

## 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

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 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.

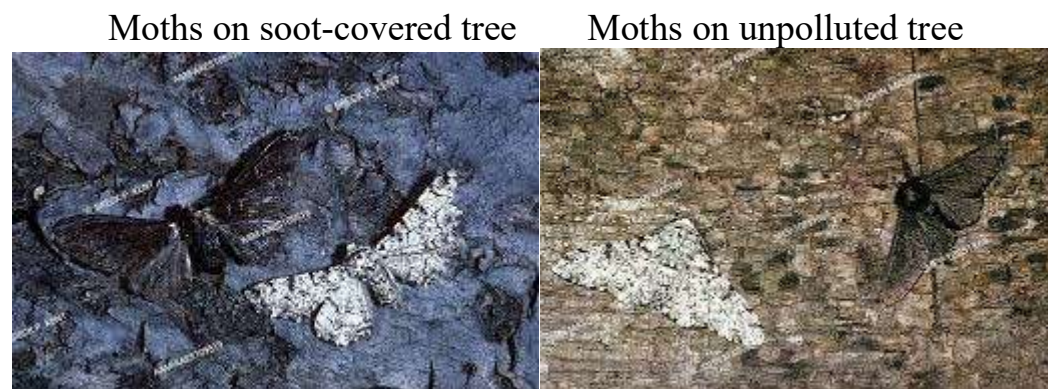
In a prospective (also called longitudinal) study we follow cases through time. We begin with cases, assign them to two or more explanatory categories, then examine whether an attribute changes. Cases are usually individuals, as in the medical applications for which this analysis was developed. Cases 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 for environmental impact assessment. In this design we take measurements before a proposed project, such as an offshore oil platform. We compare changes at the impacted site to changes at the unimpacted sites. Prospective studies control confounding variables better than cross-sectional studies.

Prospective studies often produce binomial data. We score an attribute such as survival in a cohort or acquiring a new behavior (learning) as present or absent. In binomial applications we take the number of successes as a proportion of the number of trials. We analyze the change in odds (proportion of success / proportion of failures) 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 (1955) 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. (1955). *Heredity*: 9:323–342.



Kettlewell's results placed evolutionary theory on an experimental basis, and at the same time generated considerable controversy. For an account see:

Cook LM, Grant BS, Saccheri IJ, Mallet J (2012). *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 comparing survival odds of melanistic and typical moths in a soot polluted woodland in Birmingham.

Data from Table 1 in Kettlewell (1955).

	N release	N recapt	% recapture	recapture odds	odds ratio
Typical	137	18	13%	0.151:1	
Melanic	447	123	28%	0.380:1	2.51

The odds for Typical relative to Melanic are readily obtained as the inverse:  $OR = 1/2.51 = 0.398$ . In contrast the 13% recapture for Typical is not so readily obtained from the 28% recapture for Melanic.

The non-recapture  $OR$  for Typical is  $OR = 2.634 / 6.6118 = 0.393$ .

	N release	N recapt	% recapture	nonrecapture odds	odds ratio
Typical	137	18	13%	2.634:1	
Melanic	447	123	28%	6.611:1	0.398

The non-recapture (loss) odds ratio is the same as the recapture  $OR$  for Melanic

Why do we use odds and odds ratios, rather than recapture percentages ?

McCullagh and Nelder (1989) list two reasons.

First, the analytic results are readily interpreted. The survival odds for melanic moths were 2.5 times that of typical moths. The odds ratio for Typical is the inverse of Melanic. Similar statements can be made about the odd ratios for the first and subsequent variable in an analysis. However, statements about subsequent variables become complicated if we use proportions instead of odds.

A second reason for using odds is that differences on the logistic scale can be estimated regardless of whether the data are sampled prospectively or retrospectively. This is an important property of the logistic (log odds) function not shared by probability scale functions (probit or log-log). This is an advantage in medical applications because prospective studies can take years to accumulate sufficient data for drawing inferences from the data.

## Relative Risk

The relative survival is  $0.28 / 0.13 = 2.094$ . The inverse is 0.477

The relative risk is  $RR = (1 - 0.28)/(1 - 0.13) = 0.79$

	N release	N recapt	% survive	mortality risk	relative risk
Typical	137	18	13%	0.725	
Melanic	447	123	28%	0.869	1.198

In this example the *OR* value (2.51) differs from the relative survival (2.094). When the risk is rare (typically <10%) and the value of *OR* is not too different from that of *RR*, the two can be used interchangeably. In medical applications the relative risk (of disease) is more intuitively understood. In the natural selection example survival odds are more intuitively understood.

### 1. Construct Model

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

Graphical. Plot of recapture rate of 34% (Melanic) versus 17% (Typical) is not particularly informative.

Response variable: Odds of recapture. If we use the non-recapture odds we will obtain the same goodness of fit statistic.

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

Write formal model in 3 part form.

Distribution  $N_{recapture} \sim \text{Binomial}(N_{release}, \pi)$

Link  $Odds = e^{\eta}$

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

$e^{\beta_o}$  = 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 additive on a logarithmic scale.

If we use the logit link, we have a linear model that compares odds in two categories, Typical and Melanic. It is similar to a t-test, which compares two means.

## 2. Execute analysis.

The data consist of zeros and ones. Statistical packages accept data in this form. They also accept data in condensed form, as trials and successes.

In this latter form we have 3 columns.

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 Typical)

```
Data A;  
  Input Recapt Release Type $;  
Cards;  
  137  18 typical  
  447 123 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   Melanic    18    123    0.131  
  2   Typical   123    447    0.275
```

The variable names are used to code the model statement in the statistical package.

```
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

```
MothMod <- glm(Recapt/Release ~ Type,  
  family = binomial(link="logit"),weights=Release, data = Moths
```

R code

### 3. Use residuals to evaluate the error model.

In this example we are unable to plot residuals versus fitted values.

There are two observations, two fitted values.

The residuals are zero because the two parameters fully describe the two observations. In this example we have no residuals because there are as many parameters as there are data equations. The model is “saturated.”

### 4. How good is the evidence?

#### Goodness of Fit Test

We begin with an analysis of the data as a classical goodness of fit test ( $G$ -test).

Does percent survival for melanic and typical moths (as measured by recapture) differ in woodlands with soot on trees? Overall survival =  $\hat{p} = 141/584 = 24\%$ .

We calculate the log likelihood  $\ln L$  from  $f$  and  $\hat{f}$  the observed and expected frequencies.

	$\hat{p}$	=	$(18+123) / (137+447) = (141/584)$		
$f$	=	$\hat{f}$	+	residual	$\ln L$
$f$	=	$\hat{p} \cdot N_{release}$	+	residual	$f \ln(f/\hat{f})$
18	=	$(0.241) \cdot 137$	+	-15.07	-10.952
123	=	$(0.241) \cdot 447$	+	15.07	<u>+16.084</u>
					5.132

The larger the difference in proportions, the worse the fit to the overall ratio.

The goodness of fit statistic is  $G = 2 \ln L$ .  $G = 10.26$  based on the raw residuals.

While raw residuals are appropriate for analysis with normally distributed error, they are not appropriate for other distributions (see McCullagh and Nelder 1989 p39). For binomial data, we will use analysis of deviance to evaluate the strength of evidence.

Source	df	Deviance = $G^2$	$\Delta G^2$
Intercept $e^{\beta_0}$ (typical)	1	13.033	
Type $e^{\beta_{type}}$	n-1 =1	0	13.033

The improvement in fit is  $\Delta \text{Deviance} = 13.03$

This estimate, based on the deviance, differs from that based on raw residuals.

The likelihood ratio is  $LR = e^{13.033/2} = 676$

We can say the evidence is strong. ( $100 < LR < 1000$ ).

## 5. Choose mode of inference.

At the time that Kettlewell did the study evolutionary change was thought to be a process too slow to measure 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. Nor was there any need to control Type I error in the face of economic costs or risks. The mark-recapture protocol was well defined, allowing inference to an infinite number of repeats of the protocol. However, the conditions for the study were not repeatable. Examples of uncontrolled variables include degree of soot cover and number of years moths were exposed to sooty woodlands.

Inference to all moths in this woodland at the time of the study is plausible. Population size can be estimated with mark-recapture results, given a repeated visits to recapture moths, and some assumptions (Seber, G.A.F. 1973. *The Estimation of Animal Abundance and Related Parameters*. Griffin Press). We will use the *LR* to infer from the sample to a defined population, assuming the sample is representative.

## 6. State reduced ( $H_A$ ) and unreduced ( $H_0$ ) models.

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

$$H_0: \quad dev(\beta_{Type}) = 0 \quad \text{hence:} \quad OR = e^{\beta_{Type}} = e^0 = 1$$

Statistic =  $\Delta G^2$ , often shortened to  $G$ , the improvement in fit due to the explanatory variable, a factor having two groups.

## 7. ANODEV – How certain can we be?

Many packages have logistic regression routines. Here is the output from the Minitab routine.

Logistic Regression Table							
Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI	
						Lower	Upper
Constant	-1.8888	0.2529	-7.468	0.000			
Type	0.9202	0.2742	3.356	0.000	2.51	1.5	4.42
Test that all slopes are zero: G = 13.033, DF = 1, P-Value = 0.000							

Minitab output

The table reports the change in deviance ( $G = 13.033$ ) along with odds ratio (2.51) and confidence limits.

Here is the ANODEV table from a generalized linear model routine in SAS.

LR Statistics For Type 1 Analysis					
	Source	Deviance	DF	Chi-Square	Pr > ChiSq
H <sub>0</sub>	Intercept	13.033			
H <sub>A</sub>	type	0.0000	1	13.033	<.0001

SAS output

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

Model checking with residual analysis was not possible for this example because it was a saturated model with no residuals.

## 9. Statistical conclusion.

The evidence for increased odds of survival for the melanic form relative to typical was strong (LR = 676). with a high degree of certainty ( $p < 0.001$ ) assuming inference to a finite population of moths at the time of the study,



## 10. Science conclusion. Interpreting the parameters.

Parameter	DF	Estimate	Standard Error	Confidence Limits
Intercept	1	-1.8888	0.2529	-2.4178 -1.4205
type Melanic	1	0.9202	0.2742	0.4053 1.4859

SAS output file

Generalized linear model routines differ in how the intercept is chosen.

In this case the default was Typical. Parameter estimates are back-calculated from the estimates.

$$e^{\beta_o} = e^{-1.888} = 0.151 \quad \text{Survival odds, typical moth}$$

$$e^{\beta_{Type}} = e^{0.9202} = 2.51 \quad \text{Odds ratio, melanic moth relative to typical}$$

$$e^{\beta_o + \beta_{Type}} = e^{-1.888 + 0.9202} = 0.380 \quad \text{Survival odds, melanic moth}$$

The Minitab logistic regression routine produces the same parameter estimates as those above.

The intensity of selection (measured by the odds ratio) is called the selection gradient. The selection gradient in this experiment is estimated at 2.51, with confidence interval of 0.4053 to 1.486 on the log odds scale, 1.5 to 4.42 on the odds scale.

Your turn

Kettlewell (1956) followed up the 1955 publication with selection experiments in an unpolluted wood. He released 406 *B.b.carbonaria* and 393 *B.b.typical*.

Recaptures were 19 and 54 respectively. State your research hypothesis in words and then in symbolic notation. Table the data, with odds and odds ratio.

	Nreleased	Nrecapture
B.b. carbonaria	406	19
B.b. typica	393	54

Use a binomial error to calculate the likelihood ratio, then use it to make a statement about your research hypothesis.

Kettlewell, H.B.D (1956). Further selection experiments on industrial melanism in the Lepidoptera. *Heredity* 10: 287–301.

## 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.

Ntotal	Nfree	Time
24	15	March
24	12	June
24	4	August

Does the number of plants free of damage  $N_{\text{free}}$  decline with time ?

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)  
                        $N_{\text{free}}$  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 =  $N_{\text{free}}$ , with response variable # of trees free of damage (successes)  
 Column =  $N_{\text{total}}$ , 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
  ;

```

## 2. Execute analysis.

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

SAS command file

```
AntMod <- glm(Nfree/Ntotal ~ Time,  
  family = binomial(link="logit"), weight=Ntotal, data = Ants
```

R code

The Minitab logistic regression routine uses a different format for data.

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

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

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

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

```
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.

Parameter		DF	Estimate	Standard 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								
Predictor	Coef	SE Coef	Z	P	Odds Ratio	95% CI		
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

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

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

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

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.

- No straight line assumptions, so no need to check.
- Residuals are equal to zero, so cannot check.

Data equations are too few to check assumptions.

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

Assumption of binomial error are considered appropriate for the binomial response variable.

### 4. What is the evidence?

Full (unreduced model)

Deviance = 11.77

Reduced model

Deviance = 0 (saturated model)

$\Delta$ Deviance = 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 ( $H_0$ ) pair.**

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

$$H_0: \text{dev}(\beta_t) = 0 \text{ hence: } OR = e^{\beta_t} = e^0 = 1$$

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

ANOVA table is replaced by Analysis of Deviance table.

<u>Source</u>	<u>df</u>	<u>Deviance = G</u>	<u><math>\Delta G</math></u>
Intercept $e^{\beta_0}$	1		
Time $e^{\beta_t}$	2=3-1		

LR Statistics For Type 1 Analysis					
	Source	Deviance	DF	Chi-Square	Pr > ChiSq
$H_0$	Intercept	11.7668			
$H_A$	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  $\Delta G = 11.77$

## 7. ANODEV – Poisson model

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

$$H_0: N_{\text{free, March}} = N_{\text{free, June}} = N_{\text{free, August}}$$

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

$$H_A: N_{\text{free}} \text{ not equal among months.}$$

$e^3 = (15+12+4)/(24+24+24) = 0.4306$					
$N_{\text{free}}$	$=$	$e^3 N$	$+$	residual	$\ln L = \sum f \ln(f/e^3 N)$
15	$=$	$0.43 \cdot 24$	$+$	residual	5.590
12	$=$	$0.43 \cdot 24$	$+$	residual	1.794
4	$=$	$0.43 \cdot 24$	$+$	residual	-3.796
$\sum f \ln(f/e^3 N)$					3.588
$G = 2 \sum f \ln(f/e^3 N)$					7.176

The Poisson error model is one-tenth as less likely as the binomial error model given the data.  $LR = \ln(7.176/2 - 11.77/2) = 0.1$

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

Binomial error is considered appropriate (from design).

## 9. Statistical conclusion.

Odds of no damage depend on month.  $\Delta G = 11.77$ ,  $df = 2$ ,  $LR = 359$

## 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

Parameter	DF	Estimate	Standard 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		