Water Quality in the River Clyde

A case study of statistical analysis with environmental data

Notes for teachers

Adrian Bowman

The University of Glasgow

June 2012

Background

The Scottish Environment Protection Agency has a statutory obligation to monitor the Scottish environment. This includes the River Clyde, where measurements of water quality, expressed in dissolved oxygen on a percentage scale, have been made over a long period. Measurements are available from sampling stations at two mile intervals down the river, from 0 to 26 miles from the city centre. This teaching material is based on data collected over an eight year period around 1985, when a sewage treatment plant at approximately 2 miles down the river was upgraded. The aim is to assess whether there is evidence that the water quality in the river has improved as a result of this upgrade.

The material can be accessed through the rpanel package for the widely used statistical computing environment R. Both rpanel and R are freely available from www.r-project.org.

Please note that the material is still under development. The software will, in due course, be made available inside the rpanel package.

Structure of the material

The R instructions

```
library(rpanel)
source("clyde.r")
```

launch the material described in this document. The file clyde.rda should be available in the directory from which the instructions above are given.

The starting point is a comparison of the water quality data before and after 1985, expressed in two boxplots. These show very little difference between the two groups of data.

In order to proceed, it is necessary to consider other variables which may affect the water quality in the river. This provides an opportunity for an interesting discussion. Before long, it is likely that the relevance of the time of year will emerge. This can be viewed in a scatterplot through the Day plot button.

The resulting plot shows a very strong effect of the time of year, with lower values in the summer and high values in the winter. This is due to greater rainfall and increased runoff from agricultural land in the winter months. The challenge is now to construct a model for this non-linear

relationship. At this stage, students often think if a quadratic curve as a suitable description. Tis works reasonably well, but a better model, which respects the cyclical nature of the day covariate, is to use a cosine function. This needs to be fitted to the observed data and the Day model button launches a set of sliders which allow intercept, amplitude and phase parameters to be added interactively.

It is possible at this stage to discuss the concept of least squares as a means of fitting the model. The parameter values which minimise the sum of the squared distances of the observed data points from the model provide natural estimates of the unknown, true parameter values. However, by using the trigonometric formula for cos(A-B), the model formula can be expanded and, surprisingly, shown to be linear in the covariates cos(2 pi Day/365) and sin(2 pi Day/365). Those who are familiar with regression involving more than one covariate will appreciate that this gives a very simple route to estimating the parameters.

A further button allows the cosine model to be fitted to the two separate groups of data, before and after 1985. At 2 miles down the river, there is little difference between the two fitted models, which is disappointing after all this effort! Some further discussion can be encouraged here, leading to the realisation that for any material discharged from the sewage treatment plant it will take some time for the associated biochemical processes to take place. By this time the water will here moved some distance down the river. Further buttons allow the model to be refitted to the data from other sampling stations. These show increasing separation of the two groups, with a higher mean water quality after 1985, with a maximum around 8 miles down the river. After this, the level of dissolved oxygen increases as the composition of the sample is increasingly sea water, as the river widens out into the Clyde estuary.

A final button allows a model for the whole river to be displayed, using a three-dimensional plot and a fitted model based on more sophisticated and flexible regression models. For those students who have more experience in statistical methods, this model uses local linear regression.

One of the aims of the teaching material is to illustrate that statistical modelling requires intelligent use of the context of the problem, as well as technical expertise. An effect of the sewage treatment plant upgrade has been identified, but only after some careful thinking!

Types of use

The material is suitable for use in a lecture or classroom setting, where a lecturer or teacher can lead students through a discussion of the material, with the software projected. However, it can also be used for self-study by students who have sufficient curiosity and confidence to explore the problem on their own.