Note from Toby 9-MAR-18:

I've uploaded these slides for exclusive use of NERC employees (in case they are interested). These are the GLM part of the Big Data course I did from Feb 2018 https://www.conted.ox.ac.uk/events/view/big-data-in-environmental-biology .

I'll be continuing to develop the slides during 2018 for a rerun of the course in 2019 (and possibly some training at BES Birmingham in Dec 2018, but that's dependent on funding) so I'd be very interested in any feedback / spotted errors (pls email me on tobmar@ceh.ac.uk).

I'm also in Office EF12 at CEH Wallingford if you find these methods and have difficulty applying them: my help comes at the price of however many coffees it takes us to go through things, OK?

Best,
Toby

NERC SCIENCE OF THE ENVIRONMENT CEH

Generalized Linear Models (GLMs)

Quantitative methods are evolving fast in ecology, way faster than any of us can keep up with. We lack the foundational training in mathematical, statistical or computational skills to pick these up easily: otherwise we'd work for banks, obviously. One consequence is that many of us spend a lot of our time feeling frustrated by quantitative methods.

Matthew Smith, Microsoft Research,
Introduction to the BES Computational Ecology Group
@BES_CE_SIG in the March 2014 British Ecological Society (BES) *Bulletin*.

Generalized Linear Models (GLMs)

1. Duiker in Africa

2. Recap of standard statistics

3. Validation

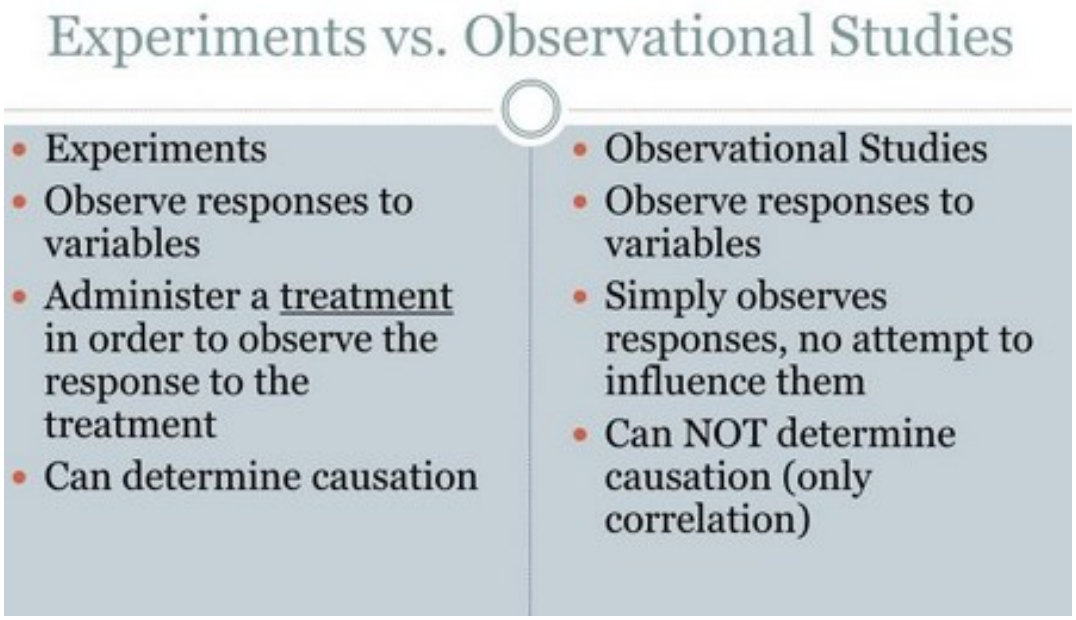
4. Interaction terms, fixed & random effects and Crawley's Two Stage Backwards Deletion Method

5. Error structure, Goodness of fit, overdispersion and using glmmPQL, dredge and glmer

Experimental design

Experimental design: There are two different classes of field experiment (Hurlbert 1984, Krebs 1999:ch.10):

- **Manipulative experiments** (or just 'experiments')
- **Mensurative experiments** ('observational studies')



FOOTNOTE (VERY IMPORTANT, BUT ONLY TO BE READ WHEN WE GET TO FIXED AND RANDOM EFFECTS):

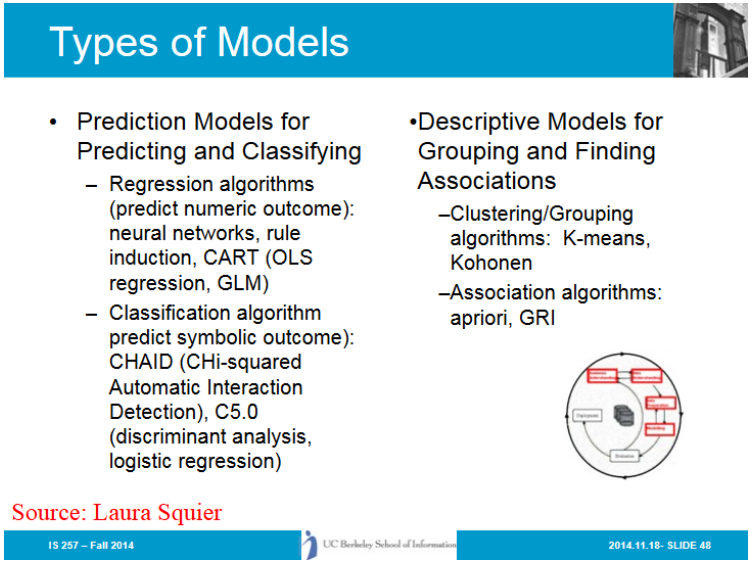
Personally, I generalise this a little: in a manipulative experiment the goal is to keep every variable constant across the design except for the 'key' variables being manipulated, but this is never 100% achievable: there is always residual/natural variation because of other known environmental factors (some of which you may have measured, some not) and residual variation (i.e. from unknown factors, sometimes called latent variables). Therefore, any manipulative experiment is in reality partly manipulative and partly mensurative.

Also, mensurative experiments can be 'approximately' manipulative, e.g. if your sites are chosen to represent the spectrum of mean annual temperatures (*MAT*), then even though you didn't manipulate the *MAT* at these sites, as long as you can argue that the sites are otherwise equivalent then you *effectively* have done so.

Two multivariate 'camps'

'process' →

(analysis of data from manipulative expts)



← 'pattern'

(analysis of data from observational studies)

For me, the world of multivariate statistics is divided into TWO main camps, which you can label (at least, I do personally) by the ecological mantra 'pattern and process'. People who are interested in **process/predicting** will gravitate towards regression and linear model techniques, those interested in **pattern/associations** will incline more towards ordination / cluster analysis techniques.

The division here isn't absolute, of course, but I find it useful. Surprisingly, very few researchers compare the two (Guisan *et al.* (1999)'s "GLM versus CCA spatial modeling of plant species distribution" in *Plant Ecology* is the only example I know of).

In this course, we cover both camps.

NERC SCIENCE OF THE ENVIRONMENT CEH

Generalized Linear Models (GLMs)

1. Duiker in Africa

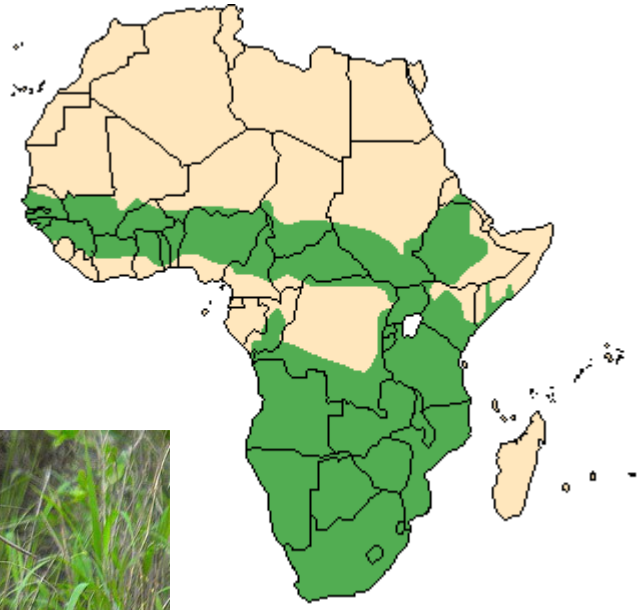
2. Recap of standard statistics

3. Validation

4. Interaction terms, fixed & random effects and Crawley's Two Stage Backwards Deletion Method

5. Error structure, Goodness of fit, overdispersion and using glmmPQL, dredge and glmer

1. Duiker in Africa



Duikers are very widespread small/medium-sized antelopes in Africa In this example you have collected population data on Common Duiker *Sylvicapra grimmia* in a savanna in Africa (warning: data are made-up just to show how this kind of analysis works).

1. Duiker in Africa

(n.b. all data are fake: it's just an example)

Usually you would have these data in a spreadsheet and get them into R by exporting to a .csv file and reading in using the read.table() or read.csv() commands, but this example is very small (*n*=27 readings only) so we can enter the data directly into R like this:

```
SiteCode=c("S1","S2","S3","S4","S5","S6","S7","S8","S9","S10","S11","S12","S13","S14","S15","S16","S17","S18","S19","S20","S21","S22","S23","S24","S25","S26","S27")  #Savanna sites S1-S27
MAT=c(25,21,25,21,29,27,25,29,29,20,23,27,29,23,25,21,29,29,20,23,27,29,21,29,27,21,20)  #Mean annual temperature at the sites
SiteLat=c(-12.3,-12.3,-12.3,-11.8,-11.7,-8.5,-8.6,-8.5,-9.4,-8.1,-7.1,-7.6,-6.6,-6.6,-6.6,-4.3,-4.3,-4.3,-2.9,-3.3,-3.3,-2.3,-2.3,-2.3,-2.3,-2.3,-2.3)  #Latitude of each site (imagining they are in a clustered transect from E. Zambia to Rwanda)
SoilTexture=factor(c("Sand","Loam","Loam","Sand","Clay","Sand","Loam","Clay","Sand","Clay","Loam","Loam","Loam","Sand","Clay","Sand","Loam","Loam","Loam","Sand","Loam","Clay","Clay","Clay","Clay","Loam","Loam"),levels=c("Clay","Sand","Loam"))  #The three basic soil textures in the area
SurvProb=c(1.00,0.21,0.76,0.75,1.00,1.00,0.76,0.95,1.00,0.25,0.76,0.76,0.76,1.00,1.00,0.75,0.76,0.76,0.01,1.00,0.46,1.00,0.70,1.00,1.00,0.51,0.06)  #Survival probability of Common Duiker to 2 yrs old (estimated from literature, say)

dataf=data.frame(SiteCode,MAT,SiteLat,SoilTexture,SurvProb)
head(dataf)  #Use head() to see the first 6 lines of the data frame
View(dataf)  #Use View() to see the whole data frame in a window
```

Data: dataf

	SiteCode	MAT	SiteLat	SoilTexture	SurvProb
1	S1	25	-12.3	Sand	1.00
2	S2	21	-12.3	Loam	0.21
3	S3	25	-12.3	Loam	0.76
4	S4	21	-11.8	Sand	0.75
5	S5	29	-11.7	Clay	1.00
6	S6	27	-8.5	Sand	1.00
7	S7	25	-8.6	Loam	0.76
8	S8	29	-8.5	Clay	0.95
9	S9	29	-9.4	Sand	1.00
10	S10	20	-8.1	Clay	0.25
11	S11	23	-7.1	Loam	0.76
12	S12	27	-7.6	Loam	0.76
13	S13	29	-6.6	Loam	0.76
14	S14	23	-6.6	Sand	1.00
15	S15	25	-6.6	Clay	1.00
16	S16	21	-4.3	Sand	0.75
17	S17	29	-4.3	Loam	0.76
18	S18	29	-4.3	Loam	0.76
19	S19	20	-2.9	Loam	0.01

Check: These data are **stacked** (/in **long-format**), i.e. ONE obs. per row only (see here). If data are unstacked (/in wide-format), stack up using reshape2 and/or perhaps tools from the Tidyverse.

NERC SCIENCE OF THE ENVIRONMENT CEH

1. Duiker in Africa

I'll be giving you a lot of R code to copy&paste in these slides. Please make sure you can do this easily.

• If you find that you tr
y to copy & paste t
he R commands bu
t they come out al
l broken over the l
ines incorrectly, ...

...then it might be because your text editor is expecting Windows text data (e.g. if you're using Windows Notepad or the script editor provided by the R GUI). Suggest to copy into MS Word first then cut & paste immediately into the R Console window from there.

Data: dataf

	SiteCode	MAT	SiteLat	SoilTexture	SurvProb
1	S1	25	-12.3	Sand	1.00
2	S2	21	-12.3	Loam	0.21
3	S3	25	-12.3	Loam	0.76
4	S4	21	-11.8	Sand	0.75
5	S5	29	-11.7	Clay	1.00
6	S6	27	-8.5	Sand	1.00
7	S7	25	-8.6	Loam	0.76
8	S8	29	-8.5	Clay	0.95
9	S9	29	-9.4	Sand	1.00
10	S10	20	-8.1	Clay	0.25
11	S11	23	-7.1	Loam	0.76
12	S12	27	-7.6	Loam	0.76
13	S13	29	-6.6	Loam	0.76
14	S14	23	-6.6	Sand	1.00
15	S15	25	-6.6	Clay	1.00
16	S16	21	-4.3	Sand	0.75
17	S17	29	-4.3	Loam	0.76
18	S18	29	-4.3	Loam	0.76
19	S19	20	-2.9	Loam	0.01

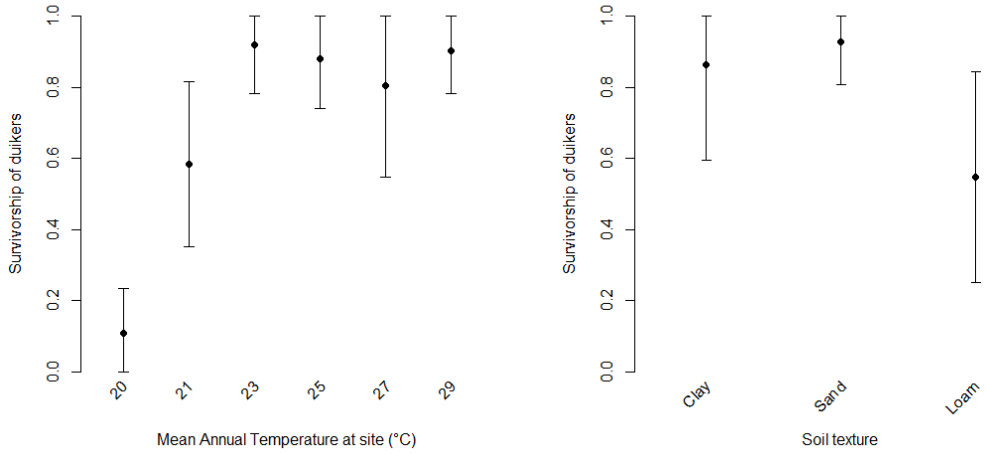
• If you get an error like "Error: unexpected input in ..." then check whether you have used 'smart quotes' anywhere (copying these into R will cause an error like that).

• Also, n.b. users of RStudio may have issues with the dev.new() commands below (suggest just to remove them: that command just opens a new plot window and isn't necessary in RStudio).

NERC SCIENCE OF THE ENVIRONMENT CEH

1. Duiker in Africa

You suspect that the survival of duiker at these sites is controlled predominantly by soil texture (which to a certain extent controls what vegetation is growing there) and mean annual temperature, but latitude is irrelevant. How to work out whether or not this idea is supported by the data? Start with a few diagnostic plots:



Check: Are you comfortable with these terms?

The **response variable** is on the y-axis (*SurvProb*, Survivorship of duikers) (aka. **dependent variable**).

The **predictor variables** go on the x-axes (aka. **explanatory variables**). We have three here:

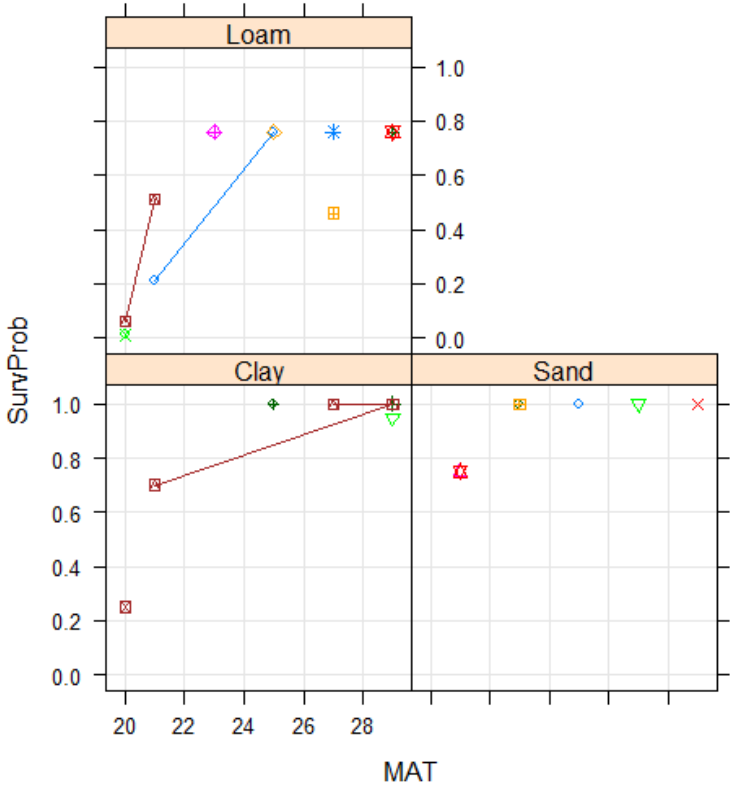
(i) Two are **continuous/numerical** variables (*MAT*, Mean Annual Temperature, and *SiteLat*, Latitude) (aka. **covariates**)

(ii) The other is a **categorical** variable (*SoilTexture*) (aka. a **factor**).

From these plots, it looks like the abundance of Common duiker increases with temperature and also seems to be slightly higher on clay and sandy soils (n.b. as I mentioned, this is a made-up example: duiker distributions in Africa do not follow these patterns, but they are not ecologically unreasonable patterns).

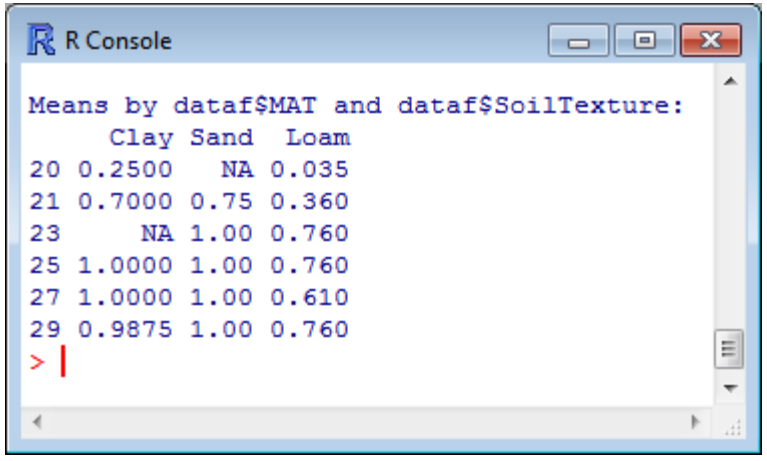
1. Duiker in Africa

Sometimes it's also useful to look at a **panel plot** like this one (which shows more or less the same message as the previous plots):



... or perhaps you prefer to just see the raw numbers, e.g. the tables of mean values generated by the following tapply() commands:

```
cat("Means by dataf$MAT and dataf$SoilTexture:\n");tapply(SurvProb,INDEX=list(dataf$MAT,dataf$SoilTexture),FUN=mean,na.rm=TRUE)
cat("Means by dataf$MAT:\n");tapply(SurvProb,INDEX=dataf$MAT,FUN=mean,na.rm=TRUE)
cat("Means by dataf$SoilTexture:\n");tapply(SurvProb,INDEX=dataf$SoilTexture,FUN=mean,na.rm=TRUE)
```



NERC SCIENCE OF THE ENVIRONMENT CEH

1. Duiker in Africa

Here are the R commands I used to make those plots (given here so that you can copy-and-paste them into R). No need to understand them right now (go through them later at your own speed): for the moment just check you can use the code to produce the diagnostic plots above for yourself.

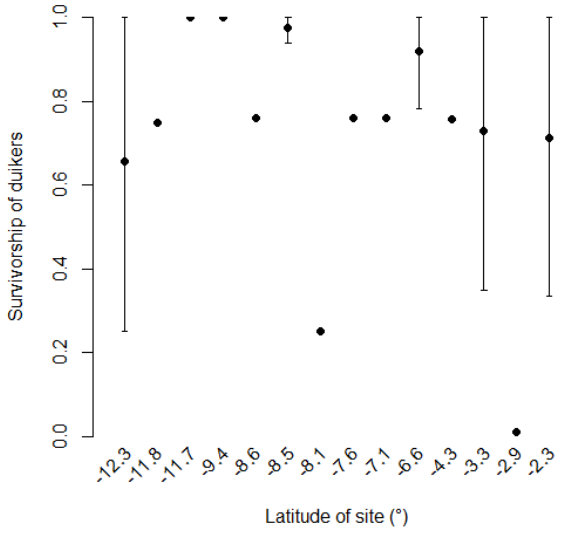
```
#Plot of SurvProb against MAT
barwidth=0.09;mr=tapply(SurvProb,INDEX=MAT,FUN=mean,na.rm=TRUE);sr=tapply(SurvProb,INDEX=MAT,FUN=sd,na.rm=TRUE)
responsevarisprobability=TRUE     #Set this to FALSE if your response is NOT a probability
dev.new();xmidpts=barplot(mr,beside=TRUE,axes=FALSE,axisnames=FALSE,ylim=c(0,1),col="white",border=NA,xlab="Mean Annual Temperature at site (°C)",ylab="Survivorship of duikers")
axis(2);text(x=xmidpts,y=par("usr")[3]-0.03,labels=names(mr),srt=45,adj=1,xpd=TRUE)
if (responsevarisprobability) {
 uppers=pmin(rep(1,times=length(mr)),mr+sr);lowers=pmax(rep(0,times=length(mr)),mr-sr)
} else {
 uppers=mr+sr;lowers=mr-sr
}
segments(xmidpts,lowers,xmidpts,uppers)
segments(xmidpts-barwidth,uppers,xmidpts+barwidth,uppers);segments(xmidpts-barwidth,lowers,xmidpts+barwidth,lowers)
points(x=xmidpts,y=mr,pch=16)


#Plot of SurvProb against SoilTexture
barwidth=0.04;mr=tapply(SurvProb,INDEX=SoilTexture,FUN=mean,na.rm=TRUE);sr=tapply(SurvProb,INDEX=SoilTexture,FUN=sd,na.rm=TRUE)
dev.new();xmidpts=barplot(mr,beside=TRUE,axes=FALSE,axisnames=FALSE,ylim=c(0,1),col="white",border=NA,xlab="Soil texture",ylab="Survivorship of duikers")
axis(2);text(x=xmidpts,y=par("usr")[3]-0.03,labels=names(mr),srt=45,adj=1,xpd=TRUE)
uppers=pmin(rep(1,times=length(mr)),mr+sr);lowers=pmax(rep(0,times=length(mr)),mr-sr)     #Remove the pmin and pmax if you are not dealing with probabilities restricted to [0,1]
segments(xmidpts,lowers,xmidpts,uppers)
segments(xmidpts-barwidth,uppers,xmidpts+barwidth,uppers);segments(xmidpts-barwidth,lowers,xmidpts+barwidth,lowers)
points(x=xmidpts,y=mr,pch=16)
```

NERC SCIENCE OF THE ENVIRONMENT CEH

1. Duiker in Africa

Check also that latitude is effectively irrelevant:

```
#Plot of SurvProb against SiteLat
barwidth=0.09;mr=tapply(SurvProb,INDEX=SiteLat,FUN=mean,na.rm=TRUE);sr=tapply(SurvProb,INDEX=SiteLat,FUN=sd,na.rm
=TRUE)
responsevarisprobability=TRUE     #Set this to FALSE if your response is NOT a probability
dev.new();xmidpts=barplot(mr,beside=TRUE,axes=FALSE,axisnames=FALSE,ylim=c(0,1),col="white",border=NA,xlab="Latitude
of site (°)",ylab="Survivorship of duikers")
axis(2);text(x=xmidpts,y=par("usr")[3]-0.03,labels=names(mr),srt=45,adj=1,xpd=TRUE)
if (responsevarisprobability) {
 uppers=pmin(rep(1,times=length(mr)),mr+sr);lowers=pmax(rep(0,times=length(mr)),mr-sr)
} else {
 uppers=mr+sr;lowers=mr-sr
}
segments(xmidpts,lowers,xmidpts,uppers)
segments(xmidpts-barwidth,uppers,xmidpts+barwidth,uppers);segments(xmidpts-barwidth,lowers,xmidpts+barwidth,lowers)
points(x=xmidpts,y=mr,pch=16)
```



... and for the panel plot:

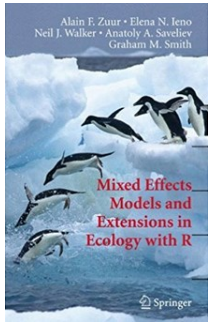
```
#The panel plot
library(lattice);library(nlme)
panlgrps=dataf$SoilTexture     #In the first plot, the panels will show the different levels of panlgrps and colours/symbols show the different levels of colgrps
colgrps=factor(dataf$SiteLat)     #If you have no need of different colours, put colgrps=rep(1,times=length(SurvProb))
numpanelsacross=2     #Try changing this to integers 1,2,3,... to make the plots look prettier
numpanelsup=ceiling(length(levels(factor(panlgrps)))/numpanelsacross)
dev.new();print(xyplot(SurvProb~MAT|panlgrps,groups=colgrps,data=dataf,type=c('g','p','l'),pch=1:length(levels(factor(colgrps))),layout=c(numpanelsacross,numpanelsup),aspect=1,key=list(poi
nts=list(pch=1:length(levels(factor(colgrps))),col=trellis.par.get("superpose.symbol")$col[1:length(levels(factor(colgrps)))]),text=list(levels(factor(colgrps))),space="right",rows=length(levels(
factor(colgrps))))))     #Add the option index.cond=function(x,y)max(y) to sort the panels by maximum y value
```

All following so far ..... ? Great.

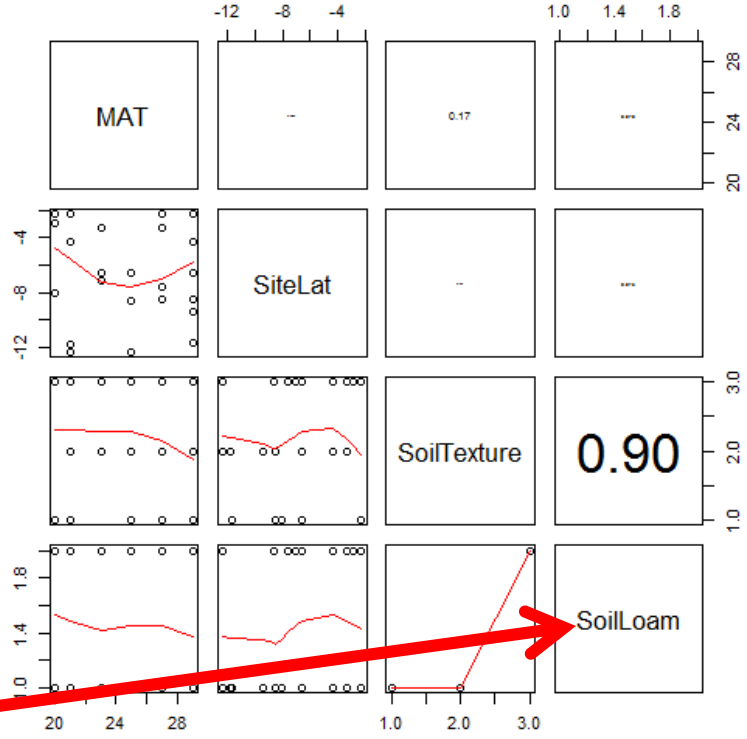
NERC SCIENCE OF THE ENVIRONMENT CEH

1. Duiker in Africa

Now we're on to the actual analysis. **STEP 1** is always to check your predictors for pairwise correlation.




```
#Function based on panel.cor in Zuur et al. (2009)
panel.cor=function(x,y,digits=2,prefix="",cex.cor,...) {
 usr=par("usr");on.exit(par(usr));par(usr=c(0,1,0,1))
 r=abs(cor(x,y)) txt=format(c(r,0.123456789),digits=digits)[1]
 txt=paste(prefix,txt,sep="")
 if (missing(cex.cor)) {cex.cor=0.8/strwidth(txt)}
 text(0.5,0.5,txt,cex=cex.cor*r)
}

pairs(~MAT+SiteLat+SoilTexture,data=dataf,lower.panel=panel.smooth,upper.p
anel=panel.cor,na.action=na.omit)
```



Check: Ignore the *SoilLoam* row for the moment: that's just there to show what the output would look like if there had been a highly correlated variable

None of *MAT*, *SiteLat* or *SoilTexture* are pairwise highly correlated, so we can continue past step 1.

NERC SCIENCE OF THE ENVIRONMENT CEH

1. Duiker in Africa

STEP 2 is to decide whether or not to include interaction terms.

I haven't removed overlap with th later interactio slides

On http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture1.htm, Univ. North Carolina gave the following good advice for this step:

When should interactions be included in models?

As a general rule, interactions should always be examined with experimental data, and rarely examined for observational data. **Observational studies** are quasi-experimental designs that fall short of being true experiments for various reasons. In a typical observational study treatments are imposed by nature rather than the experimenter. As a result there is no guarantee that treatments have been randomly assigned to subjects and rarely any balance causing some treatment combinations to be under-represented. All of this makes assessing interaction in observational studies dangerous. Main effects are hard enough to assess in such studies; interactions are truly pushing the envelope.

Based on these considerations I approach the statistical analysis of experiments and observational studies quite differently. In an experiment in which all relevant factors have been assiduously controlled and in which subjects have been randomly assigned to treatments I typically start with the most complicated interaction model possible and try to simplify it. On the other hand when I analyze observational data I start with main effects and maybe tentatively examine a few interactions that have a theoretical basis.

- I'd also add to that this advice from Thomas *et al.* (2017:56): "Never include an interaction term in a model unless you can write a sentence to explain what the interaction represents!"
- In summary, in mensurative experiments you ARE allowed to disregard the interaction terms, but YOU MUST include them when analysing a manipulative experiment.
- This duiker experiment is a mensurative/observational one, so I am going to ignore the interaction terms for now (in a later slide I'll show you what happens if you include them)

NERC SCIENCE OF THE ENVIRONMENT CEH

1. Duiker in Africa

Next: run a GLM (without interaction terms) and you get this in the summary table:

```
fm1glm=glm(SurvProb~MAT+SiteLat+SoilTexture,family=quasibinomial,data=dataf)#for now, don't worry about the 'family=quasibinomial' (and I also
summary(fm1glm)                                                              #won't explain yet why I'm excluding interaction terms here)
```

R Console

```
> fm1glm=glm(SurvProb~MAT+SiteLat+SoilTexture,family=quasibinomial,data=dataf)
> summary(fm1glm)

Call:
glm(formula = SurvProb ~ MAT + SiteLat + SoilTexture, family = quasibinomial,
    data = dataf)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6423  -0.2680   0.1818   0.2849   0.7107

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -7.63036    1.70919  -4.464 0.000194 ***
MAT              0.38039    0.07090   5.365 2.18e-05 ***
SiteLat         -0.03568    0.06269  -0.569 0.574999
SoilTextureSand  1.15088    0.74223   1.551 0.135272
SoilTextureLoam -1.73419    0.54386  -3.189 0.004244 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasibinomial family taken to be 0.1488527)

    Null deviance: 13.8985  on 26  degrees of freedom
Residual deviance:  3.6474  on 22  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5

>
```

Check: Notice the **formula syntax** that R uses here to specify the model fit:

(response variable) ~ (predictor variables separated by "+")

With interaction terms, the "+"s would all have been "*"s.

Check: Are you familiar with standard ***p*-value** levels?

>0.10 means *not significant*
0.05-0.10 means *weakly significant* (R marks this **.**)
0.01-0.05 means *significant* (R marks this ***)**
0.001-0.01 means *highly significant* (R marks this ****)**
<0.001 means *very highly significant* (R marks this *****)**

We're only interested in values <0.05 (the '5% level').

Look at the *p*-values of the predictors, which are listed in the **Pr(>|t|)** column of the Coefficients table. These values tell us that the significant predictors are *Intercept* (the constant term in the *ita* (η) equation below), *MAT* and the *Loam* level of *SoilTexture*.

NERC SCIENCE OF THE ENVIRONMENT CEH

1. Duiker in Africa

However, we have not done an important step here: we have done no MODEL SELECTION, and this is necessary because of the presence of the non-significant *SiteLat* variable.

I'll be talking a lot more about model selection shortly, but for this example all it amounts to is removing *SiteLat*, the predictor variable that the previous GLM has shown to be the least significant (*p*-value greater than 0.05 and higher than all others).

```
R Console
> fm1glm=glm(SurvProb~MAT+SiteLat+SoilTexture,family=quasibinomial,data=dataf)
> summary(fm1glm)

Call:
glm(formula = SurvProb ~ MAT + SiteLat + SoilTexture, family = quasibinomial,
    data = dataf)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6423  -0.2680   0.1818   0.2849   0.7107

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -7.63036    1.70919  -4.464 0.000194 ***
MAT              0.38039    0.07090   5.365 2.18e-05 ***
SiteLat         -0.03568    0.06269  -0.569 0.574999
SoilTextureSand  1.15088    0.74223   1.551 0.135272
SoilTextureLoam -1.73419    0.54386  -3.189 0.004244 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasibinomial family taken to be 0.1488527)

    Null deviance: 13.8985  on 26  degrees of freedom
Residual deviance:  3.6474  on 22  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5

>
```

fm2glm=glm(SurvProb~MAT+SoilTexture,family=quasibinomial,data=dataf)
summary(fm2glm)

```
R Console
> fm2glm=glm(SurvProb~MAT+SoilTexture,family=quasibinomial,data=dataf)
> summary(fm2glm)

Call:
glm(formula = SurvProb ~ MAT + SoilTexture, family = quasibinomial,
    data = dataf)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6405  -0.2869   0.1406   0.3430   0.7185

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -7.45780    1.63885  -4.551 0.000143 ***
MAT              0.38114    0.06935   5.496 1.37e-05 ***
SoilTextureSand  1.22183    0.71912   1.699 0.102795
SoilTextureLoam -1.68641    0.52460  -3.215 0.003842 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasibinomial family taken to be 0.1431616)

    Null deviance: 13.8985  on 26  degrees of freedom
Residual deviance:  3.6959  on 23  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5

>
```

Note that removing *SiteLat* has also affected the significance levels of the other predictors (I think of this as analogous to voting by single transferable vote)

NERC SCIENCE OF THE ENVIRONMENT CEH

1. Duiker in Africa

What about the non-significant *Sand* level of *SoilTexture*? Some would just leave that in, but I'm in favour of removing that too by 'recoding' the predictor (=relabeling its levels):

```
dataf$SoilLoam=factor(ifelse(dataf$SoilTexture=="Loam","Loam","Other"),levels=c("Other","Loam"))
```
#I've 'recoded' SoilTexture here as a new variable SoilLoam which is the same as SoilTexture except that it's just "Loam" and "Other" rather than the three levels of SoilTexture (i.e. I'm combining Sand and Clay sites)

```
View(dataf)
fm3glm=glm(SurvProb~MAT+SoilLoam,family=quasibinomial,data=dataf)
summary(fm3glm)
```

Data: dataf

	SiteCode	MAT	SiteLat	SoilTexture	SurvProb	SoilLoam
1	S1	25	-12.3	Sand	1.00	Other
2	S2	21	-12.3	Loam	0.21	Loam
3	S3	25	-12.3	Loam	0.76	Loam
4	S4	21	-11.8	Sand	0.75	Other
5	S5	29	-11.7	Clay	1.00	Other
6	S6	27	-8.5	Sand	1.00	Other
7	S7	25	-8.6	Loam	0.76	Loam
8	S8	29	-8.5	Clay	0.95	Other
9	S9	29	-9.4	Sand	1.00	Other
10	S10	20	-8.1	Clay	0.25	Other
11	S11	23	-7.1	Loam	0.76	Loam
12	S12	27	-7.6	Loam	0.76	Loam
13	S13	29	-6.6	Loam	0.76	Loam
14	S14	23	-6.6	Sand	1.00	Other
15	S15	25	-6.6	Clay	1.00	Other
16	S16	21	-4.3	Sand	0.75	Other
17	S17	29	-4.3	Loam	0.76	Loam
18	S18	29	-4.3	Loam	0.76	Loam
19	S19	20	-2.9	Loam	0.01	Loam

R Console

```
> fm3glm=glm(SurvProb~MAT+SoilLoam,family=quasibinomial,data=dataf)
> summary(fm3glm)

Call:
glm(formula = SurvProb ~ MAT + SoilLoam, family = quasibinomial,
    data = dataf)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8759  -0.2604   0.1763   0.3484   0.7173

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -6.84525    1.69943  -4.028 0.000491 ***
MAT           0.37936    0.07437   5.101 3.22e-05 ***
SoilLoamLoam -2.25548    0.46862  -4.813 6.68e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasibinomial family taken to be 0.1559363)

    Null deviance: 13.898  on 26  degrees of freedom
Residual deviance:  4.140  on 24  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5

>
```

All predictors in the Coefficients table are now significant so let's call this the 'all-predictors-significant' model.

NERC SCIENCE OF THE ENVIRONMENT CEH

1. Duiker in Africa

There are still a few possible models that have all their predictors significant (e.g. *MAT* or *SoilLoam* alone without the other) and this is where the **Akaike Information Criterion (AIC)** comes in. We choose the fit with the *lowest* AIC score (n.b. in linear regression the *highest* r^2 score is best, but AIC works the other way around) and we use stepAIC() to find it:

```
library(MASS)
fm4glm=glm(SurvProb~MAT+SoilLoam,family=binomial,data=dataf)        #In this example I need to change the family to "binomial" so that AIC values can be calculated (if a 'quasi-' distribution were not in use, the family would not change). Ignore the warning about "non-integer #successes".
fm4glmBestFit=stepAIC(fm4glm)
summary(fm4glmBestFit)
```

```
> fm4glmBestFit=stepAIC(fm4glm)
Start:  AIC=23.37
SurvProb ~ MAT + SoilLoam

           Df Deviance    AIC
<none>          4.1400 23.370
- SoilLoam  1   8.5923 25.823
- MAT       1   9.6442 26.874
```

The AIC of the all-predictors-significant model (confusingly, this is the row with '<none>' because no more of the predictors have been excluded) is slightly better than the AIC of simpler models based on a subset of predictors so our best-fit model is exactly the same as *fm3glm* from the last slide (n.b. not *fm4glm* because that had binomial errors). The warnings can all be ignored and that's the end of the GLM (!).

Check: Here are the steps we've followed:

1. Run a GLM with all possible predictors included. This model is the **full model**.

2. Identify predictors and predictor-levels that are non-significant, remove them *one by one* and re-run until all predictors have significant *p*-values. This step is the first stage of **model selection** and brings us to the **all-predictors-significant model**.

3. Use stepAIC() to check the AICs of simpler models based on the same set of significant predictors, choosing the lowest available AIC. This is the second stage of **model selection** and brings us to the **best-fit model**.

1. Duiker in Africa: prediction

The 'Estimate' column of the Coefficients table at the end tells us what the best-fit model is and we can generate the values of *ita* (η) from it (note that for categorical predictors like *SoilLoam* the value only applies for one category, which means I need an *ifelse* there):

```
ita = -6.84525 + (0.37936*dataf$MAT) +
ifelse(dataf$SoilLoam=="Loam",-2.25548,0.0)

Predicted_SurvProb = exp(ita)/(1+exp(ita))          #This
is 'inverting the link function', which will be explained shortly

#The best way to get the fully accurate coefficient values is:
Const=summary(fm4glmBestFit)$coefficients[1,1]
CoefMAT=summary(fm4glmBestFit)$coefficients[2,1]
CoefSoilLoamLoam=summary(fm4glmBestFit)$coefficients[3,1]
PvalConst=summary(fm4glmBestFit)$coefficients[1,4]
PvalMAT=summary(fm4glmBestFit)$coefficients[2,4]
#etc.
```

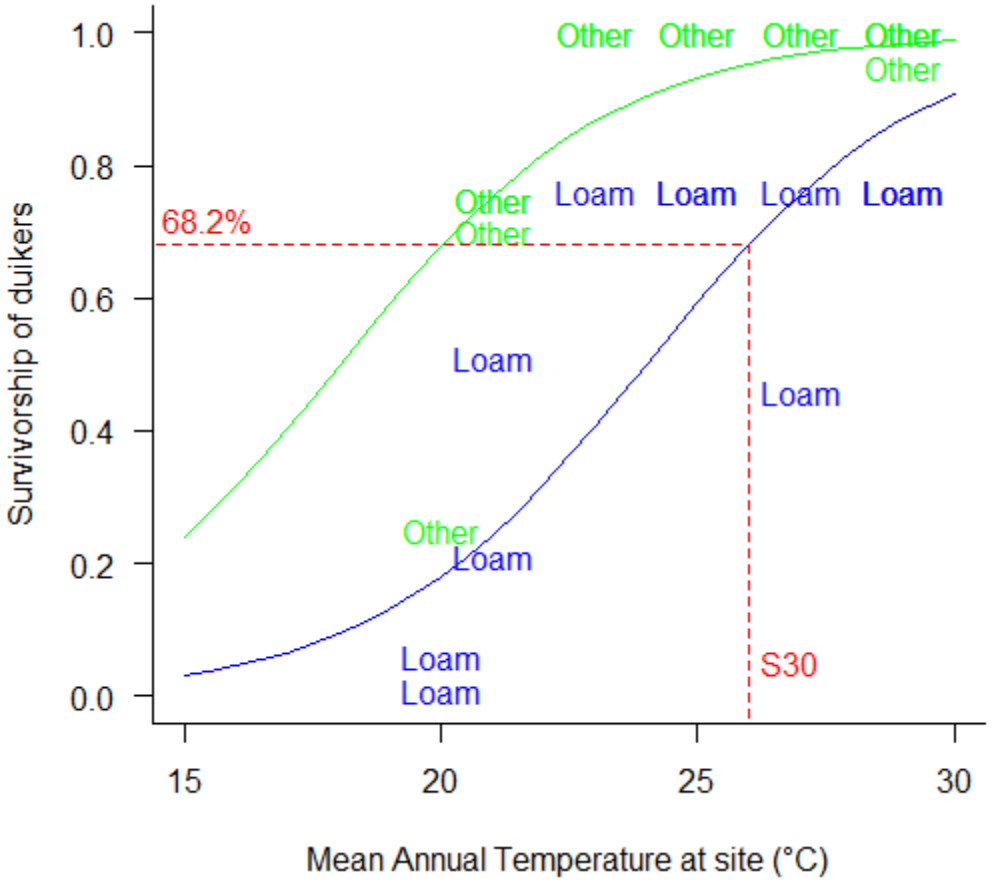
R Console

```
> summary(fm4glmBestFit)

Call:
glm(formula = SurvProb ~ MAT + SoilLoam, family = binomial, data = dataf)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.8759  -0.2604   0.1763   0.3484   0.7173

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.8452     4.3036  -1.591   0.1117
MAT           0.3794     0.1883   2.014   0.0440 *
SoilLoamLoam -2.2555     1.1867  -1.901   0.0574 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 13.898  on 26  degrees of freedom
Residual deviance:  4.140  on 24  degrees of freedom
AIC: 23.37

Number of Fisher Scoring iterations: 5

>
```

NERC SCIENCE OF THE ENVIRONMENT CEH

1. Duiker in Africa: prediction

What does that all actually mean? Well, for example if we were to make a prediction of the survival probability of duiker to expect at a new site S28, based on our dataset, with *MAT*=26°C and *SoilTexture*=Loam we would calculate:



ita = -6.84525 + (0.37936*26) - 2.25548 = 0.76263
Predicted_SurvProb = exp(ita)/(1+exp(ita)) = 0.6819

Therefore, based on our analysis, we would expect around 68.2% survival probability for duiker at the new site S28. A quicker way to calculate this is:

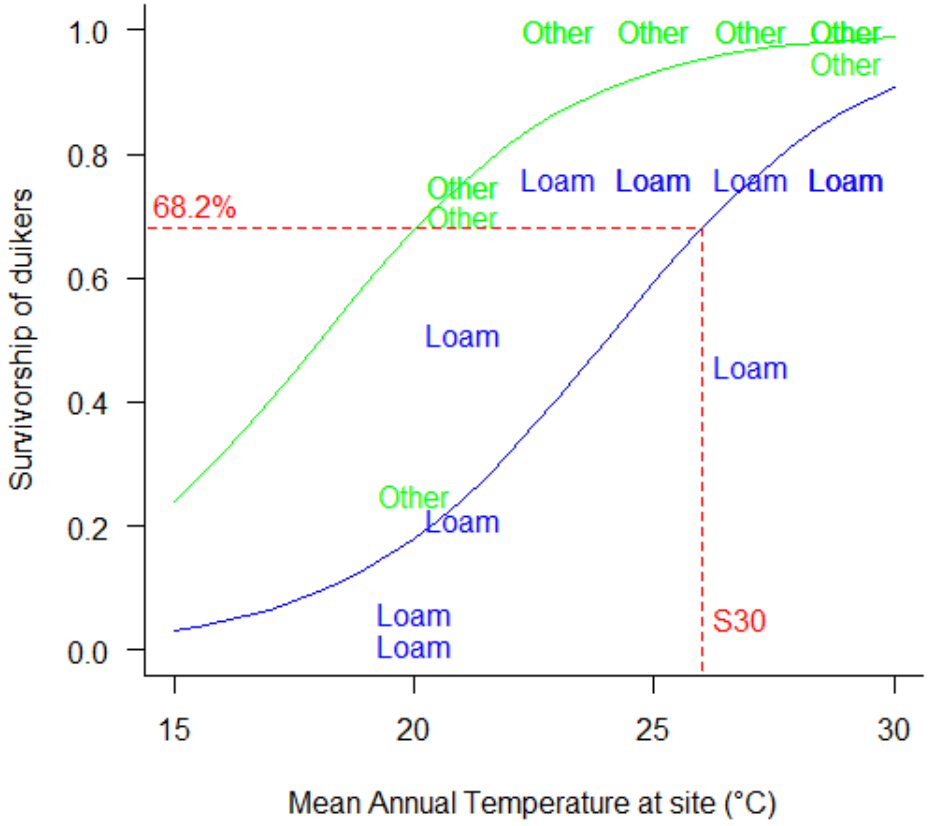
```
Predicted_SurvProb = predict(fm4glmBestFit,data.frame(MAT=26,SoilLoam="Loam"),type="response") = 0.6819
```

(n.b. if you try this, you get a slightly more accurate value from the predict() function because I only went to 5 decimal places of accuracy above).

1. Duiker in Africa: prediction

Code for the last plot (Thomas *et al.* 2017:77 gave a slightly more sophisticated version of this plot):

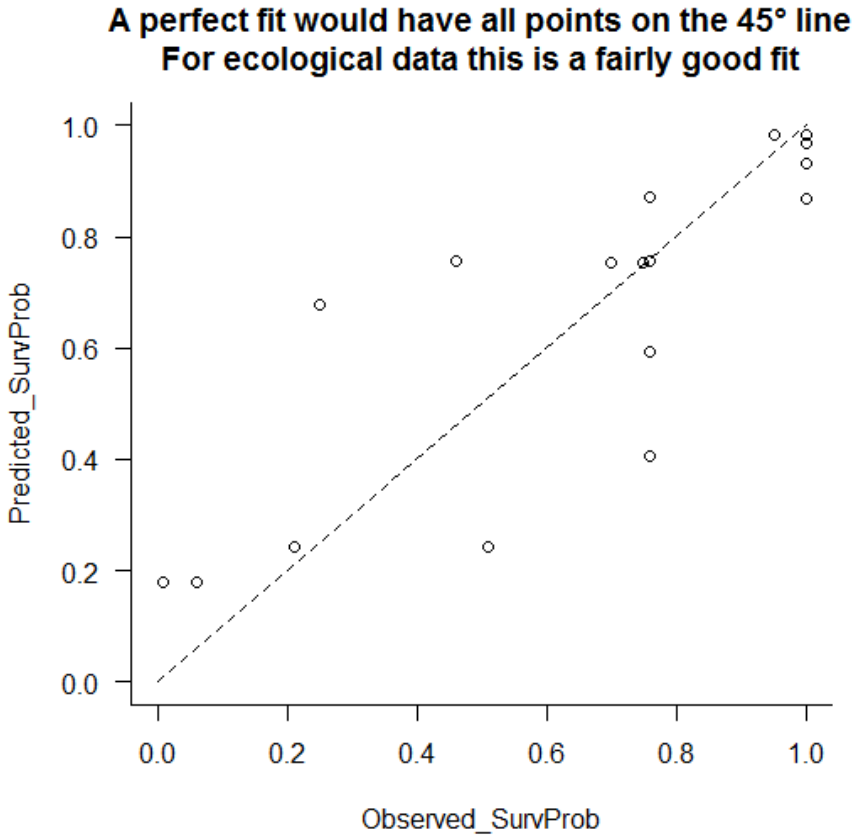
```
xvals=seq(from=15,to=30,by=0.5)
plot(x=c(min(xvals),max(xvals)),y=c(0,1),type="n",xlab="Mean Annual Temperature at site
(°C)",ylab="Survivorship of duikers",bty="l",las=1)

#OBS
datafl=subset(dataf,SoilLoam=="Loam")
text(x=datafl$MAT,y=datafl$SurvProb,labels=as.character(datafl$SoilLoam),col="blue")
datafl=subset(dataf,SoilLoam=="Other")
text(x=datafl$MAT,y=datafl$SurvProb,labels=as.character(datafl$SoilLoam),col="green")

#BEST-FIT
lines(x=xvals,y=predict(fm4glmBestFit,data.frame(MAT=xvals,SoilLoam=factor(rep("Loam
",length(xvals)),levels=levels(dataf$SoilLoam))),type="response"),col="blue")
lines(x=xvals,y=predict(fm4glmBestFit,data.frame(MAT=xvals,SoilLoam=factor(rep("Other
",length(xvals)),levels=levels(dataf$SoilLoam))),type="response"),col="green")

#S28
Predicted_SurvProb2=predict(fm4glmBestFit,data.frame(MAT=26,SoilLoam="Loam"),type
="response")
lines(x=c(min(xvals)-
5,26),y=c(Predicted_SurvProb2,Predicted_SurvProb2),col="red",lty=2)
lines(x=c(26,26),y=c(Predicted_SurvProb2,-0.5),col="red",lty=2)
text(x=26.8,y=0.05,labels="S28",col="red")
text(x=15.5,y=0.72,labels="68.2%",col="red")
```



1. Duiker in Africa: validation

Plot observed against predicted to see how well this fitted model does across all sites (n.b. this is validation not verification):

```
ita=-6.84525+(0.37936*dataf$MAT)+
ifelse(dataf$SoilTexture=="Loam",-2.25548,0.0)
Predicted_SurvProb=exp(ita)/(1+exp(ita))

plot(x=SurvProb,y=Predicted_SurvProb,xlim=c(0,1),yli
m=c(0,1),las=1,bty="l",xlab="Observed_SurvProb",ylab
="Predicted_SurvProb",main="A perfect fit would have
all points on the 45° line\nFor ecological data this is a
fairly good fit")
lines(x=c(0,1),y=c(0,1),lty=2)
```



Not done here, but at this stage it's more usual to use a *quantitative measure* of goodness-of-fit.

1. Duiker in Africa: summary

That's the end of this GLM analysis. A quick recap of the steps involved:

You do some quick plots to get a feel for your data (n.b. this is "quick" in the sense of "often takes ages" (!): it can sometimes take a long time because this step usually entails cleaning up your data (inc. gap-filling) and actually deciding what you're going to analyse).

Then you carry out a GLM analysis, following broadly the steps I outlined before (box right).

Then you need to use the best-fit GLM to make some predictions, including validating ('sanity-checking') the model you've arrived at.

Check:

1. Run a GLM with all possible predictors included. This model is the **full model**.

2. Identify predictors and predictor-levels that are non-significant, remove them *one by one* and re-run until all predictors have significant *p*-values. This step is the first stage of **model selection** and brings us to the **all-predictors-significant model** .

3. Use `stepAIC()` to check the AICs of simpler models based on the same set of significant predictors, choosing the lowest available AIC. This is the second stage of **model selection** and brings us to the **best-fit model**.

Then, of course, you need to write it up. I recommend Quesada *et al.* (2012, *Biogeosciences* paper about Amazon forest structure) as an example of how to present a GLM properly in a scientific paper.

I've glossed over details in this example, but that's the idea of a GLM analysis.

NERC SCIENCE OF THE ENVIRONMENT CEH

Generalized Linear Models (GLMs)

1. Duiker in Africa

2. Recap of standard statistics

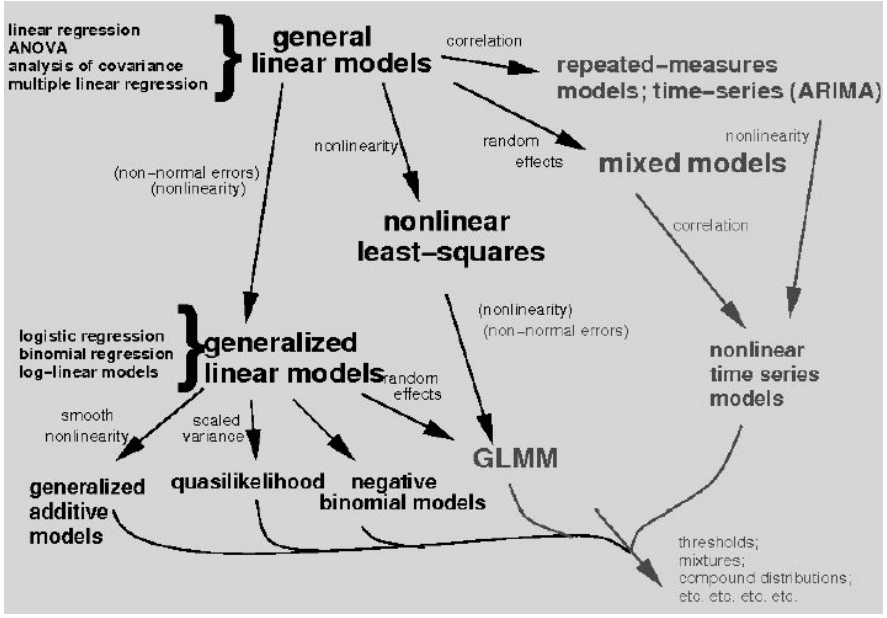
3. Validation

4. Interaction terms, fixed & random effects and Crawley's Two Stage Backwards Deletion Method

5. Error structure, Goodness of fit, overdispersion and using glmmPQL, dredge and glmer

2. Recap of standard statistics

GLMs (and GLMMs, GAMs, etc.) are part of a large array of analysis options. Which you select depends on your data and what you are trying to show from them:



Generalized Linear Models (GLMs or GLMMs)

General Linear Models (never shortened to "GLM") where "Errors" are normally distributed

ANOVA	Multiple regression
t-tests	Regression & Correlation

"Errors" are not normally distributed

e.g.

- Binomial (binomial regression)
- Negative binomial
- Poisson
- etc.
- "Mixed models" (GLMM) include a mix of "**random terms**" as well as "fixed" terms.

Left from Bolker (see http://slideplayer.com/slide/9255836/) but I prefer Thomas *et al.* (2017:51)'s diagram right. These illustrate how these forms of analysis are all related, e.g. 2-sample *t*-tests are a special case of ANOVA, while ANOVA is a special case of General Linear Model (Gelman 2005 does give exceptions, but my opinion is that you can almost always ignore them). General Linear Models are themselves a special case of Generalized Linear Models (GLMs).

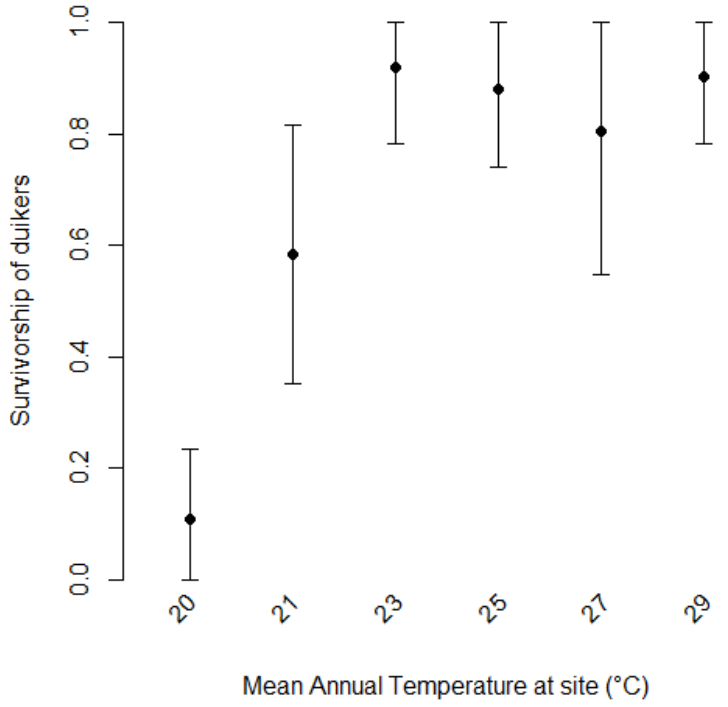
We're making our way into this array, and a good way of showing what the options mean is to carry out a few 'standard stats' analyses on the duiker data to show what you get (I'm not giving you extra work here: this is just to show that you get similar or equivalent results to the GLM we've just done, which I find reassuring).

2. Recap of standard statistics: Regression

Approximately, if I ignore the dependence on soil texture, this example is just about a relationship between survival probability and temperature.

It seems that a quick regression fit would be appropriate here, with a transformation to remove the obvious curve in the relationship.

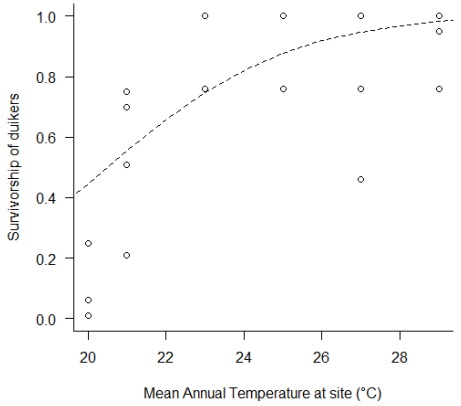
Let's try that ...



2. Recap of standard statistics: Regression

The following does a transformed linear fit like this, using R's lm() command.

```
epsilon=min(ifelse(SurvProb>0,SurvProb,1),na.rm=TRUE)    #The smallest nonzero survivorship
transformedSurvProb=log((SurvProb+epsilon)/(1-SurvProb+epsilon))    #This is an empirical logit transformation (with epsilon=0 it's a straight logit transformation). I'm following the method described under "Choice of e" on http://www.esapubs.org/archive/ecol/E092/001/appendix-B.htm
fm1lm=lm(transformedSurvProb~MAT)
plot(x=MAT,y=SurvProb,ylim=c(0,1),las=1,bty="l",xlab="Mean Annual Temperature at site (°C)",ylab="Survivorship of duikers")
xvals=19:30
fitteditavals=fm1lm$coefficients[2]*xvals+fm1lm$coefficients[1]
fittedSurvProbvals=((1+epsilon)*exp(fitteditavals)-epsilon)/(1+exp(fitteditavals))    #n.b. this reduces to Ŷ=exp(η)/(1+exp(η)) if epsilon=0
lines(x=xvals,y=fittedSurvProbvals,lty=2)    #n.b. can't just do abline(fm1lm) here because of the data transformation
summary(fm1lm)
```



R Console

```
> summary(fm1lm)

Call:
lm(formula = transformedSurvProb ~ MAT)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7005 -1.5575  0.0719  1.0331  3.5621

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -8.6416     2.9522  -2.927  0.00719 **
MAT           0.4215     0.1172   3.597  0.00138 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.066 on 25 degrees of freedom
Multiple R-squared:  0.341,     Adjusted R-squared:  0.3147
F-statistic: 12.94 on 1 and 25 DF,  p-value: 0.001384

>
```

From the summary table, the coefficient of *MAT* turns out to be 0.42150 (a bit higher than the 0.37936 we got from the glm() above, but quite close).

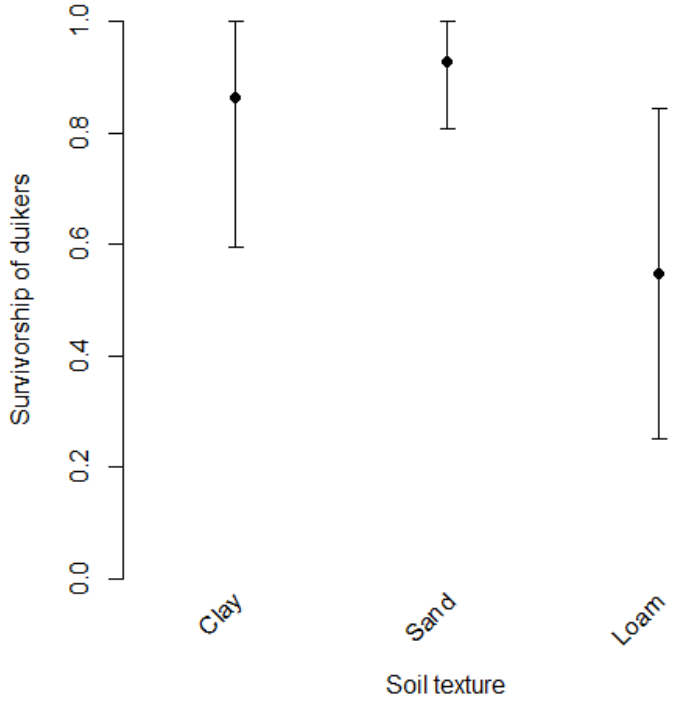
NERC SCIENCE OF THE ENVIRONMENT CEH

2. Recap of standard statistics: ANOVA

Approximately, if I ignore the dependence on temperature, this example is just about a relationship between survival probability and soil texture.

It seems that a quick ANOVA (ANalysis Of VAriance) would be appropriate here.

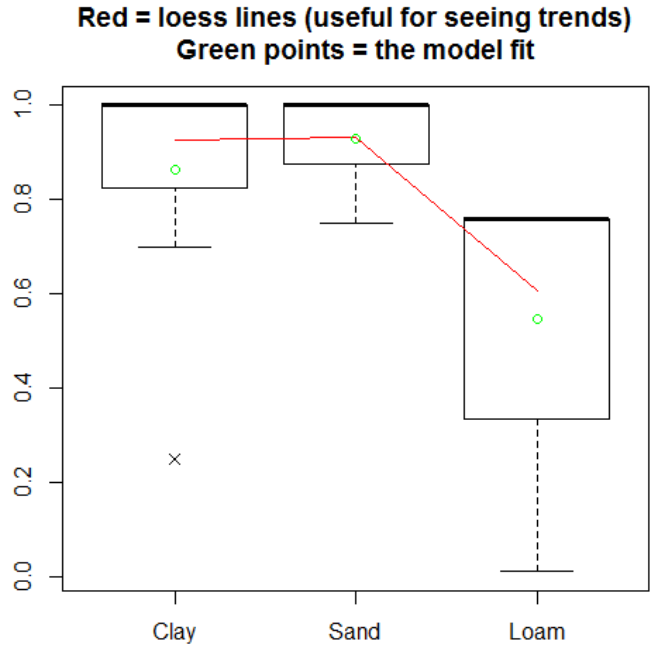
Let's try that ...



2. Recap of standard statistics: ANOVA

The following does an ANOVA analysis, using R's aov() command.

```
tmp=replications(SurvProb~SoilTexture,data=dataf);print(tmp)     #Checking the replications
if (length(unique(unlist(tmp)))==1) {cat("A balanced experimental design.\n")} else {cat("An unbalanced
experimental design (this should not be a problem for analysis, though).\n")}
library(car);testres=leveneTest(SurvProb~SoilTexture,data=dataf)     #Testing for homoscedasticity: n.b.
with another predictor XX, this formula would become SurvProb~SoilTexture*XX.
if (testres$Pr[1]>0.05) {cat("No evidence of any significant difference in variance across samples (i.e. data
may be assumed homoscedastic; variances may be assumed homogeneous)\n")} else {cat("Variances
across samples do show differences (i.e. the data may NOT be assumed homoscedastic)\n")}
fm1aov=aov(SurvProb~SoilTexture,data=dataf)
summary(fm1aov)     #n.b. fm2lm=lm(SurvProb~SoilTexture,data=dataf);anova(fm2lm) gives exactly the
same ANOVA table.
model.tables(fm1aov,"means")     #Tables of means
plot(TukeyHSD(fm1aov))     # Tukey Honest Significant Differences (see Crawley 2007:484)
dev.new();plot(x=SoilTexture,y=SurvProb,bty="l",pch=4,main="Red = loess lines (useful for seeing
trends)\nGreen points = the model
fit");lines(lowess(x=SoilTexture,y=SurvProb),col="red");points(x=SoilTexture,y=fm1aov$fit,col="green")
```




```
R Console
> summary(fm1aov)
             Df Sum Sq Mean Sq F value  Pr(>F)
SoilTexture   2 0.8136  0.4068   6.264 0.00647 **
Residuals    24 1.5587  0.0649
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>
```

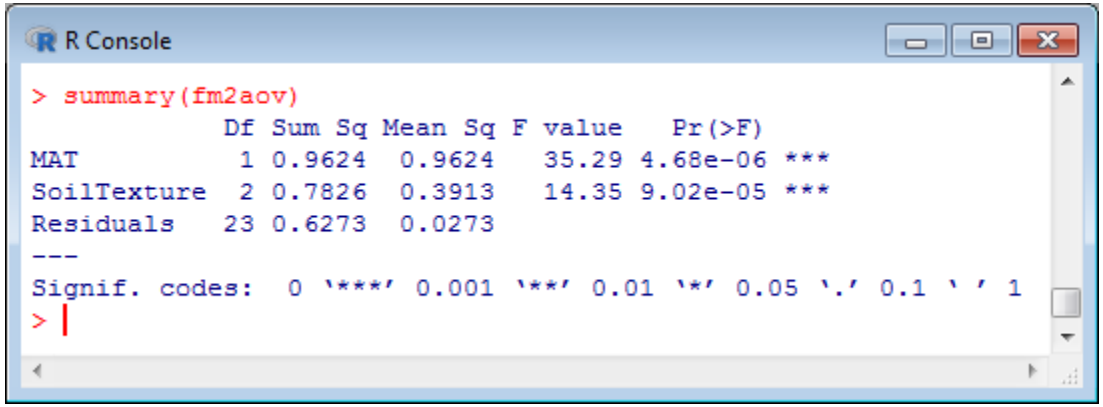
From the ANOVA table, the coefficient of *SoilTexture* is significant, indicating that changing soil texture does affect the value of *SurvProb*, which is an equivalent result to what we learnt from the GLM above.

NERC SCIENCE OF THE ENVIRONMENT CEH

2. Recap of standard statistics: ANCOVA

Finally, I can also do an ANCOVA using the aov() command:

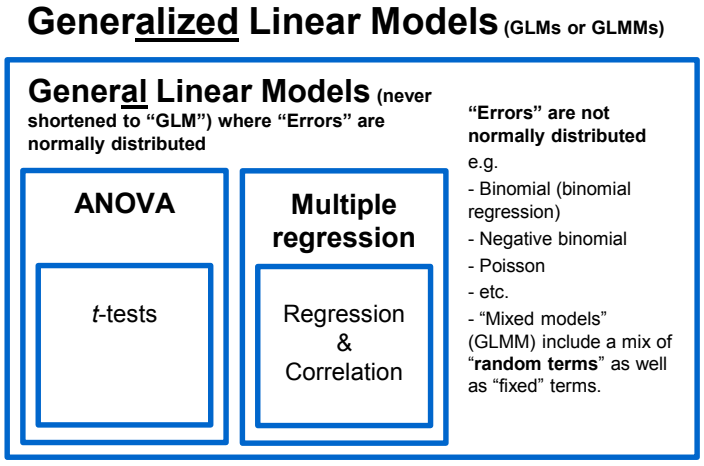
```
fm2aov=aov(SurvProb~MAT+SoilTexture,data=dataf)
summary(fm2aov)     #n.b. fm3lm=lm(SurvProb~MAT+SoilTexture,data=dataf);anova(fm3lm)
gives exactly the same ANCOVA table.
```



From the ANCOVA table, the coefficients of both *MAT* and *SoilTexture* are significant, which shows they both have a significant effect on duiker survival (but an ANCOVA doesn't quantify the *size* of the effect in the way a GLM does).

2. Recap of standard statistics

Essentially, a GLM analysis is a generalisation of doing individual regressions and ANOVAs like these. It's not just a more complicated way of doing things: it's a *more complete* way of doing things and it has the very significant advantage of allowing us to consider all predictors at the same time, whether continuous or categorical.



P.S. In 2003 John Nelder (one of the 'inventors' of GLM analysis in ~1972) was quoted as saying "I suspect we should have found some more fancy name for it that would have stuck and not been confused with the general linear model, although general and generalized are not quite the same. I can see why it might have been better to have thought of something else.".

VALIDATION

NERC SCIENCE OF THE ENVIRONMENT

CEH

3. Validation

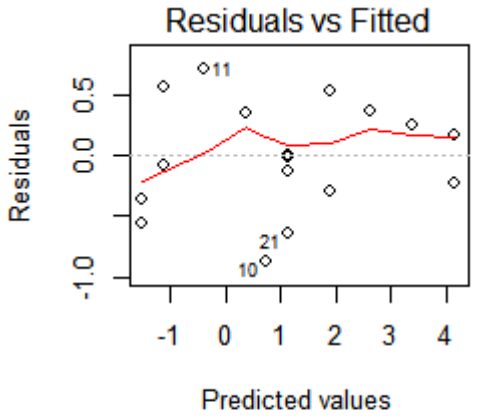
There are **Four Validation Plots** you must look at during a GLM analysis (Crawley 2007:357, Zuur *et al*. 2009:Fig.9.8):

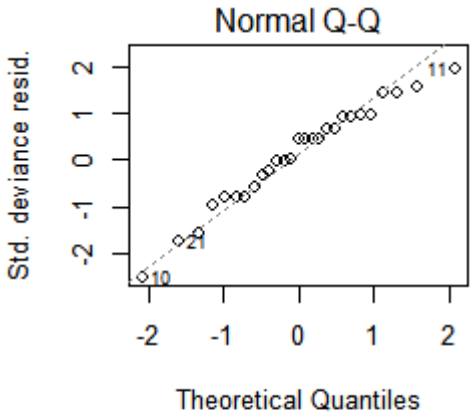
```
dev.new();par(mfrow=c(2,2));plot(fm3glm)
```

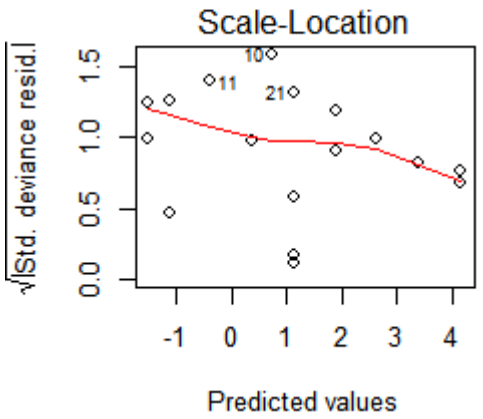
Here, what we're looking for are signs of nonlinearity and/or non-randomness or non-normality of the residuals, which can all indicate infringements of the underlying analysis assumptions (see Thomas *et al*. 2017:60-67 or perhaps Crawley 2007:ch.11 to see what to do if the assumptions are violated).

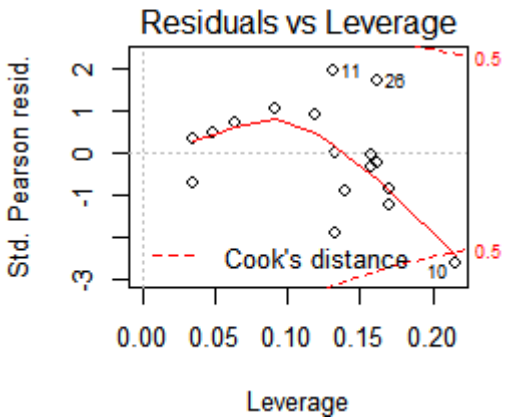
If we find any of those, then we need to look closely at each predictor and consider issues of homoscedasticity, normality, etc. (using standard stats tests).

However, note that for binomial data these plots are only of limited use (see Hector 2015:129).









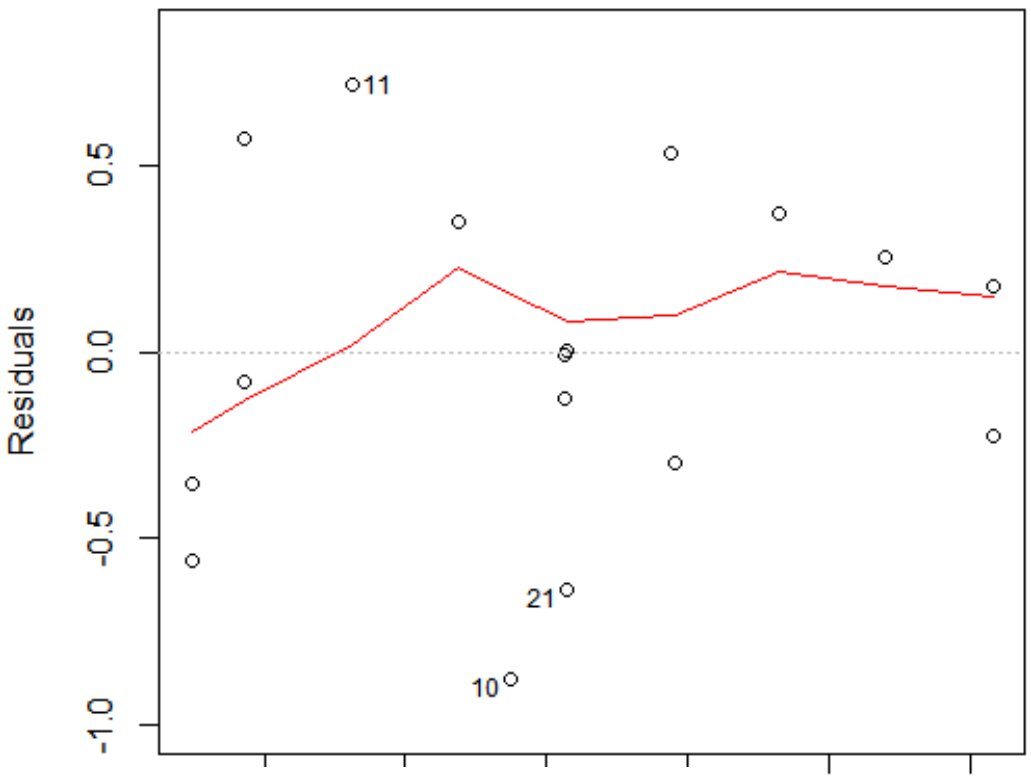
3. Validation

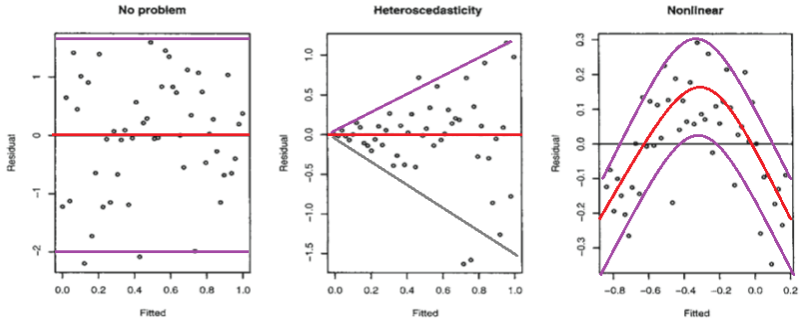
Plot #1: **Residuals *vs*. Fitted**

What is a **residual**? This is the difference between an observed value and the corresponding predicted value from a fitted model.

Pinheiro & Bates (2000:11): “In this plot we are looking for a systematic increase (or, less commonly, a systematic decrease) in the variance of the [error term] ϵ_{ij} as the level of the response increases. If this is present, the residuals on the right-hand side of the plot will have a greater vertical spread than those on the left, forming a horizontal “wedge-shaped” pattern.”.

Also, see the three illustrations below (which I took from here): we want *no* obvious trend in the residuals and also reasonable vertical symmetry in the density of points. NO wedges or nonlinearities are evident in this plot and the density looks OK too.





3. Validation

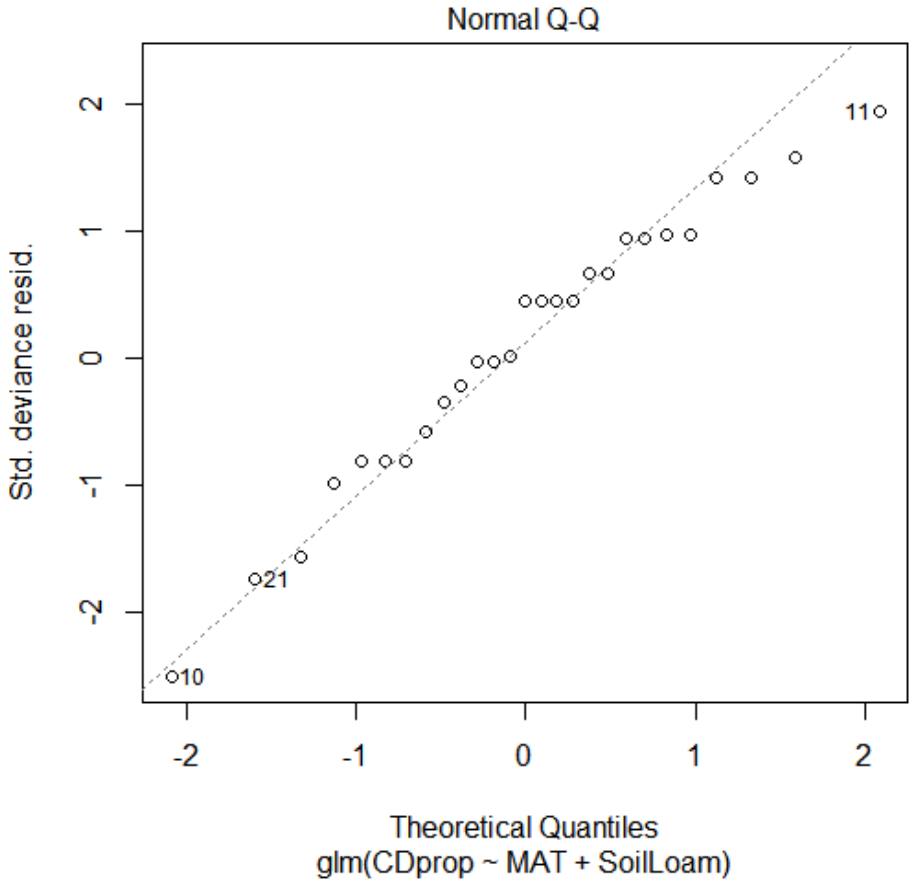
Plot #2: **Normal Q-Q**

If the residuals are normally-distributed, then the points should be reasonably close to a 45° line from bottom-left to top-right (shown). This plot can also be generated using:

```
qqnorm(residuals(fm3glm));qqline(residuals(fm3glm))
```

Here, the residuals are visually near enough to normal (see Crawley 2007:341-4 and/or here for examples of non-normality in Q-Q plots), or you can just use `shapiro.test(residuals(fm3glm))` to confirm that the *p*-value is comfortably >0.05 (I get 0.5987), indicating normality.

(the first "Q" stands for the Quantiles of your sample, the second the corresponding Quantiles of the theoretical population)

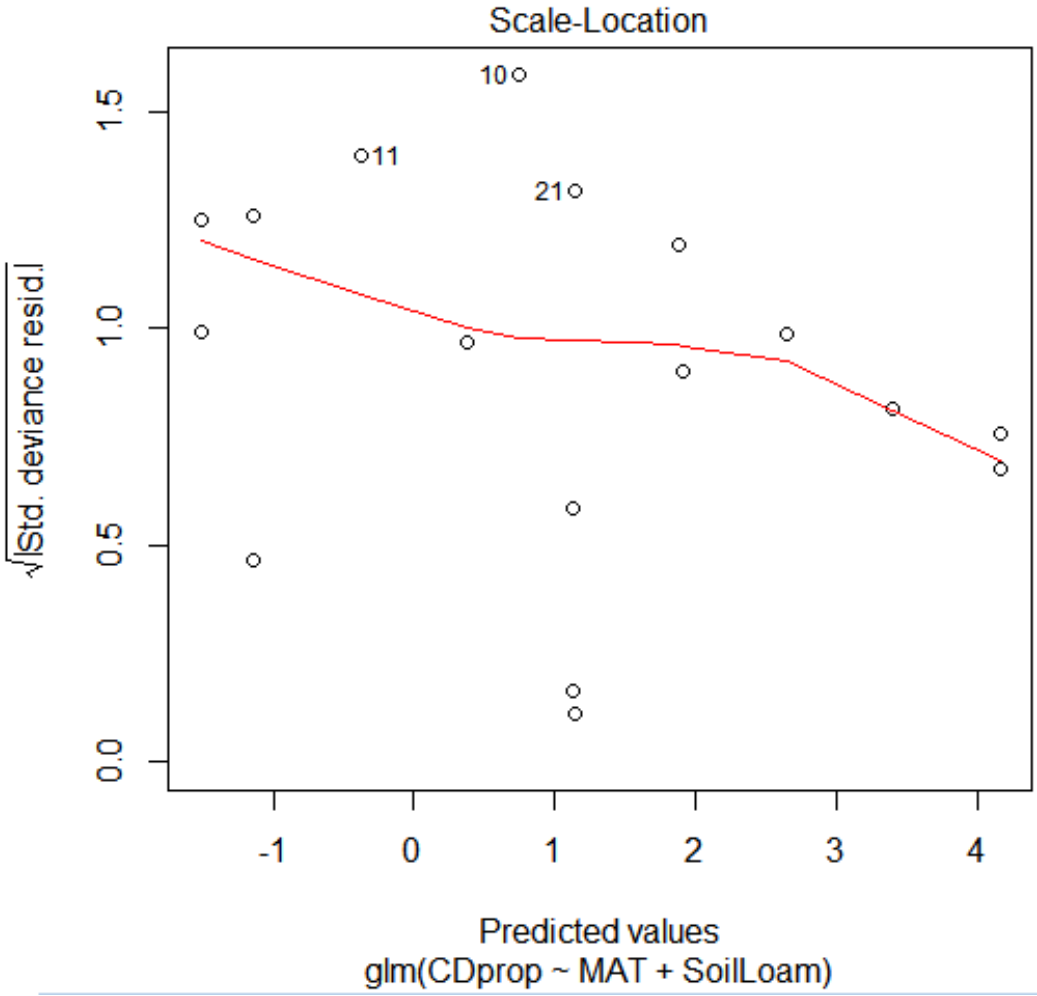


3. Validation

Plot #3: **Scale-Location**

Here, we're looking for wedges too (it's actually almost the same plot as Residuals vs. Fitted: instead of residuals against fitted values, this is a plot of sqrt(|residuals|) against fitted values so the values are all positive and it's easier to see trends).

NO wedges or nonlinearities evident here either.

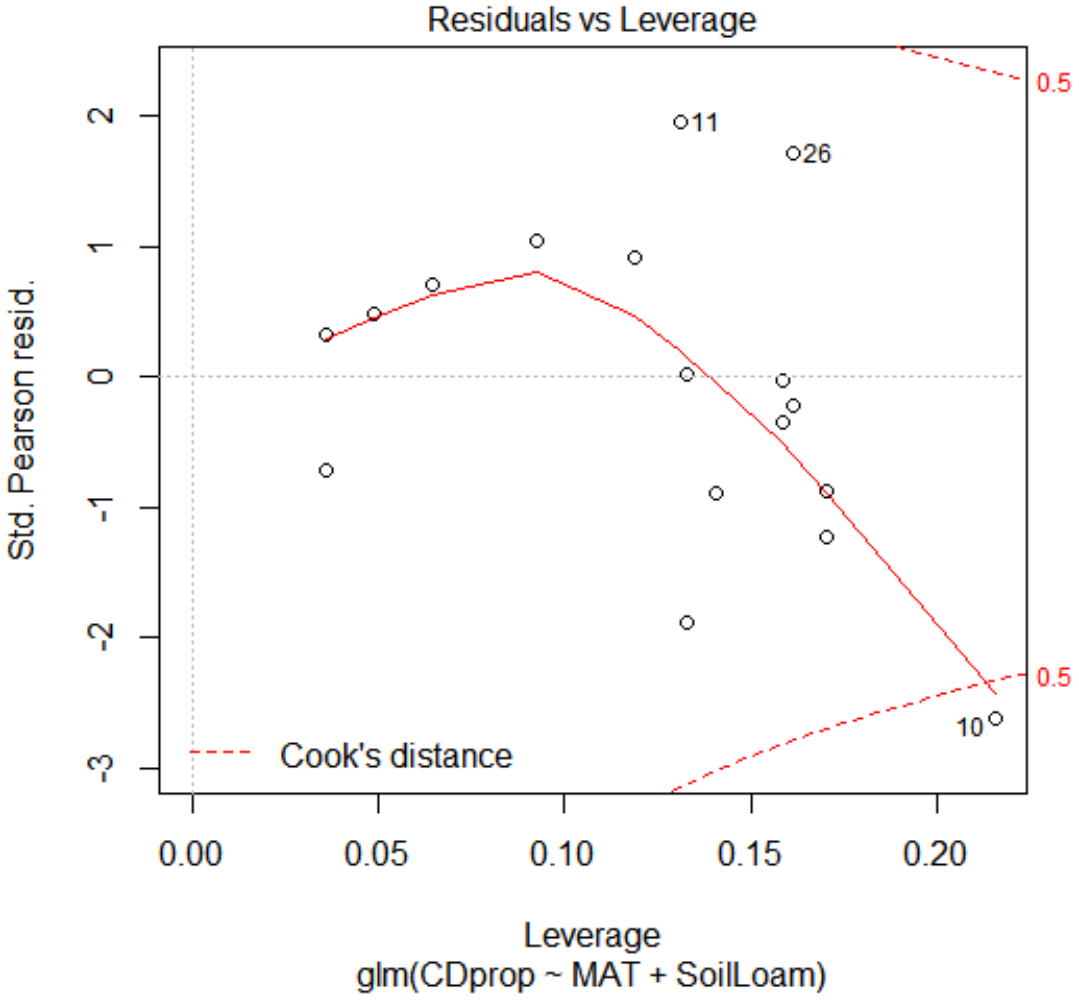


3. Validation

Plot #4: **Residuals *vs*. Leverage**

The Residuals-Leverage plot shows whether there are any outliers, which are points with **Cook's distance** ≥ approx. 1 (a conventional value). Cook's distance is a measure of how much all the residuals would change if that particular point had been omitted and **leverage** is a measure of how much that point on its own affects the model fit.

Here, Cook's distances are all <1 so no outliers (if you do have outliers, re-run the analysis with the outliers excluded (one-by-one and altogether) to see whether omitting those points would affect your conclusions).




LET'S TAKE
A
BREATHER

Interaction terms, fixed & random effects and backwards deletion

NERC SCIENCE OF THE ENVIRONMENT

CEH

4. Interaction terms

In the analysis above, some might say that I should have included **interaction terms** at the start. This involves substituting "*" for "+" on the right-hand side of the formula like this:

```
fm5glm=glm(SurvProb~MAT*SiteLat*SoilTexture,family=quasibinomial,data=dataf)
summary(fm5glm)
```

Most of you will probably have been told at some point not to forget about interaction terms. As you can see, however, in this example if I put them in then I get extra rows in the summary table and all of them are now insignificant. What's the deal???
(n.b. the ":" in an R summary table means "interacting with" and any term with a ":" in it is called an **interaction effect**, with the basic (no ":") terms above them now called **main effects**).

I avoided doing this above because there are quite a few subtleties that come into all this when there are interaction terms.

R Console

```
> summary(fm5glm)

Call:
glm(formula = SurvProb ~ MAT * SiteLat * SoilTexture, family = quasibinomial$
    data = dataf)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-0.62114  -0.07527   0.00000   0.04940   0.65363

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)               -1.831e+01  1.740e+01  -1.052    0.309
MAT                        9.360e-01  8.140e-01   1.150    0.268
SiteLat                   -3.522e-01  1.898e+00  -0.186    0.855
SoilTextureSand           -2.916e+02  1.343e+05  -0.002    0.998
SoilTextureLoam            1.466e+01  1.777e+01   0.825    0.422
MAT:SiteLat                2.643e-02  8.726e-02   0.303    0.766
MAT:SoilTextureSand        1.887e+01  6.394e+03   0.002    0.998
MAT:SoilTextureLoam       -8.197e-01  8.286e-01  -0.989    0.338
SiteLat:SoilTextureSand   -1.627e+01  1.315e+04  -0.001    0.999
SiteLat:SoilTextureLoam    1.030e+00  2.007e+00   0.513    0.615
MAT:SiteLat:SoilTextureSand 7.930e-01  6.261e+02   0.001    0.999
MAT:SiteLat:SoilTextureLoam -6.040e-02  9.176e-02  -0.658    0.520

(Dispersion parameter for quasibinomial family taken to be 0.1253748)

    Null deviance: 13.8985  on 26  degrees of freedom
Residual deviance:  1.8087  on 15  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 21

>
```

4. Fixed or random?

In order to explain this, I need to first quickly talk about **fixed and random effects** in GLMs. There are two different classes of field experiment (Hurlbert 1984, Krebs 1999:ch.10):

Manipulative experiments, where “the different experimental units receive different treatments and the assignment of treatments to experimental units is or can be randomized” (i.e. a **controlled, experimental study**), and

Mensurative experiments, which have no treatments and “involve only the making of measurements at one or more points in space or time” (i.e. an **observational study**)

However, I would generalise this a little: in a manipulative experiment the goal is to keep every variable constant across the design except for the 'key' variables being manipulated, but this is never 100% achievable: there is always residual/natural variation because of other known environmental factors (some of which you may have measured, some not) and residual variation (i.e. from unknown factors). Therefore, any manipulative experiment is really partly manipulative and partly mensurative.

Also, mensurative experiments can be ‘approximately’ manipulative, e.g. if your sites are chosen to represent the spectrum of mean annual temperatures (*MAT*), then even though you didn’t manipulate the *MAT* at these sites, as long as you can argue that the sites are otherwise equivalent then you *effectively* have done so.

NERC SCIENCE OF THE ENVIRONMENT CEH

4. Fixed or random?

Based on these definitions, I can now define what fixed and random effects are.

Fixed effects	The **'key' variables being manipulated** (or selectively chosen through site or sample selection). Called "fixed" because if you reran the experiment you would expect this variable to have the same (i.e. fixed) effect on the response. Called an "effect" because they represent a deviation from an overall mean (e.g. the "effect" of choosing a loamy soil above was to shift the duiker survivorship down by a certain amount).
Random effects	Any other **variables you have data on but are not manipulating**. n.b. in mensurative experiments where you have 'approximately' manipulated variables, you have a choice whether to treat them as fixed or random (although you need to justify the choice)
Labels **(This is <u>just my definition (!)</u>, but it reduces confusion for me to define a third one here)**	**Site codes and sample numbers** (e.g. *SiteCode* in the duiker example) Labels seem to fit the definition of fixed effects by virtue of the sites/samples having been chosen by the experimenter (are we interested in these particular sites/samples? Yes!), but also seem to fit the definition of random effects by virtue of the selection/sampling process having been random (are we trying to make general comments about all sites/samples of similar type? Yes!), which is why e.g. Crawley 2007 takes them as random effects. However, unlike fixed or random effects, I argue that labels carry no information relevant to a GLM. n.b. sometimes, variables seem not to be labels but effectively are, e.g. in a medical study where the genotype of all treatment subjects is recorded: genotype is not manipulated so you might think it's a random effect, but because every subject has a completely unique genotype it actually conveys no useful information and it becomes effectively just a label.

n.b. Some authors advise to avoid using these terms completely (e.g. Gelman (2005)'s section *Fixed and random effects* and Gelman & Hill (2007)'s *Why we avoid the term "random effects"* ; summarised at http://andrewgelman.com/2005/01/25/why_i_dont_use/): <u>I don't advise to do so</u>, but do read those sections if you have difficulty with the definitions above.

GLMs can handle both fixed and random effects, but when both are present they are called **mixed-effects models** (GLMMs; aka. **multi-level model**, **hierarchical model**). Based on these definitions, we can see that most manipulative experiments should be analysed as a GLMM including the manipulated effects as fixed (and the rest random), and most mensurative experiments as a GLMM including the 'approximately manipulated' effects as fixed (and the rest random).

NERC SCIENCE OF THE ENVIRONMENT CEH

4. Fixed or random?

The terms fixed/random effect cause HUGE confusion because every textbook seems to explain it in a very different way. Let me troubleshoot some potentially misleading definitions from the literature here:

- Krebs (1999:349) - a very standard ecological experimentalists' textbook - stated that:

fixed factors (=effects) were those where

"*(1) All levels of the classification are in the experiment, or*

(2) [only the] levels of interest to the experimenter are in the experiment, or

(3) the levels in the experiment were deliberately and not randomly chosen.".

- What does this mean? Well, for me it's right but not a very good definition (apologies to Krebs!): (1) is just a special case of (2) because the only reason all levels of the classification would be in the experiment is if they are all of interest to the experimenter. (3) really just says that a fixed effect is one that isn't a random effect so doesn't help much. Finally, (2): if you think that in manipulative experiments you choose the variable levels you are interested in testing (e.g. 7 levels of dose applied to some subjects), I think Krebs is trying to say here that a fixed effect is a manipulated/selectively chosen as I have it above.
- The way (2) is phrased seems to imply that if the experiment includes levels of the variable that are not of interest to the experimenter then it'll become a random effect, but it seems that is not the implication (and doesn't happen).
- The way (3) is phrased seems to imply that a variable like *SoilTexture* in the duiker example has to be fixed because we chose the sites and the soil textures came with the sites so, effectively, we chose those too. However, this isn't right: anything chosen only 'by extension of' choice of site/sample *doesn't* count as fixed.

NERC SCIENCE OF THE ENVIRONMENT CEH

4. Fixed or random?

- Krebs also stated that:

random factors (=effects) were those where "*All levels in the experiment are a random sample from all possible levels*".

- This is correct only if you add "... *all possible levels of the population of interest*" on the end. In the duiker example *SoilTexture* was an unmanipulated quantity (i.e. random), and therefore the values *look as if* they are just a random sample from the possible soil textures for the region where S1-S27 are located. However, if you've got something that's supposed to be random, you need to ask yourself what is it *representative of*? Perhaps this whole region is located within an unusual set of soil types (e.g. the Kalahari sands cover most of Southern Africa), in which case they might be regionally representative, but not globally representative. Does that mean we can't say this is a random variable? Also, you might be aware that the sites in use were not originally chosen randomly (field sites almost never are), so does that mean this variable is a fixed effect?
- To make Krebs's definition work, we need to define clearly the ***population of interest***, which need not be global (e.g. we might only be interested in Kalahari sites, or because you know your field sites were all built on *terra firme* forest sites in the Amazon then you'll take that forest type as your population of interest). Once defined, you will then (1) be able to say that your random variables are (more or less) randomly chosen from that population of interest (i.e. may be assumed to be representative of that population) but also remember (2) in your Discussion section when you write it up you can draw conclusions about the population of interest (e.g. survivorship of duiker in Kalahari sites), but any wider conclusions need to be justified by more than just the analysis of your data (e.g. survivorship of duiker in global savannas). Be careful about this: if you want to make global conclusions based on what are essentially data from a local study then you *can* in your Discussion, but you need to start each paragraph with "Our data suggest that / we speculate that ...", not "Our data show that ..." (and reviewers are very sensitive to this!).
- In order to label a variable as random, do I have to prove that the sites (a) were selected randomly and (b) are globally representative? NO: that is not required: in practical terms, I prefer to say that a variable is *random* if it is *not fixed* and variables are only fixed if they are manipulated (see above). Note that I'm out of step with a lot of sources in disagreeing with Krebs (1999) (e.g. ProfessorParis on https://www.youtube.com/watch?v=Jzb2tGIDEKE quotes definitions of fixed and random more or less the same as Krebs's).

NERC SCIENCE OF THE ENVIRONMENT CEH

4. Fixed or random?

- Are we interested in within-sample or within-site means/variation? We might be because we might e.g. want to check that none of our sites has suspiciously high variance in comparison to the others. Plotting/calculating this, of course, requires a label effect (like *SiteCode*) and leads to a quantity like inter-site variance or inter-subject variance σ. Obviously, a quantity like σ only has real interest if you assume idealised sampling (in which case σ has meaning beyond the particular sampling design you have used). For me, however, this is 'labels analysis' rather than random effects analysis (despite it being called random effects by many sources e.g. ProfessorParis on https://www.youtube.com/watch?v=Jzb2tGlDEKE). In environmental science, because sampling is almost never ideal, this kind of analysis achieves little more than the 'housekeeping' step of checking the assumptions of your experimental design.

- Crawley (2007:472) supplied what I think are very cryptic definitions of fixed and random:

> Mixed-effects models are so called because the explanatory variables are a mixture of fixed effects and random effects:
>
> - fixed effects influence only the *mean* of *y*;
> - random effects influence only the *variance* of *y*.

- (*y*=the response variable). I do agree with these, but I find them very unclear: Surely most predictors that change the mean of something will also change the variance as well anyway (and *vice versa*)? This isn't a definition as much as stating what we are generally most interested in for analysis.
- Imagine an experiment where we have manipulated a small no. of variables *F*, *G*, etc.. Because a fixed effect *F* is under manipulation, we are interested in seeing how the response variable *Y* changes overall as a result of changing *F* and we'd usualy plot this on a graph of *Y* against *F* and and look at the fit line.
- If we have data on any other random variables *R*, we are generally not interested in searching for any functional relationship *Y*=f(*R*), but we do want to know broadly whether or not *R* is important or not, and we can quantify that most easily by looking at the variance at each level of *R*.

NERC SCIENCE OF THE ENVIRONMENT CEH

4. Quiz: Fixed or random?

Imagine an experiment where you are looking at how fast athletes can run in a 100 m race. You select 100 athletes at random and put them on controlled diets (say *Diet* = A, B and C, with A being no change) and also give them various kinds of running shoes (say *Shoe* = S0, S1 or S2 with S0 being the shoes they start with), then measure how fast they can run 100 m (*Time*). All of this was replicated with another 2 sets of 100 athletes (groups G1, G2 and G3).

You also note other aspects of your subjects: the athletes' *height*, *shoe-size* and *sex* (M/F). As it turns out, the heights of your athletes are exactly on the UK national average in both mean and range, but the sample of athletes happens to be biased towards males and small feet (i.e. a random sample, but not representative of the UK population).

Which of these various predictors is fixed and which random? (*Hint*: some can be both depending on the kind of analysis I may try, and some can be neither).

Variable	Fixed or Random?
Diet	
Shoe	
Time	
Group	
Height	
Shoe-size	
Sex	

4. Interaction terms

OK: back to interaction terms. What is an **interaction term** in the context of a GLM? See http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture1.htm#interactions for a nice explanation of this concept (and also Fig. 4 below).

What would an interaction term have meant in this example? If changing to a site with a different soil changed the duiker survivorship by ΔA and changing to a warmer site meant ΔB then would changing to a warmer site with different soil have meant a response significantly different from ΔA+ΔB in any of the sampled cases? If yes, then there was an interaction between these two predictors.

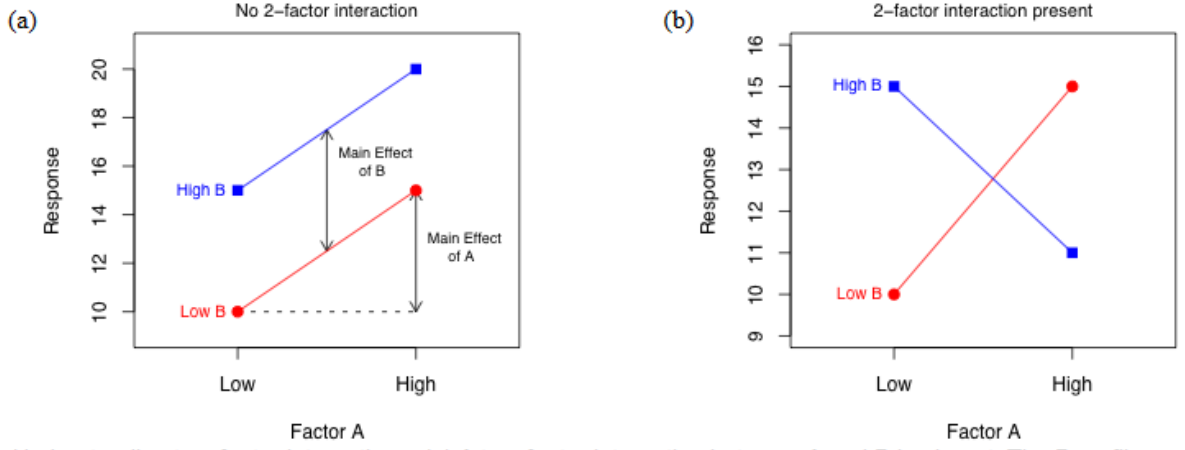


Fig. 4 Understanding two-factor interactions. (a) A two-factor interaction between A and B is absent. The B profiles are parallel and so the A main effect and B main effect are well-defined. Increasing either A or B from its low to high level has the effect of increasing the response. The change is the same regardless of the value of the other variable. (b) A substantive two-factor interaction A × B is present. The effect of each factor depends upon the level of the other. Here when B is at its low level the effect of increasing A from low to high is to increase the value of the response. When B is at its high level the effect of increasing A from low to high is to decrease the value of the response.

4. Interaction terms

Personally, I've always found that it is easier to think of **independence** rather than interactions:

The statements in these rows are equivalent:	
Predictors *A* and *B* are completely independent of each other (i.e. within experimental error, varying *A* doesn't affect the value of *B* and varying *B* doesn't affect the value of *A*)	There is no interaction between *A* and *B*
There is a dependency/correlation between *A* and *B*	The interaction term *A:B* is significant (n.b. those of you familiar with multivariate Taylor expansions will also know that a dependency between *A* and *B* will mean that you'll find the product term *A*B* is a significant predictor in your GLM as well as *A* and *B* separately)

By the way, I recommend Jim Frost's explanation of what interaction terms are in terms of hot dogs, chocolate sauce and 'it depends' questions: http://statisticsbyjim.com/regression/interaction-effects/

NERC SCIENCE OF THE ENVIRONMENT CEH

4. Interaction terms

QUESTION: “Hold on: I’ve been told to search for correlations between my predictors and exclude pairs of predictors that correlate too closely (i.e. closely enough that they’re basically collinear): doesn’t that imply I need to choose my predictors so that I don’t have any interactions? And therefore I’ll never need to include interaction terms in the final analysis?” (e.g. see https://stats.stackexchange.com/questions/52177/what-to-do-with-collinear-variables).

ANSWER: Partly right: you should definitely get rid of predictor pairs that are straight multiples of each other (e.g. height_in_feet and height_in_metres) but also look very closely for close correlates (e.g. nationality and citizenship categories, road_density_in_2012 and road_density_in_2013). There’s some nice advice about multicollinearities on https://www3.nd.edu/~rwilliam/stats2/l11.pdf. However, you should only remove variables that correlate *very* closely and there will be plenty of potential interactions remaining in the data.

How close is “*very* closely”, though? Not all agree: Thomas *et al.* (2017:53&61) put r^2>0.09 as the threshold and called this “highly correlated”, but for me that’s a lot too low: generally “highly correlated” means r^2>0.70, but actually I’d go for more like r^2>0.90 because I prefer to let the model selection rules choose what variable gets excluded if at all possible.

4. Interaction terms

QUESTION: “I’m pretty sure that variable A is an important predictor, but I’m not sure whether I’ll get the best results using A or (1/A) or ln(A) or some other function of A. Can I put them all in as predictors and let the GLM tell me which is the most significant?” (e.g. see https://stats.stackexchange.com/questions/52177/what-to-do-with-collinear-variables where they wanted to include variables *Table*, *Depth* and *Table/Depth* even though *Depth* correlated with *Table*).

ANSWER: You *can* do this, but GLMs are not designed to search all possible functional forms for the most appropriate one for your data: I recommend to do a literature search beforehand to find out which functional forms to expect in this context and only include those (e.g. in the example above, *Table* has units of length and *Table/Depth* is dimensionless, which means the coefficients for each will have different units and therefore not be directly comparable either).

QUESTION: “If I’m including predictor A then do I also have to try A^2, A^3, A^4, etc. in the GLM too to capture possible nonlinear effects? Similarly, if I include A^4 then do I need to also include all lower powers of A?”

ANSWER: The first part I’d say “no”: only go for a polynomial fit if there’s a good theoretical reason to be doing so, and stop at the power of A that theory suggests (anything more than that will be overfitting). The second part is “yes”: this is a consequence of the Principle of Marginality.

NERC SCIENCE OF THE ENVIRONMENT CEH

4. Interaction terms

On http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture1.htm, Univ. North Carolina gave the following good advice:

When should interactions be included in models?

As a general rule, interactions should always be examined with experimental data, and rarely examined for observational data. **Observational studies** are quasi-experimental designs that fall short of being true experiments for various reasons. In a typical observational study treatments are imposed by nature rather than the experimenter. As a result there is no guarantee that treatments have been randomly assigned to subjects and rarely any balance causing some treatment combinations to be under-represented. All of this makes assessing interaction in observational studies dangerous. Main effects are hard enough to assess in such studies; interactions are truly pushing the envelope.

Based on these considerations I approach the statistical analysis of experiments and observational studies quite differently. In an experiment in which all relevant factors have been assiduously controlled and in which subjects have been randomly assigned to treatments I typically start with the most complicated interaction model possible and try to simplify it. On the other hand when I analyze observational data I start with main effects and maybe tentatively examine a few interactions that have a theoretical basis.

I'd also add to that this advice from Thomas *et al.* (2017:56): "Never include an interaction term in a model unless you can write a sentence to explain what the interaction represents!"

In summary, in mensurative experiments you ARE allowed to disregard the interaction terms, but DO ALWAYS include them in the context of a manipulative experiment. Finally, this means I can answer the question a few slides back: in the duiker example it was a mensurative experiment and therefore I could have included or excluded interaction terms by choice (which means that best policy is to do it both ways to see what happens).

Next step: let's include the interaction terms and see what happens.

NERC SCIENCE OF THE ENVIRONMENT CEH

4. Crawley's Two Stage Backwards Deletion Method

The rules for **Model Selection** are modified slightly when you have interaction terms. Here is what I'm going to call **Crawley's Two Stage Backwards Deletion Method**:

1. STAGE ONE: As before, we identify predictors and predictor-levels that are non-significant, remove them one by one (this is the **backwards deletion** bit) and re-run until all predictors have significant *p*-values (n.b. to remove an (Intercept) term, add "0+" to the start of the right-hand side of the formula, i.e. straight after."~" (notice that this adds a row to any remaining categorical variables, by the way))

Do this remembering:

1a. Remove higher order terms first (i.e. terms of the form *A*:*B*:*C*:*D* before *A*:*B*:*C* before *A*:*B*). For example, here we *don't* look at *SoilTextureSand* first even though it has the highest *p*-value.

R Console

```
> summary(fm1res)

Call:
glm(formula = CDprop ~ MAT * SoilTexture, family = quasibinomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6930  -0.1336   0.0000   0.1211   0.6545

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           -15.3362     6.4251  -2.387   0.0265 *
MAT                     0.7434     0.3043   2.443   0.0235 *
SoilTextureSand      -185.3309 45046.7745  -0.004   0.9968
SoilTextureLoam         8.4850     6.7014   1.266   0.2193
MAT:SoilTextureSand     8.8644  2145.0845   0.004   0.9967
MAT:SoilTextureLoam    -0.4564     0.3140  -1.453   0.1609
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasibinomial family taken to be 0.1506076)

    Null deviance: 13.8985  on 26  degrees of freedom
Residual deviance:  2.4973  on 21  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 21

>
```

4. Crawley's Two Stage Backwards Deletion Method

1b. One case that crops up very often (though not in this example) is you find that one or both of predictors *A* and *B* are non-significant but their interaction *A*:*B* *is* significant (called a **cross-over interaction**). By **the Principle of Marginality** you must keep *all* these terms in the model, e.g. if you are keeping *A*:*B* then you must also keep both *A* and *B* even if they are non-significant (see e.g. Calcagno & de Mazancourt 2010 or the nice explanation at http://www.unc.edu/courses/2010fall/ecol/563/001/docs/lectures/lecture1.htm#interactions).

- The interpretation of **cross-over interactions** is tricky if you have many interacting terms, but basically this is happening because increasing *A* has too small an effect on the response to be significant (perhaps a small positive effect) and increasing *B* also has too small an effect to be significant (perhaps a small negative effect) but they are different enough *from each other* to give a significant interaction.

- In my duiker example the second-order interaction term is non-significant in both its levels, therefore it is excluded first and the analysis reduces to exactly what we did in the first part above without the interaction terms (this is why I ignored interaction terms the first time through this example).

4. Crawley's Two Stage Backwards Deletion Method

1c. How to deal with categorical predictor levels: e.g. what if your model has a 3-level categorical predictor *A* with *p*-values for its levels of 0.021, 0.034 and 0.092 and a continuous predictor *B* with *p*-value 0.076. What do you do?

There are two answers:

(i) Do what I did with the *SoilTexture*/*SoilLoam* variable above and combine the least significant level with another level (Crawley 2007 advises this in a few examples)

or (ii) follow the rule that you always look at the lowest significant level for categorical variables, which for *A* here would give it an overall *p*-value of 0.021 and you remove predictor *B* first (this follows the logic of the Principle of Marginality more closely)

I have seen both approaches used and, as far as I can assess from the literature I've read, both are acceptable.

2. STAGE TWO: Don't forget, you still need to do the `stepAIC()` bit after steps 1-4.

(Stages 1 and 1a are equivalent to the three Rules on Thomas *et al.* 2017:69).

NERC SCIENCE OF THE ENVIRONMENT CEH

4. Crawley's Two Stage Backwards Deletion Method

In summary, **Crawley's Two Stage Backwards Deletion Method** is:

Start with your **full model** involving all the predictors you want to use

Looking at *p*-values, but ignoring AIC

Principle of Marginality

Derive a model nested within the full model that has all its predictors statistically significant. This is the **all-predictors-significant model**.

Looking at AIC, but ignoring *p*-values

Use `stepAIC()` to check that model and all models nested within it and choose the one with the lowest AIC. This is the **best-fit model**.

• The **full model** is the model with all your proposed predictors in it. A **saturated model** is slightly different: it's one where there are as many data points as parameters.

• In GLM terms, a model *M* is **nested within** model *N* if the response variable is the same for both and all the predictors of *M* are also predictors of *N*.

• Using the Akaike information Criterion (AIC) is sometimes tricky and I recommend skimming through Anderson & Burnham's *AIC Myths and Misunderstandings* from 2006 at some point: https://sites.warnercnr.colostate.edu/anderson/wp-content/uploads/sites/26/2016/11/AIC-Myths-and-Misunderstandings.pdf.

NERC SCIENCE OF THE ENVIRONMENT CEH

4. Crawley's Two Stage Backwards Deletion Method

Where does this method come from? Basically, it's an implementation of what is suggested in Crawley (2007:324): "The best model is the model that produces the least unexplained variation (the minimal residual deviance [*Toby: usually as measured by AIC*]), subject to the constraint that all the parameters in the model should be statistically significant". Even though I feel that Crawley was less than clear about the two stage process implied by this statement, if you consider enough examples, this is what it amounts to so I've called it "Crawley's Two Stage Backwards Deletion Method" even though that name doesn't appear in Crawley (2007).

For theory see Crawley (2007:433) or the similar method in Thomas *et al*. (2017:69-74&120) (n.b. many websites portray model selection as a choice between using *p*-values or AIC - e.g. here - but both Crawley and Thomas *et al*. support a method involving both).

Crawley's ideas do not convince everyone, of course: Gelman & Hill (2007:42) stated "People sometimes think that if a coefficient estimate is not significant, then it should be excluded from the model. We disagree. It is fine to have nonsignificant coefficients in a model, as long as they make sense" *[Toby: Gelman & Hill are not referring to the predictors and predictor levels that remain because of the Principle of Marginality etc.].* I am following Crawley not Gelman & Hill on this point, but it's interesting to read Gelman & Hill (2007:§4.6) and see their reasons for taking this point of view.

NERC SCIENCE OF THE ENVIRONMENT CEH

FINAL FEW POINTS

5. Error structure

How do you know whether your response variable is Gaussian, Poisson or binomial? See this table (cf. Gelman & Hill 2007:109, Thomas *et al*. 2017:90-105).

Error distribution	The response variable contains ...		Canonical (=most usual) transformation aka. Link Function
GAUSSIAN (= NORMAL; overdispersion not possible *)	... **Measurement data** (i.e. things like people's heights, which are on a continuous scale)	This is **untransformed linear regression**	No transformation, i.e. *ita=response*
BINOMIAL (or QUASIBINOMIAL if overdispersed)	... **Binary data** (only 0 or 1 allowed, e.g. presence/absence data, or sex of duikers coded as M=0, F=1)	This is **logistic regression**	LOGIT, i.e. *ita*=ln(*response*/(1-*response*)) or ita=qlogis(response) in R
QUASIBINOMIAL	... **Proportions data** (e.g. probabilities/percentages). Like binary data, but values between 0 and 1 are allowed		LOGIT, i.e. *ita*=ln(*response*/(1-*response*)) or ita=qlogis(response) in R
POISSON (or QUASIPOISSON if overdispersed)	... **Count data** (Y must be integers ≥0, e.g. litter size of antelope which are counted, not measured on a scale).	This is **Poisson regression**	LOG, i.e. *ita*=ln(*response*)

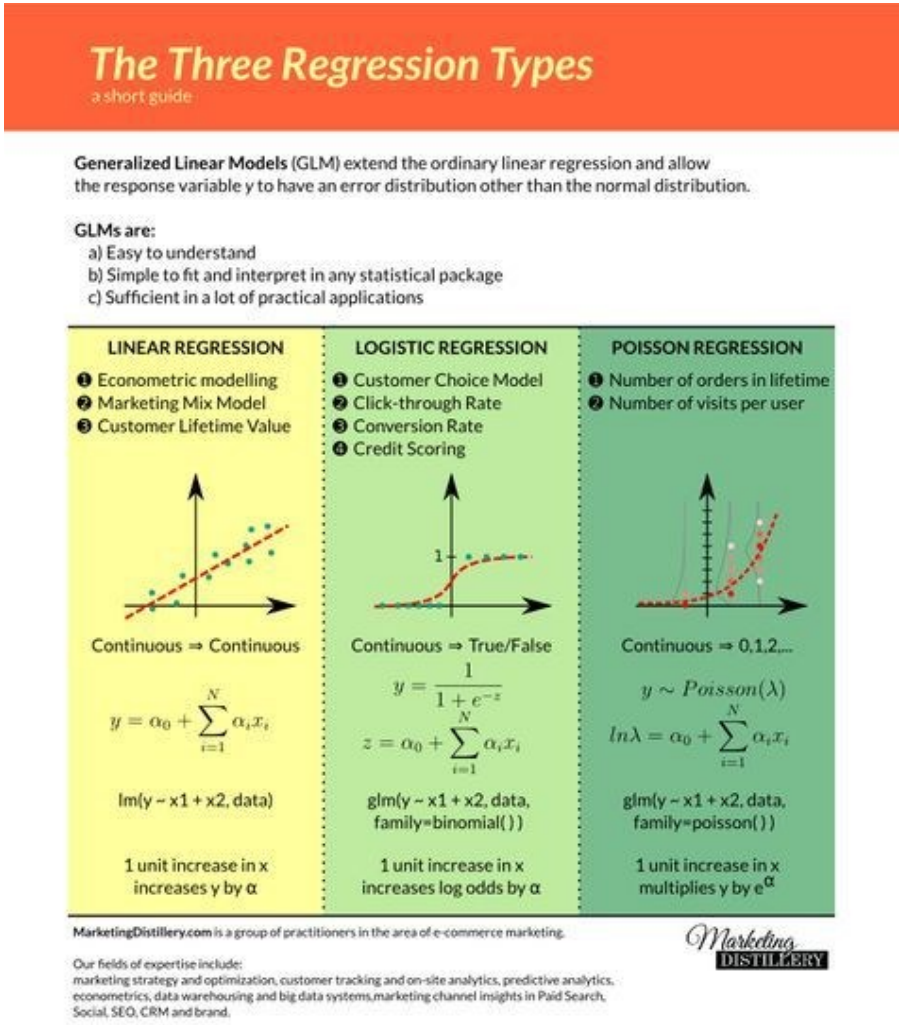
•These apply to your *response variable* only: no GLM or any other kind of regression makes any assumption about the distribution of your *predictor variables* so you should never need to transform those.
•No tests for normality, Poisson or binomial distribution here: the error structure is simply what the data are 'supposed' to go like (the link function is supposed to handle deviations from this).
•There are other kinds of errors (and there are also alternative options to those canonical link functions), but in no GLM I have used since 2005(ish) have I ever needed any but these here.

NERC SCIENCE OF THE ENVIRONMENT CEH

5. Error structure

The table above should ring some bells from non-GLM regression theory:

Error distribution	
GAUSSIAN (= NORMAL; overdispersion not possible *)	This is **untransformed linear regression**
BINOMIAL (or QUASIBINOMIAL if overdispersed)	This is **logistic regression**
QUASIBINOMIAL	
POISSON (or QUASIPOISSON if overdispersed)	This is **Poisson regression**

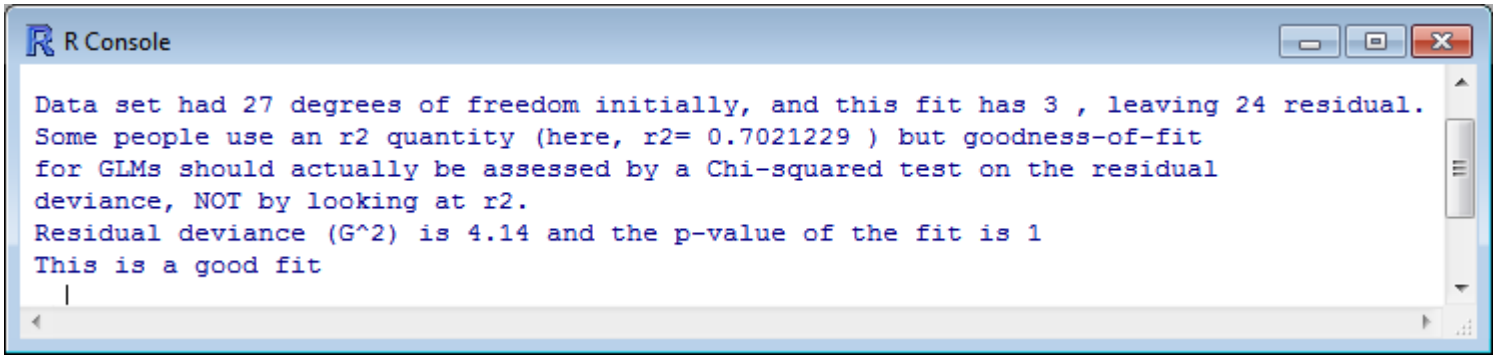


5. Goodness of fit

A quantitative measure of goodness of fit for GLMs is tricky (r^2 alone is not sufficient and the various 'adjusted r^2' measures not much better). I advise to use the following commands BUT see Burnham & Anderson (2002:37), Crawley (2007:516) and the comments on

http://bbolker.github.io/mixedmodels-misc/glmmFAQ.html#how-do-i-compute-a-coefficient-of-determination-r2-or-an-analogue-for-glmms:

```
n=length(SurvProb)
p=n-fm3glm$df.residual
cat("Data set had",n,"degrees of freedom initially, and this fit has",p,", leaving",n-p,"residual.\n")
rsquared=(fm3glm$null.deviance-fm3glm$deviance)/fm3glm$null.deviance
cat("Some people use an r2 quantity (here, r2=",rsquared,") but goodness-of-fit\nfor GLMs should actually be assessed by a Chi-squared test on the residual\ndeviance, NOT by looking at r2.\n")
cat("Residual deviance (G^2) is",round(fm3glm$deviance,4),"and the p-value of the fit is",round(1-pchisq(fm3glm$deviance,fm3glm$df.residual),4),"\n")
if ((1-pchisq(fm3glm$deviance,df=fm3glm$df.residual))>0.05) {cat("This is a good fit\n")} else {cat("This is a poor fit\n")}
anova(fm3glm,test="Chisq")
```



NERC SCIENCE OF THE ENVIRONMENT CEH

5. Overdispersion

A quantitative measure of overdispersion is also tricky (if you don't know what this is, see Thomas *et al.* 2017:94 and/or https://github.com/lme4/lme4/issues/220).

```
theta=fm3glm$deviance/fm3glm$df.residual     #From Thomas et al. (2017:section "Overdispersion and what to do about it")
toofar=2
cat("Dispersion",theta,"\n")
if (theta<1) {cat("Data are under-dispersed\n")}
if (theta>1) {
 if (theta>toofar) {cat("Data are definitely over-dispersed: see Thomas et al. (2017:section Overdispersion and what to do about it)\n")} else {cat("Data are slightly over-dispersed, but not enough to cast doubt on your fit results\n")}
}
```

See https://bbolker.github.io/mixedmodels-misc/glmmFAQ.html#overdispersion for a slightly more general version of this (e.g. the above won't work for glmer fits). Also, the ppt at http://faculty.washington.edu/heagerty/Courses/b571/handouts/OverdispQL.pdf .

5. Analysis using glmmPQL

The same analysis using glmmPQL() rather than glm():

```
library(MASS);library(MuMIn)
fm3res=glmmPQL(fixed=SurvProb~MAT+SoilTexture,random=~1|SiteCode,na.action=na.fail,family=quasibinomial,data=data)
summary(fm3res)
```

Check: Notice the formula syntax differs slightly for glmmPQL(): there is a **random term** as well now.

```
> summary(fm3res)
Linear mixed-effects model fit by maximum likelihood
 Data: dataf
  AIC BIC logLik
   NA  NA     NA

Random effects:
 Formula: ~1 | SiteCode
        (Intercept)  Residual
StdDev:   0.5713228 0.2339066

Variance function:
 Structure: fixed weights
 Formula: ~invwt
Fixed effects: CDprop ~ MAT + SoilTexture
                    Value Std.Error DF   t-value p-value
(Intercept)     -6.205590 1.6911669 23 -3.669413  0.0013
MAT              0.313885 0.0722881 23  4.342147  0.0002
SoilTextureSand  1.005808 0.7897073 23  1.273646  0.2155
SoilTextureLoam -1.289728 0.6265458 23 -2.058473  0.0510
 Correlation:
                (Intr) MAT    SlTxtS
MAT             -0.950
SoilTextureSand -0.298  0.095
SoilTextureLoam  0.070 -0.349  0.527

Standardized Within-Group Residuals:
       Min         Q1        Med         Q3        Max
-1.5283238 -0.3830878  0.4781024  0.7952929  1.3705076

Number of Observations: 27
Number of Groups: 27
>
```

```
> fm3res=glmmPQL(fixed=CDprop~1,random=~1|SiteCode,na.action=na.fail$
iteration 1
iteration 2
iteration 3
iteration 4
iteration 5
iteration 6
iteration 7
iteration 8
iteration 9
iteration 10
> summary(fm3res)
Linear mixed-effects model fit by maximum likelihood
 Data: dataf
  AIC BIC logLik
   NA  NA     NA

Random effects:
 Formula: ~1 | SiteCode
        (Intercept)  Residual
StdDev:    1.115269 0.4182258

Variance function:
 Structure: fixed weights
 Formula: ~invwt
Fixed effects: CDprop ~ 1
               Value Std.Error DF  t-value p-value
(Intercept) 1.040147 0.2811099 27 3.661072  0.0011

Standardized Within-Group Residuals:
        Min          Q1         Med          Q3         Max
-1.57098819 -0.03149951  0.04386008  0.56061152  0.56061152

Number of Observations: 27
Number of Groups: 27
>
```

Check: Estimates are a bit different compared to those from glm(): in my example for new site S28 you can work out that the prediction has increased from 68% to 74% with glmmPQL().

Check: The glmmPQL() command (from Venables & Ripley 2002) is a more general version of lme() (from Pinheiro & Bates 2000).

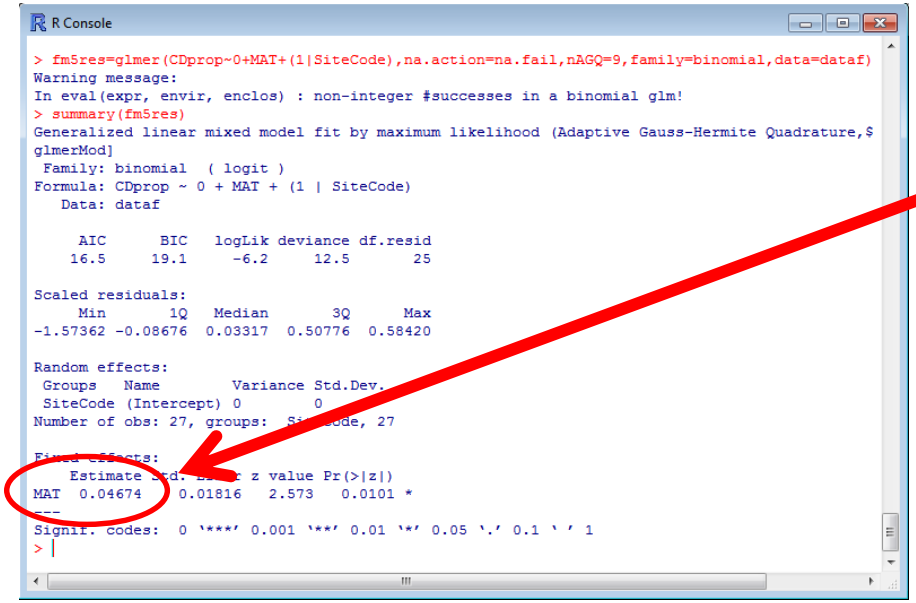
stepAIC() does not work with glmmPQL(): instead you need to use the dredge() command from package *MuMIn* instead (which is more or less equivalent but works using 'corrected AIC' AICc rather than AIC).

With glmmPQL(), however, all levels of *SoilTexture* are non-significant so the model selection step leads us to the different model: SurvProb ~ Intercept + MAT + (1 | SiteCode), and then a call to dredge(fm3res) ranks the remaining possible models by AICc (lowest on top) which leads to MAT *also* being removed (despite being significant), finally leaving the best-fit model as just ita = 1.040147

5. Analysis using glmer

The same analysis using glmer() rather than glm():

```
library(lme4);library(MuMIn)
fm5res=glmer(SurvProb~MAT+SoilTexture+(1|SiteCode),na.action=na.fail,nAGQ=9,family=binomial,data=dataf)      #Can't use quasibinomial in glmer.
Usually this would lead me to abandon glmer and rewind to using glmmPQL instead, but for demonstration purposes press on with binomial errors
summary(fm5res)
```



Check: Notice the formula syntax differs slightly for glmer(): the (1|SiteCode) term is the random term here.

Check: In my example for new site S28, the prediction is now 77% with glmer().

Check: The glmer() command is a generalisation of the lmer() command (as glmmPQL() is of lme()): when the family is Gaussian, glmer() simply calls lmer()).

stepAIC() does not work with glmer(): instead you need to use the dredge() command from package *MuMIn* instead (which is more or less equivalent but works using 'corrected AIC' AICc rather than AIC).

With glmer() all levels of *SoilTexture* are non-significant and, if you remove those from fm5res, the intercept is also non-significant so the model selection step leads to the model: SurvProb ~ 0 + MAT + (1 | SiteCode), i.e. ita = 0.04674**MAT*. (dredge() not necessary because only one remaining term). Note the slightly different conclusion this time, but we did use binomial rather than quasibinomial errors so can't be sure whether it's because of that or because of improvements in glmer().

5. The model formula

Next you need to derive your **model formula**. This is slightly different for glm(), glmmPQL() and glmer(). For the duiker example using glmer() we had:

```
SurvProb~MAT+SoilTexture+(1|SiteCode)
```

and this is a fairly standard kind of formula (for when there is no nesting in the experimental design):

```
(response)~(fixed1)+(fixed2)+...+(1|random1)+(1|random2)+...
```

The "|" symbol ("bar" or "pipe") means "grouped by" (alternatively, "given" or "conditional on").

A slash "/" means "..., within which is nested ..." (be careful: X/Y means "Y is nested within X" and *not* "X is nested within Y").

See https://stat.ethz.ch/R-manual/R-patched/library/stats/html/formula.html and/or http://cran.r-project.org/web/packages/Formula/vignettes/Formula.pdf for more information about R formula objects.

Once you have your fixed and random predictors and have worked out and your model formula, the model selection process is very similar to what we have already done above.

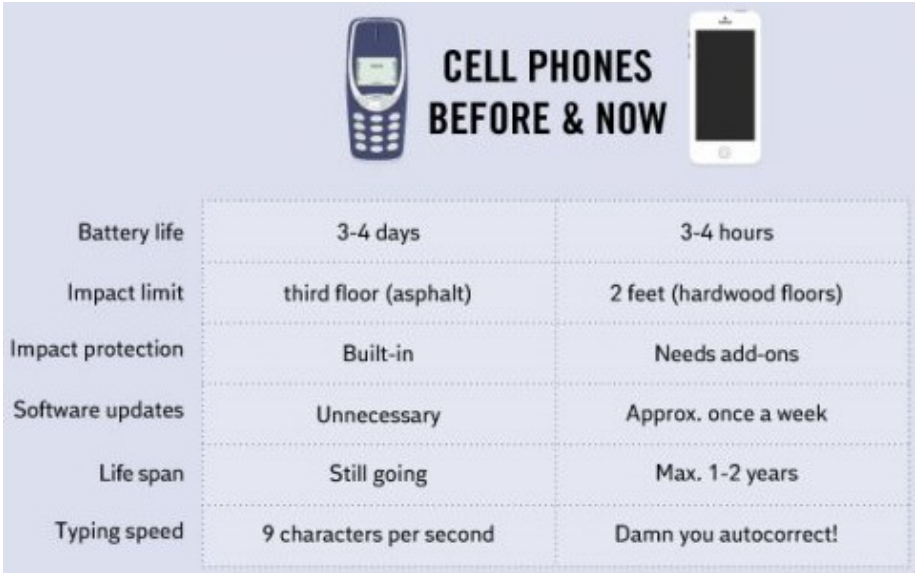
5. Which is best?

If a GLM analysis can be carried out similarly using glm(), glmmPQL() and glmer() then which is best? First thing to say is that most recent is not necessarily best (rather like with cellphones - see right from truthfacts.com):

- glm() came out 1992 (pre-dates R)

- glmmPQL() came out 2002 (package MASS)
  (based on lme() which came
  out 2000 in package nlme)

- glmer() came out 2013 (package lme4)
  (based on lmer() which came
  out 2005 in package Matrix)



My advice: When glmer works, it should be recommended over glm or glmmPQL, however it won't work with overdispersed data (quasipoisson or quasibinomial errors) and also won't work if there are any missing values (and nor will dredge). You could remove the missing values using the na.option=na.omit option (Thomas *et al.* 2017:70) but the error options you can't change so if necessary rewind to glmmPQL() which should usually work. If neither glmer nor glmmPQL converge, rewind further to the glm() command which should always work.

(see blog post https://stackoverflow.com/questions/32038355/r-glmm-glmer-vs-glmmpql which states that lme4 should do better with crossed random effects but there are other forms of error structure handled better by nlme)

Generalized Linear Models (GLMs)

1. Duiker in Africa

2. Recap of standard statistics

3. Validation

4. Interaction terms, fixed & random effects and Crawley's Two Stage Backwards Deletion Method

5. Error structure, Goodness of fit, overdispersion and using glmmPQL, dredge and glmer


HAVE A BREAK

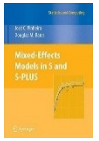
Generalized Linear Models (GLMs)

Practical

Quantitative methods are evolving fast in ecology, way faster than any of us can keep up with. We lack the foundational training in mathematical, statistical or computational skills to pick these up easily: otherwise we'd work for banks, obviously. One consequence is that many of us spend a lot of our time feeling frustrated by quantitative methods.

Matthew Smith, Microsoft Research,
Introduction to the BES Computational Ecology Group
@BES_CE_SIG in the March 2014 British Ecological Society (BES) *Bulletin*.

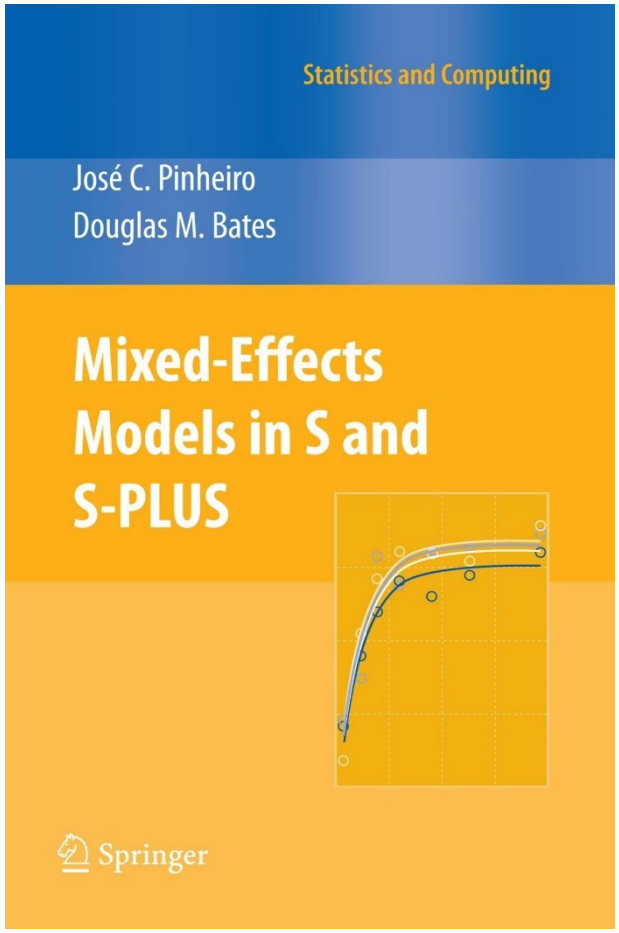
1. Examples from Pinheiro & Bates (2000)



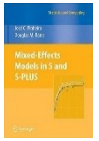
Probably the most widely-used ‘starter’ book for GLMs in ecology is Pinheiro & Bates (2000). Their Chapter 1 is where I personally started with all this in 2007 and it’s a linked set of worked examples that contain a lot of very useful information.

Luckily, you can access this book almost completely online through the PREVIEW feature on https://books.google.co.uk/books?id=ZRnoBwAAQBAJ (if that doesn’t work for you, please ask for my copy). That means you can go through P&B’s examples rather than ones I’ve put together, which has the advantage that you will find these examples all over the internet analysed by other people and can search for further details on specific points.

In the slides below I’ve presented my take (and solutions) to some of these worked examples. Going through them quickly is the best way to get a feel for how these methods work in practice.



1.1. RAILWAY RAILS example (P&B p.4)

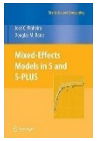


I find the presentation of the Rails example pretty confusing (I know it flummoxed me completely!):

- *Rail* is a label (as I call it), not a fixed or random effect. Therefore, although Crawley would do a random effects analysis with it, for me this should be taken with a heavy pinch of salt because all you can calculate is whether the within-rail variation is greater or less than the between-rail variation (a 'variance components analysis' or 'variance partitioning analysis'), and in my experience that is legitimate, but of very limited *practical* use.
- Also, P&B talk about the 'single-mean model' on p.5: this isn't wrong, but I find it to be an example of a simplification that just introduces confusion and is almost never used in a practical analysis, so I advise to completely ignore that bit.
- The text says "Data from a one-way classification like the rails example can be analyzed either with a fixed-effects model or with a random-effects model", which I find really misleading because a fixed-effects analysis is inappropriate here. Anyway: following a Crawley-like approach:

```
library(nlme)
View(Rail)          #Y=travel, LBL1=Rail (Rail is the classification factor)
                                #therefore formula is travel~1|Rail
plot(Rail)          #Pinheiro & Bates 2000:Fig. 1.1
plot.design(Rail)#in style of Pinheiro & Bates 2000:Fig. 1.6

#RANDOM EFFECTS ANALYSIS
fmlRail.lme=lme(travel~1,random=~1|Rail,data=Rail)              #Random-effects model on p.7
summary(fmlRail.lme)
fmlRail.lme$coefficient$fixed     #Best estimate of beta (which now takes a different value for each Rail) - see intervals(fmlRail.lme) for confidence interval
intervals(fmlRail.lme)$reStruct$Rail[2]          #Best estimate of sigma_b (between-Rail residual standard error) - see intervals(fmlRail.lme) for confidence interval
summary(fmlRail.lme)$sigma     #Best estimate of sigma (within-Rail residual standard error) - see intervals(fmlRail.lme) for confidence interval
plot(fmlRail.lme)                #Pinheiro & Bates 2000:Fig. 1.4 (the command plot(fmlRail.lme$residuals) is slightly different)
intervals(fmlRail.lme)
anova(fmlRail.lme)
plot(fmlRail.lme,form=resid(.,type="p")~fitted(.)|Rail,abline=0,lty=2)          #Pinheiro & Bates 2000:Fig. 1.8
```

1.2. ERGOSTOOL *example* (P&B p.12)



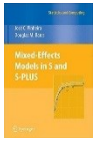
An example about the effort required to arise from various different types of stool.

Type is fixed and *Subject* is a label

```
library(nlme)
View(ergoStool) #Y=effort, XVAR=Type, LBL1=Subject (both Type and Subject are classification factors)
                #therefore formula is effort~Type|Subject
plot(ergoStool)  #Pinheiro & Bates 2000:Fig. 1.5
plot.design(ergoStool)          #Pinheiro & Bates 2000:Fig. 1.6

#MIXED EFFECTS ANALYSIS
#options(contrasts=c(factor="contr.treatment",ordered="contr.poly"))          #For me this didn't seem to be necessary ... ?
fmlergoStool.lme=lme(effort~Type,random=~1|Subject,data=ergoStool)          #model on p.14
summary(fmlergoStool.lme)
fmlergoStool.lme$coefficient$fixed          #Best estimate of beta (which now takes a different value for each Subject) - see
intervals(fmlergoStool.lme) for confidence interval
intervals(fmlergoStool.lme)$reStruct$Subject[2]          #Best estimate of sigma_b (between-Subject residual standard error) - see
intervals(fmlergoStool.lme) for confidence interval
summary(fmlergoStool.lme)$sigma          #Best estimate of sigma (within-Subject residual standard error) - see intervals(fmlergoStool.lme)
for confidence interval
plot(fmlergoStool.lme)          #Pinheiro & Bates 2000:Fig. 1.7 (the command plot(fmlergoStool.lme$residuals) is quite a lot different)
intervals(fmlergoStool.lme)
anova(fmlergoStool.lme)
plot(fmlergoStool.lme,form=resid(.,type="p")~fitted(.)|Subject,abline=0,lty=2)     #Pinheiro & Bates 2000:Fig. 1.8
```

NERC SCIENCE OF THE ENVIRONMENT CEH

1.3. MACHINES example (P&B p.21)

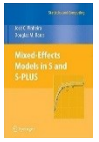


The ergoStool example plus replication

```
library(nlme)
View(Machines) #Y=score, XVAR=Machine, LBL1=Worker (both Machine and Worker are classification factors)
                #therefore formula is score~Machine|Worker
plot(Machines)  #Pinheiro & Bates 2000:Fig. 1.9
plot.design(Machines)           #in style of Pinheiro & Bates 2000:Fig. 1.6

#MIXED EFFECTS ANALYSIS
#options(contrasts=c(factor="contr.treatment",ordered="contr.poly"))        #For me this didn't seem to be necessary ... ?
fmlMachines.lme=lme(score~Machine,random=~1|Worker,data=Machines)        #model on p.23
summary(fmlMachines.lme)
fmlMachines.lme$coefficient$fixed               #Best estimate of beta (which now takes a different value for each Worker) - see
intervals(fmlMachines.lme) for confidence interval (ALWAYS CHECK THIS: abnormally wide intervals usually indicates problems with the model
definition, e.g. too little replication (see Pinheiro & Bates 2000:27))
intervals(fmlMachines.lme)$reStruct$Worker[2] #Best estimate of sigma_b (between-Worker residual standard error) - see
intervals(fmlMachines.lme) for confidence interval (ALWAYS CHECK THIS: abnormally wide intervals usually indicates problems with the model
definition, e.g. too little replication (see Pinheiro & Bates 2000:27))
summary(fmlMachines.lme)$sigma              #Best estimate of sigma (within-Worker residual standard error) - see intervals(fmlMachines.lme) for
confidence interval (ALWAYS CHECK THIS: abnormally wide intervals usually indicates problems with the model definition, e.g. too little replication (see
Pinheiro & Bates 2000:27))
plot(fmlMachines.lme)           #Pinheiro & Bates 2000:Fig. 1.7 (the command plot(fmlMachines.lme$residuals) is quite a lot different)
intervals(fmlMachines.lme)
anova(fmlMachines.lme)
plot(fmlMachines.lme,form=resid(.,type="p")~fitted(.)|Worker,abline=0,lty=2)      #Pinheiro & Bates 2000:Fig. 1.8
```

NERC SCIENCE OF THE ENVIRONMENT CEH

1.3. MACHINES example (P&B p.21)

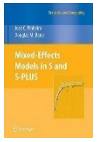


The ergoStool example plus replication

```
#MIXED EFFECTS ANALYSIS WITH NESTING
#options(contrasts=c(factor="contr.treatment",ordered="contr.poly"))          #For me this didn't seem to be necessary ... ?
fmlMachines.lme=lme(score~Machine,random=~1|Worker/Machine,data=Machines)          #model on p.23; the "1|Worker" shows that there is a single random effect for each group and the grouping is given by Worker; corresponding to subscript notation y_ij=beta_j+b_i+epsilon_ij (i=1 , ... , Worker M  ,  j=1 , ... , observation number n_i=4  ,  where y_ij is the value of score for observation j on Worker i, beta_j is the population mean of Y for stool j, b_i is a random variable representing the deviation from beta of the mean score for the ith Worker (assumed independent normally-distributed ( N(0,sigma_b^2) ); between-Worker variability) and the epsilon_ij are independent normally-distributed ( N(0,sigma^2) ) error terms (within-Worker variability). This is a hierarchical/multilevel model because there are now two sources of variation.
summary(fmlMachines.lme)
fmlMachines.lme$coefficient$fixed          #Best estimate of beta (which now takes a different value for each Worker) - see intervals(fmlMachines.lme) for confidence interval (ALWAYS CHECK THIS: abnormally wide intervals usually indicates problems with the model definition, e.g. too little replication (see Pinheiro & Bates 2000:27))
intervals(fmlMachines.lme)$reStruct$Worker[2] #Best estimate of sigma_b (between-Worker residual standard error) - see intervals(fmlMachines.lme) for confidence interval (ALWAYS CHECK THIS: abnormally wide intervals usually indicates problems with the model definition, e.g. too little replication (see Pinheiro & Bates 2000:27))
summary(fmlMachines.lme)$sigma          #Best estimate of sigma (within-Worker residual standard error) - see intervals(fmlMachines.lme) for confidence interval (ALWAYS CHECK THIS: abnormally wide intervals usually indicates problems with the model definition, e.g. too little replication (see Pinheiro & Bates 2000:27))
plot(fmlMachines.lme)          #Pinheiro & Bates 2000:Fig. 1.7 (the command plot(fmlMachines.lme$residuals) is quite a lot different)
intervals(fmlMachines.lme)
anova(fmlMachines.lme)
plot(fmlMachines.lme,form=resid(.,type="p")~fitted(.)|Worker,abline=0,lty=2)      #Pinheiro & Bates 2000:Fig. 1.8
```

NERC SCIENCE OF THE ENVIRONMENT CEH

1.4. ORTHODONT example (P&B p.30)

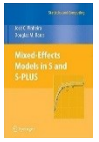


An example about orthodontic measurements in male and female subjects.

```
library(nlme)
View(Orthodont)#Y=distance, XVAR=age, LBL1=Subject (both age and Subject are classification factors)
                                #therefore formula is distance~age|Subject
                                #and we are interested in the average Y from a 'typical' LBL1 (the expected Y), the variation in average Y between LBL1
levels (the between-LBL1 variability) and the variation in the observed Y for a single LBL1 level (the within-LBL1 variability).
                                #Are we (i) interested in these particular Subjects (in which case Subject is fixed) or (ii) trying to make general
comments about all Subjects of this age, i.e. in the population from which these Subjects were drawn, (in which case Subject is random).
plot(Orthodont)  #Pinheiro & Bates 2000:Fig. 1.11
plot.design(Orthodont)          #Pinheiro & Bates 2000:Fig. 1.6

OrthoFem=Orthodont[Orthodont$Sex=="Female",]

fm1Orth.lm=lm(distance~age,data=Orthodont)    #p.135
par(mfrow=c(3,2));plot(fm1Orth.lm)
fm2Orth.lm=lm(distance~Sex*age,data=Orthodont)          #p.136 testing for differences in intercept or in slope between boys and girls
summary(fm2Orth.lm)
fm3Orth.lm=update(fm2Orth.lm,formula=.~.-Sex)
summary(fm3Orth.lm)

fm1OrthF.lis=lmList(distance~age,data=OrthoFem)
coef(fm1OrthF.lis)
intervals(fm1OrthF.lis)
plot(intervals(fm1OrthF.lis))      #Fig. 1.12 and you can use this on an lme command result too (p.156)
```

NERC SCIENCE OF THE ENVIRONMENT CEH

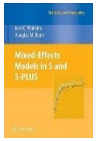
1.4. ORTHODONT example (P&B p.30)



An example about orthodontic measurements in male and female subjects.

```
pairs(fm1OrthF.lis,id=0.01,adj=-0.5)                    #Fig. 4.3
fm2OrthF.lis=update(fm1OrthF.lis,distance~I(age-11))
plot(intervals(fm2OrthF.lis))
fm1OrthF=lme(distance~age,random=~1|Subject,data=OrthoFem) #cf. fm1Orth.lis=lmList(distance~age|Subject,data=Orthodont) on p.139
summary(fm1OrthF)
#fm1OrthFM=lme(distance~age,random=~1|Subject,data=OrthoFem,method="ML")
#summary(fm1OrthFM)
fm2OrthF=lme(distance~age,random=~age|Subject,data=OrthoFem)
summary(fm2OrthF)
anova(fm1OrthF,fm2OrthF)
#ranef(fm1OrthFM)
coef(fm1OrthF)
plot(compareFits(coef(fm1OrthF),coef(fm1OrthFM)))                    #Fig. 1.15
plot(augPred(fm1OrthF),aspect="xy",grid=TRUE)  #Fig. 1.16

fm1Orth.lme=lme(distance~I(age-11),random=~I(age-11)|Subject,data=Orthodont)  #p.146
fm2Orth.lme=lme(distance~Sex*I(age-11),random=~I(age-11)|Subject,data=Orthodont)                #p.148 where they state that you can't use lmList to test for gender differences in the orthodontic growth data because it estimates individual coefficients for each Subject. "In general, we will no be able to use lmList to test for differences due to factors that are invariant with respect to the groups".
plot(comparePred(fm1Orth.lme,fm2Orth.lme,length.out=2),layout=c(8,4),between=list(y=c(0,0.5)))#Pinheiro & Bates 2000:Fig. 4.9
2000:Fig. 4.9
```

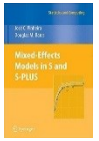
1.5. PIXEL example (P&B p.40)



An example about the pixel intensity in CT scans:

```
library(nlme)
View(Pixel)       #Y=pixel, XVAR=Side, LBL1=Dog (both Side and Dog are classification factors)
                                  #therefore formula is pixel~day|Dog/Side
                                  #and we are interested in the average Y from a 'typical' LBL1 (the expected Y), the variation in average Y between LBL1 levels (the between-LBL1 variability) and the variation in the observed Y for a single LBL1 level (the within-LBL1 variability).
                                  #Are we (i) interested in these particular Dogs (in which case Dog is fixed) or (ii) trying to make general comments about all Dogs of this Dog, i.e. in the population from which these Dogs were drawn, (in which case Dog is random).
plot(Pixel)       #Pinheiro & Bates 2000:Fig. 1.17
plot.design(Pixel)                #Pinheiro & Bates 2000:Fig. 1.6

#What they didn't tell me is that you have to have "I(day^2)" not "day^2" in the formula
fm1Pixel=lme(pixel~day+I(day^2),random=list(Dog=~day,Side=~1),data=Pixel)
intervals(fm1Pixel)
plot(augPred(fm1Pixel))           #Fig. 1.18
VarCorr(fm1Pixel)
summary(fm1Pixel)
fm2Pixel=update(fm1Pixel,random=~day|Dog)
anova(fm1Pixel,fm2Pixel)
fm3Pixel=update(fm1Pixel,random=~1|Dog/Side)
anova(fm1Pixel,fm3Pixel)
fm4Pixel=update(fm1Pixel,pixel~day+I(day^2)+Side)
summary(fm4Pixel)                 #Side isn't significant
```

1.6. OATS example (P&B p.45)



The 'Oats' example is a good one to show mixed-effects modelling. These data come from a 1935 split-plot experiment involving the yields of three varieties of oats ("Victory", "Golden Rain" and "Marvellous") and four concentrations of nitrogen (0-0.6 cwt/acre). The experimental units were arranged into six blocks, each with three plots subdivided into four subplots. The varieties of oats were assigned randomly to the plots and the concentrations of nitrogen to the subplots. All four concentrations of nitrogen were used on each plot. See Dr Scherber https://www.youtube.com/watch?v=VhMWPkTbXoY for a more in-depth video-go-through of this example:

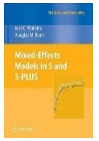
```
library(nlme)
View(Oats)        #Y=yield, XVAR=Variety, LBL1=Block (both Variety and Block are classification factors)
str(Oats)
plot(Oats)        #Pinheiro & Bates 2000:Fig. 1.20
model2=lme(yield~Variety*nitro,random=~1|Block/Variety,data=Oats)          #In this example, nitro is fixed and Variety-nested-within-Block is random
summary(model2)
plot(ranef(model2))          #You should see no trends here (random scatter horizontally) and that's what we do see.
plot(augPred(model2))


plot.design(Oats)          #Pinheiro & Bates 2000:Fig. 1.6
```

Data: Oats

	Block	Variety	nitro	yield
1	I	Victory	0	111
2	I	Victory	0.2	130
3	I	Victory	0.4	157
4	I	Victory	0.6	174
5	I	Golden Rain	0	117
6	I	Golden Rain	0.2	114
7	I	Golden Rain	0.4	161
8	I	Golden Rain	0.6	141
9	I	Marvellous	0	105
10	I	Marvellous	0.2	140
11	I	Marvellous	0.4	118
12	I	Marvellous	0.6	156
13	II	Victory	0	61
14	II	Victory	0.2	91
15	II	Victory	0.4	97
16	II	Victory	0.6	100
17	II	Golden Rain	0	70
18	II	Golden Rain	0.2	108
19	II	Golden Rain	0.4	126

In this example *Variety* and *nitro* are fixed and *Block* is random (although note *Variety* appears after random= as well). We find a significant effect of *nitro* but changing *Variety* appears not to make any systematic difference to crop *yield*.

1.6. OATS example (P&B p.45)



```
#Physically, there are three levels of grouping of the experimental units: block, plot, and subplot. Because the treatments are randomly assigned at each level of grouping, we may be tempted to associate random effects with each level. However, because there is only one yield recorded for each subplot we cannot do this as we would saturate the model with random effects. We use a random intercept at each of the block and the whole block levels",
fm1Oats=lme(yield~ordered(nitro)*Variety,random=~1|Block/Variety,data=Oats) #as explained (sort of) on p.48 the Variety %in% Block after "random=" is really a kind of plot indicator.
intervals(fm1Oats)
plot(augPred(fm1Oats))
VarCorr(fm1Oats)
summary(fm1Oats)                #the ".L", ".Q" and ".C" mean linear, quadratic and cubic terms appear because of the "ordered()" bit.
anova(fm1Oats)

fm2Oats=lme(yield~ordered(nitro)+Variety,random=~1|Block/Variety,data=Oats)
summary(fm2Oats)
anova(fm2Oats)

fm3Oats=lme(yield~ordered(nitro),random=~1|Block/Variety,data=Oats)
summary(fm3Oats)
anova(fm3Oats)

fm4Oats=lme(yield~nitro,random=~1|Block/Variety,data=Oats)
summary(fm4Oats)
VarCorr(fm4Oats)
anova(fm4Oats)
plot(augPred(fm4Oats),aspect=2.5,layout=c(6,3),between=list(x=c(0,0,0.5)))      #Fig. 1.21
```

Data: Oats

	Block	Variety	nitro	yield
1	I	Victory	0	111
2	I	Victory	0.2	130
3	I	Victory	0.4	157
4	I	Victory	0.6	174
5	I	Golden Rain	0	117
6	I	Golden Rain	0.2	114
7	I	Golden Rain	0.4	161
8	I	Golden Rain	0.6	141
9	I	Marvellous	0	105
10	I	Marvellous	0.2	140
11	I	Marvellous	0.4	118
12	I	Marvellous	0.6	156
13	II	Victory	0	61
14	II	Victory	0.2	91
15	II	Victory	0.4	97
16	II	Victory	0.6	100
17	II	Golden Rain	0	70
18	II	Golden Rain	0.2	108
19	II	Golden Rain	0.4	126

2. Deduce the equation example

Data from the Tambopata, Wayqecha and San Pedro forest plots in Peru:

```
NPP=c(11.040580,11.225498,12.615818,11.896294,8.784187,7.657955,8.349710,10.597300,13.016981,13.518252,10.547101,10.709411)
temp=c(23,25,26,23,13,14,13,13,27,27,28,29)
precipitation=c(110,100,120,115,110,89,96,140,111,120,88,77)
site=factor(c("SP","SP","SP","SP","Way","Way","Way","Way","Tam","Tam","Tam","Tam"),levels=c("Tam","SP","Way"))
elev=c(1500,1500,1500,1500,3025,3025,3025,3025,210,210,210,210)
dataf=data.frame(temp,precipitation,site,elev);head(edit(dataf))
fmXglm=glm(NPP~temp+precipitation+site)
summary(fmXglm)
```

CHALLENGE: I created those NPP numbers (they are not real) using an equation that involved the other variables somehow. What equation did I use?

ANSWER: Get the GLM table which says that the coefficient of precipitation is 0.068 and the coefficient of temp is 0.152. Only precipitation is significant, therefore the GLM analysis suggests that `NPP=0.068*precipitation` is the answer. In actual fact, I used:

```
NPP=0.25*temp+0.05*precipitation+runif(12,min=-1,max=+1)
```

(i.e. the random noise I added in (the runif) was enough to mask out the dependency on temperature I'd put in). Look what happens if I remove the noise:

```
NPP=0.25*temp+0.05*precipitation
fmXglm=glm(NPP~temp+precipitation+site)
summary(fmXglm)
```

3. Peppercorns

Here are some data about peppercorns (*Piper nigrum* fruits) that were collected from farms growing their pepper plants either on support trees, on poles or on trellises (3 types of cultivation), and collected 20-29 weeks after flowers were recorded as mature on the plant. The fruit bunches recorded were of varying lengths around 5.4 cm. What you want to know is what controls the length of the fruit bunches?

Read the data into R, then do some diagnosis plots and run a GLM analysis. I suggest you start with something like

```
setwd("D:/!.ALLFILES/NOTES/lecturing/2015.GLMpresentation/")        #You'll need to change this line to an equivalent directory on your machine.
dataf=read.table("Peppercorns.txt",header=TRUE, sep="\t",na.strings="MISSING")
```

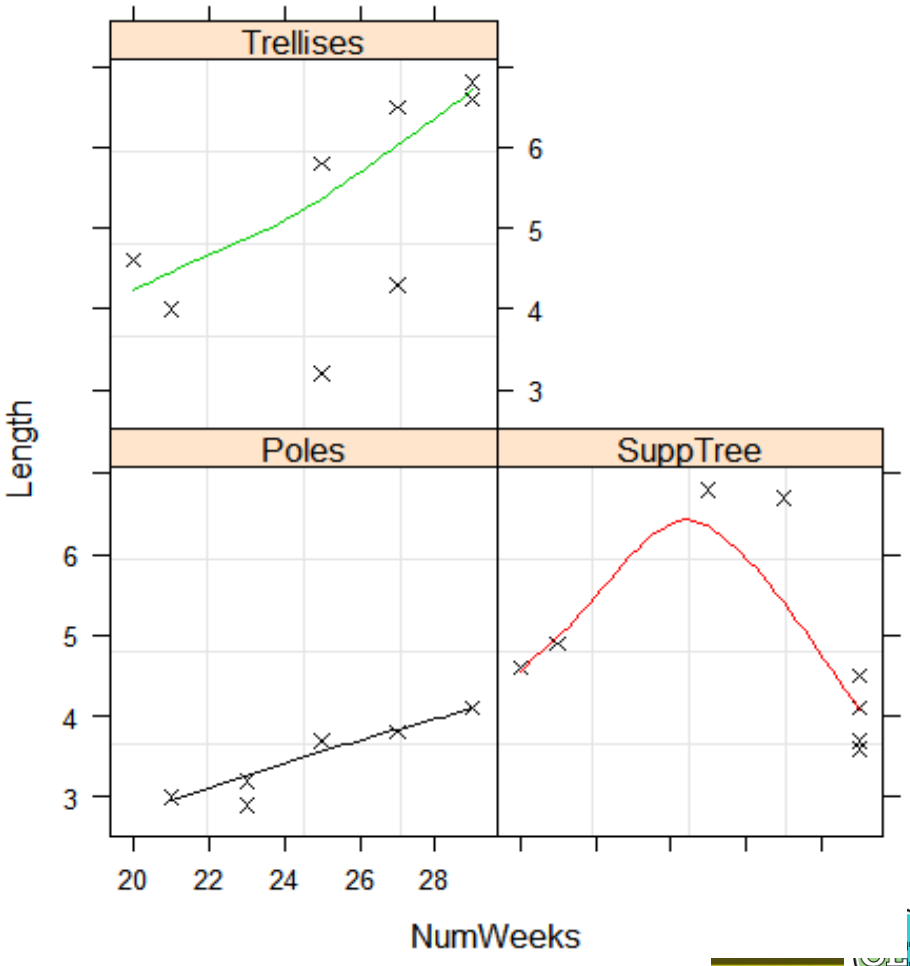
3. Peppercorns

i) Can you replicate the panel plot right? (the lines are loess lines, by the way, which I put on using panel.loess()). It looks like the fruit bunch length depends on both how long it is left to ripen and the cultivation type.

ii) Run a GLM (*with* interaction terms) with the fruit bunch length as the response variable (use the glm() command, but don't recode *CultivType* like I did *SoilLoam* above - see step 1c of Crawley's two stage method above). What is the model which best describes these data? Here's what I get:

Best_guess_of_Length.of.fruit.bunch = 3.3571, 4.8625 or 5.4000 for cultivation using poles, support trees and trellises, respectively, with no dependency on the number of weeks (despite the apparent trends in the diagnostic plot).

iii) Repeat step ii using glmmPQL(): check that you get almost exactly the same model fit.



4. Online tutorials not written by us (!)

There are many, many examples of GLMs online and it is really helpful to use them EVEN THOUGH they are of widely varying quality and also vary a lot in how well they explain the steps involved (even if those steps are correct). Here are my picks of very good examples to go through: they both take quite different approaches from what I've been advising, but going through them keeping your eye out for differences is really a good thing to do because they only differ on points that are either purely stylistic or legitimately under discussion by developers of these GLM methods. Have a go:

1) Tree abundances: http://plantecology.syr.edu/fridley/bio793/glm.html

2) Wasp aggression: http://ase.tufts.edu/gsc/gradresources/guidetomixedmodelsinr/mixed%20model%20guide.html

3) Dragons: https://ourcodingclub.github.io/2017/03/15/mixed-models.html

There are no 'solutions' for these: if you try them and find something odd, just come and find me to discuss.

ANSWERS .... Peppercorns

Data: dataf

	NumWeeks	CultivType	Length
1	25	Poles	3.7
2	21	Trellises	4
3	25	Trellises	3.2
4	21	Poles	3
5	29	SuppTree	4.1
6	27	Poles	3.8
7	25	Trellises	5.8
8	29	SuppTree	4.5
9	29	Poles	4.1
10	20	SuppTree	4.6
11	23	Trellises	NA
12	27	Trellises	4.3
13	29	Trellises	6.6
14	23	Poles	2.9
15	25	SuppTree	6.8
16	NA	Poles	2.8
17	29	Trellises	6.8
18	29	Trellises	6.8
19	20	Trellises	4.6

Here are some data about peppercorns (*Piper nigrum* fruits) that were collected from farms growing their pepper plants either on support trees, on poles or on trellises (3 types of cultivation), and collected 20-29 weeks after flowers were recorded as mature on the plant. The fruit bunches recorded were of varying lengths around 5.4 cm.

Here I read the data into R:

```
setwd("D:/!.ALLFILES/NOTES/lecturing/2015.GLMpresentation/")        #You'll need to change this line to an equivalent directory on your machine.
dataf=read.table("Peppercorns.txt",header=TRUE, sep="\t",na.strings="MISSING")
head(dataf)     #Use head() to see the first 6 lines of the data frame (for quick checks)
View(dataf)     #Use View() to see the whole data frame in a new window
attach(dataf)
```

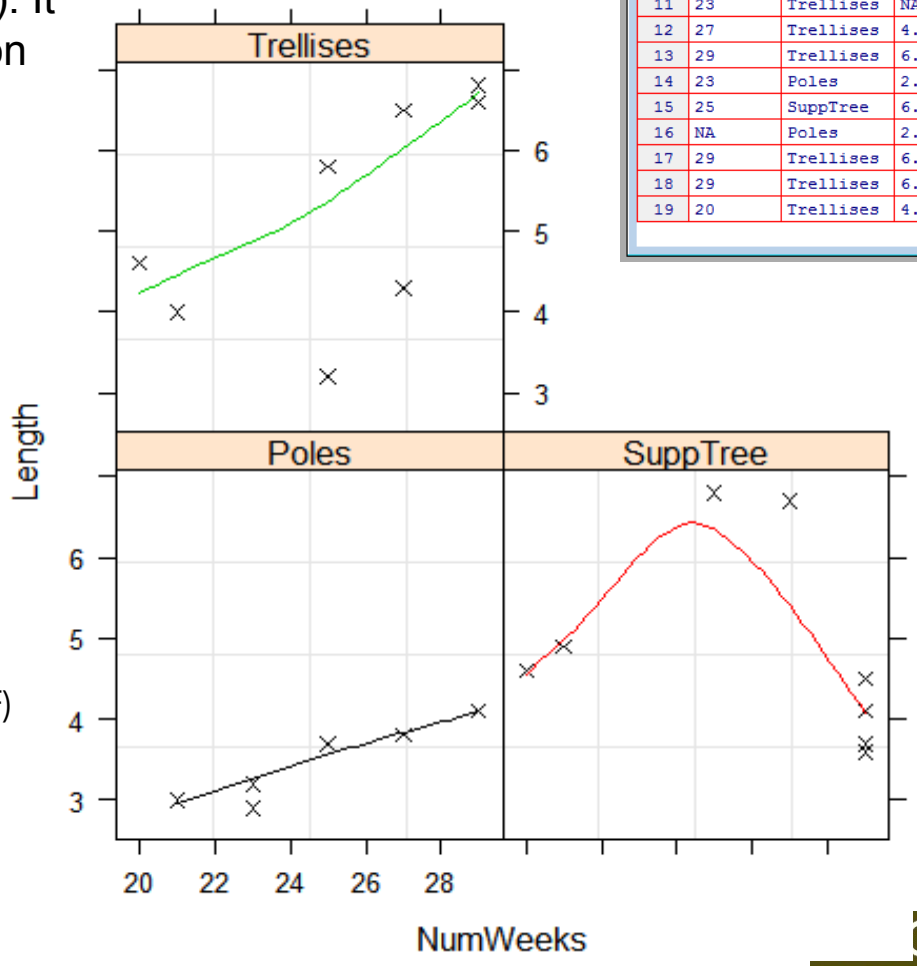
Peppercorns

Data: dataf

	NumWeeks	CultivType	Length
1	25	Poles	3.7
2	21	Trellises	4
3	25	Trellises	3.2
4	21	Poles	3
5	29	SuppTree	4.1
6	27	Poles	3.8
7	25	Trellises	5.8
8	29	SuppTree	4.5
9	29	Poles	4.1
10	20	SuppTree	4.6
11	23	Trellises	NA
12	27	Trellises	4.3
13	29	Trellises	6.6
14	23	Poles	2.9
15	25	SuppTree	6.8
16	NA	Poles	2.8
17	29	Trellises	6.8
18	29	Trellises	6.8
19	20	Trellises	4.6

Here's a panel plot (the lines are loess lines, by the way, which I put on using panel.loess()). It looks like the fruit bunch length depends on *both* how long it is left to ripen and the cultivation type.

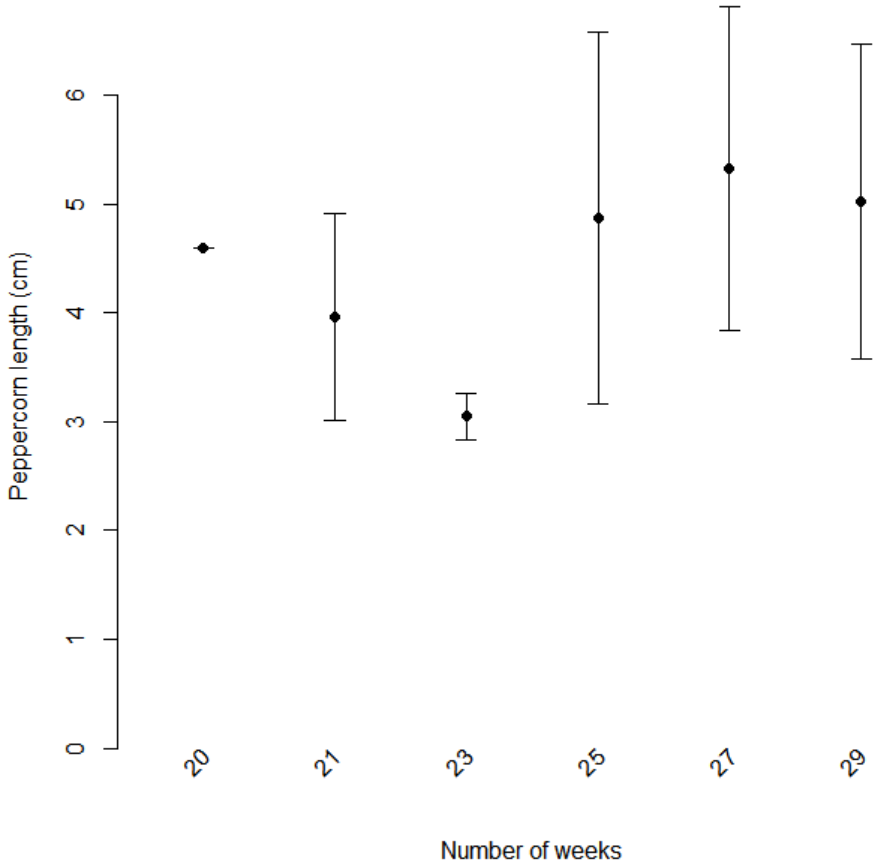
```
library(lattice);library(nlme)

pnll=function(x,y,...) {
 panel.grid(h=3,v=3)
 panel.xyplot(x,y,pch=4,col="black")
 panel.loess(x,y,span=1,lwd=1,col=panel.number())
}

grpdata=groupedData(Length~NumWeeks|CultivType,data=dataf)
plot(grpdata,layout=c(2,2),panel=pnll,asp=1)
```



Peppercorns

However for now we're going to IGNORE *CultivType* and just look at the relationship between *Length* and *NumWeeks*.

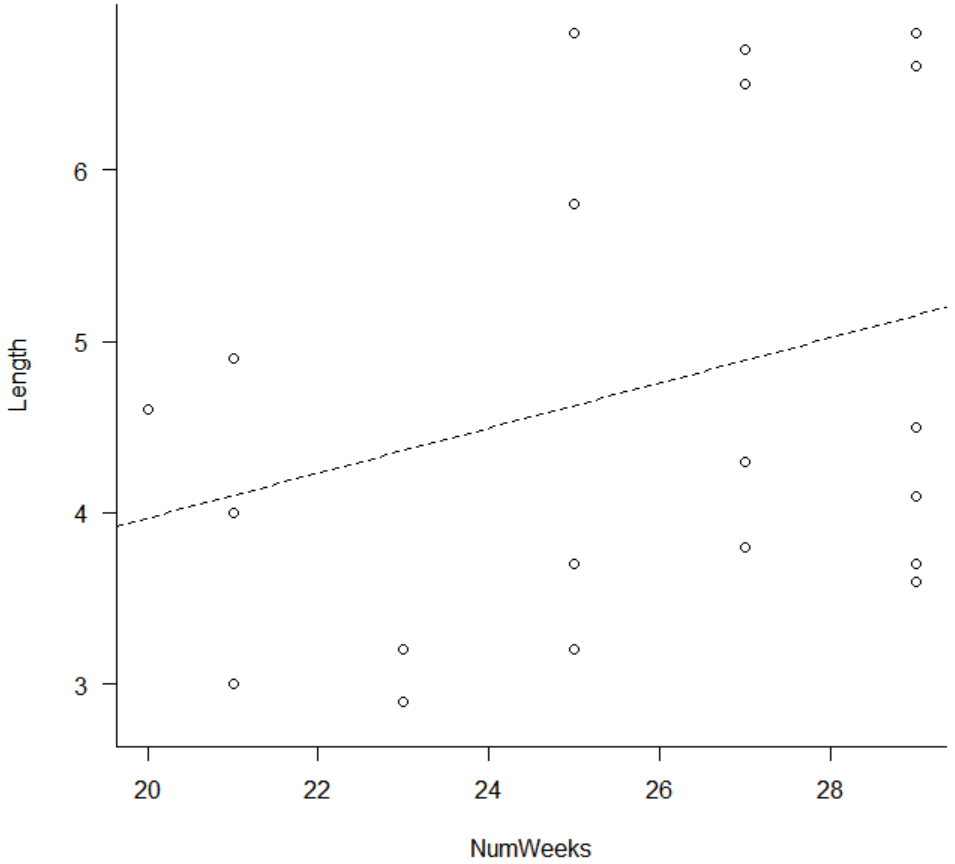
```
cat("Means by
CultivType:\n");with(dataf,tapply(Length,INDEX=CultivType,FUN=mea
n,na.rm=TRUE))

#Plot of Length against NumWeeks
barwidth=0.09;mr=tapply(Length,INDEX=NumWeeks,FUN=mean,na.r
m=TRUE);sr=tapply(Length,INDEX=NumWeeks,FUN=sd,na.rm=TRUE)
dev.new();xmidpts=barplot(mr,beside=TRUE,axes=FALSE,axisnames=
FALSE,col="white",border=NA,xlab="Number of
weeks",ylab="Peppercorn length (cm)",ylim=c(0,max(mr+sr)))
axis(2);text(x=xmidpts,y=par("usr")[3]-
0.03,labels=names(mr),srt=45,adj=1,xpd=TRUE)
uppers=mr+sr;lowers=mr-sr
segments(xmidpts,lowers,xmidpts,uppers)
segments(xmidpts-
barwidth,uppers,xmidpts+barwidth,uppers);segments(xmidpts-
barwidth,lowers,xmidpts+barwidth,lowers)
points(x=xmidpts,y=mr,pch=16)
```



Peppercorns

However for now we're going to IGNORE *CultivType* and just look at the relationship between *Length* and *NumWeeks*.

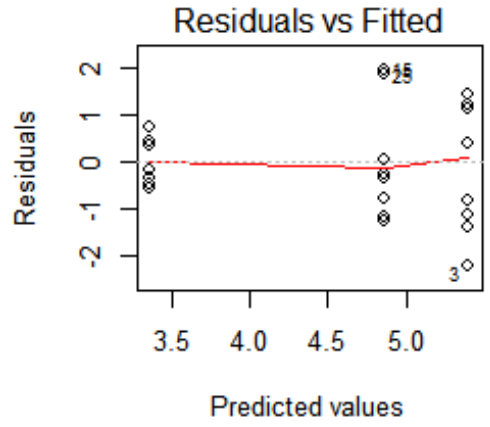
```
lmlres=lm(Length~NumWeeks,data=dataf)
summary(lmlres)
plot(x=NumWeeks,y=Length,bty="l",las=1)
abline(lmlres,lty=2)
```

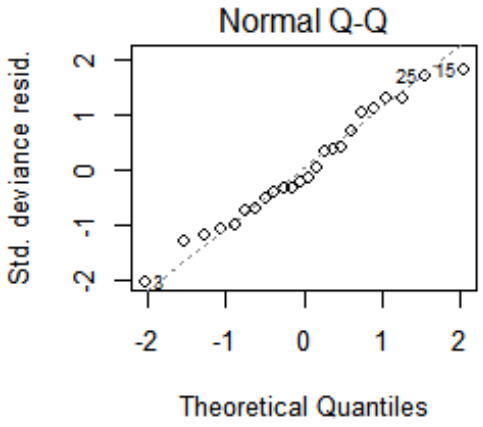


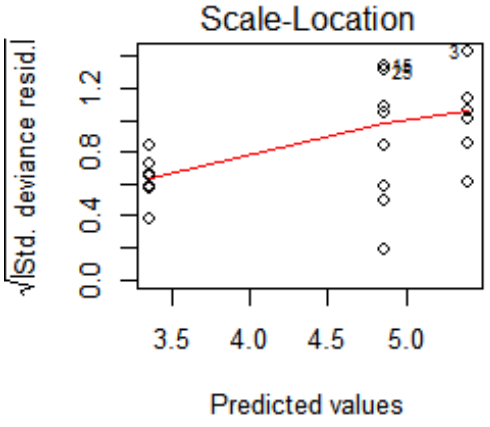
Peppercorns

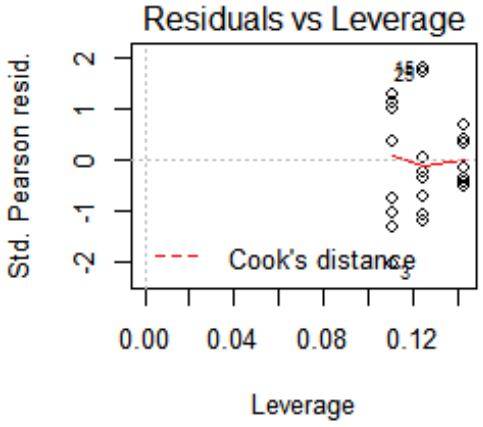

```
library(lattice);library(nlme)

pnll=function(x,y,...) {
 panel.grid(h=3,v=3)
 panel.xyplot(x,y,pch=4,col="black")
 panel.loess(x,y,span=1,lwd=1,col=panel.number())
}

grpdata=groupedData(Length~NumWeeks|CultivType,data=dataf) #There are many ways to do this plot: here's just one.
plot(grpdata,layout=c(2,2),panel=pnll,asp=1)

cat("Means by CultivType:\n");with(dataf,tapply(Length,INDEX=CultivType,FUN=mean,na.rm=TRUE))

fmlres=glm(Length~NumWeeks*CultivType,data=dataf)          #n.b. Gaussian errors here.
summary(fmlres)
fmlres=glm(Length~NumWeeks+CultivType,data=dataf)
fmlres=glm(Length~0+NumWeeks+CultivType,data=dataf)        #I said not to recode CultivType so NumWeeks is the next to be removed
fmlres=glm(Length~0+CultivType,data=dataf)
library(MASS)
stepAIC(fmlres) #(this result confirms that Length~0+CultivType is the best-fit model).
dev.new();par(mfrow=c(2,2));plot(fmlres)          #These look fine.
```

Peppercorns









Peppercorns


```
n=length(dataf$Length)
p=n-fmlres$df.residual
cat("Data set had",n,"degrees of freedom initially, and this fit has",p,", leaving",n-p,"residual.\n")
rsquared=(fmlres$null.deviance-fmlres$deviance)/fmlres$null.deviance
cat("Some people use an r2 quantity (here, r2=",rsquared,") but goodness-of-fit\nfor GLMs should actually be assessed by a Chi-squared test on the residual\ndeviance, NOT by looking at r2.\n")
cat("Residual deviance (G^2) is",round(fmlres$deviance,4),"and the p-value of the fit is",round(1-pchisq(fmlres$deviance,fmlres$df.residual),4),"\n")
if ((1-pchisq(fmlres$deviance,df=fmlres$df.residual))>0.05) {cat("This is a good fit\n")} else {cat("This is a poor fit\n")}
anova(fmlres,test="Chisq")

library(MASS)
dummy=1:nrow(dataf)
fmlres=glmmPQL(fixed=Length~NumWeeks*CultivType,random=~1|dummy,family=gaussian,data=dataf)
fmlres=glmmPQL(fixed=Length~NumWeeks+CultivType,random=~1|dummy,family=gaussian,data=dataf)
fmlres=glmmPQL(fixed=Length~CultivType,random=~1|dummy,family=gaussian,data=dataf)
summary(fmlres)
```