Workshop 9
Non-Linear Regression Models

Local regression

In Moodle there are two script files `loess1.r` and `loess2.r` containing template code. The first uses local regression to estimate the linear loess value at a specific point x_0. The second does the same over a given grid of x-values

Last week you used non-linear smoothing on the motorcycle helmet acceleration data `mcycle` in the MASS library. As a reminder, recreate the spline smoothing estimate using

```
library(MASS)
library(splines)
plot(mcycle)
fit1=smooth.spline(mcycle$times,mcycle$accel,df=10)
x.grid<-0:56
preds=predict(fit1,x.grid)
lines(x.grid,preds$y,lwd=2,col="black")
```

Use the code in `loess1.r` to obtain the local regression line at the point $x_0 = 28$ for the motorcycle data. You will need to complete some of the syntax. Once the code is working, try with span=0.75 and again with $x_0 = 14$

The code in `loess2.r` evaluates the whole loess curve over a given x-grid to get a complete loess curve.

Use the *R*-command `loess()` to replicate your algorithm in in one step. Because you have fitted a *linear* local regression, you need to specify the argument `degree=1` in the `loess` function call.

Generalised additive Models

The `Work` data

Work through Lab 7.8.3 in James et al. starting on page 294, up to the command `plot(gam.lo.i)` on page 296.

The first example uses `lm()` to fit the model. Check that the function `gam()` with the same arguments outputs the same coefficients. Hints: if you have not already done so you need to start the `gam` package, and to obtain the coefficients of a statistical model use the function `coef()`

The function `plot.gam` should be `plot.Gam`

The `College` data

The `College` dataset is available in the ISLR package, the accompanying package to James et al.
An introduction to the data set is given on page 55.

We will fit the variable `Accept`, the number of applications accepted, as the outcome variable. Obtain the mean and median and a histogram of `Accept`. What do you notice about this distribution.

Consider the following variables as potential predictor variables: `Private`, `Apps`, `F.Undergrad`, `Room.Board`, `Expend`, `PhD`, and `S.F.Ratio`.

Which variables are factor variables and which variables are numeric?

Fit a GAM to these data. Use spline smoothing with 5 degrees of freedom for the numeric predictors. Use the method used in the last section to determine which variables are significant, and for which variables a linear effect is sufficient. Obtain the response plot for each of the fitted variables including the partial residuals (`resit=TRUE`)

As both `Accept` and `Apps` have very skewed distributions repeat the GAM fitting with the logarithm of both of these variables. Are the modelling decisions the same?

Use you preferred model to obtain a prediction for `Accept` at *Harvard University* and compare this with the observed value.