Poisson Regression

Bret Larget

Departments of Botany and of Statistics
University of Wisconsin—Madison

May 1, 2007

Poisson Regression

- *Poisson regression* is a form of a generalized linear model where the response variable is modeled as having a *Poisson distribution*.
- The Poisson distribution models random variables with non-negative integer values.
- For large means, the Poisson distribution is well approximated by the normal distribution.
- In biological applications, the Poisson distribution can be useful for variables that are often small integers including zero.
- The Poisson distribution is often used to model *rare events*.

The Poisson Distribution

- The *Poisson distribution* arises in many biological contexts.
- Examples of random variables for which a Poisson distribution might be reasonable include:
  - the number of bacterial colonies in a Petri dish;
  - the number of trees in an area of land;
  - the number of offspring an individual has;
  - the number of nucleotide base substitutions in a gene over a period of time;

Probability Mass Function

- The *probability mass function* of the Poisson distribution with mean μ is
$$\mathrm{P}\{Y = k \,|\, \mu\} = \frac{\mathrm{e}^{-\mu}\mu^k}{k!} \qquad \text{for } k = 0, 1, 2, \ldots.$$
- The Poisson distribution is discrete, like the binomial distribution, but has only a single parameter μ *that is both the mean and the variance*.
- In R, you can compute Poisson probabilities with the function `dpois`.
- For example, if $\mu = 10$, we can find $\mathrm{P}\{Y = 12\} = \frac{\mathrm{e}^{-10}10^{12}}{12!}$ with the command

```
> dpois(12, 10)
[1] 0.09478033
```

Poisson approximation to the Binomial

- One way that the Poisson distribution can arise is as an approximation for the binomial distribution when p is small.
- The approximation is quite good for large enough n.
- If p is small, then the binomial probability of exactly k successes is approximately the same as the Poisson probability of k with $\mu = np$.
- Here is an example with $p = 0.01$ and $n = 10$.

```
> dbinom(0:4, 10, 0.01)
[1] 9.043821e-01 9.135172e-02 4.152351e-03 1.118478e-04 1.977108e-06
> dpois(0:4, 10 * 0.01)
[1] 9.048374e-01 9.048374e-02 4.524187e-03 1.508062e-04 3.770156e-06
```

- This approximation is most useful when n is large so that the binomial coefficients are very large.

The Poisson Process

- The *Poisson Process* arises naturally under assumptions that are often reasonable.
- For the following, think of *points* as being exact times or locations.
- The assumptions are:
  - The chance of two simultaneous points is *negligible*;
  - The expected value of the random number of points in a region is *proportional to the size of the region*.
  - The random number of points in non-overlapping regions are *independent*.
- Under these assumptions, the random variable that counts the number of points *has a Poisson distribution*.
- If the expected rate of points is λ points per unit length (area), then the distribution of the number of points in an interval (region) of size t is $\mu = \lambda t$.

Example

- Suppose that we assume that at a location, a particular species of plant is distributed according to a Poisson process with expected density 0.2 individuals per square meter.
- In a nine square meter quadrat, what is the probability of no individuals?
- **Solution:** The number of individuals has a Poisson distribution with mean $\mu = 9 \times 0.2 = 1.8$. The probability of this is

$$\mathrm{P}\{Y = 0 \,|\, \mu = 1.8\} = \frac{\mathrm{e}^{-1.8}(1.8)^0}{0!} \doteq 0.165299$$

- In R, we can compute this as

```
> dpois(0, 1.8)
[1] 0.1652989
```


Poisson Regression

- Poisson regression is a natural choice when the response variable is a small integer.
- The explanatory variables *model the mean* of the response variable.
- Since the mean must be positive but the linear combination $\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$ can take on any value, we need to use a *link function* for the parameter μ.
- The standard link function is the *natural logarithm*.

$$\log(\mu) = \eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k$$

so that

$$\mu = \exp(\eta)$$

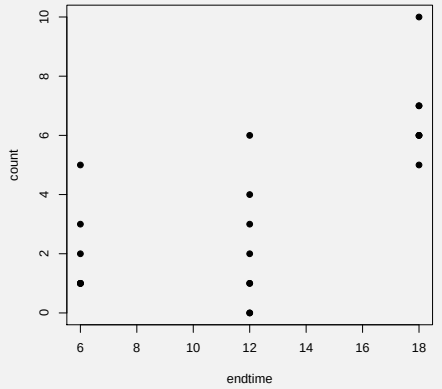
Aberrant Crypt Foci Example

- *Aberrant crypt foci* (ACF) are abnormal collections of tube-like structures that are precursors to tumors.
- In an experiment, researchers exposed 22 rats to a carcinogen and then counted the number of ACFs in the rat colons.
- There were three treatment groups based on time since first exposure to the carcinogen, either 6, 12, or 18 weeks.
- The data is in the DAAG data set ACF1 with variables `count` and `endtime`.
-
```
> library(DAAG)
> str(ACF1)
```

```
'data.frame': 22 obs. of  2 variables:
 $ count  : num  1 3 5 1 2 1 1 3 1 2 ...
 $ endtime: num  6 6 6 6 6 6 6 12 12 12 ...
```


Plot of Data

```
> attach(ACF1)
> plot(count ~ endtime, pch = 16)
```



Linear Predictor

```
> acf1.glm = glm(count ~ endtime, family = poisson)
> summary(acf1.glm)

Call:
glm(formula = count ~ endtime, family = poisson)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.46204  -0.47851  -0.07943   0.38159   2.26332

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.32152    0.40046  -0.803    0.422
endtime      0.11920    0.02642   4.511 6.44e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 51.105  on 21  degrees of freedom
Residual deviance: 28.369  on 20  degrees of freedom
AIC: 92.21

Number of Fisher Scoring iterations: 5
```

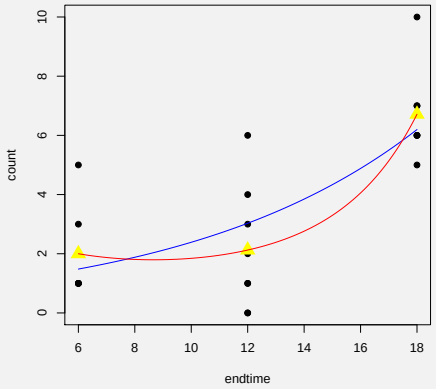

Quadratic Predictor

```
> acf2.glm = glm(count ~ endtime + I(endtime^2), family = poisson)
> summary(acf2.glm)

Call:
glm(formula = count ~ endtime + I(endtime^2), family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0616  -0.7834  -0.2808   0.4510   2.1693

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.722364   1.092494   1.577    0.115
endtime      -0.262356   0.199685  -1.314    0.189
I(endtime^2)  0.015137   0.007954   1.903    0.057 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 51.105  on 21  degrees of freedom
Residual deviance: 24.515  on 19  degrees of freedom
AIC: 90.354

Number of Fisher Scoring iterations: 5
```

Plots of Fitted Models

```
> plot(count ~ endtime, pch = 16)
> means = sapply(split(count, endtime), mean)
> unq = unique(endtime)
> points(unq, means, col = "yellow", pch = 17, cex = 2)
> u = seq(6, 18, length = 201)
> dfu = data.frame(endtime = u)
> eta1 = predict(acf1.glm, dfu)
> eta2 = predict(acf2.glm, dfu)
> lines(u, exp(eta1), col = "blue")
> lines(u, exp(eta2), col = "red")
```

- The red (quadratic) curve goes through the means exactly.



Dispersion

- The Poisson distribution assumes that the variance is equal to the mean.
- In real situations, this can fail when either:
  - There is a "random effect" for the individuals; or
  - There is a tendency for observations to cluster.
- We can examine this using a *quasipoisson* model.

Quasi-Poisson Model

```
> acf2q.glm = glm(count ~ endtime + I(endtime^2), family = quasipoisson)
> summary(acf2q.glm)

Call:
glm(formula = count ~ endtime + I(endtime^2), family = quasipoisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0616  -0.7834  -0.2808   0.4510   2.1693

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.722364   1.223433   1.408    0.175
endtime      -0.262356   0.223618  -1.173    0.255
I(endtime^2)  0.015137   0.008908   1.699    0.106

(Dispersion parameter for quasipoisson family taken to be 1.254071)

    Null deviance: 51.105  on 21  degrees of freedom
Residual deviance: 24.515  on 19  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```


- The dispersion parameter is estimated to be 1.25, more than 1.0.
- The quadratic term has lesser significance in the quasi-Poisson model.
- P-values for parameter estimates are now based on t statistics instead of z statistics.
- There is not enough information to compute the AIC.
- The residual deviance is the same for the Poisson and quasi-Poisson models, so there is not much information to allow for a comparison between the Poisson and quasi-Poisson models.