

# Poisson Regression

Bret Larget

Departments of Botany and of Statistics  
University of Wisconsin—Madison

February 26, 2008

- *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*.

1 / 26

Introduction

2 / 26

## 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  $\mu$  is

$$P\{Y = k | \mu\} = \frac{e^{-\mu} \mu^k}{k!} \quad \text{for } k = 0, 1, 2, \dots$$

- The Poisson distribution is discrete, like the binomial distribution, but has only a single parameter  $\mu$  *that is both the mean and the variance*.

## 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  $\lambda$  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

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

- 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  $\eta = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k$  can take on any value, we need to use a *link function* for the parameter  $\mu$ .
- The standard link function is the *natural logarithm*.

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

so that

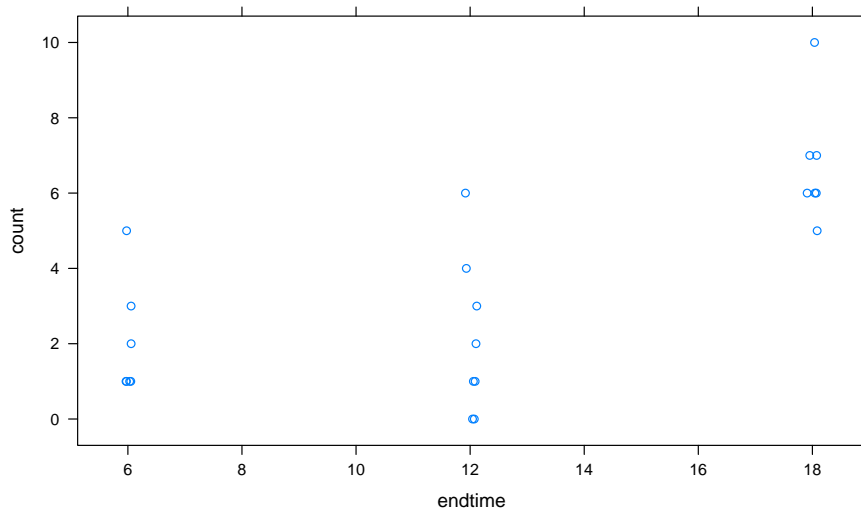
$$\mu = \exp(\eta)$$

## 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 variables are *count* and *endtime*.
- ```
> acf = read.table("acf.txt", header = T)
> str(acf)

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

## Plot of Data



## Linear Predictor

```
> library(arm)
> options(digits = 7)

> acf1.glm = glm(count ~ endtime, data = acf, family = poisson)
> display(acf1.glm, digits = 3)

glm(formula = count ~ endtime, family = poisson, data = acf)
      coef.est coef.se
(Intercept) -0.322  0.400
endtime      0.119  0.026
---
n = 22, k = 2
residual deviance = 28.4, null deviance = 51.1 (difference = 22.7)
```

## Fitted Model

- The fitted model is:

$$E_y = \exp(-0.322 + 0.119(\text{endtime})) = 0.725 \times \exp(0.119(\text{endtime}))$$

- For the three design times, 6, 12, and 18, the predicted means are:

```
> round(exp(predict(acf1.glm, data.frame(endtime = c(6,
+ 12, 18))))), 1)
```

```
 1  2  3
1.5 3.0 6.2
```

- Compare this to the sample means:

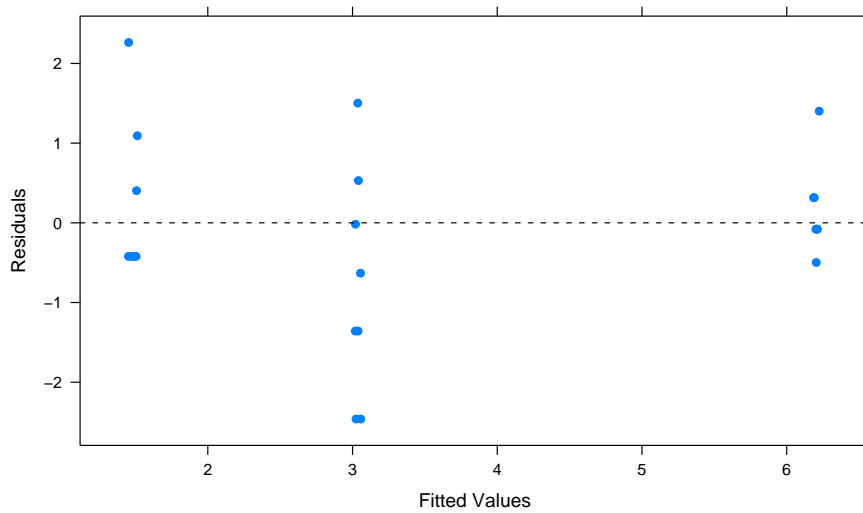
```
> with(acf, round(sapply(split(count, factor(endtime)),
+ mean), 1))
```

```
 6 12 18
2.0 2.1 6.7
```

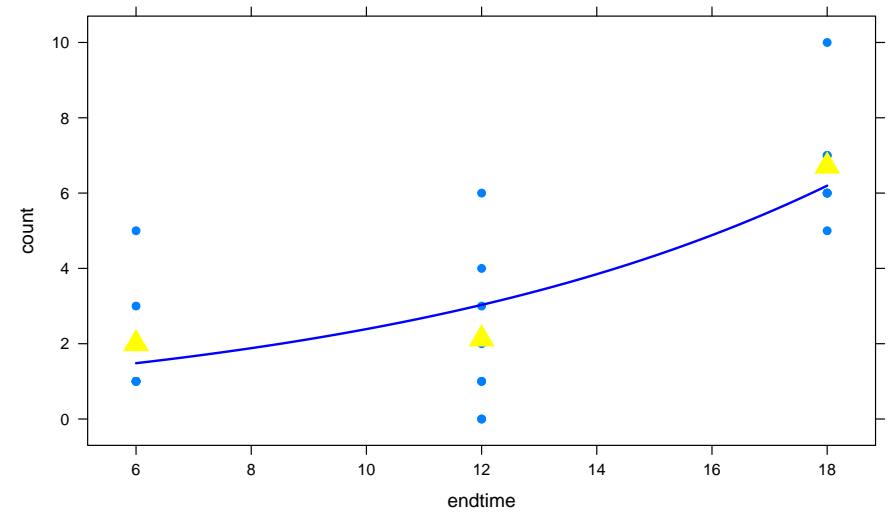
## Interpreting the Parameter

- The estimated parameter for endtime is 0.119.
- $\exp(0.119) = 1.126$ .
- An increase of one hour corresponds to an estimated increase of the count by about 12.6%.

## Residual Plot



## Plotting the Fitted Model



## Standardized Residuals

- The mean and standard deviation of the Poisson distribution are  $\mu$  and  $\sqrt{\mu}$ .
- We can standardize residuals by *subtracting the fitted values* and *dividing by the square root of the fitted values*.

$$z_i = \frac{y_i - \hat{y}_i}{\sqrt{\hat{y}_i}}$$

- We can estimate the *overdispersion* by *dividing the sum of squared standardized residuals by the degrees of freedom*.

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n z_i^2}{n - k}$$

## Numerical Example

```
> z = with(acf, (count - fitted(acf1.glm))/sqrt(fitted(acf1.glm)))  
> sum(z^2)/(22 - 2)
```

```
[1] 1.292732
```

- 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.

```
> acf1q.glm = glm(count ~ endtime, data = acf, family = quasipoisson)
> display(acf1q.glm, digits = 3)

glm(formula = count ~ endtime, family = quasipoisson, data = acf)
      coef.est coef.se
(Intercept) -0.322   0.455
endtime      0.119   0.030
---
n = 22, k = 2
residual deviance = 28.4, null deviance = 51.1 (difference = 22.7)
overdispersion parameter = 1.3
```

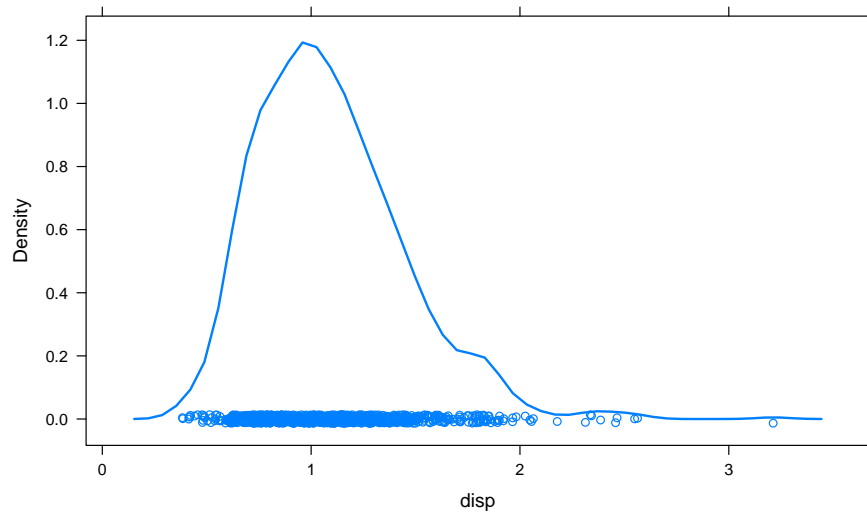
## Simulation

- The dispersion parameter is estimated to be 1.3, more than 1.0.
- We can test this with a *simulation*.

```
> mu = fitted(acf1.glm)
> disp = rep(NA, 1000)
> for (i in 1:1000) {
+   x = rpois(22, mu)
+   z = (x - mu)/sqrt(mu)
+   disp[i] = sum(z^2)/(22 - 2)
+ }
> summary(disp)

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.3838  0.8359  1.0440  1.0930  1.2930  3.2120
```

## Density Plot



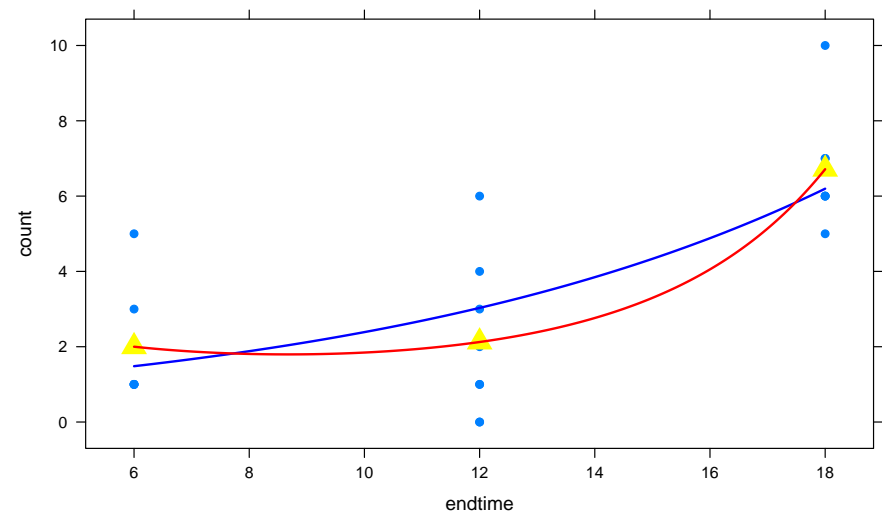
## Alternative Models

- We can also try a model with a quadratic predictor.

## Quadratic Predictor

```
> acf2.glm = glm(count ~ endtime + I(endtime^2), data = acf,  
+ family = poisson)  
> display(acf2.glm, digits = 3)  
glm(formula = count ~ endtime + I(endtime^2), family = poisson,  
     data = acf)  
             coef.est coef.se  
(Intercept)  1.722    1.092  
endtime      -0.262    0.200  
I(endtime^2)  0.015    0.008  
---  
n = 22, k = 3  
residual deviance = 24.5, null deviance = 51.1 (difference = 26.6)
```

## Plots of Fitted Models



## ANOVA

- Notice that the fitted quadratic model goes through the estimated means.
- Three points define a parabola.
- An ANOVA of the models shows the quadratic term does helps a little.

```
> anova(acf1.glm, acf2.glm)
```

Analysis of Deviance Table

Model 1: count ~ endtime

Model 2: count ~ endtime + I(endtime^2)

|   | Resid. Df | Resid. Dev | Df | Deviance |
|---|-----------|------------|----|----------|
| 1 | 20        | 28.3694    |    |          |
| 2 | 19        | 24.5146    | 1  | 3.8548   |

```
> 1 - pchisq(3.85, 1)
```

```
[1] 0.04974599
```

## Residual Plot

