Model Based Statistics in Biology.

Part V. The General Linear Model.

Chapter 17.5 Poisson ANCOVA

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 12,13,14,15) GLM with more than one explanatory variable
ReCap (Ch 16,17)

Today: Poisson response variable with one categorical and one ratio scale explanatory variable.

Wrap-up.

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.

Poisson Response Variable. Single way classification.

The classic example of Poisson data is 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 a single year in each of 4 corps. The number of deaths per year in a single corps ranged from 0 to 3.

The distribution of counts appears to fit a Poisson distribution.
The ratio of variance to mean is close to unity, as expected of Poisson distribution.

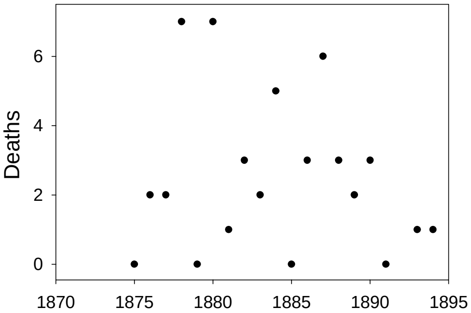
Guard	First	2nd	3rd	Total	
16	16	12	12	56	Deaths
1.00	1.39	1.1	1.1		Var/mean
				2.28	Total

The risk of death did not change over time in Guard Corps. Is there a similar lack of trend in the other units ?

We will analyze the data within the framework of the Generalized Linear Model, to show that an ANCOVA structural model can be applied to a poisson response variable.

1. Construct Model

Verbal model.
The lack of trend seen in Guard Corps will hold true over the same time for all four corps.



Graphical model
Plot of number of total deaths versus year shows no obvious pattern

Response variable: Deaths
Explanatory variable: a fixed variable, Year. Continuous variable (ratio scale)
Explanatory variable: Corps. Categorical variable (nominal scale).

To avoid negative fitted values we use log link. This will describe percent change from year to year.

Write formal model $Deaths = e^{(\mu)} + Poisson\,Error$

$$\mu = \beta_o + \beta_{Year} \cdot Year + \beta_{Corps} \cdot Corps + \beta_{Year*Corps} \cdot Year*Corps$$

The link between the Count and the structural model μ is: $Deaths = e^{\mu}$

2. Execute analysis.

Place data in model format:

Column labelled Count, with response variable # of deaths
Column labelled Year, the explanatory variable

```
data d1;
 input Year 1-4 Deaths 7 Duty $ 10 Corps $ 12-16;
cards;
1875  0  A guard
1876  2  A guard
1877  2  A guard
1878  1  A guard
1879  0  A guard
1880  0  A guard
1881  1  A guard
1882  1  A guard
1883  0  A guard
1884  3  A guard
1885  0  A guard
1886  2  A guard
1887  1  A guard
1888  0  A guard
1889  0  A guard
1890  1  A guard
1891  0  A guard
1892  1  A guard
1893  0  A guard
1894  1  A guard
1875  0  A first
1876  0  A first
1877  0  A first
1878  2  A first       (Etc for 80 observations)
;
```

SAS command file

In a package with spreadsheet format, there will be a column for each variable and 80 rows for this data set.

2. Execute analysis.
Code the GzLM model statement in statistical package

$$\mu = \beta_o + \beta_{Year} \cdot Year + \beta_{Corps} \cdot Corps + \beta_{Year*Corps} \cdot Year*Corps$$

```
Proc Genmod;
  Model Deaths = Year/
  Link=log dist=poisson type1 type3;
  output out=outB p=pred r=res;
PROC PLOT data=outB; plot res*pred/vref=0;
```

SAS command file

3. Evaluate model
a. Straight line assumption - no bowls or arches
b. Homogeneity of variables - no cones
[Note that deviance residuals are recommended for diagnosis, not simple or raw residuals]

4. State population and whether sample is representative.
Population is (?) all possible arrangements of these 80 observations in a sequence in 4 units.
Representative of (?) accidents in four military units where change in practice is suspected not to have occurred.

5. Decide on mode of inference. Is hypothesis testing appropriate?
Does death by horsekick change with year? Yes/no decision required, appropriately addressed with hypothesis testing.

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

H_A: $\beta_{Year*Corps} \neq 0$ hence: $Deaths = e^{\left(\beta_o + \beta_{Year*Corps} \cdot Year*Corps\right)} \neq constant$

H_0: $\beta_{Year*Corps} = 0$ hence: $Deaths = e^{\left(\beta_o\right)} = constant$

Statistic - Non-Pearsonian chisquare (G-statistic)
Probability distribution - chisquare
Tolerance for Type I error. $\alpha = 5\%$

7. ANODEV Table.
The F-statistic is not used for models with non-normal errors. Instead of partitioning the variance as in an ANOVA table, we will be examining the improvement in fit in one model relative to another. To do this we will construct an analysis of deviance (ANODEV) table.

For the Generalized Linear Model, step 7 is modified: we calculate the change in deviance ΔG rather than the SS for each term in the model.

```
                 LR Statistics For Type 3 Analysis

                                          Chi-
              Source              DF     Square     Pr > ChiSq

              Corps                2       1.12         0.5719
              Year                 1       0.05         0.8313
              Year*Corps           3       1.27         0.7366
```

The chisquare column is ΔG, the change in the non-Pearsonian Chisquare, G.
The improvement due to the interaction term is $\Delta G = 1.27$

7. Calculate p-value from Chisquare distribution.
Is the change in fit ΔG better than by chance ?
Equivalently, do the fitted values in each group differ from the overall mean ?
The p-value reported for $\Delta G = 1.27$ is $p = 0.7366$
The p-value is large, hence ΔG is small enough to be due to chance.
For generalized linear models, we compute a p-value on ΔG, not on the deviance G.

8. Assess p-values and estimates in light of evaluation of residuals.
Acceptable

9. Declare decision. $p = 0.7366$ hence reject H_A and accept H_o
The four corps show the same lack of trend in deaths by horsekick over two decades.

10. Analysis of parameters of biological interest.
The parameter of interest is the mean number of deaths by horsekick over 2 decades.
pr = (56 deaths / 20 years) / 4 units = 0.7 deaths/unit-year
The set of parameters describing rate of change from unit to unit $\beta_{Year*Corps}$ provides no additional information.