Model Based Statistics in Biology.

Part V. The Generalized Linear Model.

Chapter 18.4 Single Fixed Factor

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)
18 Binomial Response Variables
18.1 Logistic Regression (Dose-Response)
18.2 Single Factor. Prospective Analysis
18.3 Single Factor. Retrospective Analysis
18.4 Single Fixed Factor.
18.5 Single Explanatory Variable. Ordinal Scale.
18.6 Two Categorical Explanatory Variables
18.7 Logistic ANCOVA

Ch18.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). Generalized Linear Model. Poisson response variables.
ReCap (Ch 18). We used logistic regression to analyze the risk of cancer in smokers. This is called cross-sectional analysis because we compare individuals with different histories at a single point in time. This was called a retrospective analysis.

Today: Reformulating the response variable.
Binomial response across levels of a single fixed factor.

Wrap-up.

The analysis of counts in a two-way classification was more informative as a binomial response than as a Poisson response variable.

N.b.
Poisson analysis followed by Binomial tends to be confusing.
Replace with comparison of result to G^2 statistic from Poisson analysis, at step 10.
Replace example from Sokal and Rohlf with example from McCullagh and Nelder, using the Schoener lizard data to illustrate logistic regression for counts traditionally classified as 'log-linear'

Analysis of Binomial Response Variables. Example: Beetle counts.

Does the proportion of bright red tiger beetles change seasonally ?
(Sokal and Rohlf 1995 Box 17.8)
Data such as this are traditionally presented as a two-way classification.

	Bright red	Not bright red	
Early spring	29	11	40
Late spring	273	191	464
Early summer	8	31	39
Late summer	64	64	128
	374	297	671

We begin with the traditional approach, which stems from the two-way classification. Each of the 8 counts is considered a Poisson count in a two way design where we test for row-by column interaction.

1. Construct the model

Response variable. f = count of beetles in each season in two categories (8 such counts)

Explanatory variables. Ssn = season (4 categories, 3 parameters)
Clr = red or not red

$f = e^{(\beta_{ref})} e^{(\beta_{Ssn})} e^{(\beta_{Clr})} e^{(\beta_{Ssn*Clr})} + Poisson\,Error$ This is the alternative model

This model has two factors (season, colour) and three terms (season, colour, and season*colour).
$e^{\exists Ssn*Clr}$ is term in which we are interested, It is the change in proportion of beetles in each of 3 seasons, compared to the first season. $e^{\exists Ssn*Clr}$ stands for 3 cross-product ratios. To test this term we compare the fit to this model to the fit when the interaction term is dropped.

If we drop the interaction term, the model becomes.

$f = e^{(\beta_{ref})} e^{(\beta_{Ssn})} e^{(\beta_{Clr})} + Poisson\,Error$ This is the null model

This model has two factors and two terms. It is a null model in the sense that the interaction term is missing.

2. Execute.

The model underlying the Poisson approach is often not stated.
The analysis is traditionally completed by comparing the observed and expected count in each cell of the table.

```
MTB > set into c1
DATA> 29 273 8 64
MTB > end
MTB > set into c2
DATA> 11 191 31 64
MTB > end
```

```
MTB > let c3 = c1 + c2
MTB > let k1 = sum(c1) + sum(c2)
MTB > let c3 = c3/k1
MTB > let c4 = c3*sum(c1)/k1
MTB > let c5 = c3*sum(c2)/k1
```

Compute expected proportions in each cell of the two-way table.

```
MTB > print c4 c5

  ROW         C4         C5

    1   0.033227   0.026386
    2   0.385429   0.306076
    3   0.032396   0.025726
    4   0.106325   0.084435
```

Print the expected proportions p, one for each cell of the two-way table.

```
MTB > stack c1 c2 c6
MTB > stack c4 c5 c7
MTB > let c7 = c7*k1
MTB > let c8 = c6 - c7
MTB > name c6 'f' c7 'pN' c8 'res'
```

Compute the fitted values fits = pN.
The compute the residuals, based on the data equations:
Observed = Fits + Residual

Print the 8 data equations.

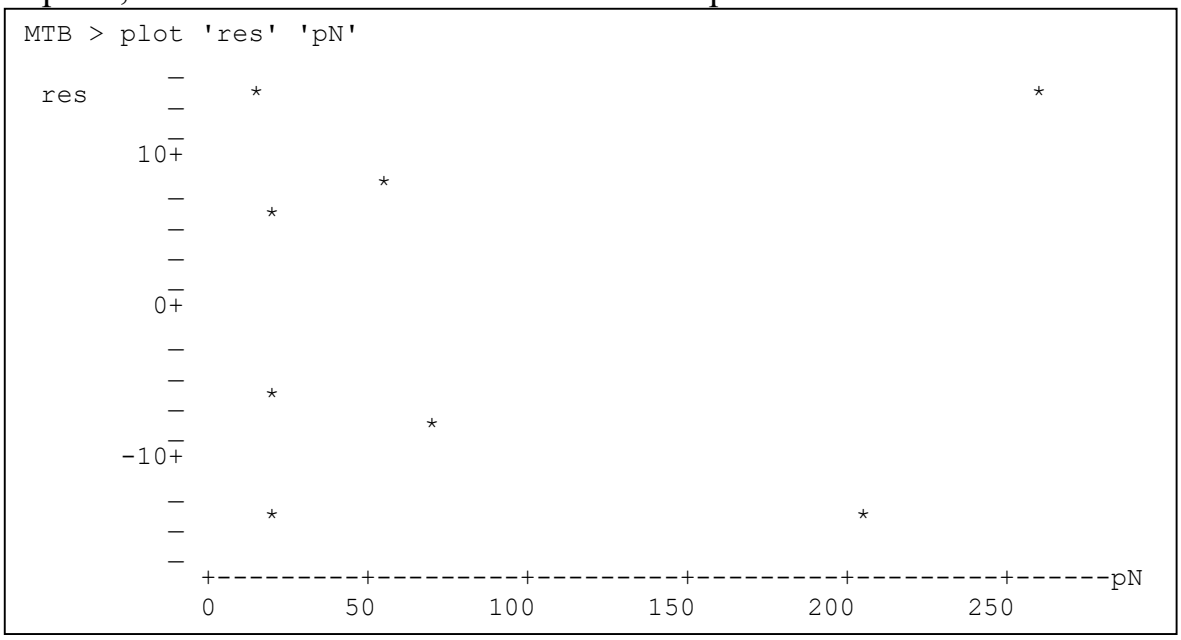
```
MTB > print c6 c7 c8

  ROW      f        pN       res

    1     29    22.295    6.7049
    2    273   258.623   14.3770
    3      8    21.738  -13.7377
    4     64    71.344   -7.3443
    5     11    17.705   -6.7049
    6    191   205.377  -14.3770
    7     31    17.262   13.7377
    8     64    56.656    7.3443
```

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

Structural model contains no regression lines, so no need to check bowl/arch. For generalized linear model we check the homogeneity assumption at this point, to make sure the binomial error assumption is correct.



We do not expect the residuals to be normal, because the data are counts, for which the variance usually increases along with the fitted value.

```
 MTB > hist 'res'

Histogram of res   N = 8

Midpoint   Count
     -15       2  **
     -10       0
      -5       2  **
       0       0
       5       2  **
      10       0
      15       2  **
```

A quick plot shows that the residuals, as expected, are not normal.

As a matter of interest, the likelihood ratios are not normal either.

```
MTB > hist '2lnL';
SUBC> increment 20.
Histogram of 2lnL   N = 8
Midpoint   Count
   -20.0       4  ****
     0.0       0
    20.0       3  ***
    40.0       1  *
```

Because we expect count data to generate non-normal errors, we have used a general linear model based on a non-normal error structure: a Poisson error structure in this case.

The plot of the residuals versus fitted values showed no pattern of expansion from left to right: no cones. Based on this, we are going to say that the residuals are acceptable for computing a p-value from the log likelihood ratio.

4. Population

We do not know the biological population from which the numbers were taken. Hence we cannot draw inferences from our sample to the entire population of tiger beetles. We are dealing with a hypothetical population consisting of all possible values that could have arisen via the stated sampling conditions. We can draw inferences only to the hypothetical population, as best we can define it.

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

Yes, we are interested in whether the interaction term (heterogeneity of proportions) is greater than by chance.

6. H_A H_o $\beta_{ssn*Clr} \neq 0$ Hence: $e^{\beta_{Ssn*Clr}} \neq 1$

$\beta_{Ssn*Clr} = 0$ Hence: $e^{\beta_{Ssn*Clr}} = 1$

statistic = G^2

probability distribution = chisquare

$\alpha = 5\%$

7. Analysis of Deviance

Compute log likelihood ratios from observed and fitted values.

```
MTB > let c9 = 'f'*log('f'/'pN')
MTB > let c9 = 2*c9
MTB > name c9 '2lnL'
MTB > print c6 c7 c8 c9

 ROW      f        pN       res      2lnL

   1     29    22.295    6.7049   15.2499
   2    273   258.623   14.3770   29.5389
   3      8    21.738  -13.7377  -15.9937
   4     64    71.344   -7.3443  -13.9051
   5     11    17.705   -6.7049  -10.4708
   6    191   205.377  -14.3770  -27.7233
   7     31    17.262   13.7377   36.2987
   8     64    56.656    7.3443   15.6019
```

Compute G^2-statistic = 2 · sum of the likelihood ratios.

```
MTB > let k2 = sum('2lnL')
MTB > print k2
K2       28.5964
```

7. Calculate improvement in fit (ΔG^2) due to explanatory variables.

ANOVA table is replaced by Analysis of Deviance table.

Source	df	Deviance = G^2	ΔG^2
Intercept $e^{\exists ref}$	1		
Ssn $e^{\exists Ssn*Clr}$	3		28.596

The degrees of freedom are computed exactly the same as was the case with the interaction term in a two-way ANOVA. The degrees of freedom for the interaction term are calculated as the product of the degrees of freedom for the main effects:

$df_{ssn*clr} = df_{ssn} * df_{clr} = (2–1)*(4–1) = 3.$

This is an example of how what we have learned is carried over to the analysis of frequencies.

The improvement is $\Delta G^2 = 28.59$ (compare Box 17.8)
Sokal and Rohlf (1995) show how to make an adjustment for continuity. Typically this adjustment will be small unless observed values less than 5 are common. There is little point in making this adjustment for continuity unless our calculated G^2-statistic has a p-value that is very close to the criterion α.

Compute p-value for the G^2-statistic.

```
MTB > cdf 28.5964;
SUBC> chisquare 3.

  28.5964     1.0000
```

The p-value is computed for three degrees of freedom. $df_{row} \cdot df_{col} = 1 \cdot 3$.

$p < 0.001$

8. Calculate randomized p-value ? Not needed if error structure appropriate.

9. Declare decision. $p < 0.001$ hence reject H_o and accept H_A

H_o: $p_{bR|early\ Sp} = p_{bR|late\ Sp} = p_{bR|early\ Su} = p_{bR|late\ Su}$

We accept the alternative hypothesis

H_A: proportions are not equal.

Frequency of bright red beetles depends on season.
($\Delta G^2 = 28.596$, df = 3, $p < 0.0001$)

Next we ask, what is the unit ?
The two columns are hard to justify as 'units.' Instead they are better viewed as binomial response (red / not red) in 4 seasons. Thus we have 671 units (beetles) scored as red or not. We rerun the analysis with binomial response variable.

1. Construct the model

Response variable. p = proportion of red beetles in each season.
Explanatory variables. Ssn = season (4 categories, 3 parameters)

$p = e^{(\beta_{ref})} e^{(\beta_{Ssn})} + Binomial\ Error$ This is the alternative model

$e^{\beta_{ssn}}$ is the change in proportion of beetles in each of 3 seasons, compared to the first season. $e^{\beta_{ssn}}$ stands for 3 cross-product ratios.
This model has 1 factor (season).

If we drop the season term, the model becomes.

$p = e^{(\beta_{ref})} + Binomial\ Error$ This is the null model (no change in proportion)

The difference between the null and alternative models is the improvement in fit.

2. Execute.

In this analysis, we can compute the residuals and the improvement in fit directly, The residuals are computed from fitted values for the second model (null model, no interaction term). These fitted values can be estimated from the marginal totals.

We will use information about the number of beetles collected in each season, and about the total number of each coloration, to estimate $\hat{f}$ for each of the 4 binomial counts.

	Bright red	Not bright red		p	p/p(ref)	f	fhat	f*ln(f/fhat)
Early sprinG²	29	11	40	0.73	1.00	29.00	22.29508	7.624962
Late spring	273	191	464	0.59	0.81	273.00	258.623	14.76946
Early summer	8	31	39	0.21	0.28	8.00	21.7377	-7.99685
Late summer	64	64	128	0.50	0.69	64.00	71.34426	-6.95257
	374	297	671	0.56				7.445001

	f	fhat	resid	f*ln(f/fhat)	G²
Early spring	29	22.3	6.7	7.625	
Late sprinG²	273	259	14.4	14.769	
Early summer	8	21.7	-13.7	-7.997	
Late summer	64	71.3	-7.34	-6.953	
				7.445	14.89

f	=	$\hat{p}$ · N	+ residual	lnL = 2 f · (f/$\hat{f}$)
29	=	0.56·40	+ 6.7	7.625
273	=	0.56·464	+ 14.4	14.769
8	=	0.56·39	− 13.7	−7.997 .
64	=	0.56·128	− 7.34	−6.953

$G^2 = 2 \sum f \cdot \ln(f/\hat{f}) = 14.89$

3. Evaluate model.

Structural model contains no regression lines, so no need to check bowl/arch.
For generalized linear model we check the homogeneity assumption at this point, to make sure the binomial error assumption is correct.
In this analysis we have 4 residuals, not enough to undertake evaluation.

4. Population As above.

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

6. **H_A / H_A** pair for season term.

H_A $\beta_{ssn} \neq 0$ Hence: $e^{\beta_{Ssn}} \neq 1$

H_o $\beta_{Ssn} = 0$ Hence: $e^{\beta_{Ssn}} = 1$

statistic = G^2
probability distribution = chisquare
$\alpha = 5\%$

7. Calculate improvement in fit (ΔG^2) due to explanatory variables.
ANOVA table is replaced by Analysis of Deviance table.

Source	df	Deviance = G^2	ΔG^2
Intercept $e^{\exists ref}$	1		
Ssn $e^{\exists Ssn}$	3 = 4–1		14.89

Calculate p-value from Chisquare distribution.
Is this improvement ΔG^2 better than by chance ?

```
MTB> cdf 14.89;
SUBC> chisquare 3.
  14.89     0.9999
```

More than 99.99% of the G^2-statistics obtained by chance will be less than our observed $G^2 = 14.89$,
The p-value reported for $\Delta G^2 = 14.89$ is $p < 0.0001$

Note ΔG^2 is less than for the log-linear analysis

8. Calculate randomized p-value ? Too few df to evaluate assumptions.
Assumptions assumed to be correct for binomial error.

9. Declare decision. $p < 0.0001$ hence reject H_o and accept H_A
H_o: $e^{\exists Ssn} = 1$
H_o: $p_{bR|early\ Sp} = p_{bR|late\ Sp} = p_{bR|early\ Su} = p_{bR|late\ Su}$
We accept the alternative hypothesis

H_o: $e^{\exists Ssn} \neq 1$
H_A: proportions are not equal.

Frequency of bright red beetles depends on season.
($\Delta G^2 = 14.89$, df = 3, $p < 0.0001$)

Next we evaluate the change in proportion in binomial data via odds ratios.

1. Construct the model

Response variable. p/(1-p) = odds of red beetles in each season.
Explanatory variables. Ssn = season (4 categories, 3 parameters)

$Odds = e^{(\beta_{ref})} e^{(\beta_{Ssn})} + Binomial\ Error$ This is the alternative model

$e^{\beta_{ssn}}$ is the change in odds of encountering a red beetles in each of 3 seasons, compared to the first season. $e^{\beta_{ssn}}$ stands for 3 cross-product ratios.
This model has 1 factor (season).
If we drop the season term, the model becomes.

$Odds = e^{(\beta_{ref})} + Binomial\ Error$ This is the null model

(no change in odds)

The difference between the null and alternative models is the improvement in fit.

2. Execute.

	Bright red	Not bright red	p/q
Early spring	29	11	2.6
Late spring	273	191	1.4
Early summer	8	31	0.26
Late summer	64	64	1

We use binomial error and logit link.

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

4. Population As above.

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

6. H_A / H_A pair for season term. As above

7. Calculate improvement in fit (ΔG^2) due to explanatory variables.

ΔG^2 from Minitab output, put into Analysis of Deviance Table.

Source	df	Deviance = G^2	ΔG^2
Intercept $e^{\beta_{ref}}$	1		
Ssn $e^{\beta_{Ssn}}$	3 = 4–1		28.596

Calculate p-value from Chisquare distribution.

Is this improvement ΔG^2 better than by chance ?
The Minitab command to compute the p-value is

```
MTB> cdf 14.89;
SUBC> chisquare 3.
  28.596   0.9999
```

More than 99.99% of the G^2-statistics obtained by chance will be less than our observed $G^2 = 28.596$,
The p-value reported for $\Delta G^2 = 28.596$ is $p < 0.0001$

Note that ΔG^2 is the same for this analysis as for the log linear model.

8. Calculate randomized p-value ? Not needed, error structure is appropriate.

9. Declare decision. $p < 0.0001$ hence reject H_o and accept H_A

10. Examine parameters of biological interest.

```
Logistic Regression Table
                                                        Odds        95% CI
Predictor           Coef    SE Coef        Z     P    Ratio    Lower    Upper
Constant          0.9694     0.3541     2.74 0.006
C2
 Early summer    -2.3239     0.5316    -4.37 0.000     0.10     0.03     0.28
 Late spring     -0.6122     0.3665    -1.67 0.095     0.54     0.26     1.11
 Late summer     -0.9694     0.3958    -2.45 0.014     0.38     0.17     0.82
```

Odds in early spring are lower than early summer (OR = 0.10) or
late summer (OR = 0.38)

Odds in early spring does not differ from late spring OR = 0.54, (confidence limits include 1.00).