Lecture 12: log-linear models/Poisson regression

Bob O'Hara

bob.ohara@ntnu.no

Before we start. . .

- This week's exercises wioll be available soon (sorry)
- I will try to get a syllabus up this week.
- Last week's exercises don't have to be handed in until Friday

Data Generating Models

Modern statistics deals much more with mechanisms

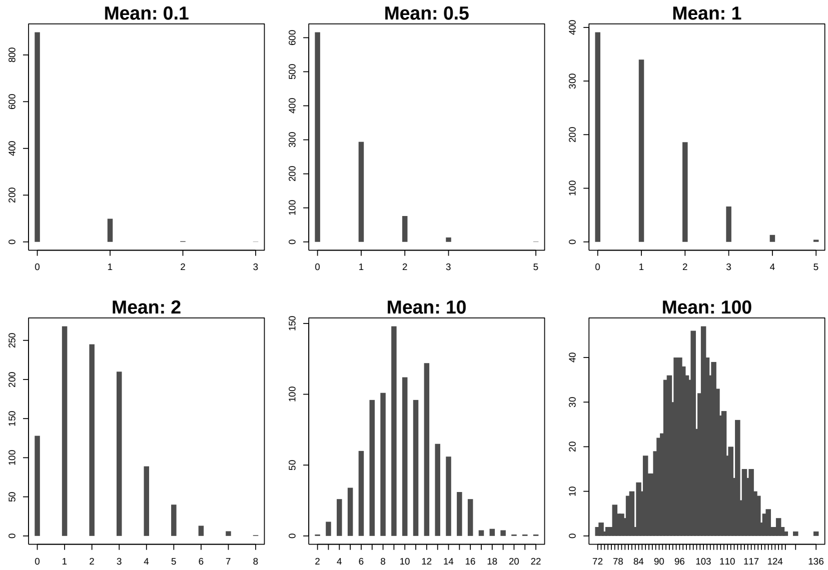
One major part: how the data were collected

Poisson Processes

Assume events happen at a constant rate, λ. If we observe for a time t then the expected number of events is $\mu = \lambda t$. The actual number varies around this, and follows a Poisson distribution:

$$Pr(N = r) = \frac{e^{-\mu}\mu^{r}}{r!}$$

A Poisson Distribution



Propoerties of the Poisson

if $r_1 \sim Poisson(mu_1)$ and $r_2 \sim Poisson(mu_2)$

- $E(r_1) = \mu_1$
- $Var(r_1) = \mu_1$
- $r_1 + r_2 \sim Poisson(\mu_1 + \mu_2)$
- $r_1 | r_1 + r_2 \sim Binomial(r_1 + r_2, \mu_1/(\mu_1 + \mu_2))$
- If $s \sim Binomial(N, p)$ with large N and small p then $s \approx Poisson(np)$

Poisson and binomial distributions are closely linked

Inference

Suppose we observe n counts from a Poisson with unknown mean μ, what is the maximum lilelihood estimate?

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

so

$$l(\mu|r) = -\mu + rlog(\mu) - log(r!)$$

DIffernetiate & set to 0:

$$0 = -1 + r\frac{1}{\mu}$$

so $\hat{\mu} = r$

Is this a GLM?

Remember that a GLM has a likelihood with the form

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

For the Poisson we have $l(\mu|r) = -\mu + rlog(\mu) - log(r!)$, so

- $\theta = log(\mu)$,
- $a(\phi) = 1$,
- and $b(\theta) = -e^{\theta}$, $c(y, \phi) = -log(r!)$

What This Means I

- $\theta = log(\mu)$

We have, naturally, a log link.

- this is the *canonical link*

Makes sense: if we are counting, the process is multiplicative (double the effort, double the counts)

This is additive on the log scale.

What This Means II

- $a(\phi) = 1$,

The dispersion is fixed, same as saying $Var(r) = \mu$

The amount of variation is determined by the mean

(we will see how to relax this later)

Interpretation

The log link means that the model is multiplicative

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

So the effect is multiplicative. For example, let x be 0 or 1, and $\beta = 0.01$. The means are

$$\mu_0 = e^{\alpha+0.010} = e^{\alpha}$$
$$\mu_1 = e^{\alpha+0.011} = e^{\alpha+0.01}$$

So the ratio μ_1/μ_0 is $e^{\alpha+0.01}/e^{\alpha} = e^{0.01} \approx 1.01$

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

Symmetry

The coefficients are symmetrical

e.g. if $\beta = -0.01$ then

$$\mu_1 = e^{\alpha - 0.011} = e^{\alpha} e^{-0.01} = e^{\alpha} / e^{0.01}$$

- reduces the mean by $e^{0.01}$ rather than increasing it by $e^{0.01}$

Hypothesis Testing and Deviance

We can use AIC/BIC just like before. But ANOVA is a bit different

- AIC = Deviance + 2*Number of parameters
- lowest is best

Deviance

From before Easter:

-2 times difference in likelihood between 2 (nested) models follows a χ^2 distribution

We call $-2l(\theta|Y)$ the *deviance*

So we can test whether a term shoud be in the model

Fitting a GLM in R

This is easy:

```r
X <- 1:100
SimR <- rpois(100, lambda = exp(1.5 + 0.0001*X))
mod <- glm(SimR ~ 1, family=poisson)
mod1 <- glm(SimR ~ X, family=poisson)

# More formally
mod <- glm(SimR ~ 1, family=poisson("log"))
```

- we use `glm` not `lm`
- `family=poisson` says to use the Poisson distribution
- `family=poisson("log")` says to use the Poisson distribution with a log link
- if we do not specify a link function, R will use the canonical link
    - i.e. the log link for the Poisson

Analysis of Deviance

```r
anova(mod, mod1, test="LRT")
```

```
Analysis of Deviance Table

Model 1: SimR ~ 1
Model 2: SimR ~ X
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1        99     95.487
2        98     94.961  1  0.52572   0.4684
```

Looking at a GLM in R

```r
summary( mod1)
```

```
Call:
glm(formula = SimR ~ X, family = poisson)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-2.9400  -0.6979  -0.1506   0.3887   2.2846  

Coefficients:
             Estimate Std. Error z value Pr(>|z|)    
(Intercept)  1.541490   0.094664  16.284   <2e-16 ***
X           -0.001198   0.001652  -0.725    0.468    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

(Dispersion parameter for poisson family taken to be 1)
```

An Example: The Hastings Rarities

British birders were worried that a lot of observations from around Hastings betwen 1890 and 1930 were frauds

John Nelder (co-inventor of GLMs) took a look at the data, and compared Hastings with two areas next to Hastings

https://en.wikipedia.org/wiki/Hastings_Rarities

Hastings Rarities Data

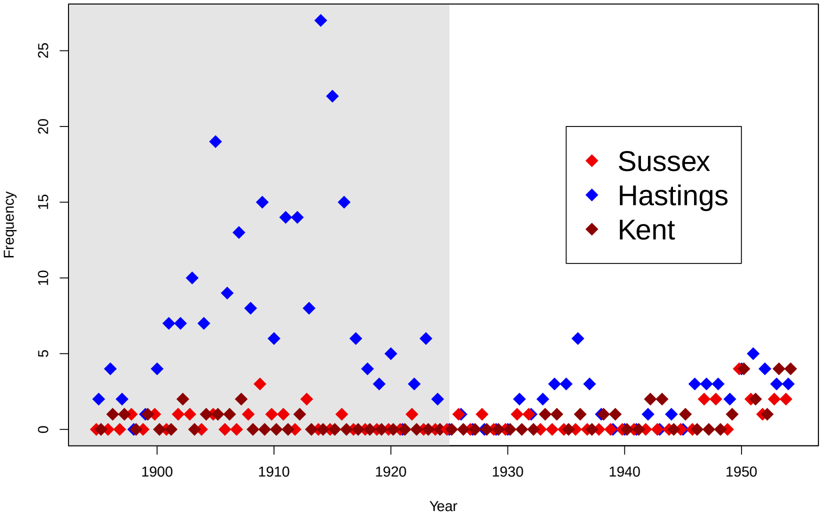
- Year (1895 to 1954)
- Area (Hastings, Sussex, Kent)
  - Hasting is a town in Sussex: Kent is next door
- Era (A, B: A is before 1925)
  - A is when the frauds were thought to occur
- Class (National rarity of species)
- Count: number of records

We will only look at the rarest species (Class I)

The problem: were there more rarities recorded around Hastings before 1925 (in Era A)?

Hastings Data

Concerned about Hastings before abnout 1925. . . .



Fitting & Testing the Model

```r
Hast.mod <- glm(Count ~ Area*Era, family="poisson",
                data=HastingsYearsI)
anova.hast <- anova(Hast.mod, test = "Chisq")
signif(data.frame(anova.hast), 3)
```

```
         Df Deviance Resid..Df Resid..Dev Pr..Chi.
NULL     NA       NA       179        848       NA
Area      2    349.0       177        499 1.41e-76
Era       1     81.6       176        417 1.67e-19
Area:Era  2     55.2       174        362 1.02e-12
```

- The test statistic for the interaction is 55.23 with 2 DF
- If we test this against a χ^2 distribution, we get p $=$ 10ˆ{-12}.

Parameter Estimates

```r
round(summary(Hast.mod)$coefficients, 2)
```

```
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)          -0.46       0.23   -1.99     0.05
AreaHastings          1.04       0.27    3.92     0.00
AreaKent              0.31       0.30    1.04     0.30
EraA                 -0.24       0.35   -0.68     0.49
AreaHastings:EraA     1.74       0.38    4.62     0.00
AreaKent:EraA        -0.62       0.50   -1.25     0.21
```

We can see that there were about $\exp(1.04) = 2.84$ times as many rare species around Hastings in Era B

But $\exp(1.74) = 5.7$ times more than *that* before 1925

95% Confidence intervals

All parameters (on log scale):

```r
round(CI <- confint(Hast.mod), 2)
```

```
                  2.5 % 97.5 %
(Intercept)       -0.94  -0.04
AreaHastings       0.54   1.59
AreaKent          -0.27   0.92
EraA              -0.93   0.44
AreaHastings:EraA  1.01   2.49
AreaKent:EraA     -1.62   0.35
```

For Hasting:AreaA on couint scale

```r
round(exp(CI["AreaHastings:EraA",]),2)
```

```
 2.5 % 97.5 %
  2.73  12.10
```

The Full Summary 1

```
summm.Hast <- paste(capture.output(print(summary(Hast.mod),
cat(summ.Hast[2:7])
```

```
Call:
 glm(formula = Count ~ Area * Era, family = "poisson", data

 Deviance Residuals:
    Min       1Q  Median       3Q      Max
  -4.02   -1.13   -0.86     0.62     5.22
```

The model (useful reminder) and summary of residuals (ignore!)

The Full Summary 2

```
Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
 (Intercept)          -0.46       0.23    -2.0     0.05 *
 AreaHastings          1.04       0.27     3.9    9e-05 ***
 AreaKent              0.31       0.30     1.0     0.30
 EraA                 -0.24       0.35    -0.7     0.49
 AreaHastings:EraA     1.74       0.38     4.6    4e-06 ***
 AreaKent:EraA        -0.62       0.50    -1.3     0.21
 ---
 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

Parameter estimates: very useful.

- also use coef() and confint()

The Full Summary 3

```
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 848.02  on 179  degrees of freedom
Residual deviance: 361.89  on 174  degrees of freedom
AIC: 639.1

Number of Fisher Scoring iterations: 5
```

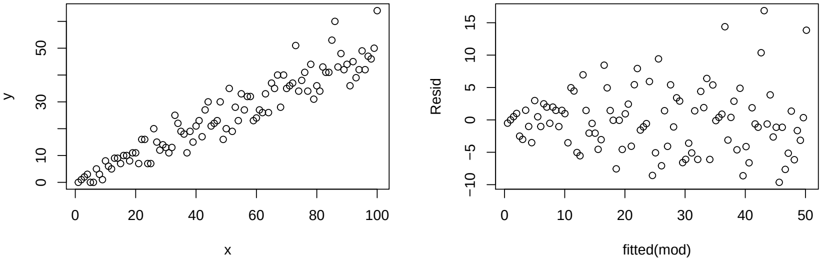
- Dispersion was mentioned last week; more on this and use of deviance later. . .
- AIC not useful without more models
- Ignore Fisher scorings: relate to efficiency in model fitting

Residuals

Our raw residuals are $y_i - E(y_i)$, i.e. Observed - Expected

We can plot these

```
x <- 1:100; y <- rpois(length(x), 0.5*x)
mod <- glm(y~log(x), family="poisson")
Resid <- y - fitted(mod)
par(mfrow=c(1,2))
plot(x, y); plot(fitted(mod), Resid)
```



Remember, $Var(x) = E(x)$

Better Residuals

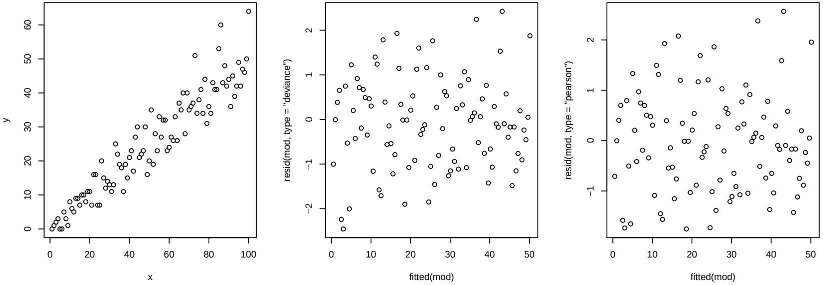
There are several solutions to this problem (none perfect)

- Pearson residuals: $(x - \mu_x)/\sigma_x$
- Deviance Residuals: $sgn(y_i - E(y_i))\sqrt{D_i}$
  - deviance for one datum is $D_i = -2 * l(y_i|\theta_i)$
  - $sgn(x)$ is 1 if $x > 0$ and -1 if $x < 0$

Deviance residuals are the default: they control for some of the variation in shape, but aren't perfect

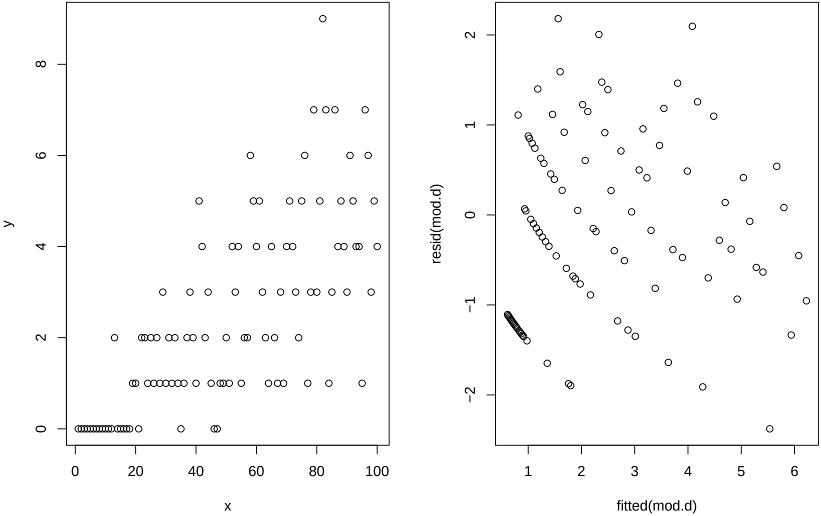
Better Residuals

```
par(mfrow=c(1,3))
plot(x, y); plot(fitted(mod), resid(mod, type="deviance"));
plot(fitted(mod), resid(mod, type="pearson"));
```



Residuals for Discrete Data

We get lines, from when $y = 0, 1, 2 \ldots$.

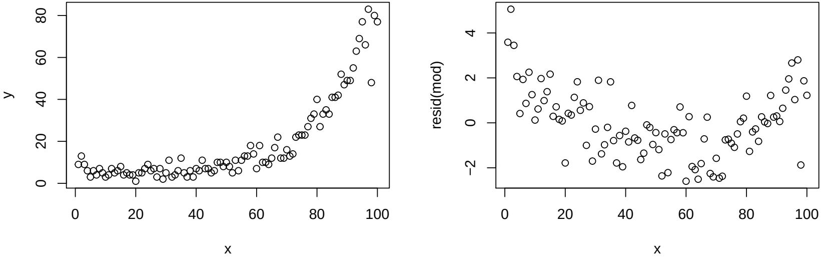


This is normal (if annoying)

Model Checking

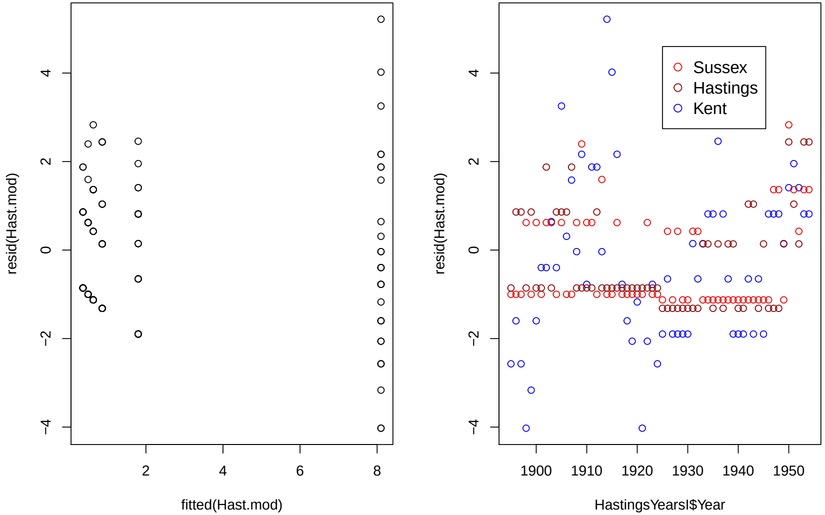
Deviance Residuals can still be informative

```r
x <- 1:100;
y <- rpois(length(x), exp(2-0.015*x + 0.0004*x^2))
mod <- glm(y~x, family="poisson")
par(mfrow=c(1,2))
plot(x, y); plot(x, resid(mod))
```



Look, curvature!

Hastings Residuals



Rise in reports after 1950?

Overdispersion

We assume that the mean controls the variance

But this is not always true: there might be extra variation

- e.g. for the Hastings data there might be more variation between years

We can check this!

Overdispersion

If the mean controls the variance, it then controls the amount of residual deviance

It turns out that the residual deviance should (asymptotically) follow a χ^2 distribution

- If there are lots of DF then residual deviance $\approx$ Deviance
- so we can test this!

Testing Overdispersion

Remember the summary?

```
cat(summm.Hast[20:24])
```

```
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 848.02  on 179  degrees of freedom
Residual deviance: 361.89  on 174  degrees of freedom
AIC: 639.1
```

We can use this to test for overdispersion

Testing Overdispersion

The residual deviance is 361.89, with 174 degrees of freedom. The p-value is

```r
pchisq(deviance(Hast.mod),
       df=df.residual(Hast.mod),
       lower.tail = FALSE)
```

```
## [1] 2.868446e-15
```

So it is unlikely that the data come from a Poisson distribution

- could be that there is another variable that should be in the model
- or there is just a lot more variation

Estimating Overdispersion

The ratio of deviance degrees of freedom is $361.89/174 = 2.08$

This is more useful: it acts like a residual variance.

If there is no overdispersion, this should be 1.

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 in the likelihood

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(Hast.mod)/df.residual(Hast.mod)
```

We can plug that into the summary:

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

```
##
## Call:
## glm(formula = Count ~ Area * Era, family = "poisson", da
##
```

Effect of Overdispersion

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

```
round(summary(Hast.mod)$coefficients[1:3,"Std. Error"],2)
```

```
##  (Intercept) AreaHastings     AreaKent
##         0.23         0.27         0.30
```

```
round(summary(Hast.mod, dispersion =
                  Dispersion)$coefficients[1:3,"Std. Error"],
```

```
##  (Intercept) AreaHastings     AreaKent
##         0.33         0.38         0.44
```

Effect of Overdispersion

Similar effect on anova():

```
cat(paste(capture.output(anova(Hast.mod, test="LRT")), "\n
```

```
##          Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
##  NULL                      179     848.02
##  Area     2   349.31       177     498.71 < 2.2e-16 **
##  Era      1    81.60       176     417.11 < 2.2e-16 **
```

```
cat(paste(capture.output(anova(Hast.mod, dispersion = 100,
```

```
##          Df Deviance Resid. Df Resid. Dev Pr(>Chi)
##  NULL                      179     848.02
##  Area     2   349.31       177     498.71   0.1744
##  Era      1    81.60       176     417.11   0.3664
```

Use a different distribution

The Negative Binomial distribution assumes that there is over-dispersion

```
Hast.NB <- MASS::glm.nb(Count ~ Area*Era, data=HastingsYear

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

```
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -0.46       0.27   -1.70     0.09
## AreaHastings          1.04       0.33    3.15     0.00
## AreaKent              0.31       0.36    0.87     0.38
## EraA                 -0.24       0.40   -0.59     0.55
## AreaHastings:EraA     1.74       0.47    3.72     0.00
## AreaKent:EraA        -0.62       0.57   -1.09     0.27
```

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.

Summary

- GLMs are like LMs
- anova() is (almost) the same
- summary() is the same
  - but parameter interpretation is important
- Because the Poisson assumes the variance, we might have to deal with that
- We should check for overdispersion & correct if necessary