Model Based Statistics in Biology.
Part V. The Generalized Linear Model.
Chapter 18.6   Two categorical explanatory variables.

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).   Response variables of interest in the natural and social sciences are often binomial: a series of trials (cases) that can be scored as yes/no, present/absent, etc.

We compared a binomial proportion to an expected proportion derived from theory.

We compared two proportions in both prospective (experimental) and retrospective (observational) analyses.

We compared several proportions in a one way design.

We can extend this to other designs, such as
  -more than on classification variables
  -more than one regression variable
  -mixtures of classification and regression variables (ANCOVA)

Today: Comparison of binomial proportions in a two-way design.

Wrap-up. The generalized linear model permits us to apply what we have learned about two-way classification of data to the analysis of binomial response variables.

Two-way classification of a binomial response variable.

Example from Sokal and Rohlf (2012)
Box 17.10, p760
Data are counts of fruit flies in an
experiment to test the effects of the
insecticide DDT added to growth medium.

Does pupation depend on pupa site (in
medium, on margin, on top of medium, away
from medium on wall)?
Do male and female flies differ in pupation
success?

```
Data a;
  Input Emerge Pupae Site $ Sex $;
  Cards;
    55 61 IM F
    34 51 IM M
    23 24 AM F
    15 20 AM M
     7 11 OW F
     3  8 OW M
     8 11 OM F
     5  8 OM M
;
```

Sokal and Rohlf analyzed this data as Poisson counts classified by 3 factors: sex, site, and healthy (pupated) or not (poisoned). The experiment appears to be cross-sectional (one point in time) in the 3-way classification of counts. They also analyzed it as a binomial response (pupation or not) versus two factors, site and sex. The data become longitudinal (prospective) if we take the time period as that from pupation to emergence. The failed and successful pupae together consist of the trials (beginning of pupation), which are then scored as healthy (emerging) or poisoned (not emerging). This assumes that all flies that fail to emerge were poisoned by DDT, rather than failing to emerge because of some other factor.

1. Construct the Model

Verbal.   Survival of fruit flies depends on pupation site, sex, and interactive effects of these two factors.

Graphical   Plot of survival rate of flies relative to site and sex.
          Response variable: odds of survival
          Explanatory variables: pupation site and sex

Write formal model

$$Odds = e^{\eta} + binomial \; \varepsilon$$
$$\eta = \beta_{intercept} + \beta_{Site}Site + \beta_{Sex}Sex + \beta_{Site \cdot Sex}Site \cdot Sex$$

$e^{\beta_{intercept}}$ = survival odds, reference group (male, on wall OW)

$e^{\beta_{site}}$ = odds ratio, other 3 sites relative to reference group (male, OW)
          other sites are at margin AM, in medium IM, on medium OM

$e^{\beta_{sex}}$ = odds ratio, female (OW) relative to ref. group (male, OW)

$e^{\beta_{sex*site}}$ = odds ratio, female relative to male at other 3 sites (not OW)

$\ln Odds = \beta_{ref} + \beta_{site} + \beta_{sex} + \beta_{site*sex}$ The model is linear on a logarithmic scale.

1. Construct the Model

The structural model is the same as that for a a 2-way ANOVA.
The model has 8 parameters and 8 data equations, hence the error term will be zero for the model with all 4 terms. This is called a saturated model. The model fits the data perfectly. This does not prevent us from evaluating the <u>improvement</u> in fit for each term.

2. Execute model.

We specify the error structure (binomial), the link (logit), and the structural model.

$$\eta = \beta_{intercept} + \beta_{Site}{\cdot}Site + \beta_{Sex}{\cdot}Sex + \beta_{Site{\cdot}Sex}{\cdot}Site \cdot Sex$$

```
Proc Genmod;
  classes Site Sex;
  model Emerge/Pupae = Site Sex Site*Sex/
  link=logit dist=binomial type1 type3;
```

<div align="right">SAS command file</div>

Fitted values from parameter estimates.

```
                     Analysis Of Parameter Estimates

                                       Standard    Wald 95% Confidence
    Parameter            DF   Estimate    Error          Limits

    Intercept             1    -0.5108    0.7303    -1.9422      0.9205
    Site        AM        1     1.6094    0.8944    -0.1436      3.3625
    Site        IM        1     1.2040    0.7884    -0.3413      2.7492
    Site        OM        1     1.0217    1.0328    -1.0026      3.0459
```

<div align="right">SAS output file</div>

$e^{\beta_{intercept}} = e^{-0.5108} = 0.6$ Survival odds, males on wall (OW) $= 3/5 = 0.6$

$e^{\beta_{site=AM}} = e^{1.6094} = 5.0$ Odds ratio, males at margin (AM) $OR = \dfrac{15}{5}{\cdot}\dfrac{5}{3} = 5$

$e^{\beta_{site=IM}} = e^{1.2040} = 3.33$ Odds ratio, males in medium (IM) $OR = \dfrac{34}{17}{\cdot}\dfrac{5}{3} = 3.33$

$e^{\beta_{site=OM}} = e^{1.0217} = 2.78$ Odds ratio, males on medium (OM) $OR = \dfrac{5}{3}{\cdot}\dfrac{5}{3} = 2.78$

2. Execute model.

Parameter estimates from statistical package, compared to calculation by hand.

```
                   Analysis Of Parameter Estimates

                                      Standard    Wald 95% Confidence
    Parameter              DF   Estimate    Error          Limits

    Sex         F          1      1.0704   0.9624    -0.8158     2.9567
    Sex         M          0      0.0000   0.0000     0.0000     0.0000
    Site*Sex    AM   F     1      0.9664   1.4954    -1.9646     3.8974
    Site*Sex    AM   M     0      0.0000   0.0000     0.0000     0.0000
    Site*Sex    IM   F     1      0.4520   1.0951    -1.6944     2.5984
    Site*Sex    IM   M     0      0.0000   0.0000     0.0000     0.0000
    Site*Sex    OM   F     1     -0.6004   1.3849    -3.3147     2.1139
```

SAS output file

Odds ratio, females relative to males (OW)

$$e^{\beta_{site=OW}} = e^{1.0704} = 2.92 \qquad OR = \frac{7}{4}\cdot\frac{5}{3} = 2.92$$

Odds ratio, females relative males (AM)

$$e^{\beta_{site=AM}} = e^{1.0704 + 0.9664} = 7.67 \quad OR = \frac{23}{1}\cdot\frac{5}{15} = 7.67$$

Odds ratio, females relative males (IM)

$$e^{\beta_{site=IM}} = e^{1.0704 + 0.4520} = 4.58 \quad OR = \frac{55}{6}\cdot\frac{17}{34} = 4.58$$

Odds ratio, females relative males (OM)

$$e^{\beta_{site=OM}} = e^{1.0704 - 0.6004} = 1.6 \quad OR = \frac{8}{3}\cdot\frac{3}{5} = 1.6$$

3. Use residuals to evaluate straight line model and error model.

There is no straight line in model, so no need to check this assumption.
Errors are zero (saturated model) so we cannot use residuals to evaluate
assumptions for estimating Type I error from χ^2 distribution.

4. ANODEV - Calculate the improvement in fit (ΔG) due to the explanatory variables.

The ANOVA table is replaced by the Analysis of Deviance table.
The AnoDev table shows the degrees of freedom for a sequence of models.
One degree of freedom is lost in fitting the intercept.
Three are lost in a model that includes *Site*
One is lost in fitting a model that includes *Sex*
Three are lost in fitting the interaction term.

4. AnoDev Table.

Source	df	Δdf	Deviance $= G$	ΔG
Intercept $e^{\$T}$	7	1		
Site	4	$3 = 4{-}1$		
Sex $e^{\$M}$	3	1		
Sex*Site	0	$3 = 3*1$		

The package computes the fit (deviance $= G$) of the data to the model as each
term is added (site, sex, and finally site*sex).

```
                 LR Statistics For Type 1 Analysis

                                              Chi-
         Source           Deviance        DF    Square     Pr > ChiSq

         Intercept         24.2976
         Site              15.3385         3
         Sex                1.3655         1
         Site*Sex           0.0000         3
```

As each term is add the deviance decreases and the fit improves.
The fit of the data to the full model is perfect $G = 0.00$ (no deviance)
The improvement in fit for the omnibus model is 24.2976 on 7 degrees of freedom.
LR $= 1.9 \times 10^5$
There is strong evidence for effects on DDT on pupation, so we proceed with
further analysis.

5. Population
The sample is count of fruit flies emerging form pupae with DDT in the growth
medium. Population to which inference is made is all possible outcomes, given the
experimental protocol. We have no basis for inference to a population of
enumerable units, such as a population of flies.

5. Decide on mode of inference. Is hypothesis testing appropriate?
The odds appear to differ, so to eliminate chance as an explanation we undertake
hypothesis testing.

6. State H_A / H_o pair, tolerance for Type I error

Site·Sex term. We begin with the interaction term, same as with two way ANOVA
The research hypothesis H_A will be that the odds ratio for *Site·Sex* will
differ from 1.

$$H_A: \quad \beta_{Site \cdot Sex} \neq 0 \qquad OR = e^{\beta_{site*sex}} \neq 1$$

The null hypothesis H_o will be that the odds ratio for *Site·Sex* will be 1.

$$H_o: \quad \beta_{Site \cdot Sex} = 0 \qquad OR = e^{\beta_{site*sex}} = 1$$

If the interaction term is not significant, then we continue by examining the
main effects, site and sex.

Site term.
The research hypothesis will be that the odds ratios differs among sites.

$$H_A: \quad \beta_{Site} \neq 0 \qquad OR = e^{\beta_{site}} \neq 1$$

The null hypothesis will be that the odds ratios do not differ among sites.

$$H_0: \quad \beta_{Site} = 0 \qquad OR = e^{\beta_{site}} = 1$$

Sex term.
The research hypothesis will be that the odds differ for males and females.

$$H_A: \quad \beta_{Sex} \neq 0 \qquad \text{hence:} \qquad OR = e^{\beta_{sex}} \neq 1$$

The null hypothesis will be that the odds do not differ for males and females.

$$H_0: \quad \beta_{Sex} = 0 \qquad \text{hence:} \qquad OR = e^{\beta_{sex}} = 1$$

statistic $= \Delta G$ where ΔG is the evidential support (improvement in fit)
for a research model compared to a simpler reference model.
For the analysis of deviance in this example the likelihood ratios are:
$LR = $ likelihood(Model$_{Site \cdot Sex}$ / Model$_{intercept}$)
$LR = $ likelihood(Model$_{Site}$ / Model$_{Site \cdot Sex}$)
$LR = $ likelihood(Model$_{Sex}$ / Model$_{Site}$)

$$LR = e^{-\Delta G / 2}$$
$$\Delta G = -2 \ln(LR)$$

Weight of evidence criteria	$LR < 20$	Insufficient evidence
	$20 < LR < 100$	Sufficient evidence
	$100 < LR$	Strong evidence

probability distribution $=$ Chisquare

Tolerance for Type I error. $\alpha = 5\%$

7. ANODEV - Calculate the improvement in fit (ΔG) due to each explanatory variables.

```
                    LR Statistics For Type 1 Analysis

                                             Chi-
         Source          Deviance      DF    Square     Pr > ChiSq

         Intercept        24.2976
         Site             15.3385      3
         Sex               1.3655      1
         Site*Sex          0.0000      3
```

SAS output

As each term is add the deviance decreases and the fit improves.
The fit of the data to the full model is perfect $G = 0.00$ (no deviance)

Next we compute the improvement in fit due to each term. This is ΔG, the change in fit after a term is introduced.
The improvement in fit due to the interaction term: $\Delta G = 1.3655 - 0.0 = 1.37$
The improvement in fit due to the sex term: $\Delta G = 15.34 - 1.37 = 13.97$
The improvement in fit due to the site term: $\Delta G = 24.298 - 15.34 = 8.96$

In the SAS output ΔG is labelled Chi-Square.

```
                    LR Statistics For Type 1 Analysis

                                             Chi-
         Source          Deviance      DF    Square     Pr > ChiSq

         Intercept        24.2976
         Site             15.3385      3     8.96         0.0298
         Sex               1.3655      1    13.97         0.0002
         Site*Sex          0.0000      3     1.37         0.7137
```

SAS output

7. ANODEV. Calculate the improvement in fit (ΔG) due to each explanatory variables.

Note 1: The deviance for the interaction term based on the logistic model will differ from the G^2 statistic for the interaction term (model of equal proportions) based on Poisson response classified by 3 factors as shown in Box 17.10 of Sokal and Rohlf 2012

Sequential (Type 1) versus adjusted (Type 3) analysis.
In looking at ANOVA tables we usually considered Type 3 analysis, where the SS is computed for each term when it is listed last. This contrasts with Type 1 partitioning, which shows the SS due to each term in the order in which it is listed in the model.

7. ANODEV table.

The SS due to a term often depends on the order in which it occurs in the model. So we do Type 3 analysis, showing the improvement due to each term if it is added last.

The same holds for the analysis of deviance. The improvement in fit due to a term often depends on the order in which it appears in the model.
For Analysis of Deviance we are going to again look at Type 3 rather than Type 1 (sequential) results.

Here is the Type 3 analysis of deviance.

```
                    LR Statistics For Type 3 Analysis

                                        Chi-
            Source             DF      Square     Pr > ChiSq

            Site               3       10.97         0.0119
            Sex                1        8.14         0.0043
            Site*Sex           3        1.37         0.7137
```

The improvement due to the interaction term is of course the same as the previous analysis because the interaction term was last in that analysis. The improvement differs for the other two terms because now each is last.

Calculate p-value from χ^2 distribution.   See output above

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

   The model is saturated model, so there are no residuals to evaluate.
   The binomial distribution is appropriate, and if observations are made independently of one another, the assumptions for computing p-value from χ^2 distribution are met.

9. Declare decision.

 Survival of females relative to males is independent of location of pupation site.

   $\Delta G = 1.3655$
   $LR = \text{likelihood}(\text{Model}_{Site \cdot Sex} / \text{Model}_{intercept})$
   $LR = \exp(-\Delta G/2) = 0.505$
   The null model is half as likely as the model with the *Site·Sex* term.
   The *Site·Sex* model is twice as likely as the null model.
   There is insufficient evidence for the model with the *Site·Sex* term.

Estimate Type I error on decision to reject the null hypothesis.
   df = 3, p = 0.7137 hence cannot reject H_o

9. Declare decision.
The null model, no interactive effect, cannot be rejected.
So we then examine the main effects that were of interest in this experiment.

There were significant differences in pupation success among sites
$\Delta G = 10.97$, df $= 3$, p $= 0.012$
There were significant differences in pupation success between males and females
$\Delta G = 8.14$, df $= 1$, p $= 0.0043$

10. Analysis of parameters of biological interest.
See estimates in step 4, with standard errors and confidence limits.
Survival odds of females are 1.6 to 7.6 times that of males

Survival odds of males (or females) are circa 3 times higher in or on medium, 5 times higher at edge of medium, relative to survival of pupae on wall.