Generalised Linear Modelling and Mixed Effects Exercise- Swedish hunting data

NES8010 Data Analaysis and Modelling

February 2021

# Statistical Background

The generalised linear model can be used to investigate the relationship between dependent and independent predictor variables. In this case we wish to investigate how numbers of animals shot change through time. Time here is less of a predictor variable, but rather an indicator of the passage of time. A more appropriate model might be a time series analysis where we quantify the patterns of variation in time. This sort of analysis can be undertaken with Generalised Least Squares (GLS) which makes fewer assumptions about the underlying data. In addition GLS allows us to investigate serial dependence. Serial dependence occurs when the dependent variable is influenced ie is dependent on previous values. In this instance it is clear that how many animals are shot at one time point will be dependent on how many were left from the time before.

A key reference is Pinheiro and Bates: Mixed Effects Modelling in S

## Analysis of Hunting Data from Sweden 1819-1919

Hunting bag data are often collected as they can be used to investigate population trends in quarry.

The data set came from a survey of parishes in Sweden over the period 1829 to 1919. They were collated by Elmhagen in 2007. The data were summarised at County level at 5 year intervals. The records are unique in that they reflect Swedish society over the 100 year period ~ absolute census of socio-economics (number of cattle, amount of seed sown) etc were collected by parish priests.These data form a valuable data set with which to explore trends in human populations, agriculture and charismatic biodiversity (wolf, lynx and red fox) over the period.

## Objective

The aim is to investigate trends in bags of top predators and mesopredators through time. You will use appropriate statistical methodology to investigate these trends.

## Load Libraries for analysis

wants <- c("ggfortify", "here", "nlme", "lattice","ggplot2", "MASS")  
has <- wants %in% rownames(installed.packages())  
if(any(!has)) install.packages(wants[!has])  
  
  
library(nlme)  
library(lattice)  
library(MASS)

## Load the Data

The data are called ‘sweden\_simple.csv’. First, read in the data. Ensure you are either using RProj or have the correct filepath to where your data are stored using the Session> set working >directory> choose working directory from the Rstudio menus.

The code in these notes uses the here package and the data are in the ‘Data’ directory.

library(here)

## here() starts at /Users/aileenmill/OneDrive - Newcastle University/Documents/MSc/NES8010 Quantitative Skills/NES8010 Quantitative Skills

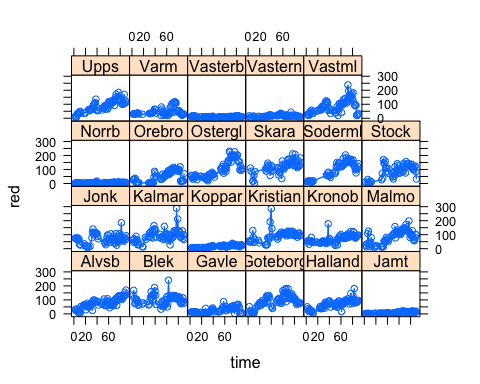
rawd<-read.csv(here("Data", "sweden\_simple.csv"))

Note these data describe the number of red fox, lynx and wolf shot over a 90 year period in each county of Sweden (1829-1919)

## Data Exploration

Plot the number of red fox shot by county, either using lattice or facets in ggplot

xyplot(red~time|name, rawd, typ=c("a","b"))



Plot the number of wolf shot by county. It is conventional to model these as log…

## Analysing the Fox Bag

Are there any trends with time and county?

fox\_trends.gls<-gls(log(red)~time+name, rawd)

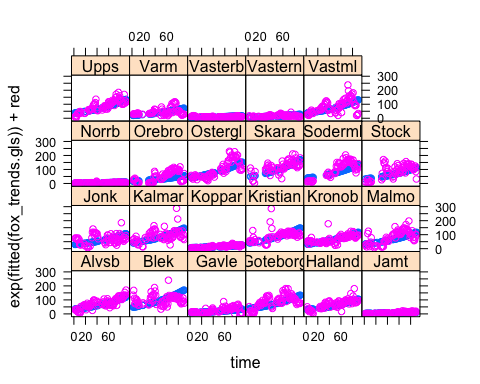
now get the output

summary(fox\_trends.gls)

Are any predictors significant?

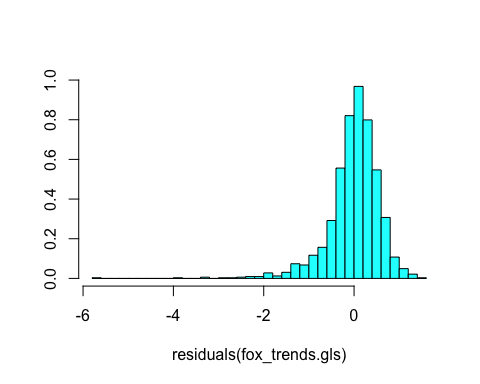
Look at the fitted model by county

xyplot(exp(fitted(fox\_trends.gls))+red~time|name, rawd)



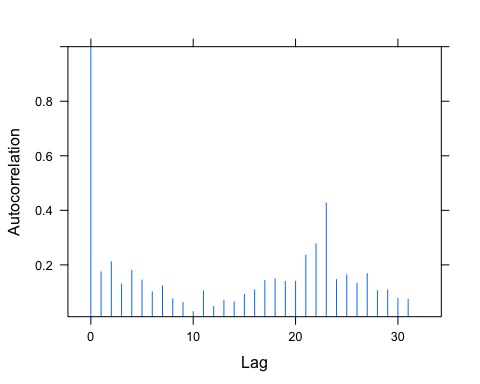
What about the residuals for the model?

truehist(residuals(fox\_trends.gls))



These look reasonable BUT we have not allowed for autocorrelation in the model. This is basically the dependence of the response on previous years. We can investigate the autocorrelation with the acf function:

plot(ACF(fox\_trends.gls))



You can see that there is considerable autocorrelation. This means that the estimates of significance will be biased upwards. We adjust using the correlation statement. We have to define what sort of correlation exists. Here we will use the simple autoregressive which assumes that part of the observed pattern is driven by the last measurement (here the previous bag). We will also assume that this varies by county.

fox\_trends\_auto.gls<-gls(log(red)~time+name,correlation=corAR1(form=~1|name), rawd)

summary(fox\_trends\_auto.gls)

Note that *phi* gives the estimate of the serial dependence between measures.

## Comparing Models

But is this a better model? We use AIC to compare models ~ the model with the lowest is better. The likelihood ratio test gives us how much better!

anova(fox\_trends.gls,fox\_trends\_auto.gls)

## Model df AIC BIC logLik Test L.Ratio  
## fox\_trends.gls 1 25 2955.791 3090.281 -1452.8953   
## fox\_trends\_auto.gls 2 26 1989.432 2129.303 -968.7161 1 vs 2 968.3583  
## p-value  
## fox\_trends.gls   
## fox\_trends\_auto.gls <.0001

Note that *phi* gives the estimate of the serial dependence between measures.

## Mixed Effect Modelling

If we look at the coefficients and predicted curves for each county we can see that there are vast differences between counties. In some cases there is a rapid increase in fox bags and in others the increase is slower. We have unmeasured phenomena in counties that will be influencing what is bagged. When we have unmeasured phenomena on one of the sample units we can partial out this unmeasured error as an effect specific to the unit. Here we can say that there is unmeasured variation do to the county. You can hypothesise about what that might be ~ better guns, better habitat etc.

We use the lme function to estimate the error due to county. This error is called a **random effect**.

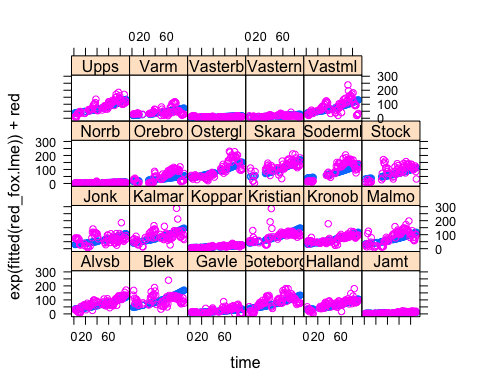
red\_fox.lme<-lme(log(red)~time, random=~1|name, rawd)

summary(red\_fox.lme)

The coefficients are interpreted as before. BUT what about the mixed effect or error due to county? To assess this we look at the *Standard Deviation* for the random effect and compare it to that of the remaining residuals. In this model the STdev for county is 0.898 whereas the residual is 0.600. This indicates that we can account for 0.898/(0.898+0.600) ie 59% of the error/ residuals because we have accounted for random effects in the county measurements of bag.

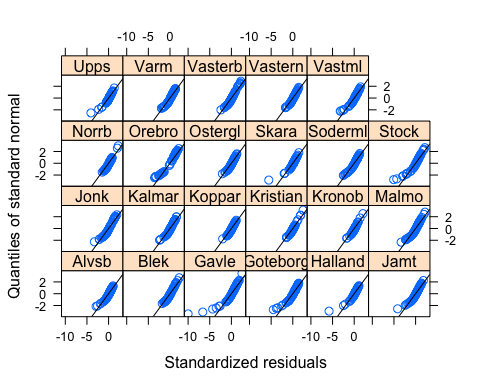
The model is very good! Look at predicted versus observed:

xyplot(exp(fitted(red\_fox.lme))+red~time|name, rawd)



Investigate the residuals. You will have done this before in simple regression analyses. Here the qqnorm function is used to abstract the residuals by county and plot a straight line representing that on which the points should fit.

qqnorm(red\_fox.lme, ~resid(.,type="p")|name, abline=c(0,1))



BUT we have forgotten the autocorrelation which the GLS identified this can be added to the model by specifiying and autorcorrelation function - there are many options, corAR1 is the simplist.

red\_fox\_auto.lme<-lme(log(red)~time, random=~1|name,correlation=corAR1(form=~1|name), rawd)

Now test to see if the model is better for inclusion of autocorrelation

anova(red\_fox.lme,red\_fox\_auto.lme)

## Model df AIC BIC logLik Test L.Ratio p-value  
## red\_fox.lme 1 4 2975.367 2996.940 -1483.6836   
## red\_fox\_auto.lme 2 5 2008.974 2035.941 -999.4872 1 vs 2 968.3927 <.0001

## Repeat the analysis for the wolf and the lynx bags

How do these results compare with those for the red fox? Note that you will have to add +1 to logged data for wolves/lynx as there are many zeroes. You may choose to merge lynx and wolf into one variable top predators.