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 (Ch 13, 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

Ch17.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) 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).

Introduction.
Many of the analyses undertaken in biology are concerned with counts that are small, with values near enough zero that errors based on using a normal error 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 because the explanatory variables are additive (linear) on a log scale.

Example.
We will use a 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, assembled and published by Bortkiewicz (1898). Guard corps duties differed from other units, so accidents might be expected be less than other corps.

The unit of analysis is now a single corp over 20 years.
The counts across many corps can be shown to fit 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 GzLM, to show that the G-test is a based on a model with a structure similar to a one-way ANOVA. The comparisons will be ratios, not differences in means, as with a GLM.

We begin with the computation of the goodness of fit of observed to expected frequency, i.e. all units with the same number of deaths.

Introduction. Calculation from formula.

$$G = 2*\Sigma\left(f \cdot \ln\left(\frac{f}{\hat{f}}\right)\right) \qquad \hat{f} = 56/4 = 14$$

	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 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.

$f = e^{\mu} + PoissonError$ Unscaled error used in iterative estimation
Scaled errors are used for model checking
$\mu = \beta_{ref} + \beta_{Corps} \cdot Corps$ Log link with Poisson error

2. Execute analysis.

Arrange data into model format. SAS command file

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

2. Execute analysis.

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+

Here are the parameter estimates.

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

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

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

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

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

4. What is the evidence?

The improvement in fit for the research model relative to the null model (shown above) is twice the loglikelihood ratio. G = 2 ln(LR).
The output from a generalized linear model is shown as an analysis of deviance table (ANODEV table) instead of an ANOVA table.

output from R/S+

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

The improvement in fit is G = 1.147 – 0 = 1.147, due to a model with 3 fewer degrees of freedom.
The likelihood ratio is LR = $e^{G/2}$ LR = exp(1.147/2) = 1.8
There is insufficient evidence (LR<10) for the alternative to the null model.

5. Decide on mode of inference. Is hypothesis testing appropriate? State population and whether sample is representative.

Population could be taken as all possible arrangements of these 16+16+12+12 = 56 deaths into 4 units. A randomization test could be conducted in this way. The inference would be only to the data at hand, not to other corps or other mounted cavalry.

This is an observational study that cannot be repeated. Soldiers are no longer recruited to cavalry corps, or die of horse kicks. There are many uncontrolled variables that limit inference beyond the data at hand. We will report only the strength of evidence for this data.

Here is the AnoDev table, from SAS.

LR Statistics For Type 1 Analysis

Source	Deviance	DF	Chi-Square	Pr > ChiSq
Intercept	98.5389			
Corps	97.3921	3	1.15	0.7658

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 is: $G = 97.3921$ ($df = 0$)
The improvement is: $\Delta G = 1.1468$ ($\Delta df = 3$)

10. Evaluate parameter estimates.

There insufficient evidence of differences in accident frequency 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