:Age  1 2.336721         4   43.43269
402 >
```

That is pretty boring. We have learned that `Age` might be appropriate, but we have no strategy. Instead, let us try to actually struggle our way through this, like we would an ANOVA.

```
402 > anova(mymod <- glm( Resp ~ Dose*Age + Age*Cancer + Dose*Cancer, family=poi>
Analysis of Deviance Table

Poisson model

Response: Resp

Terms added sequentially (first to last)
            Df Deviance Resid. Df Resid. Dev
       NULL                     7   49.47734
       Dose  1  0.00000         6   49.47734
        Age  1  3.70793         5   45.76941
     Cancer  1 32.82534         4   12.94407
   Dose:Age  1  2.33672         3   10.60735
 Age:Cancer  1  0.00739         2   10.59995
Dose:Cancer  1  7.60034         1    2.99961
```

Even though things are **NOT** orthogonal, it looks like one might wish to remove the `Age` times `Cancer` interaction (leaving the main effects)

```
402 > anova(mymod <- update(mymod, . ~ . - Age:Cancer))
Analysis of Deviance Table
```

```
Poisson model

Response: Resp

Terms added sequentially (first to last)
            Df Deviance Resid. Df Resid. Dev
       NULL                     7   49.47734
       Dose  1  0.00000         6   49.47734
        Age  1  3.70793         5   45.76941
     Cancer  1 32.82534         4   12.94407
   Dose:Age  1  2.33672         3   10.60735
Dose:Cancer  1  7.31735         2    3.29000
```

Now things are not clear, one may—or may not—remove the `Dose` times `Age` interaction. The upper 95% point of χ^2_1 is

```
402 > qchisq(.95,1)
[1] 3.841459
```

so maybe we should just remove it.

```
402 > anova(mymod <- update(mymod, . ~ . - Age:Dose))
Analysis of Deviance Table

Poisson model

Response: Resp

Terms added sequentially (first to last)
            Df Deviance Resid. Df Resid. Dev
       NULL                     7   49.47734
       Dose  1  0.00000         6   49.47734
        Age  1  3.70793         5   45.76941
     Cancer  1 32.82534         4   12.94407
Dose:Cancer  1  7.31735         3    5.62672
```

```
402 > Pearson(mymod)
Degrees of Freedom: 8 Total; 3 Residual
Residual Pearson X^2 4.228423
```

Now it is clear we can't proceed very much. Either of these last two models seems adequate and both are comparatively easy to interpret.

Oh yeah? How does one interpret these models, and what does it all mean. The final model (the one I would probably end up with) has an overall goodness of fit statistic of (likelihood ratio) 5.62 or (Pearson) 4.22, on 3 degrees of freedom. The upper 95% point of the appropriate chi-square is

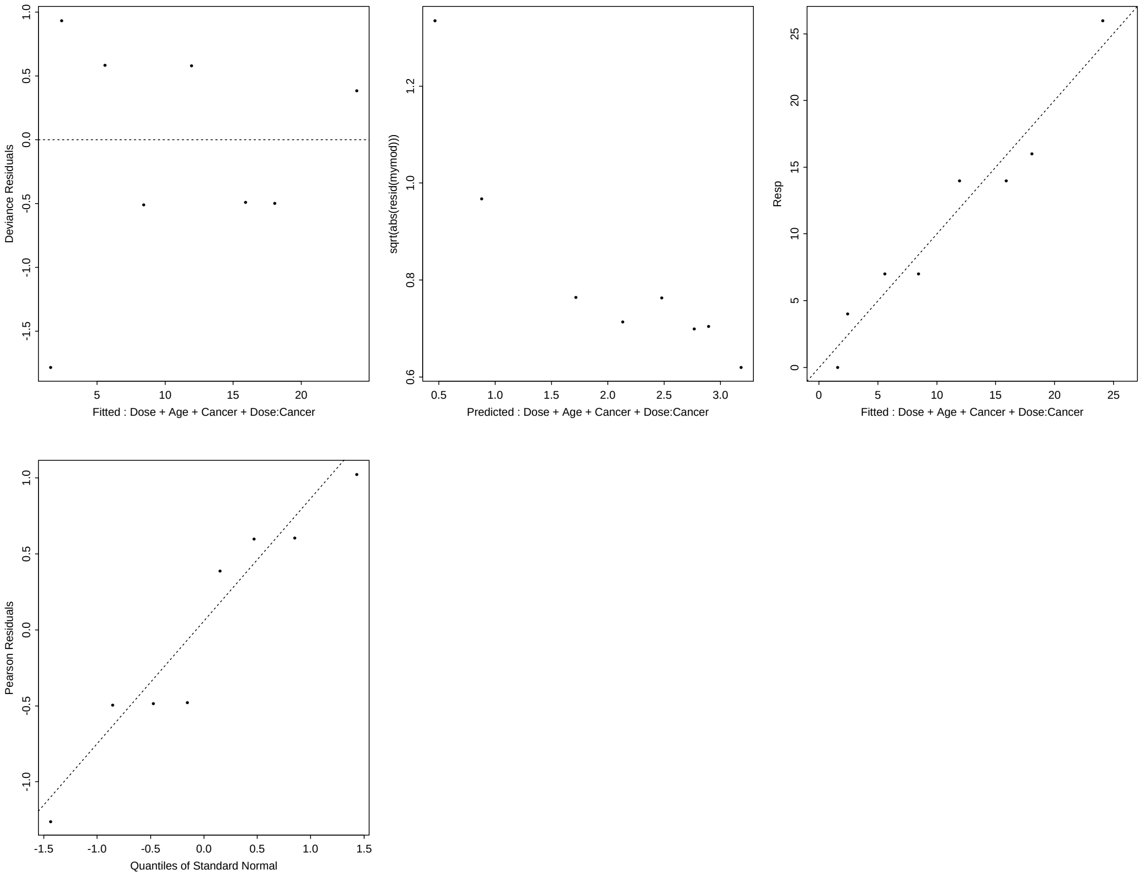
```
402 > qchisq(.95,3)
[1] 7.814728
```

So, on an overall basis, the model fits very well. The model itself says that `Age` is indepedent of everything, but that `Cancer` does depend on `Dose`. We could look at the actual parameter estimates, but would not learn very much.

However, we are not yet done. We certainly want to look at some plots and diagnostics, even though we only have 8 data points. We *think* we have a good model, do the diagnostics confuse the issue.

Start with a plot of the model/residuals.

```
402 > par(mfrow=c(2,3))
402 > plot(mymod)
```



This doesn't show anything at all untoward.

Goodness of Fit Statistics

There are two common goodness-of-fit statistics for log-linear models. We have looked at both the Pearson Chi-squared statistic,

$$\mathrm{X}^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

and the log likelihood ratio Chi-squared statistic,

$$\mathrm{G}^2 = 2 \times \sum \text{Observed} \times \log\left(\frac{\text{Observed}}{\text{Expected}}\right)$$

This statistic is really a version of

$$\mathrm{G}^2 = 2 \times \log \left\{ \frac{\text{likelihood}(\text{data}|\text{Model}_1)}{\text{likelihood}(\text{data}|\text{Model}_2)} \right\}$$

where $\text{Model}_2 \subset \text{Model}_1$.

If Model_1 is $\mathbf{R}^n$ (i.e., no model, or the full data), then the MLE is just the data itself (x), and if for Model_2 the MLE is $\hat{m}$, then the likelihood ratio statistic above reduces to

$$\mathrm{G}^2 = 2 \times \sum x_{i,j} \times \log \left(\frac{x_{i,j}}{\hat{m}_{i,j}} \right)$$

at least for Poisson likelihoods.

Otherwise, if MLE_1 is $\hat{m}^1$ and Otherwise, if MLE_2 is $\hat{m}^2$, then the general likelihood ratio statistic becomes

$$\mathrm{G}^2 = 2 \times \sum \hat{m}^1_{i,j} \times \log \left(\frac{\hat{m}^1_{i,j}}{\hat{m}^2_{i,j}} \right)$$

This statistic allows you to compare two models, rather than just a model with the fixed data. In other words, this allows us to look at the improvement of one model over another as the reduction in G^2.

Either way the distribution of G^2 tends towards χ^2_k where $k = \text{dimen}(\text{Model}_1) - \text{dimen}(\text{Model}_2)$.

The X^2 statistic can be similarly used to compare models, e.g.,

$$\mathrm{X}^2_* = \sum \frac{(\hat{m}^1_{i,j} - \hat{m}^2_{i,j})^2}{\hat{m}^2_{i,j}}$$

However G^2 has an advantage. Consider three models, $\mathcal{M}_3 \subset \mathcal{M}_2 \subset \mathcal{M}_1$.

Then

$$\mathrm{G}^2 \text{ for comparing } \mathcal{M}_3 \text{ to } \mathcal{M}_2 \quad \text{is } 2 \times \log \left\{ \frac{\text{lik}(x|\mathcal{M}_2)}{\text{lik}(x|\mathcal{M}_3)} \right\} \tag{1}$$

$$\mathrm{G}^2 \text{ for comparing } \mathcal{M}_2 \text{ to } \mathcal{M}_1 \quad \text{is } 2 \times \log \left\{ \frac{\text{lik}(x|\mathcal{M}_1)}{\text{lik}(x|\mathcal{M}_2)} \right\} \tag{2}$$

$$\mathrm{G}^2 \text{ for comparing } \mathcal{M}_3 \text{ to } \mathcal{M}_1 \quad \text{is } 2 \times \log \left\{ \frac{\text{lik}(x|\mathcal{M}_1)}{\text{lik}(x|\mathcal{M}_3)} \right\} \tag{3}$$

Now notice that (1) + (2) = (3), i.e., the G^2 tests add up, just like SS add up in regression. This is not true for X^2.

Both G^2 and X^2 are asymptotically equivalent—both have χ^2_k distributions where k is the number of parameters in $\mathcal{M}_1$ minus the number of parameters in $\mathcal{M}_2$. Both measures are useful, but it appears that X^2 is a little more accurate when the counts in all cells are quite small.