Log-linear model for melanoma data

Contents

1. Introduction

In this example we want to fit a model to the data published by Roberts et. al. (1981) on the occurrence of malignant melanoma, a form of skin cancer classified by the type of tumor. The data shown in the table below was obtained from Dobson and Barnett (2008, Table 9.4).

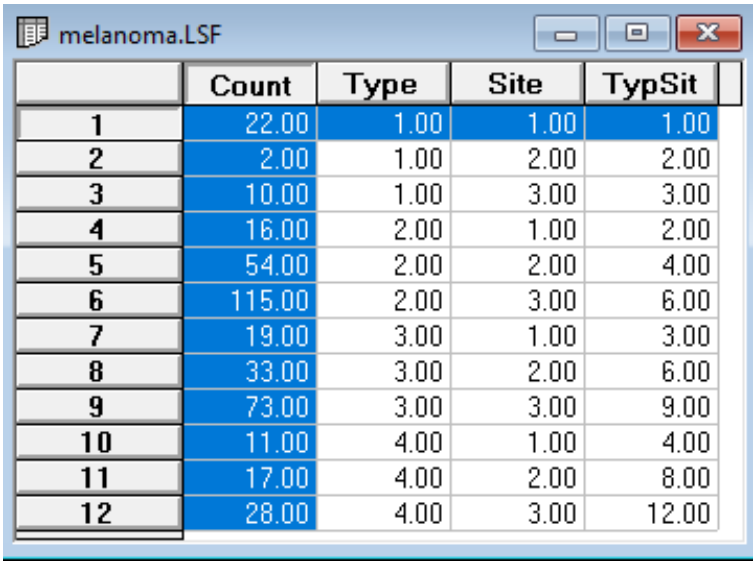
	Tumor site			
Type of tumor	**HNK**	**TNK**	**EXT**	**Total**
HMF	22	2	10	34
SSM	16	54	115	185
NOD	19	33	73	125
IND	11	17	28	56
Total	68	106	226	400

The variables cross-classified in this table are:

- Type of tumor, where HMF indicates the tumor to be a Hutchinson's melanotic freckle, SSM denotes a superficial spreading melanoma, NOD a nodular tumor, and IND an indeterminate type of tumor.
- Tumor site, where HNK indicates the tumor site to be on the head or neck, TNK that it is on the trunk, and EXT that the tumor occurred on an extremity.

Instead of having an outcome variable of interest, such as whether a death penalty is handed down or not, we wish to examine the possibility of an association between the two variables in the table. Is there a relationship between the type of tumor and the place it occurs on the body?

These data are given in **melanoma.lsf**, in a format more appropriate to a LISREL analysis. Data and syntax files can be found in the **MVABOOK\Chapter3** folder. Note the addition of another variable, TypSit, which represents the interaction between Type (of tumor) and Site (of the same).



	Count	Type	Site	TypSit
1	22.00	1.00	1.00	1.00
2	2.00	1.00	2.00	2.00
3	10.00	1.00	3.00	3.00
4	16.00	2.00	1.00	2.00
5	54.00	2.00	2.00	4.00
6	115.00	2.00	3.00	6.00
7	19.00	3.00	1.00	3.00
8	33.00	3.00	2.00	6.00
9	73.00	3.00	3.00	9.00
10	11.00	4.00	1.00	4.00
11	17.00	4.00	2.00	8.00
12	28.00	4.00	3.00	12.00

We fit a loglinear model to these data, in which the two variables are treated as equal. In other words, there is no dependent or independent variable per se.

If we define y_{ij} as the frequency of a cell formed by the *i*-th row and *j*-th column of a $I \times J$ contingency table, the sum of these will be a Poisson variable with $E(N) = \sum\sum(N) = \sum\sum(\mu_{ij}) = \mu$ if each of the y_{ij} is an independent Poisson variable with $E(y_{ij}) = \mu_{ij}$.

The conditional distribution of the y_{ij} 's, given their sum is the multinomial distribution, can be written as

$$f(\mathbf{y} \mid N) = N! \prod\prod \pi_{ij}^{\,y_{ij}} / y_{ij}!$$

where $\pi_{ij} = \mu_{ij} / \mu$ is the probability of cell (i, j). Since $E(y_{ij}) = \mu_{ij} = N\pi_{ij}$ the log link function can be used for the Poisson distribution, so that

$$\ln(\mu_{ij}) = \ln(L) + \ln(\pi_{ij})$$

with an offset set to the same for all *ij*.

To test the independence of the rows and columns in a contingency table such as the one considered in this example, the model

$$\ln(\mu_{ij}) = \ln(N) + \ln(\pi_i) + \ln(\pi_j)$$

If we denote the row effects by α_i and the column effects by β_j , we can rewrite the model as

$$\ln(\mu_{ij}) = \mu + \alpha_i + \beta_j$$

of, if the independence between rows and columns does not hold, the extended model

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

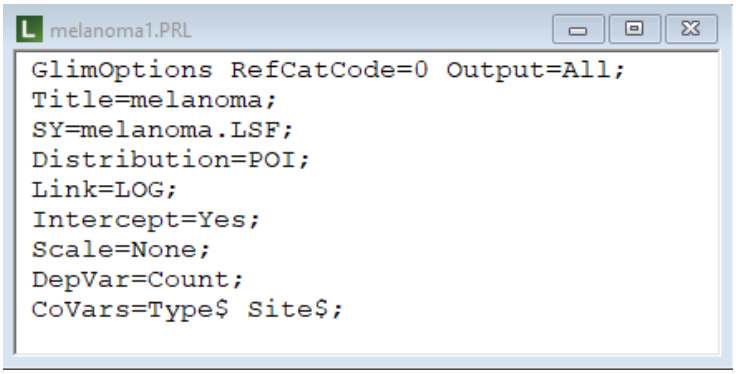
where $(\alpha\beta)_{ij}$ denotes the interaction effects.

2. Testing independence of row and column effects

We test the hypothesis of independence of tumor type and tumor site

$$H_0 : (\alpha\beta)_{ij} = 0$$
$$H_1 : (\alpha\beta)_{ij} \neq 0$$

using the syntax given in **melanoma1.prl**:



Note that in this syntax the Refcat option, representing the reference category used in the analysis, is set to 0. We opt to set the reference categories for both Type and Site to the first category by adding the syntax

```
Refcats = 1 1;
```

So all effects of tumor type will be measured relative to that of HMF, and all effects of tumor site relative to HNK. The adjusted syntax is echoed at the start of the output file

```
GlimOptions RefCatCode=0 Output=All;
 Title=melanoma;
 SY=melanoma.LSF;
 Distribution=POI;
 Link=LOG;
 Intercept=Yes;
 Scale=None;
 DepVar=Count;
 CoVars=Type$ Site$;
 Refcats = 1 1;
```

For this model, we obtain the estimates

Goodness of Fit Statistics

Statistic	Value	DF	Ratio
Likelihood Ratio Chi-square	51.7950	6	8.6325
Pearson Chi-square	65.8129	6	10.9688
-2 Log Likelihood Function	-2248.6543		
Akaike Information Criterion	-2236.6543		
Schwarz Criterion	-2233.7449		

Estimated Regression Weights

Parameter	Estimate	Standard Error	z Value	P Value
intcept	1.7544	0.2040	8.6000	0.0000
Type2	1.6940	0.1866	9.0787	0.0000
Type3	1.3020	0.1934	6.7313	0.0000
Type4	0.4990	0.2174	2.2951	0.0217
Site2	0.4439	0.1554	2.8573	0.0043
Site3	1.2010	0.1383	8.6834	0.0000

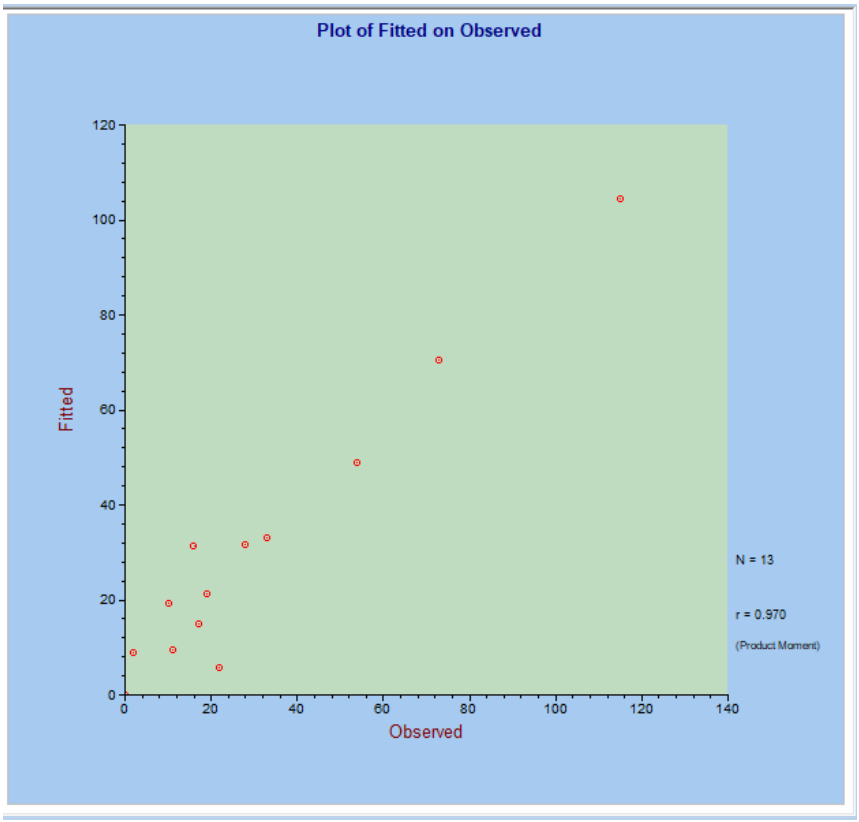
We see evidence of dependence when looking at both the statistical significance of the estimated effects and the chi-squares given in the goodness-of-fit section.

Given that the Output option on the GlimOptions line was set to All, we also have a file containing the residuals obtained under this model. The output file will always be an LSF file with the same name of the LSF file used in analysis, with "_res" added to the filename to distinguish between the two files. Contents of the file **melanona1_RES.LSF** are shown below. From the residual file we note that, under the assumption of independence, the expected prevalence of Hutchinson's freckle is only 5.78. From the observed data, we see that this tumor occurs most frequently on the head and neck of patients.

melanoma1_RES.LSF

	Observed	Fitted	Raw	Pearson	Deviance	Likelihd	SPearson	SDevianc
1	22.00	5.78	16.22	6.75	5.14	6.39	7.74	5.89
2	2.00	9.01	-7.01	-2.34	-2.83	-3.26	-2.85	-3.45
3	10.00	19.21	-9.21	-2.10	-2.32	-3.47	-3.33	-3.67
4	16.00	31.45	-15.45	-2.75	-3.05	-4.32	-4.12	-4.56
5	54.00	49.02	4.98	0.71	0.70	1.12	1.13	1.11
6	115.00	104.53	10.47	1.02	1.01	2.11	2.12	2.08
7	19.00	21.25	-2.25	-0.49	-0.50	-0.65	-0.65	-0.66
8	33.00	33.12	-0.12	-0.02	-0.02	-0.03	-0.03	-0.03
9	73.00	70.63	2.37	0.28	0.28	0.52	0.52	0.51
10	11.00	9.52	1.48	0.48	0.47	0.56	0.57	0.55
11	17.00	14.84	2.16	0.56	0.55	0.70	0.71	0.69
12	28.00	31.64	-3.64	-0.65	-0.66	-1.07	-1.06	-1.08
13	0.00	0.00	0.00	65.81	51.80	90.52	104.54	89.11

A scatterplot of the fitted against observed values is given below. While generally following the diagonal, there seems to be bigger differences when the observed frequency was lower.



In general, we conclude that there is no evidence to support the hypothesis that the type and site of tumors are independent of each other.

3. Including interaction terms

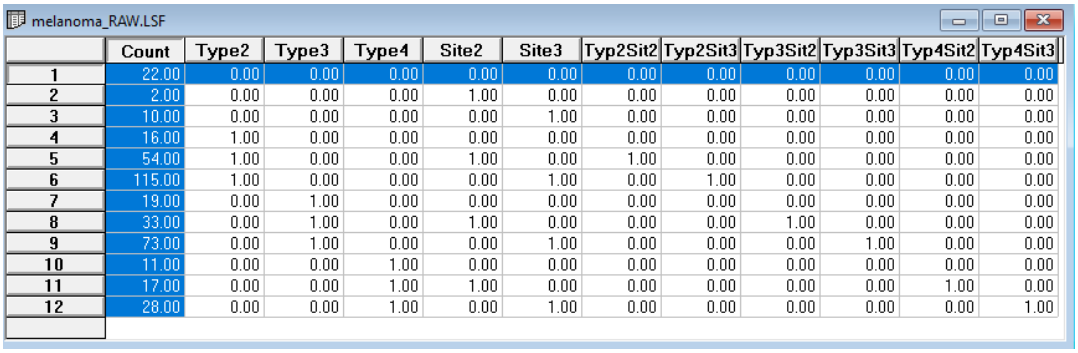
We now extend the model fitted in the previous section to include an interaction term, as represented by the variable TypSit in the LSF file.

melanoma2.PRL

```
GlimOptions;
Title=Melanoma with Interaction;
SY=melanoma_RAW.LSF;
Distribution=POI;
Link=LOG;
Intercept=Yes;
Scale=None;
DepVar=Count;
CoVars=Type2 Type3 Type4 Site2 Site3 Typ2Sit2 Typ2Sit3 Typ3Sit2 Typ3Sit3
       Typ4Sit2 Typ4Sit3;
```

In the previous analysis, dummy variables were created for the two variables listed in the Covars statement. Three were created for the 4 category variable Type, and 2 for the 3 category variable Site. Assuming that as before we use the first category of each as reference category, the inclusion of an interaction term means the evaluation of all effects created by Type = 2, 3, or 4 and Site = 2, 3. These dummies are shown on the Covars statement in the syntax file above.

To create the dummy variables we need for this analysis, we amend our data file to look as shown below. The variable Typ2Sit2, for example, represents the potential interaction between the second type of tumor and the second site. Note that this can be done using the **Compute** option from the **Transformation** menu on the LISREL window's main menu bar.



melanoma_RAW.LSF

	Count	Type2	Type3	Type4	Site2	Site3	Typ2Sit2	Typ2Sit3	Typ3Sit2	Typ3Sit3	Typ4Sit2	Typ4Sit3
1	22.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	2.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	10.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
4	16.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	54.00	1.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
6	115.00	1.00	0.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00
7	19.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
8	33.00	0.00	1.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
9	73.00	0.00	1.00	0.00	0.00	1.00	0.00	0.00	0.00	1.00	0.00	0.00
10	11.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
11	17.00	0.00	0.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
12	28.00	0.00	0.00	1.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	1.00

Results for the model are given below.

Goodness of Fit Statistics

Statistic	Value	DF
Likelihood Ratio Chi-square	0.0000	0
Pearson Chi-square	0.0000	0
-2 Log Likelihood Function	-2300.4493	
Akaike Information Criterion	-2276.4493	
Schwarz Criterion	-2270.6304	

The goodness-of-fit statistics indicate that this is a saturated model, i.e., a model in which all possible effects have been included.

Estimated Regression Weights

Parameter	Estimate	Standard Error	z Value	P Value
intcept	3.0910	0.2132	14.4983	0.0000
Type2	-0.3185	0.3286	-0.9692	0.3324
Type3	-0.1466	0.3132	-0.4681	0.6397
Type4	-0.6931	0.3693	-1.8771	0.0605
Site2	-2.3979	0.7378	-3.2499	0.0012
Site3	-0.7885	0.3814	-2.0674	0.0387
Typ2Sit2	3.6143	0.7908	4.5702	0.0000
Typ2Sit3	2.7608	0.4655	5.9314	0.0000
Typ3Sit2	2.9500	0.7920	3.7245	0.0002
Typ3Sit3	2.1345	0.4602	4.6381	0.0000
Typ4Sit2	2.8332	0.8332	3.4006	0.0007
Typ4Sit3	1.7228	0.5216	3.3028	0.0010

The regression results show that the interaction terms are all statistically significant. This is not the case for some of the "main" effects: for example, neither Type2 (representing a superficial spreading melanoma) or Type3 (representing a nodular tumor) is significant. These results offer further supporting evidence that the hypothesis of independence between the type and site of tumors is not realistic.