Notes for Tuesday February 24, 1998

Some References

Around 1972, the analysis of categorical data was revolutionized by Nelder and Wedderburn who recognized that many disparate models for discrete data could be unified as special cases of the *generalized linear model (GLM).* A big advantage of GLM's is that they make the analysis of discrete categorical data (which had been a bag of mostly seemingly-unrelated tricks) look much more like linear regression and ANOVA. For example, we can use our intutions about fitting models with `lm()` and `aov()` to inform our analysis using `glm()`[1] in SPLUS.

You can learn about log-linear models and logistic regression models from the following sources (among others!).

Agresti, A. (1990). *Categorical data analysis.* New York: Wiley.

A fully-detailed textbook on categorical data, generalized linear models, etc.

Chambers, J. M. and Hastie, T. J. (1992). *Statistical models in S.* Pacific Grove, CA: Wadsworth & Brooks/Cole. *[Especially Chapter 6.]*

A nice summary of what `glm()` can do. I may xerox this chapter for you at some point.

Fienberg, S. E. (1987). *The analysis of cross-classified categorical data, 2^{nd} Edition.* Cambridge MA: MIT Press.

Can be thought of as an abbreviated version of Agresti (1990). Much more emphasis on tables of counts, much less on logistic regression and generalized linear models. Many of the examples and intuitions for this part of the course will come from this book.

McCullagh, P. and Nelder, J. A. (1989). *Generalized linear models, 2^{nd} edition.* London: Chapman and Hall.

[1] Not to be confused with PROC GLM and the "general linear model" in SAS. PROC GLM simply fits linear regression and ANOVA models. The SAS procedure that corresponds to `glm()` in SPLUS is called PROC GENMOD.

This book really codified the revolution in discrete data analysis, by two of the main developers of the generalized linear model approach.

Venables, W. N. and Ripley, B. D. (1994). *Modern applied statistics with Splus.* New York: Springer-Verlag. *[Especially Chapter 7.]*

As usual, Venables and Ripley is better after you sort of know what is going on. But then it is a great reference and reminder of what is possible.

Log-linear Models

Let's return to the 2×2 examples from last time.

Example 1: Vitamin C Data

	Cold	No Cold	Total
Placebo	31	109	140
Vit C (1 gm/day)	17	122	139
Total	48	231	279

The sampling model was product binomial, and the expected table is

$$\begin{array}{|c|c|}\hline n_1\hat{p} & n_1(1-\hat{p}) \\ \hline n_2\hat{p} & n_2(1-\hat{p}) \\ \hline \end{array} = \begin{array}{|c|c|}\hline 24.1 & 115.9 \\ \hline 23.9 & 115.1 \\ \hline \end{array}$$

Example 2: Lizard Perches

Height	Diameter (in)		
(feet)	<= 4.0	> 4.0	Total
> 4.75	32	11	43
<= 4.75	86	35	121
Total	118	46	164

Here the sampling model could be either Poisson or Multinomial, and the expected table is

$$\begin{array}{|c|c|}\hline np_{+,1}p_{1,+} & np_{+,2}p_{1,+} \\ \hline np_{+,1}p_{2,+} & np_{+,2}p_{2,+} \\ \hline \end{array} = \begin{array}{|c|c|}\hline \frac{x_{+,1}x_{1,+}}{x_{+,+}} & \frac{x_{+,2}x_{1,+}}{x_{+,+}} \\ \hline \frac{x_{+,1}x_{2,+}}{x_{+,+}} & \frac{x_{+,2}x_{2,+}}{x_{+,+}} \\ \hline \end{array} = \begin{array}{|c|c|}\hline 30.9 & 12.1 \\ \hline 87.1 & 33.9 \\ \hline \end{array}$$

Notice that for both of these 2×2 table examples the fitted values were of the form

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

even though the sampling models were very different. For example, for the first table,

$$48 \cdot 140/279 = 24.1 \qquad 231 \cdot 139/279 = 115.1 \qquad \textit{etc.}$$

and for the second table,

$164{\cdot}(118/164){\cdot}(43/164) = 30.9 = 43{\cdot}118/164 \qquad 164{\cdot}(46/164){\cdot}(121/164) = 33.9 = 36{\cdot}121/164 \qquad$ *etc.*

Now, take logarithms of this expression and notice that

$$\log(\hat{m}_{i,j}) = -\log(x_{+,+}) + \log(x_{i,+}) + \log(x_{+,j})$$

In other words, the fitted values are *linear* functions of the margins on the logarithmic scale.

Now lets consider the model for the fitted values which says,

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

with the constraints that $\sum_i u_{1(i)} = 0$, and $\sum_j u_{2(j)} = 0$, and $\sum_i u_{12(i,j)} = \sum_j u_{12(i,j)} = 0$.

These models are the analogue of ANOVA models

$$\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

and are "natural" models for Poisson/Multinomial data. In particular, the notions of independence and conditional independence can be easily considered using these models. *The shift in focus, from linear models for the mean μ_{ij}, to linear models for some* **function** *of the mean such as* $\log(m_{ij})$*, is the central idea beind* **generalized linear models (glm's)**.

Another way of viewing the loglinear model is to say that it represents the logarithm of the expected values $(m_{i,j})$ in terms of main effects and interactions—which is just what the ANOVA models did.

A third, and most general, way of describing the models is to say that the loglinear models posit that

$$\log(m) \in \mathcal{M}$$

where $\mathcal{M}$ is a linear space which describes the hypothesis of interest. That is about the same as saying that an ANOVA model is just a linear model, and just as in the ANOVA situation it ignores the usefulness of various forms for the linear space $\mathcal{M}$.

The first formulation, in terms of the u-term parameterization, is one useful way of describing the linear space $\mathcal{M}$.

To make this a little more concrete consider a 3×3 table,

$$\begin{array}{|c|c|c|}\hline x_{1,1} & x_{1,2} & x_{1,3} \\ \hline x_{2,1} & x_{2,2} & x_{2,3} \\ \hline x_{3,1} & x_{3,2} & x_{3,3} \\ \hline \end{array}$$

where the two categorizing variables are both responses. (as in the lizard example). Let's assume that the expected values of the response (under some model to be determined) are

$$m = \begin{array}{|c|c|c|}\hline m_{1,1} & m_{1,2} & m_{1,3} \\ \hline m_{2,1} & m_{2,2} & m_{2,3} \\ \hline m_{3,1} & m_{3,2} & m_{3,3} \\ \hline \end{array}$$

The model which asserts independence of the two categories is

$$ln(m) \in \text{span}\left\{ \underbrace{\begin{array}{|c|c|c|}\hline 1&1&1\\\hline 1&1&1\\\hline 1&1&1\\\hline\end{array}}_{u}, \underbrace{\begin{array}{|c|c|c|}\hline 1&1&1\\\hline 0&0&0\\\hline 0&0&0\\\hline\end{array}}_{u_{1(1)}}, \underbrace{\begin{array}{|c|c|c|}\hline 0&0&0\\\hline 1&1&1\\\hline 0&0&0\\\hline\end{array}}_{u_{1(2)}}, \underbrace{\begin{array}{|c|c|c|}\hline 0&0&0\\\hline 0&0&0\\\hline 1&1&1\\\hline\end{array}}_{u_{1(3)}}, \underbrace{\begin{array}{|c|c|c|}\hline 1&0&0\\\hline 1&0&0\\\hline 1&0&0\\\hline\end{array}}_{u_{2(1)}}, \underbrace{\begin{array}{|c|c|c|}\hline 0&1&0\\\hline 0&1&0\\\hline 0&1&0\\\hline\end{array}}_{u_{2(2)}}, \underbrace{\begin{array}{|c|c|c|}\hline 0&0&1\\\hline 0&0&1\\\hline 0&0&1\\\hline\end{array}}_{u_{2(3)}} \right\}$$

That is, any set of fitted values which can be written as a loglinear combination of the above vectors satisfies independence. One can also see from this discussion where the constraints of our particular parameterization come from.

In general, we will oscillate between talking about abstract loglinear models, specific parameterizations, and the compromise model formulae of a package like S-Plus. Mostly (hopefully all the time, but I'll hedge a little), the model embodies some underlying notion and one should be careful to talk about the underlying model and not over-interpret the specific parameter estimates.

Now that we have written a model in this general (loglinear) form, an obvious question arises.

- Why not model $m \in \mathcal{M}$, a linear model?

or on some other scale?

The main reason is that if the loglinear model is written as a model for probabilities (which come directly from the fitted values), then we can be sure that the estimated probabilities will lie between 0 and 1. Without a lot of extra constraints and complications, linear models can not guarantee this. A more (or maybe less) compelling reason is that under the loglinear model the likelihood function factorizes in a natural way.

Bigger Two-way Tables

Consider the following cross-classification of 1019 school children into a teachers rating of schoolwork and an "independent" rating of homework conditions. (Independent in the non-statistical sense!) We would like to ask whether the two responses are statistically independent of each other.

	Homework Conditions					
Rating	A	B	C	D	E	
A	141	67	114	79	39	440
B	131	66	143	72	35	447
C	36	14	38	28	16	132
	308	147	295	179	90	1019

Under (statistical) independence the fitted values are generated by multiplying the row margins but the column margins and dividing by the grand total. This doesn't have to be obvious, but it is true.

Looking at the data we see that as homework conditions get worse (A to E) more people get a bad teachers rating. Is this significant, or am I imagining it? The table of fitted values, under independence is

133	63.5	121.4	77.3	38.9
135.1	64.5	129.4	78.5	39.5
39.9	19.0	38.2	23.2	11.6

with $\mathrm{X}^2 = 9.09$ on $(3-1) \times (5-1) = 8$ degrees of freedom. The value of X^2 tells us that there is not strong reason to reject the assumption of independence. What about looking at the residuals (generally a good thing to do), but which residuals? (there are at least two commonly used residuals . . .).

The observations $x_{i,j} - \hat{m}_{i,j}$ have different variances (which is approximately $\hat{m}_{i,j}$) so it makes sense to look at the Pearson standardized residuals

$$r_{i,j} = \frac{x_{i,j} - \hat{m}_{i,j}}{\sqrt{\hat{m}_{i,j}}}$$

For these data these numbers are

0.69	0.44	-0.67	0.19	0.02
-0.35	0.19	1.20	-0.73	-0.72
-0.62	-1.15	0.0	1.0	1.29

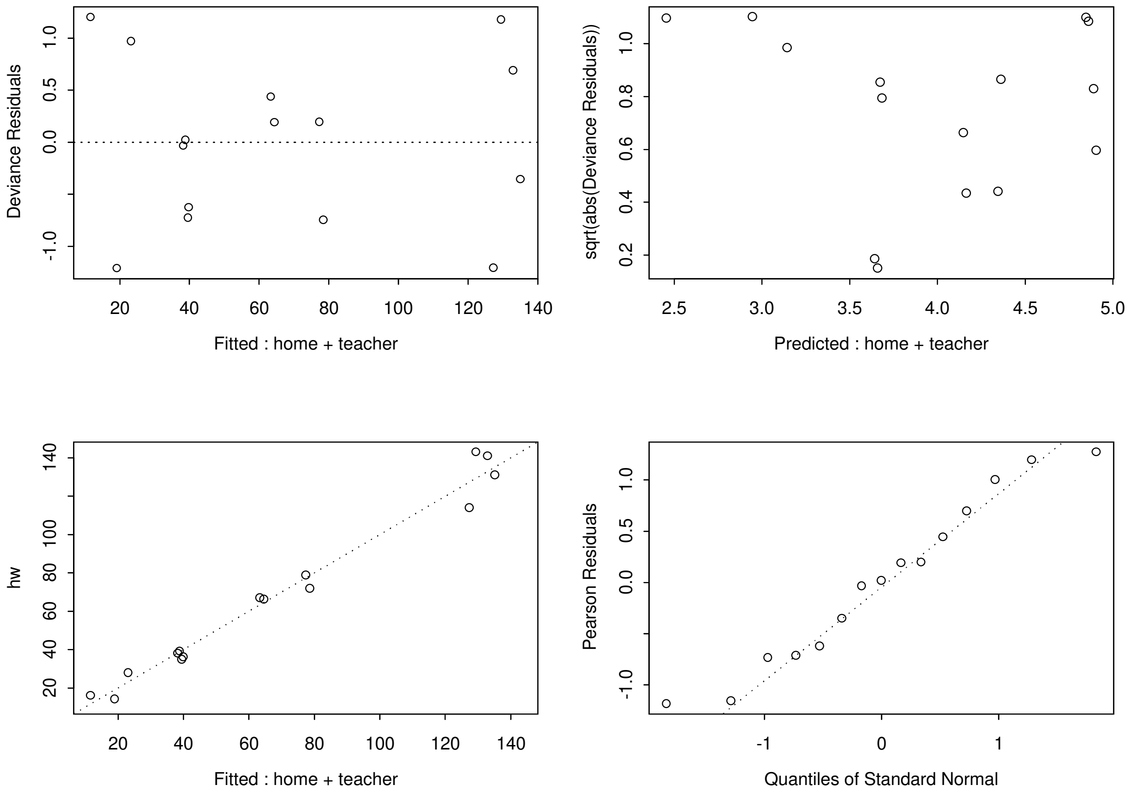
None of these residuals is very large (compare to $\mathrm{N}(0,1)$ nor does there appear to be any strong pattern to them. It seems that it is pretty safe to assume that the homework conditions and teachers ratings are independent.

Here is how this analysis would look in SPLUS:

```
402 > hw _ scan(sep="&")
1: 141 & 67 & 114 & 79 & 39
6: 131 & 66 & 143 & 72 & 35
11: 36 & 14 &  38 & 28 & 16
17:
402 > factors _ fac.design(c(5,3),list(LETTERS[1:5],LETTERS[1:3]))
402 > ratings _ data.frame(hw,factors)
402 > names(ratings) _ c("hw","home","teacher")
402 > ratings
    hw home teacher
 1 141    A       A
 2  67    B       A
 3 114    C       A
 4  79    D       A
 5  39    E       A
 6 131    A       B
 7  66    B       B
 8 143    C       B
 9  72    D       B
10  35    E       B
```

```
11  36    A       C
12  14    B       C
13  38    C       C
14  28    D       C
15  16    E       C

402 > crosstabs(hw ~ home + teacher , data=ratings)

Call:
crosstabs(hw ~ home + teacher, data = ratings)
1019 cases in table
+----------+
|N         |
|N/RowTotal|
|N/ColTotal|
|N/Total   |
+----------+
home   |teacher
       |A      |B      |C      |RowTotl|
-------+-------+-------+-------+-------+
A      |141    |131    | 36    |308    |
       |0.458  |0.425  |0.117  |0.302  |
       |0.320  |0.293  |0.273  |       |
       |0.138  |0.129  |0.035  |       |
-------+-------+-------+-------+-------+
B      | 67    | 66    | 14    |147    |
       |0.456  |0.449  |0.095  |0.144  |
       |0.152  |0.148  |0.106  |       |
       |0.066  |0.065  |0.014  |       |
-------+-------+-------+-------+-------+
C      |114    |143    | 38    |295    |
       |0.386  |0.485  |0.129  |0.289  |
       |0.259  |0.320  |0.288  |       |
       |0.112  |0.140  |0.037  |       |
-------+-------+-------+-------+-------+
D      | 79    | 72    | 28    |179    |
```

```
       |0.441  |0.402  |0.156  |0.176  |
       |0.180  |0.161  |0.212  |       |
       |0.078  |0.071  |0.027  |       |
-------+-------+-------+-------+-------+
E      | 39    | 35    | 16    |90     |
       |0.433  |0.389  |0.178  |0.088  |
       |0.089  |0.078  |0.121  |       |
       |0.038  |0.034  |0.016  |       |
-------+-------+-------+-------+-------+
ColTotl|440    |447    |132    |1019   |
       |0.43   |0.44   |0.13   |       |
-------+-------+-------+-------+-------+
Test for independence of all factors
Chi^2 = 9.092812 d.f.= 8 (p=0.334527)
Yates' correction not used

402 > hw.glm1 _ glm(hw ~ home + teacher , family=poisson, data=ratings)

402 > fitted(hw.glm1)
        1        2        3        4        5        6        7        8       9
 132.9931 63.47399 127.3798 77.29146 38.86163 135.1089 64.48381 129.4063 78.5211
       10       11       12       13       14       15
 39.47988 39.89794 19.0422 38.21394 23.18744 11.65849
402 > (hw-fitted(hw.glm1))/sqrt(fitted(hw.glm1))
         1         2         3         4          5          6         7       8
 0.6943016 0.4425732 -1.185493 0.1943385 0.02219649 -0.3534978 0.1888117 1.19498
          9         10         11        12          13        14      15
 -0.7359149 -0.7129822 -0.6171061 -1.155477 -0.03460765 0.9994242 1.27151
402 > sum(.Last.value^2)
[1] 9.092812

402 > par(mfrow=c(2,2))
402 > plot(hw.glm1)
```



The bottom two plots show how well the model is fitting, and how normal the Pearson residuals $(x_{ij} - m_{ij})/\sqrt{m_{ij}}$ are (they should be). The top two plots use a different kind of residual, called a "deviance residual", which still looks like $(x_{ij} - m_{ij})/(something)$, but is set up so that the sum of the squares is the log-likelihood chi-squared statistic, instead of the Pearson chi-squared statistic (in most well behaved problems the two kinds of residuals are approximately equal).

Informal Exercise

1. Verify that the table of fitted values can be written as $\log(m_{i,j}) = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{12(i,j)}$ with the constraints that $\sum_i \mu_{1(i)} = 0$, and $\sum_j \mu_{2(j)} = 0$.

2. Use the residuals to calculate X^2.

Example This example is drawn from Problem 2.2 in Fienberg (1987). The data below, from the National Crime Survey, represent repeat-victimization information (i.e., successive pairs of victimization for households in the survey). If, over th course of the period in question, a household is victimized m times, then these m victimizations will lead to $m-1$ entries in the table, one for each successive pair of victimizations. Here A is assault; B is burglary; HL is household larceny; MV is motor vehicle theft; PL is personal larceny; PP/PS is pocket picking and purse snatching; Ra is rape; and Ro is robbery.

	Second Victimization								
First	Ra	A	Ro	PP/PS	PL	B	HL	MV	Totals
Ra	26	50	11	6	82	39	48	11	273
A	65	2997	238	85	2553	1083	1349	216	8586
Ro	12	279	197	36	459	197	221	47	1448
PP/PS	3	102	40	61	243	115	101	38	703
PL	75	2628	413	229	12137	2658	3689	687	22526
B	52	1117	191	102	2649	3210	1973	301	9595
HL	42	1251	206	117	3757	1962	4646	391	12372
MV	3	221	51	24	678	301	367	269	1914
Total	278	8645	1347	660	22558	9565	12394	1960	

There are many things to ask of these data, including:

- Test these data for homogeneity of proportions
- Examine a table of the square roots of the contribution from each cell to X^2, i.e. (Observed – Expected)/$\sqrt{\text{Expected}}$ Which cell stands out?
- Provide an interpretation to your findings above.

- Why might you not have bothered to do the full calculation of X^2 after you found the contribution of the $(1, 1)$ cell?

Here is a cursory set of answers to these questions. A quick examination of the data shows that the row and colun margins are very similar and the diagonal cells of teh table (corresponding to repeat victimizations of the same type) clearly dominate the table.

Thus prior to formal modelling we have good reason to expect that the model of homogeneity of row proportions will *not* be appropriate.

Under the model of homogeneity of row proportions, with data $x_{i,j}$, and estimated expected values

$$\hat{m}_{i,j} = (x_{i,+} \cdot x_{+,j})/x_{+,+}$$

then the fitted values are as follows.

```
402 > Reiss <- fac.design(c(8,8),factor.names=c("Second","First"))
402 > Reiss$Resp <- scan("reiss.dat")
402 > Reiss[1:10,]
   Second First Resp
 1  Scnd1 Frst1   26
 2  Scnd2 Frst1   50
 3  Scnd3 Frst1   11
 4  Scnd4 Frst1    6
 5  Scnd5 Frst1   82
 6  Scnd6 Frst1   39
 7  Scnd7 Frst1   48
 8  Scnd8 Frst1   11
 9  Scnd1 Frst2   65
10  Scnd2 Frst2 2997
```

Now, fit the model and look at the Analysis of Deviance statistics,

```
402 > mymod<-glm(Resp ~ First + Second, family=poisson, data=Reiss)
402 > anova(mymod)
Analysis of Deviance Table
```

```
Poisson model

Response: Resp

Terms added sequentially (first to last)
       Df Deviance Resid. Df Resid. Dev
  NULL                   63   127032.9
 First  7 58883.15       56    68149.7
Second  7 59387.44       49     8762.3
```

What do the fitted values look like?

```
402 > t(matrix(round(fitted(mymod),digits=3),8,8))
        [,1]     [,2]    [,3]    [,4]     [,5]     [,6]     [,7]    [,8]
[1,]   1.322   41.111   6.406   3.139  107.275   45.487   58.940   9.321
[2,]  41.579 1292.978 201.462  98.712 3373.857 1430.576 1853.692 293.145
[3,]   7.012  218.056  33.976  16.647  568.990  241.262  312.619  49.438
[4,]   3.404  105.866  16.495   8.082  276.243  117.132  151.776  24.002
[5,] 109.036 3390.716 528.316 258.863 8847.631 3751.555 4861.137 768.745
[6,]  46.465 1444.924 225.137 110.312 3770.342 1598.693 2071.532 327.594
[7,]  59.913 1863.117 290.297 142.239 4861.560 2061.389 2671.078 422.407
[8,]   9.269  288.232  44.910  22.005  752.104  318.906  413.227  65.348
402 >
```

Notice the "trick" to get the data to print out in a nice tabular format.

The goodness-of-fit statistics for the model of homogeneity of proportions is $\mathrm{X}^2 = 11,189$ and $\mathrm{G}^2 = 8,762$, both of with should be asymptotically distributed as a χ^2_{49} random variable. The likelihood ratio statistic comes directly from the output,

```
402 > anova(mymod)
Analysis of Deviance Table

Poisson model
```

```
Response: Resp

Terms added sequentially (first to last)
       Df Deviance Resid. Df Resid. Dev
  NULL                    63   127032.9
 First  7 58883.15        56    68149.7
Second  7 59387.44        49     8762.3
402 >
```

The Pearson statistic requires a custom function. I wrote a little function, called Pearson, to print the statistics and degrees of freedom.

```
402 > Pearson(mymod)
Degrees of Freedom: 64 Total; 49 Residual
Residual Pearson X^2 11189.88
```

The Pearson function itself if,

```
402 > Pearson
function(mymod)
{
# This is not completely bullet proof.
        nobs <- length(mymod$residuals)
        rdf <- mymod$df.resid
        if(is.null(rdf))
                rdf <- nobs - rank
        cat("Degrees of Freedom:", nobs, "Total;", rdf, "Residual\n")
        cat("Residual Pearson X^2", format(sum((mymod$y - fitted(mymod))^2/
                fitted(mymod))), "\n")
        invisible(mymod)
}
```

It also makes a lot of sense to examine the residuals from this model. There are many different residuals one could consider, but a natural candidate is the so-called pearson residual which is the component of Pearson's X^2,

$$(\text{Observed} - \text{Expected})/\sqrt{\text{Expected}}$$

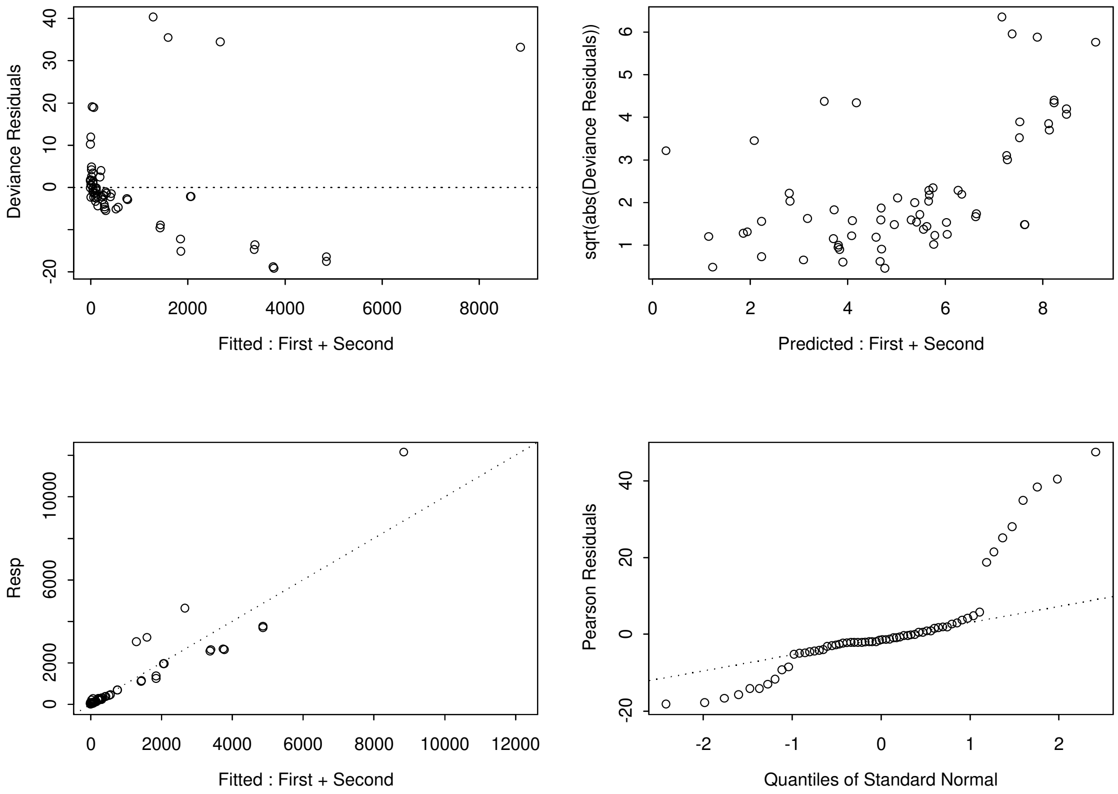
For these data, the numbers are

```
402 > t(matrix(round( residuals(mymod,type="pearson"), dig=3), 8,8))
        [,1]    [,2]    [,3]    [,4]    [,5]    [,6]    [,7]    [,8]
[1,]  21.457   1.386   1.815   1.615  -2.440  -0.962  -1.425   0.550
[2,]   3.632  47.388   2.574  -1.380 -14.132  -9.189 -11.722  -4.506
[3,]   1.883   4.127  27.967   4.743  -4.611  -2.850  -5.182  -0.347
[4,]  -0.219  -0.376   5.787  18.613  -2.000  -0.197  -4.121   2.857
[5,]  -3.259 -13.098  -5.017  -1.856  34.970 -17.854 -16.812  -2.948
[6,]   0.812  -8.627  -2.275  -0.791 -18.262  40.299  -2.165  -1.469
[7,]  -2.314 -14.181  -4.947  -2.116 -15.842  -2.189  38.213  -1.528
[8,]  -2.059  -3.960   0.909   0.425  -2.702  -1.003  -2.274  25.192
```

The contribution to the overall X^2 from just the first component of the residuals is $(21.457)^2 = 460.40$, which is much bigger than any upper bound of a χ^2_{49}.

Finally, here are some diagnostic plots

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



The high and low outlying residuals, suggesting deviation from the model of statistical independence, can clearly be seen in the upper left and lower right plots.