Model Based Statistics in Biology.

Part V. The Generalized Linear Model.

Chapter 18.2 Single Factor. Prospective Analysis

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 Random Factor.
18.5 Single Explanatory Variable. Ordinal Scale.
18.6 Two Categorical Explanatory Variables
18.7 Logistic ANCOVA

on chalk board

Ch18.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). Generalized Linear Model. Poisson response variables.
ReCap (Ch 18) We analyze dose-response data with logistic regression, in which the response variable is the odds, and relation of odds to dose is exponential.

Today Binomial response variables. Single factor.
Prospective analysis of natural selection intensity.

Wrap-up.
We use the Generalized Linear Model (logit link, binomial error) to analyze count data consisting of units scored as 1 or 0 (e.g. present / absent, live / dead, *etc*).

The GzLM expands our analytic capacity by allowing explanatory variables beyond just the explanatory variable of interest.

We use the improvement in fit (ANODEV table) to evaluate GzLM, instead of the ANOVA table for sums of squares.

Prospective Analysis.

In a prospective analysis, we begin with cases, assign them to two or more explanatory categories, then at a later time examine whether an attribute changes. Cases are usually individuals in the medical applications for which this analysis was developed. Cases can also be spatial units. An example is crop productivity over time in farms that differ in agricultural practices. Another example is the BACI (before after control impact) design. In this design we take measurements at several sites before a proposed project, such as offshore oil platform. We compare changes at the impacted site to changes at unimpacted sites. Prospective studies are longitudinal. By following cases, they control confounding variables better than cross-sectional studies, which do not.

Prospective studies often produce binomial data. We score as present or absent an attribute, such as survival in a cohort or acquiring a new behaviour (learning). In binomial applications we analyze the change in odds across categories of the explanatory variable.

Example – Natural selection.

The strength of natural selection is measured by a selection gradient defined as the regression of a component of fitness (such as survival) on a trait (Lande 1979, Lande and Arnold 1983). Kettlewell (1956) used a mark recapture study to demonstrate natural selection on typical and melanic moths released in a woodland with soot on trees from local industry.
Kettlewell, H.B.D. (1956). Further selection experiments on industrial melanism in the Lepidoptera. *Heredity* 10: 287-301.

Kettlewell's results placed evolutionary theory on an experimental basis, and at the same time generated considerable controversy. For a recent account see:
Cook LM, Grant BS, Saccheri IJ, Mallet J (2012). "Selective bird predation on the peppered moth: the last experiment of Michael Majerus". Biology Letters 8 (4): 609–612. doi:10.1098/rsbl.2011.1136. PMC 3391436. PMID 22319093.

Survival Odds and Risk

Do melanic moths have a higher survival rate in a woodland where trees are covered with industrial soot? We begin by looking at % survival and risk (1- % survival)

	N release	N recapt	% recapture	recapture odds	odds ratio
Typical	201	34	17%	0.2036:1	
Melanic	601	205	34%	0.5177:1	2.54

How strong is the evidence? Is the difference too large to be due to chance?
We will use the GzLM to address this question. We will use the recapture odds instead of %recapture. The recapture odds are higher for melanic moths in the sooty woodland. Recapture odds = 205 / (601-205) = 0.5177:1 for melanic

Why do we use odds and odds ratios, rather than recapture % ?
We use the odds ratio because it is not affected by whether we look at survival (as measured by recapture) or look at risk (1 - survival).
The relative survival was 0.34 / 0.17 = 2.016.
The relative risk is RR = (1 - 0.34)/(1 - 0.17) = 0.6589 / 0.83085 = 0.79

	N release	N recapt	% survive	mortality risk	relative risk
Typical	201	34	17%	0.83085	
Melanic	601	205	34%	0.65890	0.79

Goodness of Fit Test

How strong is the evidence ? What is the Type I error on concluding that the observed selection gradient is more than just chance?
We begin with an analysis of the data as a classical goodness of fit test (*G*-test)

Do the odds of survival (as measured by recapture) increase for melanic moths in woodlands with soot on trees?

p = (34+205) / (201+601) = (239/802)

f	=	p	·	$N_{release}$+		residual	lnL
34	=	(239/802)	·	201	–	25.9	–19.25
205	=	(239/802)	·	601	+	25.9	+27.687
							8.437

The larger the difference in proportions, the worse the fit to the single ratio.
$LR = \exp^{(8.437)} = 4614$

Goodness of Fit Test

The evidence is strong. The difference in proportion is 4000 times more likely than no difference in proportion.
During the 20th century it became customary to evaluate the evidence with a probability statement.
$G^2 = 2 \cdot\ 8.437 = 16.87$
p = 1- cdf(16.87, chisquare, df = 1) = (1- 0.99996)
p = 0.00004. We reject H_0 of equal proportion.
We found that the G^2 -statistic was too large to be due to chance.
We concluded that the proportions differ.

Because they analysis is based on percentages, there is no simple relation between relative survival (2.016) and relative risk (0.79)

Odds Ratios

There is a simple relation between odds of recapture and odds of non-recapture
The non-recapture odds are Odds = Nnonrecapture / Nrecapture

```
             N           N        %          nonrecapture      odds
          release      recapt   recapture    odds              ratio
Typical     201          34       17%        4.9118:1
Melanic     601          205      34%        1.9317:1          0.393
```

The odds ratio is 1.9317 / 4.9118 = 0.393
The odds of nonrecapture (presumably due to mortality) are lower for the melanic form. They are 0.393 times the odds for the typical form.
The odds ratio for recapture (presumably due to better survival) is the inverse of the odds ratio for loss: $0.393^{-1} = 2.54$
The odds ratio is a convenient measure because there is a simple relation between odds of nonrecapture and odds of recapture.

However, we have analyzed the change in proportion, not the change in odds. The analysis of proportions will not necessarily match the analysis of odds.
To demonstrate the analysis of odds we will use the Generalized Linear Model.

1. Construct Model

Verbal. Recapture rate (survival) of melanic moths higher than typical moths in a woodland where trees are covered with soot.

Graphical Plot of recapture rate of 34% (melanic) versus 17% (typical)

Response variable: odds of recapture (same results if we use nonrecapture odds)

Explanatory variable: moth phenotype (2 levels in factor called Type)

Write formal model

Distribution $Nrecapture \sim Binomial(Nrelease, \pi)$

Link Odds = e^{η} .

$$\eta = \beta_o + \beta_{Type} Type$$

e^{β_o} = survival odds, typical form

$e^{\beta_{Type}}$ = odds ratio, melanic form relative to typical

$e^{(\beta_o + \beta_{Type})}$ = survival odds, melanic form

Note $\ln Odds = \beta_o + \beta_{Type}$ The model is linear on a logarithmic scale.

$\ln Odds$ This is the logit transformation of the proportion p

If we use the logit transformation, we have a linear model that looks like a t-test.We are comparing two categories, typical and melanic.

2. Execute analysis.

Place data in model format:

Binomial response variable in two columns, success and trials

Column labelled Recapt, with response variable # of recaptures (successes)

Column labelled Release, with response variable # of releases (trials)

Column labelled Type, with explanatory variable Type (melanic or not)

```
Data A;
  Input Recapt Release Type $;
  Cards;
    201  34 typical
    601 205 melanic;
```

SAS command file

In a package with spreadsheet format, there will be two lines with three variables.

```
MTB > print c1-c4
 Row   Type  Success  Trial   %Success
   1      1       34    201   0.169154
   2      2      205    601   0.341098
```

Minitab command lines

2. Execute analysis.

Code the model statement in statistical package according to the GzLM

```
MTB > BLogistic 'Recapt' 'Release' = Type;
SUBC>   ST;
SUBC>   Logit;
SUBC>   Brief 2.
```

Minitab command lines

```
Click Stat
   Click Regression
     Click Binary Logistic Regression
        Click Success, place column of recaptures,
        Click trials, place column of releases
        Click Model, place column with categories
        Click Storage (optional) Click Pearson residuals, Event
probability, ok
```

Minitab sequence to produce line commands

```
Proc Genmod; Classes Type;
  Model Recapt/Release = Type/
  Link=logit dist=binomial type1 type3;
```

SAS command file

3. Use residuals to evaluate model.

We cannot plot residuals versus fitted values. In this example there are two observations, two fitted values, and two residual values.
The residuals are zero because the two parameters fully describe the two observations.

Straight line assumption not applicable.

Error assumptions when using χ^2 distribution. In this example we have no residuals because there are as many parameters as there are data equations. The model is "saturated."

4. What is the evidence?

Source	df	Deviance = G^2	ΔG^2
Intercept $e^{\$o}$ (typical)	1	22.98	
Type $e^{\$type}$	n–1 =1	0	22.98

The improvement in fit is ΔDeviance = 22.98
The likelihood ratio is $e^{22.98/2} = 9.8 \times 10^4$ The evidence is very strong.

5. Choose mode of inference.

At the time that Kettlewell did the study evolutionary change was thought to be a slow process that did not lend itself to measurement during a short period. With only a few exceptions there were no measurements of the strength of natural selection. Nor were there any theoretical models upon which to establish inference from a prior probability. Because this was a field study, manipulative control of variation was not possible. Nor was there any need to control Type I error in the face of economic costs or risks. The protocol was well defined, allowing inference to an infinite number of repeats of the protocol. Inference to all moths in this woodland was also possible. Mark-recapture results as in this study can be used to estimate the population size, given a second visit to recapture moths, and some assumptions (Seber, G.A.F. 1973. *The Estimation of Animal Abundance and Related Parameters.* Griffin Press). We will use direct likelihood inference in conjunction with estimates of the effect size (strength of natural selection).

6. State reduce ($\mathbf{H_A}$) and unreduced ($\mathbf{H_0}$) models.

H_A: $dev(\beta_{Type}) > 0$ hence: $OR = e^{\beta_{Type}} \neq 1$

H_o: $dev(\beta_{Type}) = 0$ hence: $OR = e^{\beta_{Type}} = e^0 = 1$

Statistic = ΔG^2, the improvement in fit due to explanatory variable (two groups)

7. ANODEV - Calculate change in fit (ΔG^2) due to explanatory variables.

For the generalized linear model, we calculate the deviance rather than the variance. The deviance is measured by the *G*-statistic.

We examine whether the deviance is reduced by adding an explanatory variable to the model. The change in deviance ΔG is tabled for each explanatory variable in the model. Here is the Anodev table from SAS.

```
                         LR Statistics For Type 1 Analysis

                                                       Chi-
           Source          Deviance        DF        Square     Pr > ChiSq
Ho         Intercept        22.9767
HA         type              0.0000         1         22.98         <.0001
```

SAS output

The AnoDev table shows 1 df for the intercept and 1 df for the model term.

The chisquare column is Δ*G*, the change in the non-Pearsonian Chisquare, *G*.
The goodness of fit of the data to the null model is $G^2 = 22.9767$
The fit of the data to the alternative model is perfect $G^2 = 0.00$
The improvement is $\Delta G^2 = 22.98$

7. ANODEV

Not all packages have a generalized linear model routine but many have a logistic regression routine. Here is the output from the Minitab logistic regression routine.

```
Logistic Regression Table
                                                     Odds        95% CI
Predictor        Coef    SE Coef          Z      P   Ratio    Lower    Upper
Constant     -1.5916     0.1881     -6.54 0.000
Type          0.9332     0.2069      4.51 0.000      2.54     1.70      3.81

Log-Likelihood = -477.062
Test that all slopes are zero: G = 22.977, DF = 1, P-Value = 0.000
```

Minitab output

Instead of an analysis of deviance table, we see the odds ratio (2.54) and the change in deviance (*G* = 22.977) associated with this term.

The more the odds ratio differs from 1, the larger the value of *G*.
The Minitab output shows Type I error at less than 10^{-4}, assuming an infinite number of repeats and from that the use of a normal error as an approximation.

8. If assumptions not met, decide whether to recompute p-value.

Type I error can also be calculated by randomization.

9. Statistical conclusion.

The odds of survival for the melanic form are 2.54 times that of the typical form.

10. Science conclusion. Interpreting the parameters.

```
                                         Standard
 Parameter                DF    Estimate     Error     Confidence Limits

 Intercept                 1     -1.5916    0.1881     -1.9604    -1.2229
 type          melanic     1      0.9332    0.2069      0.5277     1.3387
```

SAS output file

Generalized linear model routine (SAS) treats the first class listed as the intercept.

$e^{\beta_o} = e^{-1.5916} = 0.2036$ Survival odds, typical moth

$e^{\beta_{Type}} = e^{0.9332} = 2.54$ Odds ratio, melanic moth relative to typical

$e^{\beta_o + \beta_{Type}} = e^{-1.5916+0.9332} = 0.51767$ Survival odds, melanic moth

Minitab logistic regression routine produces the same parameter estimates (above).

The intensity of selection (measured by the odds ratio) is called the selection gradient. The selection gradient in this experiment is estimated at 2.54, with confidence interval of 1.7 to 3.81 ($e^{0.5277} = 1.7$, $e^{1.3387} = 3.81$)

Binomial Frequencies -- Prospective Analysis.
Comparison of three proportions

The example comes from data in Box 17.16 (p 782) of Sokal and Rohlf (1995). The response variable is the number of acacia plants free of recent damage (scoring positive) in three successive months, after the removal of ants that normally protect the plants from phytophagous insects.

Does the number of plants free of damage N_{free} decline with time (T = March, June, August)?

The explanatory variable is month, in three classes. This is a prospective analysis. It is an experiment where we start with a known number of cases, then score those cases as having or lacking some attribute.

1. Construct Model

Verbal. N is number of acacia trees (24)

N_{free} is number free of pest damage in March, June, and August

The number of plants without damage will decrease after removal of ants.

The odds of being free of damage will decrease at later times.

Graphical Plot of percent trees scoring positive, against time.

Response variable: odds of having damage

Explanatory variable: time (3 categories).

Write formal model $Odds = e^{(\beta_0)}e^{(\beta_t)}$

$e^{\beta_0} =$ odds having damage, at time zero.

$e^{\beta_t} =$ odds ratio, at later times

$e^{(\beta_0+\beta_t)} =$ odds of having damage at later times

2. Execute analysis.

Place data in model format for generalized linear model routine:

Binomial response variable in two columns, success and trials
Column = Nfree, with response variable # of trees free of damage (successes)
Column = Ntotal, with response variable # of trees (trials)
Column labelled Time, with explanatory variable Time = March, June, August

```
Data A;
  Input Ntotal Nfree Time $;
  Cards;
    24 15 March
    24 12 June
    24  4 August
 ;
```

SAS command file

2. Execute analysis. Place data in model format for logistic regression routine:

Column for success (Nfree)
Column for trials (Ntot)
Column for 2 of the 3 levels of the categorical variable time.

```
MTB > print c1-c4

 Row  Nfree   Ntot   June  August

   1     15     24      0       0
   2     12     24      1       0
   3      4     24      0       1
```

Minitab format

Code the model statement in statistical package according to the GzLM

$$\ln Odds \ = \beta_o + \beta_t$$

$$\ln\left(\frac{p}{1-p}\right) \ = \beta_o + \beta_t$$

$$\ln\left(\frac{Nfree}{N - Nfree}\right) = \beta_o + \beta_t$$

```
Proc Genmod; Classes Time;
  Model Nfree/Ntotal = Time/
  Link=logit dist=binomial type1 type3;
```

SAS command file

```
MTB > BLogistic 'Nfree' 'Ntot' =  'June' 'August';
SUBC>   ST;
SUBC>   Logit;
SUBC>   Brief 2.
```

Minitab command lines

2. Execute analysis

Fits and residuals.
In this example there are three fitted values (one for each of three observations)
The residuals are zero (the three parameters describe the three observations).
Fitted values from model output.

```
                                             Standard
 Parameter             DF    Estimate         Error      Confidence Limits

 Intercept              1      0.5108        0.4216     -0.3156      1.3372
 Time         August    1     -2.1203        0.6912     -3.4750     -0.7655
 Time         June      1     -0.5108        0.5869     -1.6611      0.6395
 Time         March     0      0.0000        0.0000
```

SAS output

```
Logistic Regression Table
                                                       Odds         95% CI
Predictor          Coef     SE Coef          Z       P     Ratio     Lower     Upper
Constant         0.5108      0.4216       1.21 0.226
June            -0.5108      0.5869      -0.87 0.384      0.60      0.19      1.90
August          -2.1203      0.6912      -3.07 0.002      0.12      0.03      0.47
```

Minitab output

$e^{\beta_0} = e^{0.5108} = 1.67$ Odds of no damage, March

$e^{\beta_0+\beta_T} = e^{0.5108-2.12026} = 0.2$ Odds ratio, June relative to March

$e^{\beta_0+\beta_T} = e^{0.5108-2.12026-0.5108} = 0.12$ Odds ratio, August relative to March

Fitted values by direct computation

```
Ntot  Nfree                      Odds         OR      lnOR
24    15     (15/24) / (9/24) = 1.67         1        0.0        March
24    12     (12/24) / (12/24) = 1.0         0.6     !0.5108     June
24    4      (4/24) / (20.24) = 0.2          0.12    !2.12026    August
```

3. Evaluate model.

a. No straight line assumptions, so no need to check.

b. Residuals are equal to zero, so cannot check.

Too few data equations to check assumptions.

Residuals equal to zero because there are as many parameters as observations (rows in the spreadsheet).

Assumption of binomial error considered appropriate for binomial response variable.

4. What is he evidence?

Full (unreduced model)	Deviance = 11.77
Reduced model	Deviance = 0 (saturated model)
	ΔDeviance = 11.77
LR = $e^{11.7668/2}$ = 359	Strong evidence for reduction in number of leaves free of damage

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

The odds change from month to month. In the absence of a defensible prior probability, or a need to control Type I error, we will use direct likelihood inference.

Population.

All possible measurements on acacia trees in the study area during 3 months.

6. State reduced (H_A) and unreduced (H_o) pair.

H_A: $dev(\beta_t) > 0$ hence: $OR = e^{\beta_t} \neq 1$

H_o: $dev(\beta_t) = 0$ hence: $OR = e^{\beta_t} = e^0 = 1$

7. ANODEV - Calculate improvement in fit (Δ*G*) due to explanatory variables.

ANOVA table is replaced by Analysis of Deviance table.

Source	df	Deviance = G	ΔG
Intercept e^{β_o}	1		
Time e^{β_t}	2=3–1		

```
                         LR Statistics For Type 1 Analysis

                                                          Chi-
         Source             Deviance          DF        Square     Pr > ChiSq
Ho       Intercept           11.7668
HA       Time                 0.0000           2         11.77         0.0028
```

SAS output

The goodness of fit of the null model to the data is G = 11.77

The fit of the alternative model to the data is perfect G = 0.00

The improvement is ΔG = 11.77

7. ANODEV – Poisson model

The change in deviance ΔG for the binomial model (logit link) differs from ΔG statistic for equal proportions (Poisson error with log link).

H_o is $N_{free,March} = N_{free,June} = N_{free,August}$

Equivalently: $p_{March} = p_{June} = p_{August}$

H_A: N_{free} not equal among months.

$e^{\$} = (15+12+4)/(24+24+24) = 0.4306$

N_{free}	=	$e^{\$}N$	+	residual	$\ln L = f \ln(f/e^{\$}N)$
15	=	$0.43^{\cdot}24$	+	residual	5.590
12	=	$0.43^{\cdot}24$	+	residual	1.794
4	=	$0.43^{\cdot}24$	+	residual	–3.796
				$\sum f \ln(f/e^{\$}N)$	3.588
$G = 2 \sum f \ln(f/e^{\$}N)$				7.176	

The Poisson error model is less likely than the binomial error model, given the data

8. If assumptions not met, decide whether to recompute p-value.

Binomial error considered appropriate (from design).

9. Statistical conclusion.

Odds of no damage depend on month. ΔG = 11.77, df = 2, LR = 359

10. Analysis of parameters of biological interest

The following table shows that the decrease is from March to June (G = 9.41, df = 1) with no change from June to August (G =1.47)

SAS output

```
                                         Standard
Parameter              DF    Estimate       Error   Chisquare   Pr > ChiSq

Intercept               1      0.5108      0.4216    1.47           0.2257
Time        August      1     -2.1203      0.6912    9.41           0.0022
Time        June        1     -0.5108      0.5869    0.76           0.3841
Time        March       0      0.0000      0.0000
```