Lab 2: Generalised Additive (Mixed) Models (for Location, Scale and Shape)

Please log into the computer using the credentials given to you. Please note that each account is specific to one computer. A link to R *Studio can be found in the 'Maths and Stats' folder on the desktop.*

In this lab you will need a data object which you can load using the command:

```
load(url(''http://www.stats.gla.ac.uk/~claire/aptslab2.RData''))
```

This handout can be downloaded from: `http://www.stats.gla.ac.uk/~claire/aptslab2.pdf`

Data and R objects

Use `ls()` in R to explore the objects available in this `RData` file.

The available datasets are:

SO2 in the Czech Republic

This data set (called `CZ03`) contains measurements of SO_2 (on a log scale) from a monitoring station in the Czech Republic. There are also 7 additional columns, the year of the measurement (on a decimal scale), the week the measurement was taken, rainfall, temperature and humidity - all recorded on the same date.

The Dutch Boys BMI data

This is the data (called `dbbmi`) and contains two variables: age and BMI for a sample of 1000 data points from the original study. This was a cross-sectional study that measures growth and development of the Dutch population between the ages 0 and 21 years. The study measured, among other variables, height, weight, head circumference and age for 7482 males and 7018 females.

Swiss lakes data

This data set (called `swisslakes`) contains depth, water temperature and air temperature recorded for lakes in Switzerland at a particular point in time.

Tasks:

1. **SO_2 in the Czech Republic - GAMs**

   (a) Produce plots of SO_2 against each of the potential other covariates: year, week , rainfall, temperature and humidity.

   (b) Use `library(mgcv)` to fit an additive model which relates SO_2 to the covariates Year and Week.

   e.g. `model1 <- gam(y~s(x1)+s(x2))`

   Plots of the fitted smooth components with partial residuals can be produced using:

   `plot(model1, residuals=T)`

   There is a clear downward trend over the years and, as expected a strong seasonal effect.

   (c) There is interest in whether having information on local meteorology will be important in estimating the effects of Year and Week. Add the variables: Rain, Temp, and Humidity as further flexible terms in the additive model, plot the smooth fitted components and interpret what you see.

   (d) The command `anova(model1)` can be used to assess the statistical significance of fitted effects in a model, and the two models can be compared formally using `anova(model1, model2, test='F')`. Explore the results here.

   (e) Within this `gam` function, penalised regression splines are being used with the smoothing parameter (λ) automatically selected using generalised cross validation (GCV). This can be altered to use, for example, REML (to select λ) using: `model1 <- gam(y~s(x1)+s(x2), method='REML')`.

2. **The Dutch Boys BMI data - GAMLSS**

   (a) Plot the data to investigate the relationship between BMI and age.

   (b) Use `library(mgcv)` to fit a model for mean BMI as a function of age and plot the fit.

   (c) Use `library(quantreg)` to fit an additive quantile regression model for BMI as a function of age for $\tau = 0.05, 0.25, 0.5, 0.75, 0.95$. This can be done using the `rqss()` function. Plot the fitted conditional quantile functions.

   (d) Use the package `gamlss` to fit a model that assumes the Box-Cox Cole and Green distribution (BCCG) mentioned in Chapter 5 of the notes. The commands below fit penalised B-spline terms for each of the parameters in the model.

```
m1 <- gamlss(bmi~pb(age), sigma.formula=~pb(age),
             nu.formula=~pb(age), family=BCCGo, data=dbbmi)
```

   (e) Use the `fittedPlot()` command on the model to plot the parameters μ, σ, and ν.

   (f) The code below can be used to plot the estimated centile curves from the GAMLSS. Compare these with the additive quantile regression fit from above.

```
centiles(m1,dbbmi$age,cent=c(5,25,50,75,95), ylab='BMI', xlab='Age',
         col.centiles = c(2,6,1,6,2), lty.centiles = c(2,3,1,3,2),
         lwd.centiles =c(2,2,2.5,2,2))
```

3. **Swiss Lakes Data - GAMMs**

For these data, we are initially interested in the model:

$$\text{WaterTemp}_i = \beta_0 + f_1(\text{AirTemp}_i) + \gamma_{ji} + \epsilon_i, \quad i = 1, \ldots, n,\ j = 1, \ldots, 5(\text{depths})$$

and the errors are normally distributed with mean 0 and constant variance.

(a) Use boxplots to examine the relationship between water temperature and depth (`boxplot(y~x)`) and plot the relationship between water temperature and air temperature.

(b) Fit an additive model, (using `library(mgcv)`) to explore the relationship between water temperature, as the response, air temperature as a smooth covariate and depth as a factor. To do this use the following code appropriately adjusted for y, x_1 and x_2:

```
library(mgcv)
m1 <- gam(y~s(x1)+factor(x2), data=swisslakes)
```

(c) Provide a summary of the model output and plot the fitted smooth function of air temperature using the `summary()` and `plot()` commands.

(d) Use residual plots to assess the assumptions for this model e.g.

```
gam.check(m1)
```

(e) From this model, predict the water temperature when the air temperature is 15^{o}C and the Depth is 7.5m using:

```
predict(m1, data.frame(AirTemp=15, Depth=7.5))
```

(f) The default in `gam` uses thin plate regression splines for the smoothing - alternative spline bases can be used by adding the term `bs` to the smooth e.g. for p-splines:

```
m2 <- gam(y~s(x1, bs='ps')+factor(x2), data=swisslakes)
```

(g) Finally, Depth could be investigated as a random effect in the model instead of as a fixed effect. The different depths could be thought of as a random sample of all possible depths. An example of this could be to include a random intercept term for depth in the model:

```
m3 <- gamm(WaterTemp~s(AirTemp, bs='ps'), random=list(Depth=~1))
```

If you would like to check your answers or would prefer to work through a preprepared script, then this is available from:

`http://www.stats.gla.ac.uk/~claire/aptslab2.R`