Model Based Statistics in Biology.

Part V. The Generalized 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 (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 12,13,14,15) 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 the biological, environmental, and social sciences are concerned with data that include zero counts per unit. Consequently, the variability around fitted valued near zero will be small, compared to variability around larger fitted values. A plot of errors (residuals versus fits) will fan out, widening out to the right at larger fitted values. If a normal error is used, the lower confidence will sometimes be negative—it will include impossible values. We certainly do not want to produce a confidence limits on fish counts than include negative fish!

A commonly used solution to this problem is to transform the response variable to a log scale. Unfortunately this remedy has bad side effects. First of all, we have to convert zeros to "almost zero" (0.01 or 0.001). The choice of "almost zero" yields different estimates. Worse yet, log transformation results in biased estimates of parameters--means, contrasts between means, and slopes. To remedy these problems log-linear models were developed in the latter quarter of the 20th century. Bishop *et al* (1975) provided the first comprehensive text. Log-linear models are covered in many subsequent texts, where they are called G-tests, contingency tests, or sometimes log-linear analyses. In almost every case log-linear models refer to categorical variables. Within the framework of the generalized linear model we can include regression variables (covariates) in conjunction with one or more categorical variables.

Poisson ANCOVA.

Log linear models and Poisson regression are special cases of the generalized linear model. Having learned how to write Ancova models, with at least one categorical variable and one covariate, we can extend these models to Poisson variables. Poisson Ancova is an apt name. This makes it clear we are not using a normal error structure, which is implied by the term Ancova when it was proposed.

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, assembled and published by Bortkiewicz (1898).

The unit of analysis is a single year in each corps. The number of deaths per year in the four corps shown here ranged from 0 to 3.

Year	Guard	First	2nd	3rd
1875	0	0	0	0
1876	2	0	0	0
1877	2	0	0	0
1878	1	2	2	2
1879	0	0	0	0
1880	0	3	2	2
1881	1	0	0	0
1882	1	2	0	0
1883	0	0	1	1
1884	3	0	1	1
1885	0	0	0	0
1886	2	1	0	0
1887	1	1	2	2
1888	0	1	1	1
1889	0	0	1	1
1890	1	2	0	0
1891	0	0	0	0
1892	1	3	2	2
1893	0	1	0	0
1894	1	0	0	0
	16	16	12	12

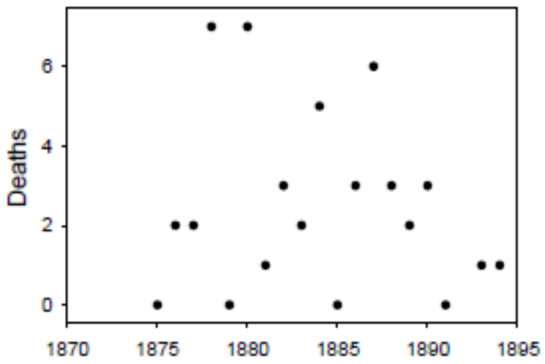
1. Construct Model

Verbal model.

Safety of conscripts into the Prussian army was hardly a concern in the late 19th century. Thus we expect no change over time in number of deaths by horsekick in 4 different corps, Guard, First, Second, and Third. Duties differed in Guard and First compared to Second and Third, so we might see some differences among corps because of different in exposure of conscripts to horses.

Graphical model

A plot of number of total deaths in all 4 corps shows little or no change by year. Draw your estimate of the best fitting trend line through the data.



Formal model

Response variable: Deaths/year, each corps

Explanatory variables:

Year. Continuous variable (ratio scale)

Corps. Categorical variable.

Choice of error model.

As a rule of thumb, a Poisson error model is appropriate where counts are produced at low rates by a random process, resulting in mean values less than 10 in order of magnitude. That is, counts are 'rare and random.' The technical criterion is that the variance in counts is roughly the same as the mean count. The distribution of counts appears to fit a Poisson distribution. The variance to mean ratio 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

Choice of link. For log-linear models we use a log link. This describes percent change from year to year and from corps to corps. An identity link (additive effects) is possible but inconsistent with the idea that we take the product or ratio of percentages.

1. Construct Model

$Deaths = E(Deaths) + \varepsilon$ Expected value is exponential relation.

$Deaths = \exp(\beta_{ref} + \beta_{Yr} \cdot Yr + \beta_{Corps} \cdot Corps + \beta_{Y \cdot C} \cdot Yr \cdot Corps) + \varepsilon$

Transform the data

$log_e(Deaths) = \beta_{Yr} \cdot Yr + \beta_{Corps} \cdot Corps + \beta_{Y \cdot C} \cdot Yr \cdot Corps + \varepsilon$

Transform expected value *E(Deaths)*

$Deaths = e^{\eta} + \varepsilon$

$\eta = \beta_{ref} + \beta_{Yr} \cdot Yr + \beta_{Corps} \cdot Corps + \beta_{Y \cdot C} \cdot Yr \cdot Corps$

Distribution $Deaths \sim Poisson(\lambda)$

Link $Death = e^{\eta}$

Structural model

$\eta = \beta_{ref} + \beta_{Yr} \cdot Yr + \beta_{Corps} \cdot Corps + \beta_{Y \cdot C} \cdot Yr \cdot Corps$

2. Execute analysis.

The data in table format need to be re-organized to model format.

Column labeled Count, with response variable # of deaths

Column labeled 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
.
.
1893  0  A guard
1894  1  A guard
1875  0  A first
1876  0  A first       (Etc for 80 observations)
;
```

SAS command file

Data appear in a similar format in packages with a spreadsheet interface.

2. Execute analysis.

Code the GzLM model statement in statistical package

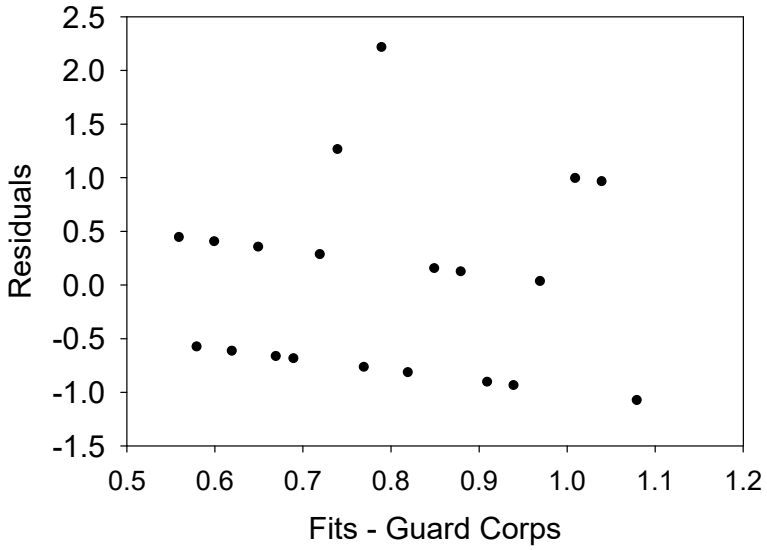
$\eta = \beta_{ref} + \beta_{Yr} \cdot Yr + \beta_{Corps} \cdot Corps + \beta_{Y \cdot C} \cdot Yr \cdot Corps$

```
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

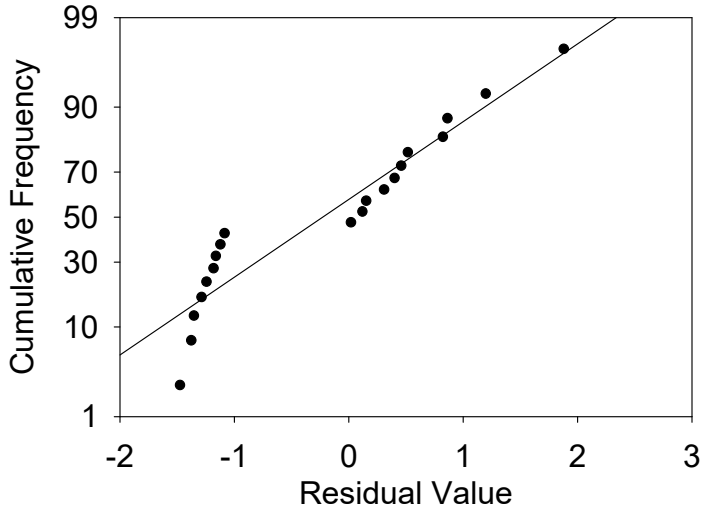
A. A straight line was used in the model. Acceptable? Yes. No bowls or arches



B1. A Poisson error was used to eliminate change in variability (vertical spread) with change in the fitted values. Acceptable? Yes. Fans and spindles not evident.

Deviance residuals are used for diagnosis, not simple or raw residuals.

B2. A Poisson error structure was used to fit the model. This was expected to remove any change in vertical spread in the residual vs fit plot. Poisson error acceptable? Yes. Vertical spread in residuals from left to right in the graph shows no fan or spindle pattern.



B3. Confidence limits and p-values assume normal errors. The quantile-quantile plot shows acceptably normal errors.

4. What is the evidence?

LR Statistics For Type 1 Analysis

Source	Deviance	DF	Chi-Square	Pr > ChiSq
Intercept	98.5389			
Corps	97.3921	3	1.15	0.7658
Year	97.3577	1	0.03	0.8529
Year*Corps	96.0891	3	1.27	0.7366

The improvement in fit for the fully reduced model (all three terms) relative to the null model, is 98.5389 - 96.0891 = 2.45.

$LR = \exp(2.45/2) = 3.4$

There is insufficient evidence (LR<10) for the omnibus model (all three terms). Hence no evidential support for detailed analysis of the three components.

5. Mode of inference.

Priorist? No. We have a suitable probability model, but we lack a definite prior value of the Poisson parameter.

Evidentialist? Yes. We have a suitable probability model,

Frequentist? Perhaps. The number of deaths per year falls within a range that suggests a stable value for the rate parameter λ, deaths/year.

Decision theoretic? No. We have no way of gauging Type I versus Type II error in declaring a decision on this data. Nor can we define the costs and risks of Type I error.

6. Population, sample, hypotheses.

This is an observational study, with many uncontrolled sources of variation. The population can be defined by a Poisson model of chance outcomes from the conditions that prevailed in an army consisting of conscripts, and where military mobility was by horses.

The sample consists of four fully censused military units where change in practice is suspected not to have occurred over 2 decades. We have insufficient evidence for change in death rate over time or across corps, so we form no hypotheses about each term.

7. ANODEV Table.

In the absence of evidence for the omnibus model there is no logic in proceeding to analysis of individual terms. For comparison, here is the ANODEV table for adjusted (Type 3) deviances. They are the same as Type 1. However, they cannot be summed to back calculate the overall change.

LR Statistics For Type 3 Analysis

Source	DF	Chi-Square	Pr > ChiSq
Corps	3	1.12	0.5719
Year	1	0.05	0.8313
Year*Corps	3	1.27	0.7366

10. Analysis of parameters of biological interest.

The parameter of interest is λ, the number of deaths per year by horsekick over 2 decades. The estimate of this parameter is

$\hat{\lambda}$= (56 deaths / 20 years) / 4 corps) = 0.7 deaths/corps-year

The 6 parameters in the Poisson Ancova provide no additional information.