

## Model Based Statistics in Biology.

### Part V. The Generalized Linear Model.

#### Chapter 17.1 Poisson Regression

ReCap. Part I (Chapters 1,2,3,4), Part II (Ch 5, 6, 7)  
ReCap Part III (Ch 9, 10, 11), Part IV (Ch 12, 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)

Many of the analyses undertaken in biology are concerned with frequencies.

Frequencies are also analyzed by the *Generalized* Linear Model, which compares observed to expected (model) values.

We use the Analysis of Deviance to calculate the improvement in fit, the Likelihood ratio, and the likelihood ratio statistic.

Today: Poisson Regression.

#### Wrap-up. .

The example today demonstrated Poisson regression. The response variable has a variance that increases with the mean. The explanatory variable is numeric. The link between the response and explanatory variable is logarithmic, hence the analysis considers percent change in the response variable with change in the explanatory variable.

## Poisson Regression.

Example: Death by Horsekick

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 data were assembled and published by Ladislaus Bortkiewicz in his book, *The Law of Small Numbers* (1898). Bortkiewicz was an economist and statistician who taught at Berlin University (1901–1931). Bortkiewicz showed that the horsekick data fit a Poisson distribution, which was introduced 1837 by Siméon Poisson in *Récherches sur la Probabilité des Jugements en Matière Criminelle et en Matière Civile*.

The unit of analysis is a single army corps in a single year. The number of deaths per year in a single corps ranged from 0 to 3. The deaths occurred because most of the soldiers knew nothing about horses. They were conscripts from cities and did not know about why one does not stand behind a horse. This is a Poisson count because we do not know the number of trials (kicks) that resulted in death.

With this data we ask: Was there any trend in the number of deaths?  
Poisson regression as a special case of the generalized linear model.

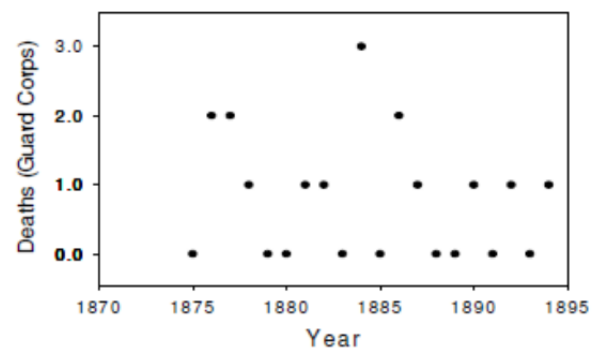
### 1. Construct Model

#### Verbal model.

Does number of deaths by horsekick in Guard corps show a trend from 1875 to 1894?

#### Graphical model

If asked to draw a line on the graph, many people might draw a slight downward trend. How strong is the evidence for such a trend?



### 1. Construct Model

Response variable: Deaths/year

Explanatory variable: a numeric variable, Year.

Choice of error structure. The data consist of counts in definable units, a single army corps over a year. This points at one of the probability models for discrete variables. These include the Poisson distribution for rare a random events, or the negative binomial distribution for clustered or contagious events, such as disease cases. We will start with the Poisson distribution. Then we will use a residual fit plot to check this choice.

## 1. Construct Model

Choice of link. The preferred link for the Poisson error is the log link, where effects are multiplicative (additive on a log scale).

Write formal model:

$$\begin{aligned} \text{Distribution} \quad & \frac{\text{Deaths}}{\text{year}} \sim \text{Poisson}(\lambda) \\ \text{Link} \quad & \frac{\text{Deaths}}{\text{year}} = e^\eta \\ & \eta = \beta_0 + \beta_{Yr} Yr \end{aligned}$$

The link between the response variable and the linear predictor  $\eta$  is a log link  $\beta_{Yr}$  quantifies annual change. with units of %/year.

## 2. Execute analysis.

Place data in model format:

Column labeled Count,

with response variable # of deaths/year

Column labeled Year,

the explanatory variable

Data definition, SAS input file

In a package with spreadsheet format, there will be two columns (variables) and 20 rows for this data set.

The data set includes name of the corps and the type of duty for that corp.

The numbers are for Guard Corp.

In the next section of this chapter we will compare death rates among different corps.

```
Data Horsekick;
  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
```

## 2. Execute analysis.

We use the model statement to code the model in our statistical package

$$\text{Deaths} = e^{\beta_0 + \beta_{Yr} Yr}$$

```
Proc Genmod;
  Model Deaths = Year/
  Link=log dist=poisson type1 type3;
  Obstats;
PROC PLOT data=; plot res*pred/vref=0;
SAS
```

```
PoissonMod <- glm(formula = Deaths ~ Year,
  family = poisson(link = log), data = Horsekick)
anova(PoissonMod)
R
```

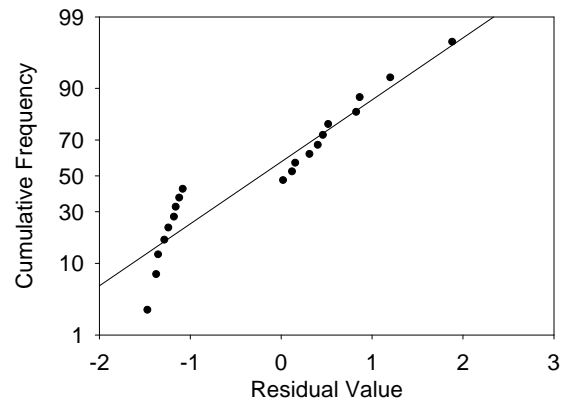
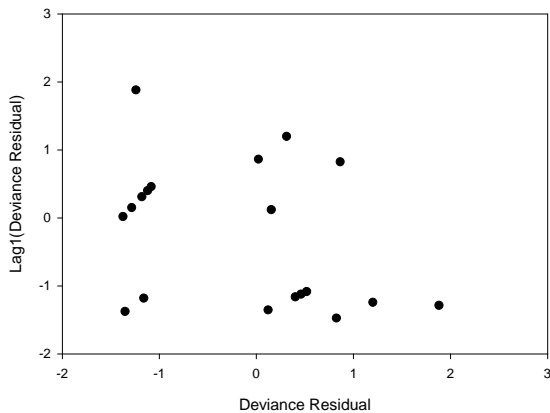
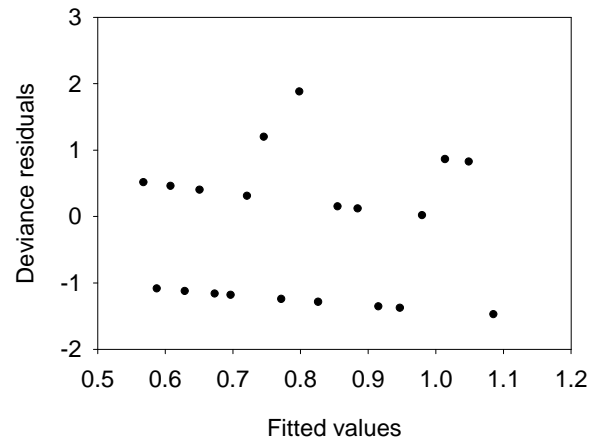
### 3. Evaluate model

#### A1. Straight line assumption .

While there is pattern in the residuals, the pattern is not a curve (no bowls or arches). So the straight line assumption on a log scale is acceptable.

#### A2. Distributional assumptions. Homogeneity. Judged acceptable

Normal. The zero observations, stacked on the left side of the plot show deviation from the normal line. The stack centers on the normal line, and so the deviations are judged not serious.



Independent. We know the temporal sequence of the observations so we check this assumption. Assumption judged acceptable – no trends up or down across the entire graph.

### 4. What is the evidence?

For the Generalized Linear Model we calculate the change in deviance  $\Delta G$  due to a term in the model, rather than the Sum of Squares for each term in the model. With 20 observations, we have 19 df after fitting the intercept  $\beta_0$  and 18 df after estimating  $\beta_{Yr}$ .

The goodness of fit of the data to the full (null) model is	$G = 22.0500$
The fit of the data to the reduced model is	$G = 21.4387$
The improvement due to the regression term is	$\Delta G = 0.61$

Here are the df and deviance calculations, aligned with the model.

	Deaths	=	$\exp(\beta_o + \beta_{Yr} Yr)$	+	$\epsilon$
residual df	19	=	1	+	18
$G^2$	22.05	=	0.61	+	21.4387
$\Delta G^2$			0.61		

The results are displayed in an analysis of deviance (ANODEV) table.

Source	df	Deviance = $G^2$	$\Delta G^2$	Pr > ChiSq
Intercept	1	22.0500		
Year	1	21.4387	0.61	0.4343

Reduced model  $\beta_o + \beta_{Year}$  hence:  $Deaths = e^{(\beta_o + \beta_{Year} \cdot Year)} \neq constant$

Full model  $\beta_o$  hence:  $Deaths = e^{(\beta_o)} = constant$

$$LR = \frac{L(\beta_o, \beta_{Year} | Data)}{L(\beta_o | Data)}$$

$$\Delta G = -2 \ln LR \quad LR = e^{0.61/2} = 1.36$$

## 5. Analytic Mode.

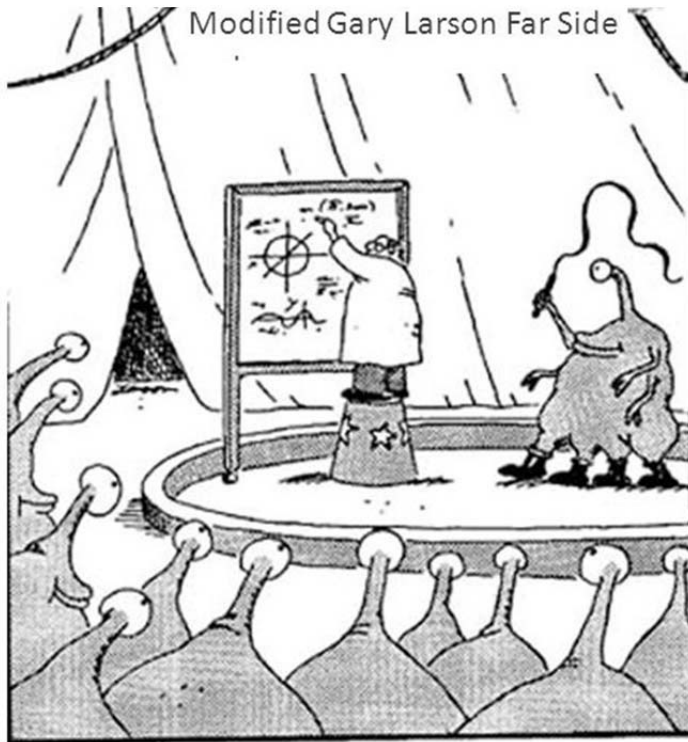
We can use our measure of evidence in any of several analytic modes.

Bayesian? No. We have insufficient prior information to set up a defensible prior probability.

Frequentist? No. We have no survey or measurement protocols that are repeatable.

Decision theoretic? No. We have no way of gauging Type I versus Type II error.

Evidentialist? Yes. We have little need of probabilities to temper judgement based on likelihood ratios because our comparison is on single degree of freedom.



Abducted by an alien circus company, Professor Tversky is forced to calculate p-values in center ring.

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, 76, 105-110.

This is an observational study with many sources of uncontrolled variability. We will avoid using probabilities (p-values) as evidence. We will avoid declaring decisions against a fixed Type I error rate.

Tversky and Kahnemann (1971) noted that p-values on data with uncontrolled variability tend to support belief in what they called the “law of small numbers,” the judgmental bias that occurs when it is assumed that the characteristics of a population can be estimated from a small number of observations or data points.

### 10. Science conclusion. Analysis of parameters of biological interest.

The parameter describing rate of change from year to year was small in magnitude  $\hat{\beta}_{Year} = -0.0341$ . The estimate was no more likely than not,  $\beta_{Year} = 0$ . The rate parameter provides no additional information beyond the mean number of 1.25 years/death.  $\text{Mean}(\text{Deaths/year}) = 16 \text{ deaths} / 20 \text{ years} = 0.8 \text{ deaths/year}$ . Deaths are not decimal numbers so alternatively, we re-express the mean as an average time per event, 1.25 years/death.