Model Based Statistics in Biology.

Part V. The Generalized Linear Model.

Chapter 17.4 Two or More Categorical Explanatory Variables.

ReCap. Part I (Chapters 1,2,3,4), Part II (Ch 5, 6, 7)
ReCap Part III (Ch 9, 10, 11), Part IV (Ch 13, 14)
17 Poisson Response Variables
17.1 Poisson Regression
17.2 Single Categorical Explanatory Variable
(Log-linear Model)
17.3 Single Categorical Explanatory Variable
(Sensitivity Analysis)
17.4 Two or More Categorical Explanatory Variables
Classical two way contingency test
Model based analysis
BACI design (to be added)
17.5 Poisson ANCOVA
17.6 Model Revision

Ch17.xls

on chalk board

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 variables.

First example.	Two-way Contingency Test (Ant Acacia)
Second example:	Multiway Contingency Test (trees * woodlands)
Third example:	Environmental Impact Assessment
	Before-After Control-Impact (BACI) Design.
	Example from text by Green (Fixed * fixed)

Wrap-up.

Count data are frequently presented as cross-classified counts.
We use a log linear model (Poisson response variable, multiple factors) to test for independent effects of two (or more) factors on counts.
No causal ordering of cross classified counts.

Background

Contingency tests appear in Fisher's (1925) text in statistics and have remained a regular feature in subsequent texts. They remain a staple practice in both the natural and social science literature. Fisher's analysis used a Poisson error structure. Fisher's analysis was extended from 2 x 2 tables to 2 x 2 x 2 x 2 tables later in the 20th century (Bishop *et al.* 1975 *Discrete Multivariate Analysis Theory and Practice,* MIT Press). Nearly all of the examples in Bishop *et al* are better considered as binomial analyses of one of one the "factors" relative to other factors. Here is an example where the total number of units was fixed, allowing a contingency test. Had the number of units been fixed by factor (leaf type or soil type) a binomial analysis would be more appropriate.

Example: Tree counts. Classical two-way contingency test.

Does leaf type (pubescent or smooth) depend on soil (serpentine or not)? Serpentine minerals form when rocks are extruded from deep in the earth's interior, then metamorphosed at convergent plate boundaries where an oceanic plate is pushed down into the mantle. Due to their origin, serpentine soils have low calcium-to-magnesium ratio and lack many essential nutrients, notably nitrogen, phosphorus, and potassium. Serpentine soils contain high concentrations of metals, including chromium, cobalt, and nickel. These challenging conditions restrict the number of plant species that can persist in serpentine soils. Serpentine soils have a high rate of rare endemic species, species found only in serpentine soils. Morphological adaptations to dry conditions and high light environments (and possible nutrient deficiencies) include fleshy leaves, leaf hairs, and protective pigments. Here are data on one such adaptation, leaf type. 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. Leaves are classified as pubescent or smooth.

In this example the number of trees examined was fixed at 100. Leaf is scored by type, not by presence or absence. Soil is scored by type, not by presence or absence.

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

Example: Tree counts. Classical two-way contingency test
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.

Cross product ratios: $\frac{a}{b} \div \frac{c}{d} = \frac{a \cdot d}{b \cdot c}$ $\frac{12}{22} \div \frac{16}{50} = 1.7045$

The odds ratio by rows $\frac{12}{22} \div \frac{16}{50} = 1.7045$

The odds ratio by columns has the same value.

We begin with the contingency test using the textbook formula.

Here is the formula for the G-statistic.

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

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

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

f	=	$\hat{p}_{Lype\text{-}Stype}$	· N	+	residual	lnL	= f(ln(f/fhat)
12	=	0.0952	· 100	+	2.48	2.78	
22	=	0.2448	· 100	−	2.48	−2.31	
16	=	0.1848	· 100	−	2.48	−2.35	
50	=	0.4752	· 100	+	2.48	2.54	

0.67 · 2 = 1.33
G = 1.33

$LR = e^{1.33/2} = 1.94$
The observed cross product ratio is no more likely than CPR = 1.
There is no evidence for a difference in proportions.

Tree counts. Generalized linear model

Now, for comparison, we analyze the same data as a generalized linear model with a Poisson error and log link.

1. Construct the model

Verbal model. Leaf type depends on soil type.

Graphical model. Ratio of pubescent to smooth, plotted against soil type.

List variables

C = count of trees in a class.

Ltype = leaf type (2 categories)

Stype = soil type (2 categories)

List dependencies or each pair of variables.

Count depends on leaf type and soil type.

Soil type does not depend on the other two variables

Leaf type depends on soil but not count.

The verbal model is Count (Poisson) depends on soil and leaf type.

For count data we use the 3 part notation for a generalized linear model.

Distribution $C \sim Poisson(\lambda)$ where λ is mean count per unit.

Link $C = e^{\eta}$

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

$$\hat{C} = \hat{p}_{Ltype}\hat{p}_{Soil} \cdot N$$

In order to construct a model having multiplicative effects of the explanatory variables we use a log link between the response variable (Count) and the structural model.

$$\hat{C} = e^{\beta_{Ltype} \cdot Ltype + \beta_S \cdot Soil}$$

where: $\hat{p}_{Ltype} = e^{\beta_{Ltype} \cdot Ltype}$

$\hat{p}_{Soil} = e^{\beta_S \cdot Soil}$

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

$$\eta = \beta_0 + \beta_{Soil}Soil + \beta_{Ltype}Ltype + \beta_{S \cdot L}Soil \cdot Ltype$$

1. Construct the model

The expected values in the 2 way table will be:

e^{β_o}	= count in reference class
$e^{\beta_L Ltype}$	= relative frequency * leaf type = 0 or 1
$e^{\beta_S Soil}$	= relative frequency * soil type = 0 or 1
$e^{\beta_{S \cdot L} Soil \cdot Ltype}$	= 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

$$\eta = \beta_0 + \beta_{Soil} Soil + \beta_{Ltype} Ltype + \beta_{S \cdot L} Soil \cdot Ltype$$

Use model to execute analysis

```
CountMod <- glm(formula = Count ~ Soil * Ltype,
      family = poisson(link = log),
      data = AHtable55)
```

R script

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

SAS command file

2. Execute analysis.

Obtain parameter estimates.

Analysis of Parameter Estimates

Parameter			DF	Estimate	Standard Error	Wald 95% 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

$e^{\beta_o} = e^{3.091} = 22$ 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. What is the evidence?

The ANODEV table shows the change in deviance (ΔG = 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

Source	Deviance	DF	Chi-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

$LR = e^{1.33/2} = 1.94$ The model with the interaction term is just as likely as the model without the term.

5. Choice of inferential Mode.

Priorist?	No. We have insufficient prior information to set up a defensible prior probability.
Frequentist?	Yes. We have a measurement protocol that is repeatable.
Evidentialist ?	Yes. This is an observational study with many uncontrolled sources of variance. We have no way of gauging Type I versus Type II error. Nor any consideration of risk or cost of Type I error to use in declaring a decision against a fixed Type I error.
Decision Theoretic?	We will use a fixed Type I error to illustrate an analysis of sample size needed to detect an effect at a rate of $\alpha = 0.05$.

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 similar locations outside the area.

Likelihood ratios.

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

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

We compare the likelihood of two models, one with the interaction term, one without.

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

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

Note that the models differ by a single term. This is called a nested analysis.
If there is no interactive effect, the interaction term is $e^0 = 1$.
The likelihood ratio is $LR = H_A / H_o = 1$.

9. Statistical conclusion.

The fit to the saturated model (4 parameters) is perfect. $G = 0$.
The fit to the model without the interaction term is $G = 1.33$.
The change in fit is $\Delta G = 1.33$
From this we calculate $LR = e^{(1.33 / 2)} = 1.95$
$LR < 10$.
The model with no interactive effect is just as likely as the model with an interactive effect.

10. Science conclusion. 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.
The ratio in non-serpentine soil was just as likely as that in serpentine soil.

10. Prospective power analysis.

We begin with the Type I error.
$\Delta G = 1.33$ df = 1 $p = 0.2484$ from $\chi 2$ distribution.
How large a sample would we need to detect a difference, at a fixed error rate of $\alpha = 5\%$?
To find out we increase the frequencies by successively greater multiples until ΔG reaches 3.84, the critical value of G (df = 1) at $\alpha = 5\%$.

ΔG 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 $100 \cdot 2.88 = 288$ trees to detect a change in proportion at an error rate fixed at 5%. Note that with hypothesis testing, we cannot add trees until $p < 0.05$. We are committed to undertake the entire sample.

A similar calculation can be done with respect to a likelihood ratio considered a some evidence, LR =10
We would need 100*3.46 = 346 trees to attain a likelihood ratio greater than $LR = 10$ ($\Delta G = 4.61$). With likelihood, we are not committed to undertaking the full sample. We update our prior to a posterior probability.

Second Example.
Poisson Response Variable. Two way classification.

Does relative abundance of sycamores and birches depend on woodland?

Data from Andrews and Herzberg.
A&H Table 55.txt

1. Construct the model

Variables. Count. Number of trees in each cell
Sp Species (2)
Loc Location (8)

Evaluate dependencies to identify response variable.
Count depends on species and location.
Species depends on location, but not count.
Location does not depend on count or species.
Count is the response variable

Distribution $C \sim Poisson(\lambda)$
Link $C = e^{\eta}$
$\eta = \beta_0 + \beta_{Loc} Loc + \beta_{Sp} Sp + \beta_{L \cdot Sp} Loc \cdot Sp$

Distribution: Poisson distribution for counts where neither explanatory variable clearly depends on the other.
Link: Log link. Multiplicative effects of explanatory factors
Structural model: Same structure as 2-way ANOVA and previous example

2. Execute model

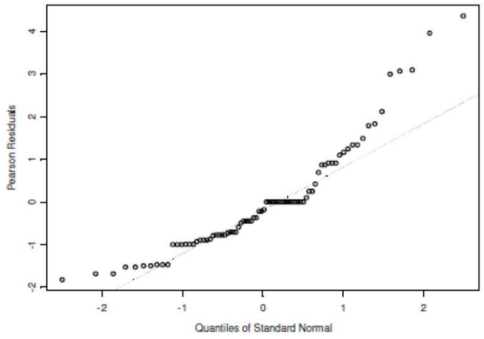
```
> modell <- glm(formula = Count ~ Location * TreeSp,
      family = poisson(link = log),
      data = AHtable55)
```

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

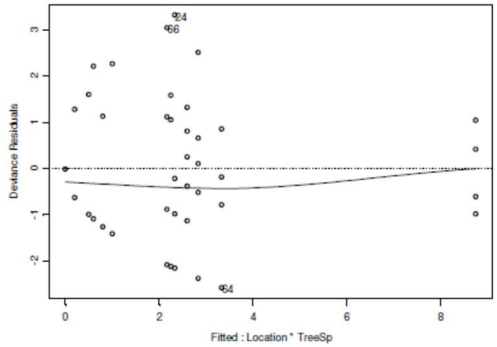
Loc	Syc	Birch

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

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



The diagnostic plots show heterogeneity and deviation from normal residuals. Sample size is large so estimates of likelihood ratios and p-values will be robust to violation of assumptions.



4. What is the evidence?

	Df	Deviance	seq df	Resid. Dev	Pr(Chi)
NULL			79	283.3562	
Location	7	71.34234	72	212.0139	7.90E-13
TreeSp	1	32.44743	71	179.5665	1.22E-08
Location:TreeSp	7	46.78186	64	132.7846	6.16E-08

$LR = e^{132.7846/2} > 10^{28}$ kilogigabyte scale evidence

4. Population. Units are plots, not trees. If plots were placed randomly in a defined area (the frame) then the population is all of the plots in the frame. If the plots were placed haphazardly, then the sample might still be taken as representative of the population in that area (frame). We may wish to infer, informally, to other locations (frames) with similar growing conditions.

5. Choose mode of inference. Is hypothesis testing appropriate?

Hypotheses testing appropriate, based on research questions, does relative abundance of sycamores and birch differ among woodlands? We report the evidence as a likelihood ratio, followed by a Type I error.

6. **State H_A / H_o**

   H_A : $\exists_{Loc*Sp} \neq 0$ $e^{\beta_{Loc*Sp}} \neq 1$ Count depends on interactive effect

   H_o : $\exists_{Loc*Sp} = 0$ $e^{\beta_{Loc*Sp}} = 1$ Count does not depend on interactive effect.

7. **ANODEV Table.** As above

8. **Assess p-values and estimates in light of evaluation of assumptions.**
   Assumptions were not met. Sample size was large, LR was very large, Type I error was very small. Randomization will produce a Type I error rate that does not depend on assumptions of a distributional model. We could also seek a better distributional model, such as negative binomial. Randomization or a better distributional model will change the *LR* and the Type I error estimate. They will not change a conclusion at conventional levels of Type I error.

9. **Statistical conclusion.**
   A difference in abundance of sycamores and birches among woodlands is far more likely than no difference. With a very low Type I error we can reject the hypothesis of difference due only to chance.
   ($\Delta G = 71.34$, df = 7, $p < 10^{-12}$)

10. **Science conclusion. Evaluate parameters of biological interest.**
    In this analysis only the interaction term was of interest. The test is indifferent to whether differences in woodlands result in differences in the relative abundance of the two tree species, whether differences in tree species result in different woodlands, or whether some other factor produces the observed relation. The analysis is thus similar to correlation, where the statistical analysis is mute on causal ordering.

Third Example. Poisson Response Variable. Two way classification. BACI design.

The BACI design (before-after control-impact) design is widely used in environmental impact assessment. Comparing a measured variable at a site impacted by some activity, such as release of effluent from a paper mill, is a natural approach. But unfortunately this approach is less then rigorous. This is because difference at the impacted site might be due to some peculiaritie of the site. This is called confounding. Eberhart (1976) suggested a paired approach, where before-after measurements at an impacted site are paired with before-measurements at a control (unimpacted) site with similar characteristics. Time (before versus after) and space (impacted, unimpacted) are both fixed factors. The approach was popularized by Stewart– Oaten *et al.* (1986) and became known as the BACI model, although it is better to refer to it as a BACI paired (BACIP) model to avoid confusion with the unpaired design. The design was extended to a Before-After Gradient design by Ellis and Schneider (1997).

For a physical variable, such as parts per million of a contaminant, we would use a two-way ANOVA design. For a count variable, such as number of organisms or number of species in sampled areas, we would use the same two-way design, but might well find that the residuals fan out in the cone shape characteristic of count data. We would then use the two way design in a GzLM with an appropriate error structure. Where variance in counts is approximately equal to the mean count, our first choice would be a Poisson error structure. Where the variance exceeds the mean, other error structures are appropriate. These include over dispersed Poisson, and negative binomial.

Bishop et al (1975) *Discrete Multivariate Analysis. Theory and Practice*, MIT Press.

Eberhardt, L.L. (1976). Quantitative ecology and impact assessment, Journal of Environmental Management 4: 27–70.

Fisher, R.A. 1925. Statistical Methods for Research Workers. Hafner.

1997 Ellis, J.I., D.C. Schneider. Evaluation of a gradient sampling design for environmental impact assessment. Environmental Monitoring and Assessment 48: 157-172.

McCullagh, P. and J. Nelder 1983. Generalized Linear Models. CRC Press.

Stewart-Oaten, A., Murdoch, W.W. & Parker, K.R. (1986). Environmental impact assessment: Pseudoreplication in time? Ecology 67: 929–940.