GLMs with a Poisson

Bob O'Hara

March 8, 2019

Last Week: The binomial distribution

What is the binomial GLM (logistic regression)?

When and why to use a binomial GLM?

Link Functions

Categorical and continuous variables

Overdispersion

This Week: log-linear models

This week you will:

- learn about log-linear models
- learn about over-dispersion
    - when there is more error than you expect

A Typical Problem: Count data

Numbers of

- fish caught
- murders
- offspring
- bacterial/fungal colonies
- deaths due to lip cancer

A Model for Counts: Fishing

Figure 1: Anglers by Raoul Dufy

A Model for Counts: Fishing

We sit by the Seine, fishing. We catch fish at a constant rate

If we catch fish for an hour, how many fish do we catch?

A Model for Counts: Fishing

If we catch fish at reate λ, the meran number we catch in time t will be λt

The actual number will vary, and will follow a Poisson distribution:

$$Pr(N = r|\lambda) = \frac{\lambda^r e^{-\lambda}}{t!}$$

A Model for Counts

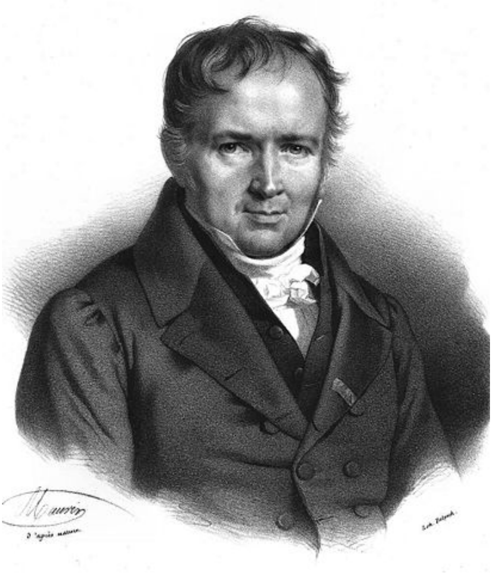


Figure 2: Siméon Denis Poisson

The Poisson distribution

Look at simulations of the Poisson distribution for different means (`plot(table(...))` is nicer than `hist()`)

What happens to the shape of the distribution when

- the mean is less than 1?
- the mean equals 1
- the mean is above 1
- the mean gets large?

```r
Pois <- rpois(1e3, 1)
plot(table(Pois), lwd=8, lend=3)
```

Is the Poisson a GLM?

The log-likelihood:

$$l(N = r|\lambda) = r \log \lambda - \lambda - log(t!)$$

GLM likelihood:

$$l(\theta|y) = \frac{y\theta - b(\theta)}{a(\phi)} - c(\phi, y)$$

Is the Poisson a GLM?

The log-likelihood:

$$l(N = r|\lambda) = r \log \lambda - \lambda - \log(t!)$$

GLM likelihood:

$$l(\theta|y) = \frac{y\theta - b(\theta)}{a(\phi)} - c(\phi, y)$$

So $a(\phi) = 1$

and $\theta = \log \lambda$

Interpretation

This is a GLM

The natural link function is a log link

- very rare something else is used

If we are counting, the process is multiplicative (double the effort, double the counts)

This is additive on the log scale.

Interpretation

The log link means that the model is multiplicative

$$\log(\lambda) = \alpha + \beta x$$
$$\lambda = e^{\alpha+\beta x} = e^{\alpha}e^{\beta x}$$

(we'll assume x is a dummy variable, i.e. 0 or 1)

e.g. if $\alpha = 0$, $e^{\alpha} = 1$. Then if β doubles the mean (i.e. $\lambda = 2e^{\alpha}$), $\beta = \log(2) = 0.69$

Some claims

If a coefficient is small, it is (approximately) the percent increase

- $e^{\alpha+\beta}$ means an increase by $e^{\beta} \approx 1 + \beta$ times (if β is small)

The coefficients are symmetrical

- a value of $+0.01$ increases the mean by $e^{0.01}$ times
- a value of -0.01 *decreases* the mean by $e^{0.01}$ times

Some claims: An Exercise

Write some questions to illustrate these claims:

- If a coefficient is small, it is (approximately) the percent increase
- The coefficients are symmetrical

e.g. “If we have a Poisson process with a mean of 1, what value of β would we need to double the mean?”

Model Fitting

Model fitting is easy:

```r
mu <- seq(1,2, length=10)
Count <- rpois(length(mu), mu)
m1 <- glm(Count ~1, family=poisson("log"))
m1a <- glm(Count ~1, family="poisson")
```

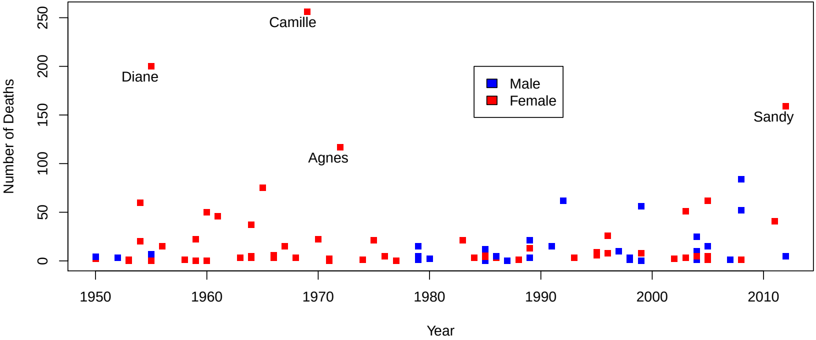
An Example: Himmicanes

A few years a go a strange paper appeared in PNAS that suggested that hurricanes in the USA with female names caused more deaths than those with male names.

```r
Stem <- "https://www.math.ntnu.no/emner/"
Fl <- "ST2304/2019v/Week13/Himmicanes.csv"
Data <- read.csv(paste0(Stem,Fl), stringsAsFactors=FALSE)
# Select hurricanes with > 100 deaths
BigH <- which(Data$alldeaths>100)
```

Himmicane Data

```r
plot(Data$Year, Data$alldeaths, col=Data$ColourMF,
     type="p", pch=15, xlab="Year",
     ylab="Number of Deaths")
text(Data$Year[BigH], Data$alldeaths[BigH],
     Data$Name[BigH], adj=c(0.8,1.5))
legend(1984,200,c("Male","Female"),fill=c("blue","red"))
```



The Data

The variables in the data are:

- Year: Year
- Name: Hurricane's name
- Gender: Gender (0: Male, 1: Female)
- MasFem: A scoring of how feminine the name sounds (we won't use this here)
- Minpressure: minimum air pressure in the hurricane (a measure of stength)
- Category: Category of hurricane (larger is more severe)
- NDAM: Normalised damage (i.e. how much the hurricane cost, corrected for inflation etc.)
- alldeaths: Number of deaths

The aim is to predict the number of deaths.

The Modelling Step 1: Chose a model

Your task (to discuss in your group):

- what distribution should we use?
- what link function?
- which variables do you want to consider to explain the number of deaths?
  - should we only use Gender, or do we need other variables?

Step II: Get Estimates of the Parameters

Fit the model!

Does there seem to be an effect of gender?

Step III: Model Checking

(we could also do model selection now)

The big problem: over-dispersion

Overdispersion

Too much variation!

Mean = Variance for a Poisson

But the data may have more variation (“extra-Poisson variation”)

What happens when we have overdispersion?

Let's simulate some data without over-dispersion

```r
alpha <- 1.5;   beta <- 0.1
X <- rnorm(1e3)
lambda <- exp(alpha + beta*X)
Y <- rpois(length(lambda), lambda)
var(Y)
```

```
## [1] 4.575532
```

. . . and simulate some data with over-dispersion

```r
eps <- rnorm(length(X), 0, 0.5)
lambda2 <- exp(alpha + beta*X + eps)
Y2 <- rpois(length(lambda2), lambda2)
var(Y2)
```

```
## [1] 11.05616
```

What happens when we have overdispersion III?

Fit the model without over-dispersion

```r
mod.noOD <- glm(Y ~ X, family = poisson())
summary(mod.noOD)
```

. . . and with overdispersion

```r
mod.OD <- glm(Y2 ~ X, family = poisson())
summary(mod.OD)
```

What happens when we have overdispersion IV?

Now, we can do these simulations lots of times

```r
SimWithOD <- function(alpha, beta, sigma, N) {
  X <- rnorm(N)
  eps <- rnorm(N, 0, sigma)
  lambda <- exp(alpha + beta*X + eps)
  Y <- rpois(length(lambda), lambda)
  mod <- glm(Y ~ X, family = poisson())
  coef(mod)["X"]
}
Repbeta0 <- replicate(1e2, SimWithOD(alpha=1.5, beta=0.1, s
Repbeta1 <- replicate(1e2, SimWithOD(alpha=1.5, beta=0.1, s
c(var(Repbeta0), var(Repbeta1))
```

The effects of Over-dispersion

The uncertainty in the estimates increases **but** this isn't seen in the confidence intervals

However, we can see it in the residual deviance increase

- this should (roughly) equal the residual degrees of freedom

Dealing With Overdispersion

There are a few ways to deal with overdispersion

- Correct in the likelihood
- Use a mixed model (later?)
- Use a different distribution

Correct the likelihood I

The likelihood is

$$l(\theta|y) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)$$

So we can estimate ϕ, the dispersion. We can use the deviance ratio.

Deviance/Degrees of Freedom

```
Dispersion <- deviance(mod.OD)/df.residual(mod.OD)
```

We can plug that into the summary:

```
summary(mod.OD, dispersion = Dispersion)
```

Correct the likelihood II

Effect is to increase standard errors by sqrt(Dispersion):

```
summary(mod.OD)$coefficients[,"Std. Error"]
```

```
## (Intercept)           X
##  0.01438724  0.01424773
```

```
summary(mod.OD, dispersion =
                Dispersion)$coefficients[,"Std. Error"]
```

```
## (Intercept)           X
##  0.02118581  0.02098038
```

Use a different distribution I

The Negative Binomial distribution assumes that there is over-dispersion

```
mod.NB <- MASS::glm.nb(Y2 ~ X)

round(summary(mod.NB)$coefficients, 2)
```

```
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)     1.58       0.02   75.21        0
## X               0.12       0.02    5.52        0
```

Use a different distribution: long version

Our model is $\log(\mu_i) = \sum_j X_{ij}\beta_j$. But we could add a random term, so it becomes $\log(\mu_i) = \sum_j X_{ij}\beta_j + \varepsilon_i$

If we use $\varepsilon_i \sim N(0, \sigma^2)$ this is like a regression

- need a Generalised Linear Mixed Model to estimate it

We could also use $e^{\varepsilon_i} \sim \chi^2_\nu$. This is the same as assuming a negative binomial distribution.

Back to Himmicanes

Is there evidence of over-dispersion?

What happens if you correct it?

- either using a neative binomial distribution or correcting the likelihood

Summary

We have seen a log-linear model, using a Poisson distribution

- use with counts

Interpret the parameters on the log scale

- multiplicative on the scale of the data

Easy to fit the model

Overdispersion is a common problem

- can now solve in 2 ways

Tomorrow

More of the same

- more models checking

Some links to binomial distributions