Model Based Statistics in Biology.

Part V. The Generalized Linear Model.

Chapter 17.2 Single Categorical Explanatory Variable

ReCap. Part I (Chapters 1,2,3,4), Part II (Ch 5, 6, 7)
ReCap Part III (Ch 9, 10, 11), Part IV (Ch13, 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
17.5 Poisson ANCOVA
17.6 Model Revision

on chalk board

Ch17.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) GLM with more than one explanatory variable
ReCap (Ch 15) GLM review
ReCap (Ch 16) The generalized linear model.
ReCap (Ch 17) Poisson regression. Variance of response variable increases as the square of the fitted value. We use the generalized linear model to take this into account.

Today: Poisson response variable with single categorical explanatory variable.

Wrap-up.
The General Linear Model is a special case of the Generalized Linear Model. Consequently, we can carry out any GLM as a GzLM.

The example today demonstrated log-linear analysis for Poisson counts. The response variable has a variance that increases with the mean. There is a single explanatory variable, which is categorical. The link between the response and explanatory variable is logarithmic, hence the analysis considers percent change in the response variable across levels of the categorical variable (factor).

Log link. Analysis of multiplicative effects (changes in proportion) without having to resort to log transform.

Often called G-tests or log-linear models.

Introduction.

Many of the analyses undertaken in biology are concerned with counts that are small, with values near enough zero that deviations from any model parameter won't be normal and homogeneous. A plot of errors (residuals versus fits) will look like a cone, widening out to the right at larger fitted values.

The generalized linear model based on Poisson errors is covered under the heading of G-tests in many texts, including Sokal and Rohlf (1995). In this course we will treat G-tests as still another special case of the generalized linear model, rather than treating them as a separate topic.

Poisson response variables (counts) are analyzed in relation to categorical variables. These are called log-linear models.

In this course we will treat log linear models as a special case of the generalized linear model.

Example.

We return to the classic example of Poisson data, the number of deaths by horse kick, for each of 16 corps in the Prussian army, from 1875 to 1894.

The unit of analysis is now a single corp over 20 years.
The distribution of counts fits a Poisson distribution.

Does the risk of death due to horsekick depend on corps within an army?

Here we will analyze the data within the framework of the Generalized Linear Model, to show that the G-test is a based on a model similar to a one-way ANOVA.

We begin with the computation of the goodness of fit of observed to expected.

fhat = 56/4 = 14

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

```
           f  fhat       Dev = f*ln(f/fhat)
  Guard   16   14          2.14
  First   16   14          2.14
  2nd     12   14         -1.8
  3rd     12   14         -1.8

          56   56          0.57
                             x2
                      G=   1.15
```

Next, analysis of the same data as a generalized linear model with a poisson response variable.

1. Construct the model

Verbal model. Number of deaths depends on corps (Guard, 1st, 2nd, 3rd).
Graphical model.
Formal model.
Response variable. f = deaths.
Explanatory variable. Corps

We will treat the number of deaths as the result of probabilities $f = (p1\ p2\ ...)(N)$.
We are interested in whether the probability differs among corps.
Hence we will use a logarithmic scale for our model of frequency f.
Here is the model.

$$f = e^{\mu} + PoissonError$$

log link, Poisson error

$$\mu = \beta_{ref} + \beta_{Corps} \cdot Corps$$

2. Execute analysis.

Arrange data into model format.

```
Data Hkick;
  Input Count Corps $ ;
  Cards;
    16 Guard
    16 First
    12 Second
    12 Third
;
```

SAS command file

Use model to execute analysis.

$$f = e^{\left(\beta_{ref}\right)} e^{\left(\beta_{Corps} \cdot Corps\right)} + error$$

```
Proc Genmod;  Classes Corps;
  Model Count = Corps/
  Link=log dist=poisson type1 type3;
```

SAS command file

```
>   glm(formula = Count ~ Corps,
     family = poisson(link = log),
     data = Hkick)
```

R/S+

$\beta_{ref} = \beta_{Guard} = 2.7726 \qquad e^{\left(\beta_{Guard}\right)} = 16$

$\beta_{First} = 0 \qquad e^{\left(\beta_{Guard} + \beta_{First}\right)} = e^{(2.7726 + 0.0)} = 16$

$\beta_{Second} = -0.2877 \qquad e^{\left(\beta_{Guard} + \beta_{Second}\right)} = e^{(2.7726 - 0.2877)} = 12$

$\beta_{Third} = -0.2877 \qquad e^{\left(\beta_{Guard} + \beta_{Third}\right)} = e^{(2.7726 - 0.2877)} = 12$

In this example we have only 4 observations, and 4 parameters. This is called a saturated model. There are no residuals.

4. State population and whether sample is representative.

Population is (?) all possible arrangements of these 16+16+12+12 = 56 deaths into 4 units.

Representative of (?) accidents in four military unit that are suspected of having similar practices and accident rates over 2 decades.

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

Does death by horsekick depend on unit? Yes/no decision appropriately addressed with hypothesis testing.

The units differ in observed deaths. Are these differences greater than chance ?

6. State H_A H_o etc

H_A : $\beta_{Corps} \neq 0$ $e^{\beta_{Corps}} \neq 1$ frequency depends on both leaf type and soil type, hence cross-product ratio differs from unity.

H_o : $\beta_{Corps} = 0$ $e^{\beta_{Corps}} = 1$ frequency does not depend on both leaf type and soil type, hence cross-product ratio equal to unity.

H_A : $f = e^{(\beta_{ref})} e^{(\beta_{Corps} \cdot Corps)}$

H_o : $f = e^{(\beta_{ref})}$

Statistic: G Distribution: chisquare $\alpha = 0.05$

7. ANODEV Table

ANOVA table is replaced by Analysis of Deviance table.

The improvement in fit is Δ G = 1.147

```
      Df Deviance   Resid. Df     Resid. Dev
 NULL                       3      1.146776
Corps  3 1.146776           0      0.000000
```

output from R/S+

Here is the AnoDev table, from SAS.

```
                      LR Statistics For Type 1 Analysis

                                                 Chi-
          Source            Deviance        DF     Square    Pr > ChiSq

          Intercept          98.5389
          Corps              97.3921         3       1.15        0.7658
```

SAS output file

The goodness of fit of the data to the null model is G = 98.5389 (df = 3)
The fit of the data to the alternative model G = 97.3921 (df = 0)
The improvement is Δ G = 1.1468 (Δ df = 3)

This measure (G = 1.15) is exactly the same as that computed by the goodness of fit test.

7. Calculate improvement in fit due to explanatory variables.
Calculate p-value from Chisquare distribution.

Is this improvement ΔG better than by chance ?

The p-value reported for $\Delta G = 1.15$ is $p = 0.7658$

The p-value computed from the chisquare distribution is reliable for ΔG, but not necessarily for G.

8. Recompute p-value if warranted.
Residual deviance = 0, so assumptions cannot be evaluated from residuals.

9. Declare decision. $\Delta G = 1.15$, df = 3, $p = 0.7658$ hence accept H_o (reject H_A)
The frequency of death was independent of corps.

10. Evaluate parameter estimates.
There was no significant difference on counts among corps, so the parameter of interest is the mean number of deaths by horsekick over 2 decades in all 4 units. pr = (56 deaths / 20 years) / 4 units = 0.7 deaths/unit-year