More on GLMs

Akaike (again), Overdispersion

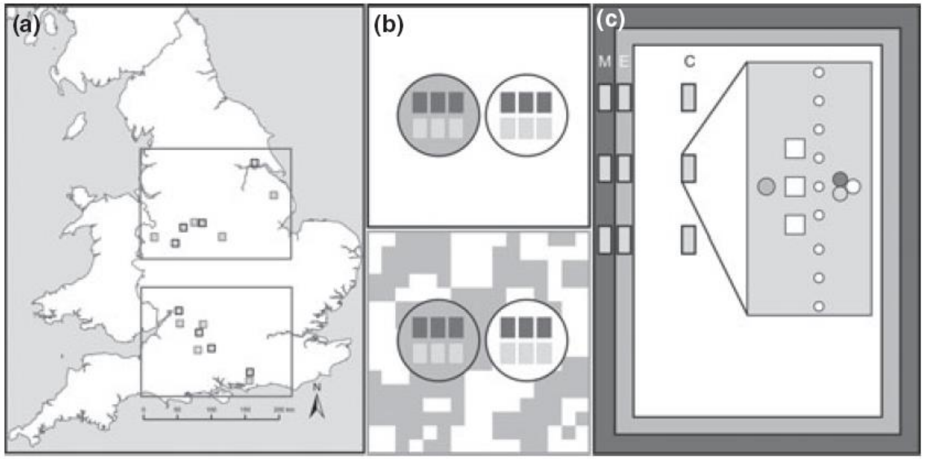
Standard approach to model simplification

Step	Procedure	Explanation
1	Fit the maximal model	Fit all the factors, interactions and covariates of interest. Note the residual deviance If you are using Poisson or binomial errors, check for overdispersion and rescale if necessary (see GLM notes later)
2	Begin model simplification	Inspect the parameter estimates Remove the least significant terms first, starting with the highest order interactions, progressing on to lower order interaction terms and then main effects. Remember that main effects that figure in significant interactions should not be deleted
3	If the deletion causes an insignificant increase in deviance	Leave that term out of the model Inspect the parameter values again Remove the least significant term remaining
4	If the deletion causes a significant increase in deviance	Put the term back in the model These are the statistically significant terms as assessed by deletion from the maximal model
5	Keep removing terms from the model	Repeat steps 3 or 4 until the model contains nothing but significant terms This is the minimal adequate model If none of the parameters is significant, then the minimal adequate model is the null model
6	When you have the MAM: diagnostics	Plot residuals against: fits, explanatory variables, sequence of data collection and as a histogram/normal plot. If problems: is model mis-specified? Is transformation necessary? Is error structure or link function mis-specified (see GLM section)
7	When you have the MAM: simplification	Are any coefficients close to "sensible" values and can be simplified? Consider using offsets to test (i.e. setting up a column with required values and constraining coefficient = 1).

Revisiting MAM

- With many factors in a model, model simplification can be complicated
  - Unstable MAM
  - Interpreting complex interactions etc
- Pragmatism
  - Forward and backwards methods to check model robustness
- AIC and Aikaike weights

example



- Hierarchical sampling design:
- Samples taken at different places (L)
- Within fields of different crops (CT)
- Within farms, managed as O or C (M)
- Within landscapes that are “hot or cold” (HC)
- Within regions (R)
- Huge number of potential models

Gabriel, D., et al. (2010). "Scale matters: the impact of organic farming on biodiversity at different spatial scales." Ecology Letters **13**(7): 858-869.

Density of butterflies individuals n=856

Model	Parameters	*K*	AICc	ΔAICc	w_i
111	R*CT*L+HC+M	14	2156.68	0.00	0.82
99	R*CT*L+M	13	2160.18	3.50	0.14
Null		5	2460.42	303.74	0.00
Global	R*HC*M*CT*L	36	2180.01	23.33	0.00

Density of hoverfly individuals n=4354

Model	Parameters	*K*	AIC	ΔAIC	w_i
99	R*CT*L+M	14	9297.06	0.00	0.39
111	R*CT*L+HC+M	15	9297.26	0.20	0.35
54	R*M*CT*L	21	9298.40	1.34	0.20
Null		6	9475.88	178.82	0.00
Global	R*HC*M*CT*L	37	9317.22	20.16	0.00

Density of bumblebee individuals n=4354

Model	Parameters	*K*	AIC	ΔAIC	w_i
45	R*CT*L	13	7087.20	0.00	0.14
48	HC*CT*L	13	7087.64	0.45	0.11
98	R*CT*L+HC	14	7088.36	1.16	0.08
99	R*CT*L+M	14	7088.74	1.55	0.07
54	R*M*CT*L	21	7088.76	1.56	0.07
44	M*CT*L	13	7088.83	1.64	0.06
40	CT*L	9	7089.08	1.88	0.06
88	HC*CT*L+M	14	7089.22	2.02	0.05
89	HC*CT*L+R	14	7089.22	2.03	0.05
111	R*CT*L+HC+M	15	7089.91	2.72	0.04
93	M*CT*L+HC	14	7089.97	2.78	0.04
68	CT*L+HC	10	7090.25	3.05	0.03
53	HC*M*CT*L	21	7090.26	3.06	0.03
92	M*CT*L+R	14	7090.42	3.22	0.03
69	CT*L+M	10	7090.62	3.43	0.03
67	CT*L+R	10	7090.66	3.47	0.03
114	HC*CT*L+M+R	15	7090.80	3.60	0.02
Null		6	7129.40	42.20	0.00
Global	R*HC*M*CT*L	37	7095.18	7.98	0.02

Model Simplification using AIC

Example for practical

```
> s<-read.csv2("soay2.csv")
> attach(s)
> m1<-glm(WEIGHT~factor(AGE)*STR*SEX)
> m2<-glm(WEIGHT~factor(AGE)*SEX+STR)
> m3<-glm(WEIGHT~factor(AGE)*SEX)
>
> aics<-data.frame(paste("m",1:3,sep=""),c(m1$aic,m2$aic,m3$aic),row.names=NULL)
>
> colnames(aics)<-c("model","AIC")
>
> aics<-aics[order(aics$AIC),]
>
> for(i in 1:dim(aics)[1]){aics$diff[i]<-aics$AIC[1]-aics$AIC[i]}
>
> aics$wi<-2.71828182845904523536^(0.5*aics$diff)
> aics$aic.weights<-aics$wi/sum(aics$wi)
> aics
```

```
   model      AIC       diff           wi  aic.weights
 2    m2 2142.962   0.000000 1.000000e+00 7.638520e-01
 1    m1 2145.309  -2.347831 3.091541e-01 2.361480e-01
 3    m3 2180.494 -37.532642 7.077671e-09 5.406293e-09
```

Model averaging

- Akaike weights show us how important each model is, within your specified set of models:

$$\frac{r_i}{\Sigma r_i} \quad \text{where} \quad r_i = \exp(-\tfrac{1}{2}\Delta\mathrm{AIC}_\mathrm{i})$$

- These can be used for averaging coefficients over several models.
- Now we can handle uncertainty among models as well as within models!

```
  model      AIC       diff           wi  aic.weights
2    m2 2142.962   0.000000 1.000000e+00 7.638520e-01
1    m1 2145.309  -2.347831 3.091541e-01 2.361480e-01
3    m3 2180.494 -37.532642 7.077671e-09 5.406293e-09
```

Getting GLMs to fit awkward (but real) data

Dispersion

- With a Poisson or Binomial model, the pdf implies that the dispersion index (residual deviance/residual df)=1
  - This is because variance in these models is a function of the mean (unlike in Gaussian models where Var is a free parameter)
  - So, a well fitted model has appropriate error deviance
- If the “empirical scale parameter” is not approx 1 then there is more (or less) deviance than you would expect given the error structure (and therefore the assumptions you are making),

If the Scale parameter is not approx 1?

(a) because you have not measured factors which are important

- residuals are not randomly distributed, but are affected by the unmeasured variable
- e.g. Poisson processes - require that there is a constant probability of an event happening in time or space and you may need to account for a "masking factor"

Or, (b), the underlying error structure is not correct

- e.g. parasites are aggregated, so the error structure is not Poisson but negative binomial.

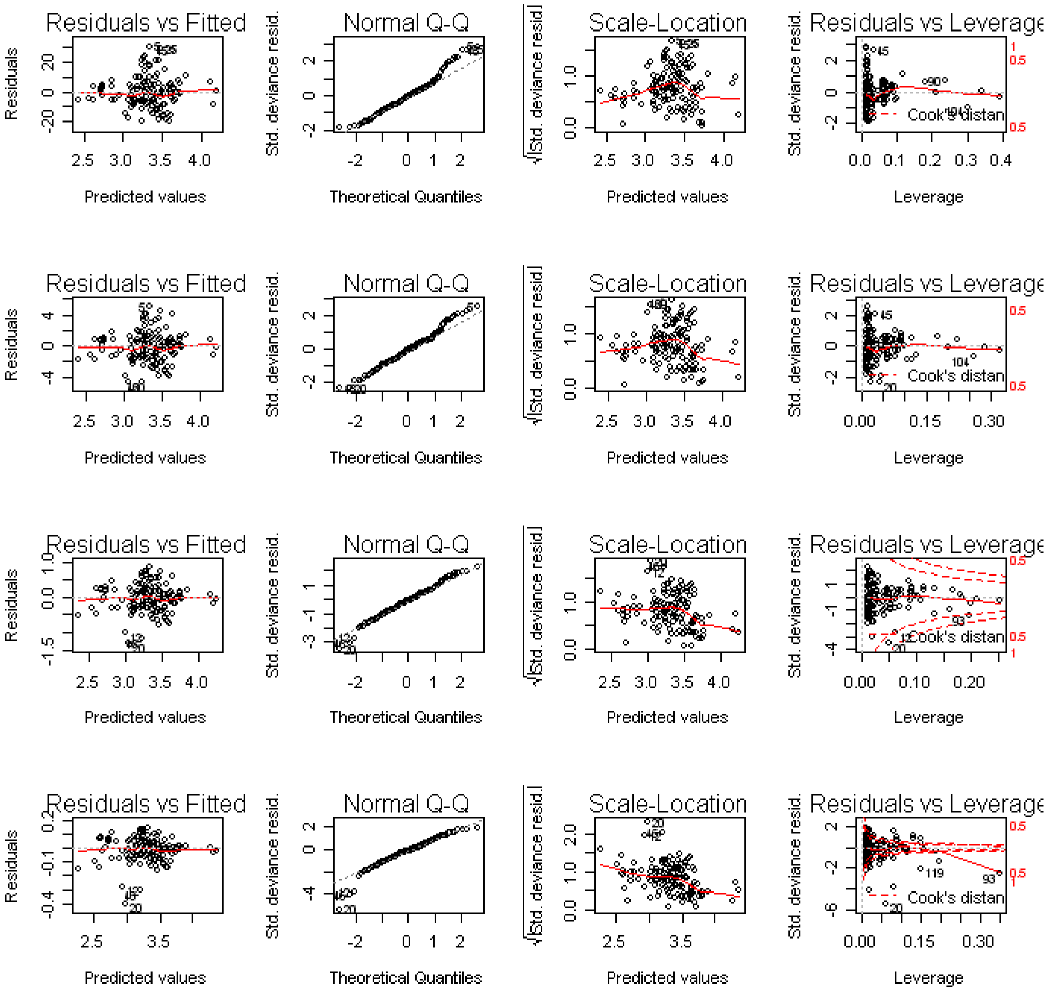
Fixing Overdispersion (1): *scaling deviances*

- For Poisson distribution the variance should equal the mean, and for a binomial the variance should equal np(1-p).
- *If* we *assume* that the variances are *proportional* to theoretical variances (rather than equal) by a factor s (scale parameter=dispersion index= resid dev/df)
- **then we can use F tests rather than chi-square tests.**
  - this measures the change in variance rather than the change in deviance:
  - it is the deviance scaled to units of the "error mean square" and so corrects for the over dispersion by essentially scaling down the sample size (fiddling the ratio of resid deviance to resid df).

Fixing Over-dispersion (2): *quasi-likelihood*

- *Quasi-likelihood* is the standard mechanism for scaling deviances
- allows the data to specify the variance of the error distribution.
  - Rather than supplying the error distribution and link, one specifies the variance (e.g. proportional to the mean) and the link.
  - In practice, try different recipes for the mean-variance relationship to minimise heteroscedasticity

```
brown2<-glm(Number~Area*Country,family=quasi(link=log,var="constant"))
brown3<-glm(Number~Area*Country,family=quasi(link=log,var="mu"))
brown4<-glm(Number~Area*Country,family=quasi(link=log,var="mu^2"))
brown5<-glm(Number~Area*Country,family=quasi(link=log,var="mu^3"))
```



```
glm(y~x,family=quasi(link=“log”, var=“mu”)), or
glm(y~x,family=quasipoisson)
```

Use F-tests in quasi models

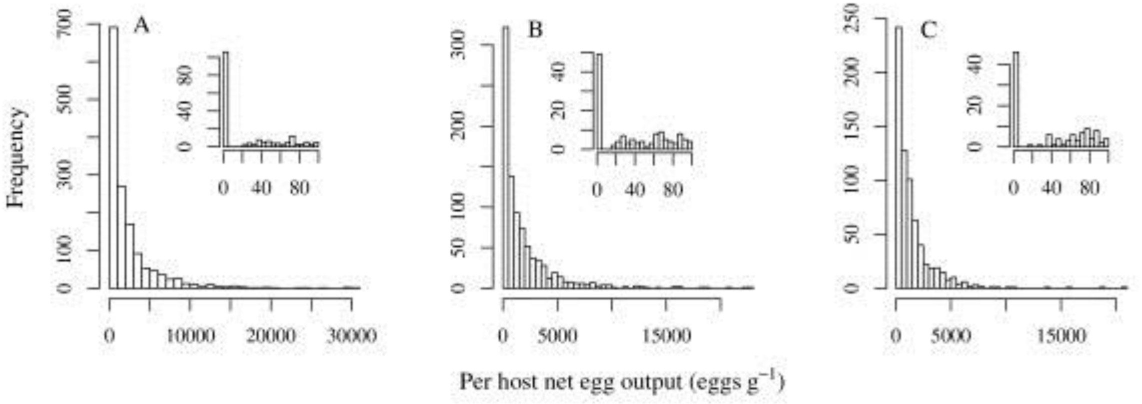
- In this process, the SEs of the parameter estimates (from `summary()`) are multiplied by √(*scale parameter)*, but parameter estimates are unaffected
- This procedure works well (though type II errors are more likely) but breaks down when sample sizes vary widely between groups.

		Significance:					
		Found			**Not found**		
		Result	Prob	Name	Result	Prob	Name
Effect	**Exists**	**Good**	**1-β**		**Error**	**β**	**Type II**
	Doesn't	**Error**	**α**	**Type I**	**Good**	**1-α**	

Fixing over-dispersion (3): getting the model right!

- Include missing explanatory variables
- Good Experimental Design
- Get the right error-distribution
- e.g. neg bin for count data
  - – Dogs in cars
  - – parasites

(4) If problems persist:



The distribution of per host egg output. Histograms depicting the distribution of the per host egg output in the baseline (A), 1st (B) and 2nd (C) re-infection populations. The insets are histograms of the distribution between 0–100 eggs gram-1 highlighting the high proportion of zero counts.
Walker *et al. Parasites & Vectors* 2009 **2**:11 doi:10.1186/1756-3305-2-11

- Think about a different model
  - – Mixture models such as zero-inflated Poisson or NB (e.g. R package "pscl")
  - – "hurdle approach" two stage models (zero vs non-zero and then "if non-zero how big/many")
  - – Random effects (gls or lmer) if you can see that there may be a cause (e.g. different areas have different variances)
  - – Fiddle the data

“Fiddle the data” (is 0 really 0, or NA?)

Zeroes could be due to:

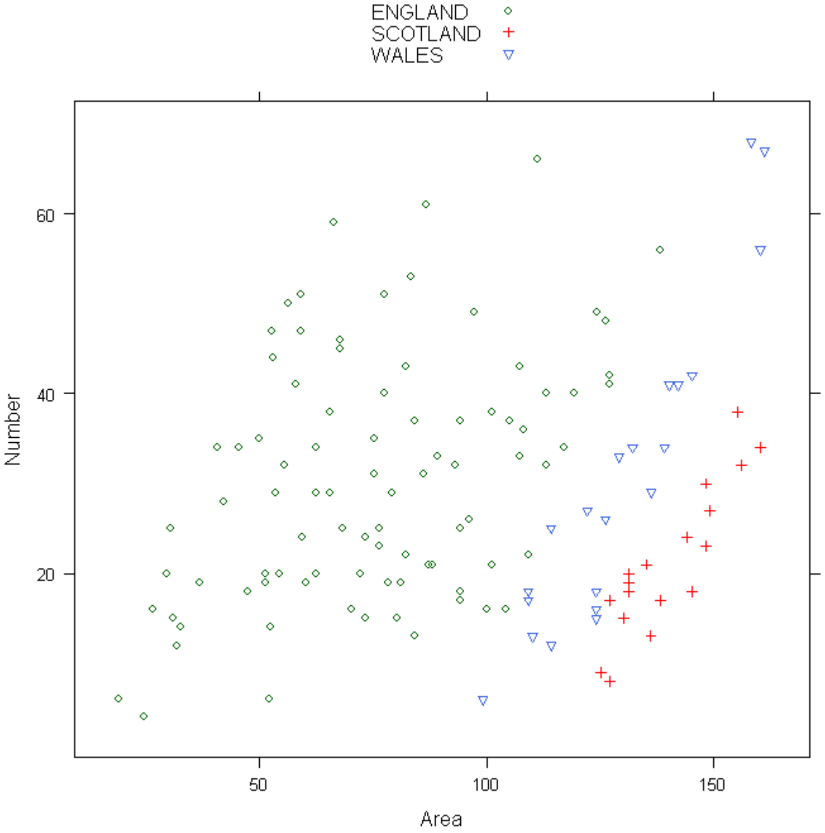
- Structural errors/ design errors (looking at wrong time or place – e.g. no farms in sea, no growth in winter)
- Process error (zero count because subject made mistake or didn’t need to be there)
- Observer error (zero count because you fell asleep)

Solutions: examine zeros and drop some

- preferably finding some evidence to do so – e.g. subsetting data to construct a robust analysis and using this to identify dodgy data)

An example

- Bird counts (species=LBJ = “little brown job”) per area according to country



1. Model without all variables

```
>brown0<-glm(Number~1.,family=poisson)
> anova(brown0)

Analysis of Deviance Table
Model: poisson, link: log
Response: Number
Terms added sequentially (first to last)

     Df Deviance Resid. Df Resid. Dev
NULL                       126     833.35
```

Over dispersed with s approx 7: reason not enough explanatory variables
as country and area are creating lots of variation not being taken into account

2. Poisson model

```
> brown1<-glm(Number~Area*Country,family=poisson)
> anova(brown1)

Analysis of Deviance Table
Model: poisson, link: log
Response: Number
Terms added sequentially (first to last)


              Df Deviance Resid. Df Resid. Dev
NULL                            126     833.35
Area           1    33.26       125     800.10
Country        2   202.68       123     597.42
Area:Country   2   112.90       121     484.52
```

Over dispersed with s approx 4

3. Poisson and F-tests

```
> anova(brown1,test="F")

Analysis of Deviance Table
Model: poisson, link: log
Response: Number
Terms added sequentially (first to last)


              Df Deviance Resid. Df Resid. Dev       F    Pr(>F)
NULL                            126     833.35
Area           1    33.26       125     800.10  33.257 8.076e-09 ***
Country        2   202.68       123     597.42 101.339 < 2.2e-16 ***
Area:Country   2   112.90       121     484.52  56.451 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Warning message:
using F test with a poisson family is inappropriate in: anova.glm(brown1, test = "F")
>
```

- **NEED TO MULTIPLY SEs of coefficients IN COEFs TABLE BY SQRT(484/121)**

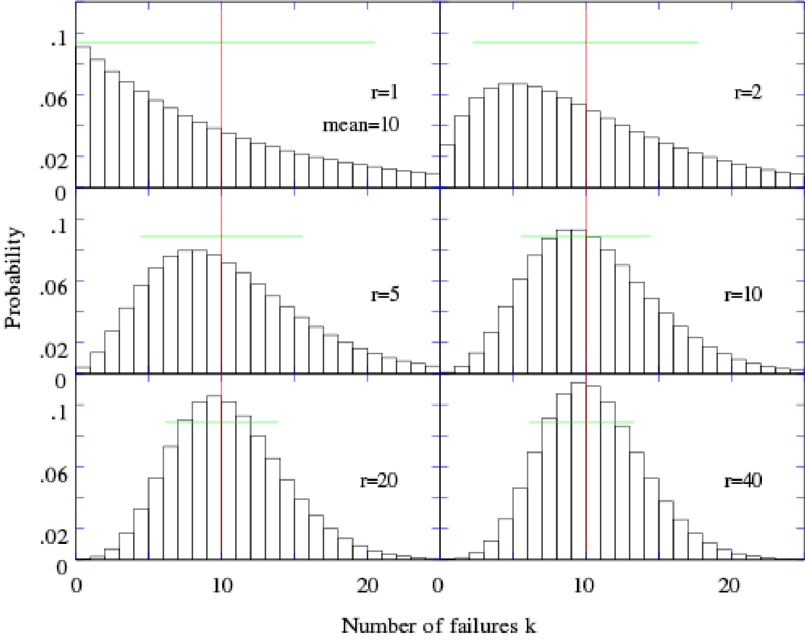
4. Quasi-likelihood

```
> brown2<-glm(Number~Area*Country,family=quasipoisson)
> anova(brown2,test="F")
```

```
Analysis of Deviance Table
Model: quasipoisson, link: log
Response: Number
Terms added sequentially (first to last)


             Df Deviance Resid. Df Resid. Dev       F    Pr(>F)
NULL                           126     833.35
Area          1    33.26       125     800.10  8.2754  0.004751 **
Country       2   202.68       123     597.42 25.2165 7.011e-10 ***
Area:Country  2   112.90       121     484.52 14.0470 3.267e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

5. Neg. Bin. errors



```
> brown3<-glm.nb(Number~Area*Country)
> anova(brown3)

Analysis of Deviance Table
Model: Negative Binomial(9.7805), link: log
Response: Number
Terms added sequentially (first to last)


              Df Deviance Resid. Df Resid. Dev P(>|Chi|)
NULL                            126    220.651
Area           1    8.461       125    212.189     0.004
Country        2   53.134       123    159.056 2.898e-12
Area:Country   2   28.214       121    130.842 7.473e-07
Warning message:
tests made without re-estimating 'theta' in: anova.negbin(brown3)
```

P:	3.045e-25
P(F):	2.2e-16
QP:	3.267e-06
NB:	7.473e-07

Plus any “fitted” (i.e. line or mean or prediction) will have different SEs

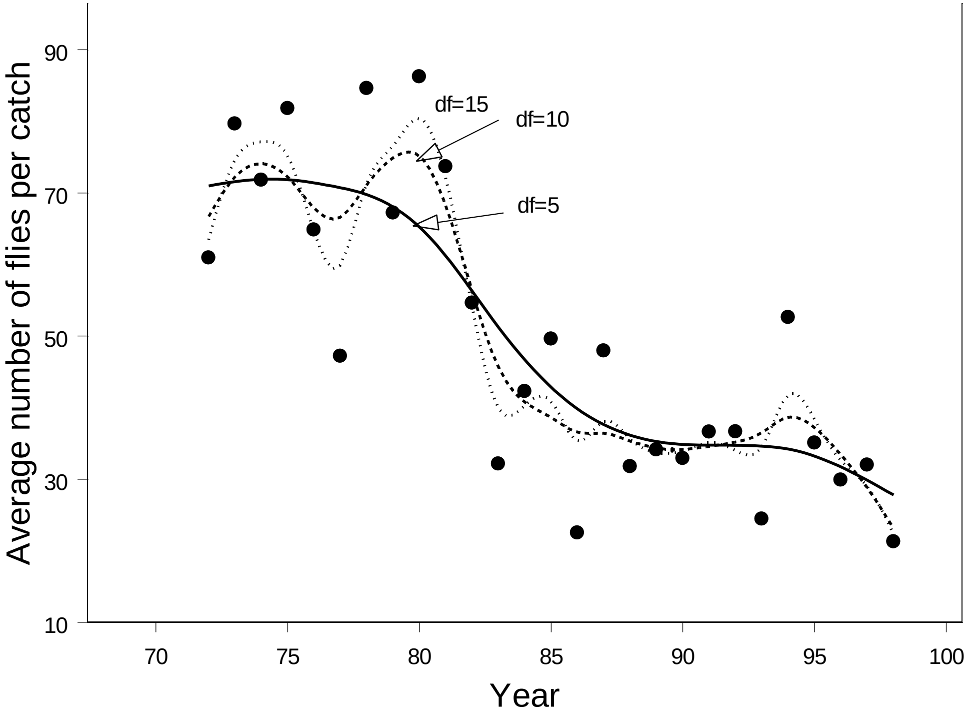
Why bother with doing it right?

Analysis of fish sex-ratios in relation to distance from pollution source *(from K. Wilson 2002 in ICW Hardy ed Sex Ratios CUP)*

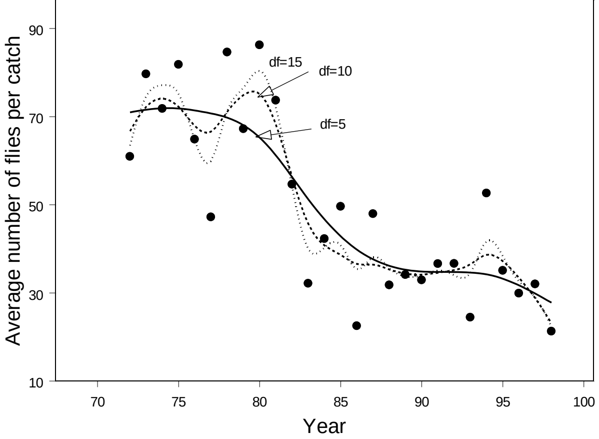
Model	Test	Type of model	***P*-value** (Distance)
1	linear model (unweighted, untransformed data)	Parametric – normal errors	P = 0.013 *
2	linear model (unweighted, arcsine-transformed data)	Parametric – normal errors	P = 0.012 *
3	linear model (unweighted, logit-transformed data)	Parametric – normal errors	P = 0.012 *
4	linear model (weighted, arcsine-transformed data)	Parametric – normal errors	P > 0.16 ns
5	Generalised linear model (weighted, untransformed data)	Parametric – binomial errors	P > 0.21 ns

GAMs: generalised additive models

- Sometimes parametric models are too restrictive
  - Parametric regressions have a fixed shape that may not match data
- Non-parametric regressions possible
  - Splines, loess etc

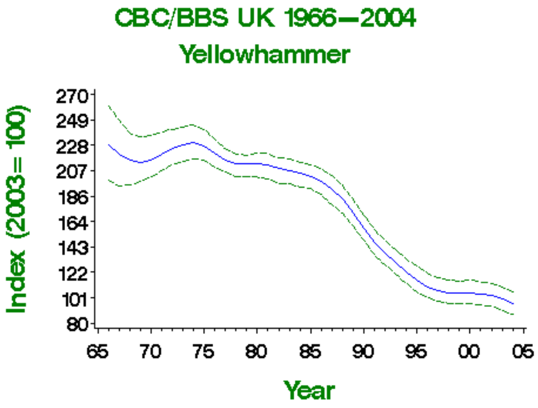


GAMs



- Data is a decline in fly numbers with time
  - appropriate GLM would be `numbers ~ year`.
  - GAM **`numbers ~ s(year,df)`** (where s stands for spline).
- As with GLM, GAMs require the relevent error structure and link function to be specified.
- The Chi-square/F that results will measure the change in deviance associated with the spline (i.e. is the curviness significant?)

GAMs

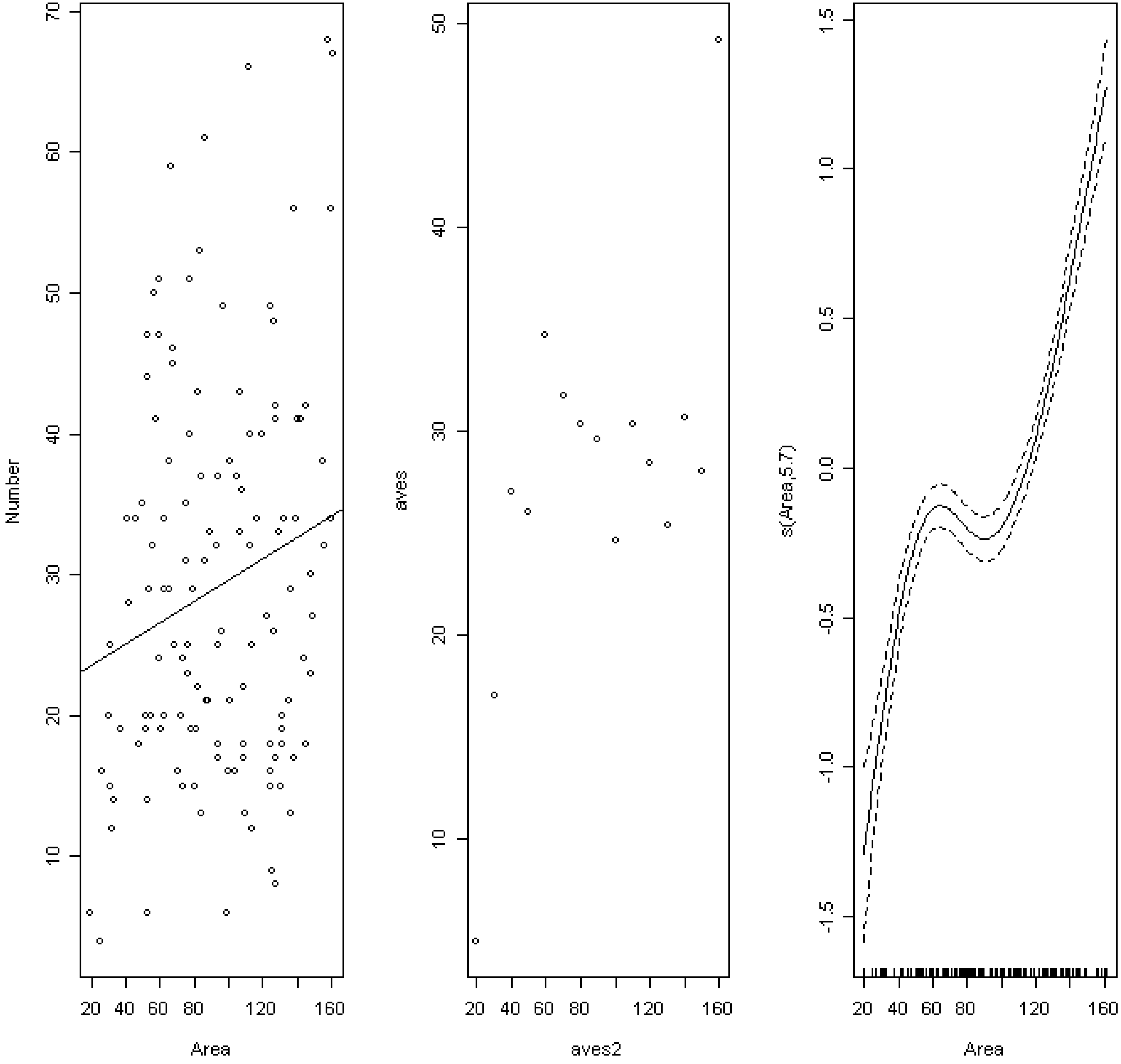


- As with GLMs, GAMs can contain multiple explanatory variables and can have a mixture of parametric and non-parametric terms

  (e.g. `log numbers ~ s(year,10)+site`)

- Non-parametric means no parameters!
  - CIs or changes in gradient need to be bootstrapped

```
>library(mgcv)
> brown5<-gam(Number~s(Area)+Country, family=poisson)
```


Number
70
60
50
40
30
20
10
20 40 60 80 120 160
Area
aves
50
40
30
20
10
20 40 60 80 120 160
aves2
s(Area,5.7)
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
20 40 60 80 120 160
Area

Practical

- Generalised LM part 2
  - This afternoon finish GLMs, go back over previous work, etc
  - Start thinking about your own data with our input?