Beyond GLM

Recap

Random effects, mixed models and beyond GLM

1. Generalised linear models
2. Random effects and mixed effects models
3. Polynomial regression
4. Generalised additive models

Generalised linear models

NB: beware. Some (perhaps many, even most) textbooks use GLM to refer to *Generalised* linear models, rather than *General* linear models.

Here, I'll abbreviate Generalised Linear Model as **GLIM**

Generalised linear models are a natural extension of GLMs – all GLMs that we have seen are also GLIMs, and GLIMs also generalise some other methods you may have come across – logistic regression, poisson regression

Essentially, GLIM attempts to relax the GLM assumption of normal errors and homogenous variance, while keeping the general framework of linear combinations of effects

Generalised linear models

A GLIM consists of three elements:

1. A distribution function *f*, from the exponential family, which specifies the shape of the error variance.

2. A linear predictor η, which is a linear combination of categorical and continuous explanatory variables, in exactly the same way as a model formula in a GLM

3. A link function *g* such that for fitted value y, g(y) = η. This controls how the combined explanatory variables are related to the fitted values for the response.

Generalised linear models

Note that for a link function g:

$g(y) = \eta$ is the same as: $y = g^{-1}(\eta)$

The equation for a GLIM looks like this:

e.g. For 2 categorical factors

$y_{i,j,k} = g^{-1}(\alpha_i + \beta_j) + \varepsilon_{i,j,k}$

Note that g is **almost** the same as transforming the response variable, except that it does not affect the error term: this means you can alter distribution of the residuals separately from the variance

GLM as Generalised linear model

To help you get your bearings, a GLM is a GLIM with

the normal distribution as its error distribution function

and the 'identity link' – i.e. The fitted values are exactly equal to the values of the linear predictor

Assumptions

GLM

1. Independence
2. Normality of error
3. Homogeneity of variance
4. Linearity/additivity

GLIM

1. Independence
2. Error follows distribution function
3. Variance distribution depends on distribution function
4. Linearity/additivity

Going from GLM to GLIM..

1. Lots of things stay familiar:
   - Model formulae
   - Partition variation in the data into different effects
   - Fit mixtures of categorical and continuous explanatory variables
   - Main effects and interactions
   - Coefficients, fitted values, residuals
   - Still need to check assumptions

- Some things change:
  - Error not normal, variance not constant
  - In general, no more F, no more SS or MS. Instead we have deviance (-2 * log likelihood) and a test using (usually) a χ^2 distribution

Binomial (binary) data and logistic regression

So far, all of our **response** variables have been measured on a continuous scale. GLIMs can be used to fit other kinds of responses, for example binomial data (taking only the values 0 or 1).

For this kind of data, we attempt to model the proportion of 1s and 0s i.e. a continuous variable between 0 and 1.

The appropriate tool for regression with this kind of data is called **logistic regression.** It is a GLIM with:

distribution function: binomial

link function: logit $g(p) = \ln\left(\frac{p}{1-p}\right)$.

Count data

Data based on counts is notoriously hard to analyse by ANOVA, GLM etc. Typically:

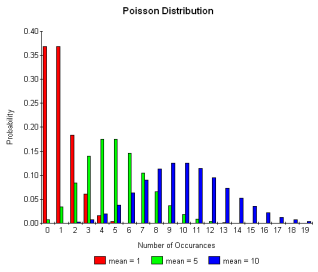
Variance increases with the mean

Zero counts are present – make transformations difficult

Errors are not normally distributed

Linear models will predict negative or fractional counts – these are nonsensical

Poisson errors



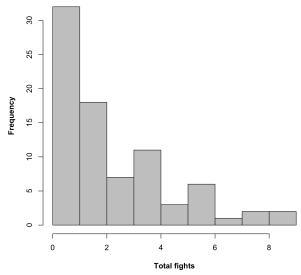
Poisson errors in a GLIM

This uses

the poisson distribution as its distribution function

and a log link: the fitted values are $e^{(\text{linear predictor})}$

Poisson regression example



Poisson regression example

```
>
> mod1<-glm(Fights~Density+I(Density^2),poisson)
> drop1(mod1,test="Chisq")
Single term deletions

Model:
Fights ~ Density + I(Density^2)
             Df Deviance    AIC    LRT  Pr(>Chi)
<none>            184.89 366.98
Density       1   198.41 378.50 13.518 0.0002364 ***
I(Density^2)  1   199.82 379.91 14.926 0.0001118 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Poisson regression example

```
> summary(mod1)

Call:
glm(formula = Fights ~ Density + I(Density^2), family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.53838  -1.53273  -0.09201   0.85437   2.63347

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.41988    0.43904  -0.956 0.338891
Density       0.57528    0.16402   3.507 0.000453 ***
I(Density^2) -0.05172    0.01409  -3.671 0.000241 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 200.18  on 81  degrees of freedom
Residual deviance: 184.89  on 79  degrees of freedom
AIC: 366.98

Number of Fisher Scoring iterations: 5

>
```

Poisson regression example

```
> summary(mod1)

Call:
glm(formula = Fights ~ Density + I(Density^2), family = poisson)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.53838  -1.53273  -0.09201   0.85437   2.63347

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.41988    0.43904  -0.956 0.338891
Density       0.57528    0.16402   3.507 0.000453 ***
I(Density^2) -0.05172    0.01409  -3.671 0.000241 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 200.18  on 81  degrees of freedom
Residual deviance: 184.89  on 79  degrees of freedom
AIC: 366.98

Number of Fisher Scoring iterations: 5

>
```

Poisson regression example

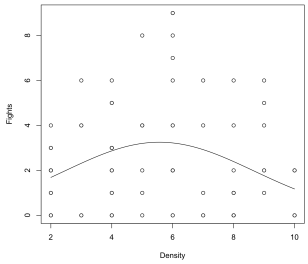
```
> mod2<-glm(Fights ~ Density + I(Density^2),family=quasipoisson)
> drop1(mod2,test="F")
Single term deletions

Model:
Fights ~ Density + I(Density^2)
             Df Deviance F value  Pr(>F)
<none>            184.89
Density       1   198.41  5.7757 0.01859 *
I(Density^2)  1   199.82  6.3777 0.01356 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Poisson regression example

```
> summary(mod2)

Call:
glm(formula = Fights ~ Density + I(Density^2), family = quasipoisson)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.53838  -1.53273  -0.09201   0.85437   2.63347

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.41988    0.61971  -0.678   0.5000
Density       0.57528    0.23152   2.485   0.0151 *
I(Density^2) -0.05172    0.01988  -2.601   0.0111 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 1.992334)

    Null deviance: 200.18  on 81  degrees of freedom
Residual deviance: 184.89  on 79  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

Poisson regression example

```
>
>
> X1<-seq(2,10,length=100)
> Y1<-exp(-0.41988+0.57528*X1-0.05172*(X1^2))
> plot(Density,Fights)
> points(X1,Y1,type="l")
>
```

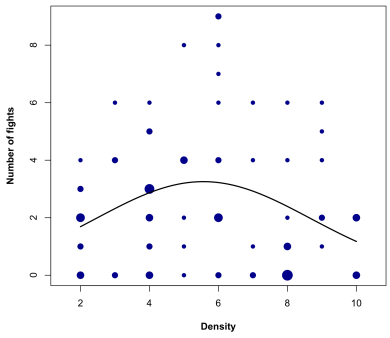


Poisson regression example

```
>
> bubble<-expand.grid(1:10,0:9,0)
> for (i in 1:100) {bubble[i,3]<-length(josh$Fights[josh$Density==bubble[i,1] & josh
$Fights==bubble[i,2]])}
> head(bubble)
  Var1 Var2 Var3
1    1    0    0
2    2    0    3
3    3    0    2
4    4    0    3
5    5    0    1
6    6    0    2
>
```

```
>
> plot(bubble[,1],bubble[,2],cex=sqrt(bubble[,3]),pch=16,xlab="Density",ylab="Number of
fights",col="darkblue",font.lab=2)
> points(X1,Y1,type="l",lwd=2)
>
```

Poisson regression example



Random effects and mixed effects models

What are random effects?

So far, all the factors we have examined have been **fixed effects**. For a fixed effect, we assume that we have included all levels of interest in the experiment, and our conclusions will be restricted to those levels

A **random effect** is a categorical variable where the levels can be thought of as random samples from a population of levels that we wish to reach conclusions about

What are random effects?

Another way to identify random effects is to think about repeating the experiment:

Would we consider the experiment meaningfully repeated only if we used the same levels of the factor, or would different levels be acceptable?

What are random effects?

Criterion:	*Repetition:* If the experiment were repeated:	*Desired inference:* The conclusions refer to:
Fixed effects	Same levels would be used	Just the actual levels used
Random effects	Different levels would be used	A population from which the levels used are just a (random) sample

Common random effects

- Brood
- Block within a field
- Household
- Individuals in repeated measures designs
- Parent

Random effect models

ANOVA: for a response *weight_gain* and explanatory factor *diet*, with a random effect being the individual animal chosen for the experiment from some population

$Weight\ gain_{i,j} = diet_i + individual_{i,} + \varepsilon_{i,j}$

In a random effects model, both *individual* and the error term are random variables, drawn independently for each observation from a normal distribution with mean 0.

So both the random effect term and the error have variance

Different hypotheses

For fixed effects, we're interested in the effect different levels have on the mean response, so e.g. For a four-level factor, a null hypothesis would be

$\mu_1 = \mu_2 = \mu_3 = \mu_4$

For a random effect, this question is of no interest – we don't care about the mean effect of choosing a particular leaf to measure. What we're interested in is the *variance* between different leaves. The null hypothesis is:

$\sigma_l^2 = 0$

Nesting

Random effects models are often *nested*, rather than *crossed/factorial*.

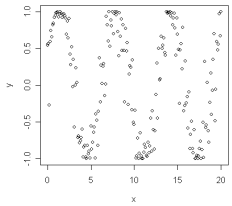
so each level (e.g. Biological Individual) of the random factor is only exposed to a single combination of factors

Denominators for F ratios

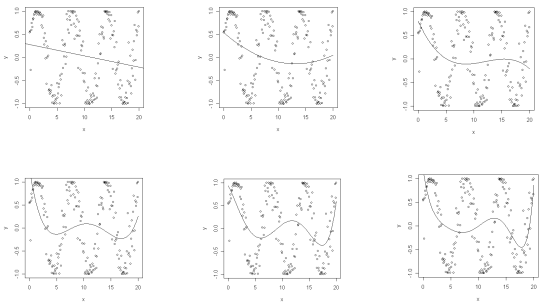
1. So far, the error MS has been the denominator for all the F ratios we have calculated.
2. This is also true for the lowest level of factors in a nested design, but not for others..
3. For some designs, no denominator exists.. 'synthetic denominators' are calculated in a way I don't really understand!

Highly non-linear data

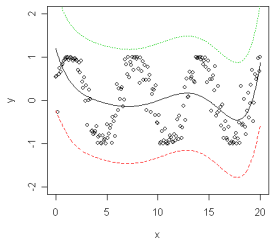
Sometimes we get data that is inherently non-linear, and difficult to fit with polynomials:



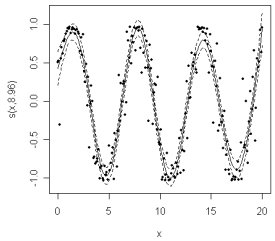
Polynomial regression



Polynomial regression is doing a terrible job!



General additive models



What is a General additive model?

GLM single-factor regression equation: $height_i = \alpha + \beta.y + \varepsilon_i$

GAM single-factor regression equation: $height_i = \alpha + S(y) + \varepsilon_i$

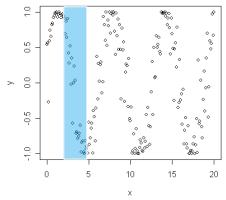
Where *S(y)* is a non-parametric smoother fitted to *y*

NB: *Generalised* additive models also exist

Non Parametric Smoothing

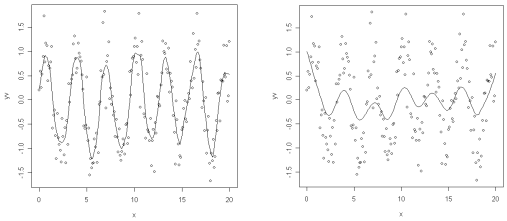
e.g. Running means, Lowess regression, cubic splines

Perform averaging/linear regression/polynomial regression
On a window sliding along the data:



Non Parametric Smoothing

Effect of Window size



General additive model

Just like with GLM, you can also do hypothesis testing with GAMs

The basic idea is familiar: measure SS of points away from the mean (total SS), the SS of points away from the smoothing line (error SS) and the SS of the line from the mean (effect SS)

Calculate an F-ratio as normal.

The difficulty is in calculating how many degrees of freedom the smoothed line has. This has to be estimated - usually, this isn't an integer, and is approximate

General additive models