Model Based Statistics in Biology.

Part V. The General Linear Model.

Chapter 17.4 Two or More Categorical Explanatory Variables.

on chalk board

Ch17.xls

ReCap Part I (Chapters 1,2,3,4) Quantitative reasoning
ReCap Part II (Chapters 5,6,7) Hypothesis testing and estimation
ReCap (Ch 9, 10,11) The General Linear Model with a single explanatory variable.
ReCap (Ch 12,13,14,15) GLM with more than one explanatory variable
ReCap (Ch 16,17)

Today: Poisson response variable with two or more categorical explanatory variable.

Wrap-up.

Trees classified in a two-way table.

Counts are often presented in a contingency table. Here is an example.
Data are from Box 17.6 in Sokal and Rohlf (2012)

A plant ecologist examines 100 trees of a rare species from a 400 square mile area. Each tree is recorded as rooted in serpentine soil or not. Its leaves are classified as pubescent or smooth. Does leaf type depend on soil type?

In this example the number of trees examined was fixed at 100. We assume that the number of trees found in each soil type was free to vary. And we assume that leaf type was free to vary.

	Leaf Type		
Soil	Pubescent	Smooth	Totals
Serpentine	12	22	34
Not serpentine	16	50	66
Totals	28	72	100

Because leaf type and soil type are free to vary, we have two factors. We are interested in the interaction term (does leaf type depend on soil type?). Note the resemblance to the two-way ANOVA.

In a two way table, the interaction term is computed as the cross-product ratio, which measures the equality of proportions.

$$\frac{a}{b} \div \frac{c}{d} = \frac{a \cdot d}{b \cdot c} \qquad \frac{12}{22} \div \frac{16}{50} = 1.7045 \qquad \text{Odds ratio} \quad \frac{22}{50} \div \frac{12}{16} = 0.587 = 1.7045^{-1}$$

The G-test for independence in this two-way table addresses whether leaf type depends on soil type.

We begin with the computation of the goodness of fit of observed to expected, using the classical two-way contingency test. We will then analyze the data within the framework of the Generalized Linear Model, to show that the G-test for independence is the same as testing the interaction term in an ANOVA, except that now we will be using a Poisson error.

Example: Tree counts. Classical two-way contingency test.
Here is the goodness of fit of observed to expected, using the classical two-way contingency test.

$$G = 2 * \Sigma\left(f \cdot \ln\left(\frac{f}{\hat{f}}\right)\right)$$

	Pubescent	Smooth	
Serpentine	12	22	34
NonSerpentine	16	50	66
	28	72	100

f	=	$\hat{p}_{Ltype} \cdot \hat{p}_{Stype}$	$\cdot$ N	+	residual	lnL	= f(ln(f/fhat))
12	=	(28/100)(34/100)	$\cdot$ 100	+	2.48	2.78	
22	=	(72/100)(34/100)	$\cdot$ 100	−	2.48	−2.31	
16	=	(28/100)(66/100)	$\cdot$ 100	−	2.48	−2.35	
50	=	(72/100)(66/100)	$\cdot$ 100	+	2.48	2.54	

f	=	$\hat{p}_{Ltype \cdot Stype}$	$\cdot$ N	+	residual	lnL	= f(ln(f/fhat))
12	=	0.0952	$\cdot$ 100	+	2.48	2.78	
22	=	0.2448	$\cdot$ 100	−	2.48	−2.31	
16	=	0.1848	$\cdot$ 100	−	2.48	−2.35	
50	=	0.4752	$\cdot$ 100	+	2.48	2.54	
						0.67 $\cdot$ 2	= 1.33
							G = 1.33

Here is an equivalent formula to compute the row by column contingency statistic (Sokal and Rohlf 2012).

```
G =    2Σf ln(f)     =  12 ln12 + 22 ln22 + 16 ln16 + 50 ln50
      -2Σ(Σf) ln(Σf)      = -34 ln34 - 66 ln66 - 28 ln28 - 72 ln72
      +2 n ln(n)     - 100 ln100

G = 2( 337.78438 - 797.63516 + 460.51702 ) = 1.33249
```

Tree counts. Model-based Analysis
Now, for comparison, we analyze the same data as a generalized linear model with a poisson error.

1. Construct the model

Verbal model. Leaf type depends on soil type.
Graphical model. Ratio of pubescent to smooth, plotted against soil type.
Formal model

Response variable	f = count of trees in a class.
Explanatory variables	Ltype = leaf type (2 categories)
	Stype = soil type (2 categories)

1. Construct the model
In the previous analysis we saw that the main effects (leaf type and soil type) were multiplicative.

$$f = \hat{p}_{Ltype}\hat{p}_{Stype} \cdot N + residual$$

In order to construct a model having additive effects of the explanatory variables we use a log link between the parameters and the proportions *p*.

$$f = e^{\beta Ltype + \beta Stype} \cdot N + residual$$

$$\text{where} \quad \hat{p}_{Ltype} = e^{\beta Ltype} \quad \text{and} \quad \hat{p}_{Stype} = e^{\beta Stype}$$

$f = \hat{f} + PoissonError$ Poisson error for counts

$f = e^{\eta} + PoissonError$ log link with Poisson error

$\eta = \beta_{ref} + \beta_{Ltype} \cdot Ltype + \beta_{Stype} \cdot Stype$ additive effects of explanatory factors

To evaluate interactive effect (does leaf type depend on soil type?) we add the interaction term to the list of explanatory terms.

$$\eta = \beta_{ref} + \beta_{Soil} \cdot Soil + \beta_{Leaf} \cdot Leaf + \beta_{Soil*Leaf} \cdot Soil \cdot Leaf$$

The expected values in the 2 way table will be:

$$\hat{f} = e^{(\beta_{ref})} e^{(\beta_{Leaf} \cdot Leaf)} e^{(\beta_{Soil} \cdot Soil)} e^{(\beta_{Leaf*Soil} \cdot Leaf \cdot Soil)}$$

$e^{\beta_{ref}}$ = count in reference class

$e^{\beta_{Leaf} \cdot Leaf}$ = relative frequency * leaf type = 0 or 1

$e^{\beta_{Soil} \cdot Soil}$ = relative frequency * soil type = 0 or 1

$e^{\beta_{Leaf*Soil} \cdot Leaf \cdot Soil}$ = cross-product ratio

2. Execute analysis.

Arrange data into model format.

```
Data A;
  Input Count Leaf $ Soil $;
  Cards;
    12 Pbsc   Serp
    22 Smooth Serp
    16 Pbsc   NonSerp
    50 Smooth NonSerp
;
```

SAS command file

$$f = e^{(\beta_{ref})} e^{(\beta_{Leaf} \cdot Leaf)} e^{(\beta_{Soil} \cdot Soil)} e^{(\beta_{Leaf*Soil} \cdot Leaf \cdot Soil)} + \textit{Poisson error}$$

Use model to execute analysis.

```
Proc Genmod;  Classes Leaf Soil;
  Model Count = Leaf Soil Leaf*Soil/
  Link=log dist=poisson type1 type3;
  Output out=B p=fit resdev=res;
```

SAS command file

Obtain parameter estimates.

```
                          Analysis Of Parameter Estimates

                                               Standard      Wald 95%
Parameter                    DF   Estimate       Error   Confidence Limits

Intercept                     1     3.0910      0.2132     2.6732    3.5089
leaf        Pbsc              1    -0.6061      0.3589    -1.3095    0.0972
leaf        Smooth            0     0.0000      0.0000     0.0000    0.0000
soil        NonSerp           1     0.8210      0.2558     0.3195    1.3224
soil        Serp              0     0.0000      0.0000     0.0000    0.0000
leaf*soil   Pbsc     NonSerp  1    -0.5333      0.4597    -1.4342    0.3676
```

SAS output file

$e^{\beta_{ref}} = e^{3.0910} = 22$ count, reference group

$e^{\beta_{Ltype}} = e^{-0.6061} = 0.545 = \frac{12}{22}$ relative frequency, leaf type

$e^{\beta_{Stype}} = e^{0.8210} = 2.27 = \frac{50}{22}$ relative frequency, soil type

$e^{\beta_{L*S}} = e^{-0.5333} = 0.587 = \frac{22}{50} \div \frac{12}{16}$ cross-product ratio

3. Use parameter estimates to calculate residuals, evaluate model.

We cannot evaluate assumptions from the residuals. This is a saturated model, there are as many parameter estimates as observations (rows of data) and so there are no residuals.

4. Population. If the trees were sampled randomly, then the population is all of the trees of that species in the 400 square mile area. If the trees were sampled haphazardly, then the sample might still be taken as representative of the population in that area. We may wish to infer, informally, to other locations.

5. Decide on mode of inference. Is hypothesis testing appropriate?

The cross product ratio is 0.587, which is less than 1.
But this might be due to chance. So we undertake hypothesis testing.

6. State H_A / H_o *etc.*

H_A : $\beta_{Leaf*Soil} \neq 0 \quad e^{\beta_{Leaf*Soil}} \neq 1$ frequency depends on both leaf type and soil type, hence cross-product ratio differs from unity.

H_o : $\beta_{Leaf*Soil} = 0 \quad e^{\beta_{Leaf*Soil}} = 1$ frequency does not depend on both leaf type and soil type, hence cross-product ratio equal to unity.

H_A : $$f = e^{(\beta_{ref})} e^{(\beta_{Leaf} \cdot Leaf)} e^{(\beta_{Soil} \cdot Soil)} e^{(\beta_{Leaf*Soil} \cdot Leaf \cdot Soil)}$$

H_o : $$f = e^{(\beta_{ref})} e^{(\beta_{Ltype} Leaf)} e^{(\beta_{Stype} Soil)}$$

We will test whether the cross-product ratio differs from unity. This is equivalent to testing whether including the interaction term improves the fit. In the model format, H_A and H_o differ by a single term.

Statistic: G

Probability distribution: chisquare

α = 0.05

7. ANODEV Table.

Analysis of Deviance table is set up in much the same was as the ANOVA table.

Source	df	Deviance = G	Δ G
Intercept	1		
Leaf	1 = 2 - 1		
Soil	1 = 2 - 1		
Leaf*Soil	1 = 1*1		

The AnoDev table shows the deviance of the data from the model, for a sequence of models. It also shows the change in deviance (Δ G = improvement in fit) due to each term in the model. This is labelled Chi-square in the SAS output.

```
                 LR Statistics For Type 1 Analysis

                                                   Chi-
Source               Deviance          DF        Square     Pr > ChiSq

Intercept             31.7936
leaf                  11.7548           1         20.04         <.0001
soil                   1.3325           1         10.42         0.0012
leaf*soil              0.0000           1          1.33         0.2484
```

SAS output file

The improvement in fit due to the interaction term is Δ G = 1.3325 (df = 1)
This is the same value we obtained from the classical analysis of contingency in a two-way table.

Calculate p-value from Chisquare distribution.

Δ G = 1.3325, df = 1 --> p = 0.2484

8. Assess p-values and estimates in light of evaluation of assumptions.
Because this is a saturated model, residuals cannot be used to evaluate model.

9. Declare decision. p = 0.2484 hence accept H_o (reject H_A)
Frequency of leaf type does not depend on soil type.
(Δ G = 1.33, df = 1, p = 0.2484)

10. Evaluate parameters of biological interest.
In this analysis only the interaction term was of interest.

The ratio of pubescent to smooth was 12 / 22 = 0.545 in serpentine soil
The ratio was 16 / 50 = 0.32 in non serpentine soil.

We cannot dismiss chance as an explanation of the observed difference in ratios.
How large a sample would we need for these ratios to differ significantly?
To find out, we increase the frequencies by successively greater multiples until G reaches 3.84, the critical value of G (df = 1) at p = 0.05.
G reaches 3.84 when all 4 frequencies have been multiplied by 2.88. This results in a table with 288 trees, in the same proportions as the table with 100 trees.

f	Pubescent	Smooth	
Serpentine	34	64	98
NonSerpentine	46	144	190
	80	208	288

We would need 100*2.88 = 288 trees for the observed difference in proportion to be significant.

Poisson Response Variable. Two way classification.

Does relative abundance of sycamores and birches depend on woodland ?

Data from Andrews and Herzberg.
A&HTable55.dat

```
1      7      2
1      10     0
1      12     0
1      6      0
2      4      0
2      5      4
2      0      0
2      0      0
3      4      0
3      1      0
3      1      0
3      5      0
3      2      0
4      2      0
4      0      0
4      0      0
4      0      0
4      2      0
5      2      5
5      0      2
5      1      5
5      0      3
5      2      5
5      9      0
6      3      4
6      0      8
6      4      1
6      0      0
6      2      0
6      8      0
7      0      0
7      0      0
7      0      0
7      0      0
7      3      0
8      1      0
8      3      0
8      3      0
8      2      0
8      4      1

Loc    Syc    Birch

Location
1 Dungoon (DU)
2 Northcliffe West (NW)
3 Northcliffe Middle (NM)
4 Northcliffe East (NW)
5 Low Wood (LW)
6 Dixon's Wood (DW)
7 Royd's Cliffe (RC)
8 Weather Royd's (WR)
```