APTS course 1st-5th September 2014

Nonparametric Smoothing

Preliminary Material

Adrian Bowman & Ludger Evers

University of Glasgow | School of Mathematics & Statistics

This APTS course will cover a variety of methods which enable data to be modelled in a flexible manner. As preparation, it would be helpful to revise the following topics covered in earlier APTS courses:

- linear models, including the Bayesian version;
- generalised linear models;
- R programming;
- Taylor series expansions and standard asymptotic methods.

The main emphasis will be on regression settings, because of the widespread use and application of this kind of data structure. However, the preliminary material covers various aspects of *density estimation*, to introduce some of the main ideas of nonparametric smoothing and to highlight some of the main issues involved. It is likely that many people will have come across these ideas in one form or another. The preliminary material aims to

- explore simple kernel methods;
- show how these can be used to construct smooth density estimates;
- investigate some simple theoretical properties;
- experiment with software available in R;
- consider some illustrations of their use in analysing data.

Exercises are provided to assist in engaging with the material.

1 The idea of density estimation

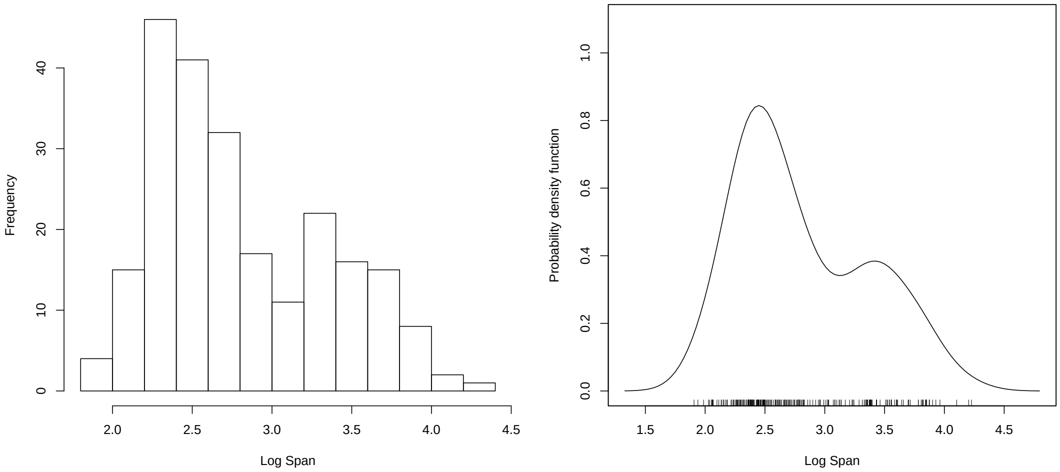
A probability density function is a key concept through which variability can be expressed precisely. In statistical modelling its role is often to capture variation sufficiently well, within a model where the main interest lies in structural terms such as regression coefficients. However, there are some situations where the shape of the density function itself is the focus of attention. The example below illustrates this.

Example: Aircraft data

These data record six characteristics of aircraft designs which appeared during the twentieth century. The variables are:

Yr: year of first manufacture
Period: a code to indicate one of three broad time periods
Power: total engine power (kW)
Span: wing span (m)
Length: length (m)
Weight: maximum take-off weight (kg)
Speed: maximum speed (km/h)
Range: range (km)

A brief look at the data suggests that the six measurements on each aircraft should be expressed on the log scale to reduce skewness. Span is displayed on a log scale below, for Period 3 which corresponds to the years after the second World War.



The pattern of variability shown in both the histogram and the density estimate exhibits some skewness. There is perhaps even a suggestion of a subsidiary mode at high values of log span, although this is difficult to evaluate.

The histogram is a very familiar object. It can be written as

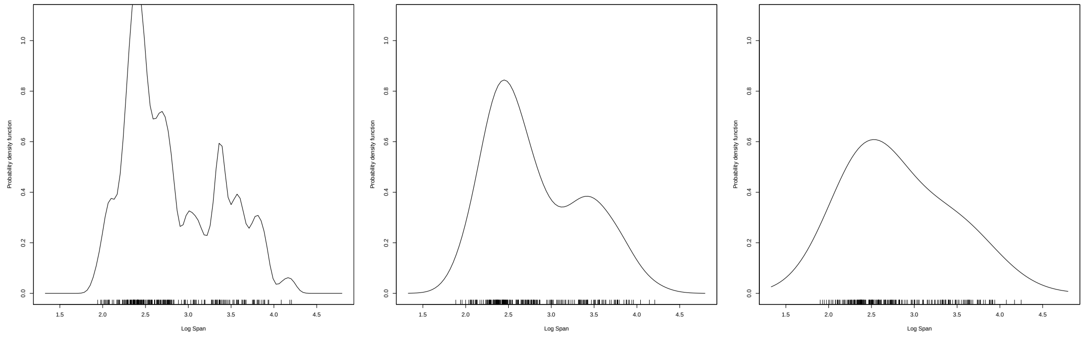
$$\tilde{f}(y) = \sum_{i=1}^{n} I(y - \tilde{y}_i; h),$$

where $\{y_1, \ldots, y_n\}$ denote the observed data, $\tilde{y}_i$ denotes the centre of the interval in which y_i falls and $I(z; h)$ is the indicator function of the interval $[-h, h]$. The form of the construction of $\tilde{f}$ highlights some feature which are open to criticism if we view the histogram as an estimator of the underlying density function. Firstly the histogram is not smooth, when we expect that the underlying density usually will be. Secondly, some information is lost when we replace each observation y_i by the bin mid-point $\tilde{y}_i$. Both of the issues can be addressed by using a density estimator in the form

$$\hat{f}(y) = \frac{1}{n} \sum_{i=1}^{n} w(y - y_i; h),$$

where w is a probability density, called here a *kernel function*, whose variance is controlled by the *smoothing parameter h*.

Large changes in the value of the smoothing parameter have large effects on the smoothness of the resulting estimates, as the plots below illustrate.

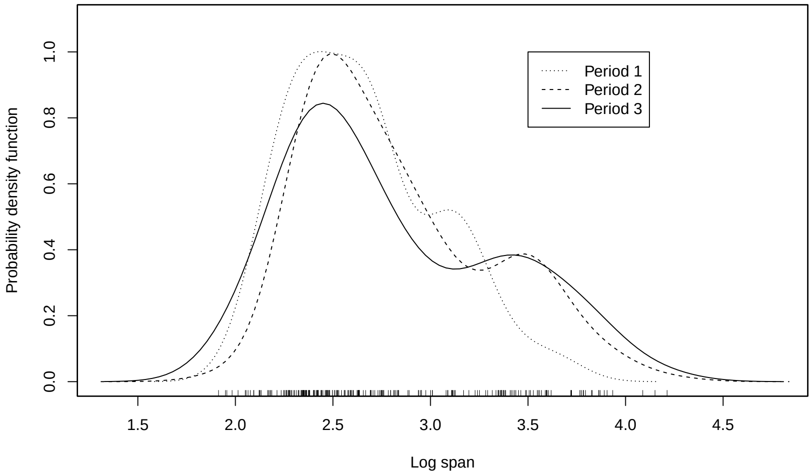


Exercise: The effect of the smoothing parameter

To experiment with density estimates, download the `sm`, `rpanel` and `tkrplot`packages for R. The code below should launch a new window with interactive controls. Try altering the smoothing parameter through the slider. Does this help you assess whether the subsidiary mode is a genuine feature or an artefact of random variation?

```
library(sm)
y <- log(aircraft$Span[aircraft$Period == 3])
sm.density(y, panel = TRUE)
```

One advantage of density estimates is that it is a simple matter to superimpose these to allow different groups to be compared. Here the groups for the three different time periods are compared. It is interesting that the 'shoulder' appears in all three time periods.



Exercise: Computing density estimates

Write some R code which will take a vector of data y and return a density estimate. See if you can do this without using a for loop!

2 Extending the idea

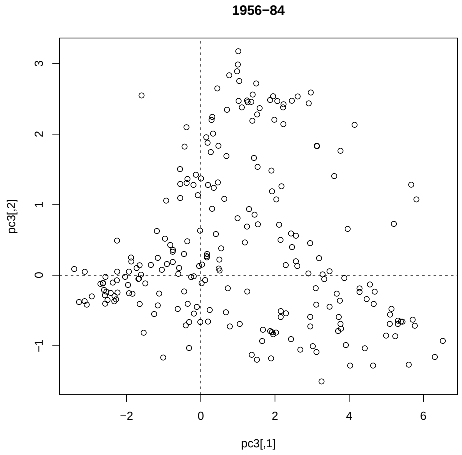
The simple idea of density estimation is to place a kernel function, which in fact is itself a density function, on top of each observation and average these functions. This extends very naturally to a wide variety of other types of data and sample spaces.

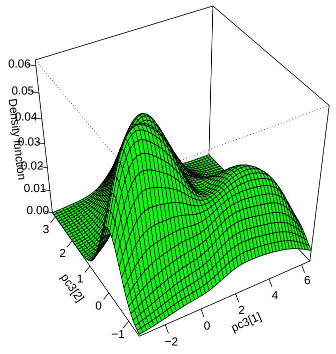
For example, a two-dimensional density estimate can be constructed from bivariate data $\{(y_{1i}, y_{2i}) : i = 1, \ldots, n\}$ by employing a two-dimensional kernel function in the form

$$\hat{f}(y_1, y_2) = \frac{1}{n} \sum_{i=1}^{n} w(y_1 - y_{1i}; h_1)\, w(y_1 - y_{2i}; h_2).$$

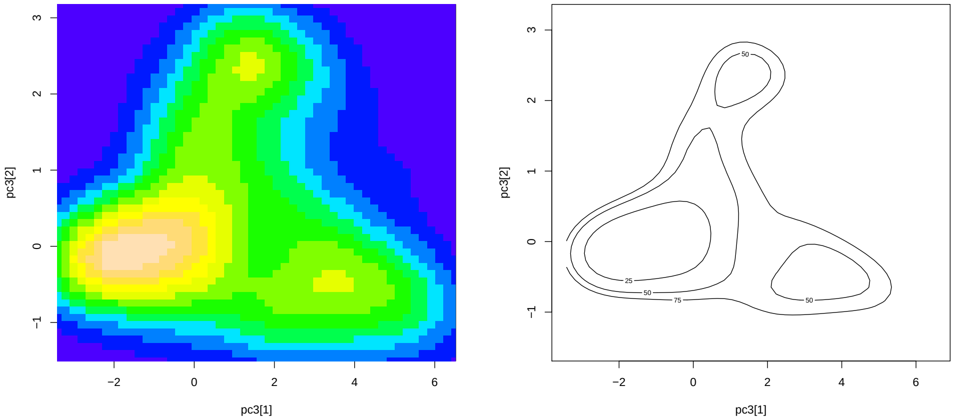
Notice that there are now two smoothing parameters, (h_1, h_2). A more general two-dimensional kernel function could be used, but the simple product form is very convenient and usually very effective.

Here is an example which uses the scores from the first two principal components of the aircraft data, again focussing on the third time period. The left hand scatterplot shows the individual scores while the right hand plot shows a density estimate, from which suggests three separate modes. This feature is not so easily seen from the raw scatterplot.





Here are alternative ways of presenting a two-dimensional estimate, using a coloured image on the left and contour lines on the right. Notice that the contours on the right have been chosen carefully to contain the quarters of the data with successively higher density, in a manner which has some similarities with a box plot.



This principle extends to all kinds of other data structures and sample spaces by suitable choice of an appropriate kernel function.

Exercise: Circular data

If you were given a sample of circular data, consisting of angles (in radians) between 0 and 2π (where, of course, these two end points are coincident), how could you construct a smooth density estimate by choice of an appropriate kernel function? (*Hint*: the *von Mises* distribution provides a natural model for a distribution on a circle.) Can you write some code in R which will construct a density estimate from a sample of circular data. (*Hint*: you might find the R function `besselI` useful.)

3 Simple properties of density estimates

Without any real restriction, we can assume that the kernel function can be written in the simple form $w(y - y_i; h) = \frac{1}{h} w\left(\frac{y-y_i}{h}\right)$. The mean of a density estimator can then be written as

$$\mathbb{E}\{\}\,\hat{f}(y) = \int \frac{1}{h} w\left(\frac{y-z}{h}\right) f(z)dz = \int w(u;1) f(y-hu)du,$$

where the last expression simply involves a change of variable $u = \frac{y-z}{h}$. A Taylor series expansion of the term involving f in the last expression gives

$$f(y-hu) = f(y) - huf'(y) + \frac{1}{2}h^2u^2f''(y) + o(h^2)$$

and, on insertion into the expression for the mean, this produces the approximation

$$\mathbb{E}\{\}\,\hat{f}(y) \approx f(y) + \frac{h^2}{2}\sigma_w^2\, f''(y),$$

where we assume that the kernel function is symmetric so that $\int uw(u)du = 0$, and where σ_w^2 denotes the variance of the kernel, namely $\int u^2w(u)du$.

The variance of the density estimate can be written as

$$\begin{aligned} \mathsf{var}\{\}\,\hat{f}(y) &= \frac{1}{n}\mathsf{var}\{Y\}\,\frac{1}{h}w\left(\frac{y-Y}{h}\right) \\ &= \frac{1}{n}\left\{\mathbb{E}\{Y\}\left[\frac{1}{h}w\left(\frac{y-Y}{h}\right)\right]^2 - \mathbb{E}\{Y\}\,\frac{1}{h}w\left(\frac{y-Y}{h}\right)^2\right\} \end{aligned}$$

A similar change of variable and Taylor series expansion produces the approximation

$$\mathsf{var}\{\}\,\hat{f}(y) = \frac{1}{nh}f(y)\,\alpha(w) + o\left(\frac{1}{nh}\right),$$

where $\alpha(w) = \int w^2(u)du$.

These expressions capture the essential features of smoothing. In particular, bias is incurred and we can see that this is controlled by f'', which means that where the density has peaks and valleys the density estimate will underestimate and overestimate respectively. This makes intuitive sense.

Similar expressions can be derived in higher dimensions.

Exercise: Asymptotic properties

A useful global measure of performance is the *mean integrated squared error* (MISE) which balances squared bias and variance.

$$\begin{aligned} \text{MISE}(\hat{f}) &= \mathbb{E}\{\} \int [\hat{f}(y) - f(y)]^2 dy \\ &= \int \left[\mathbb{E}\{\}\, \hat{f}(y) - f(y)\right]^2 dy + \int \mathsf{var}\{\}\, \hat{f}(y) dy. \end{aligned}$$

(i) Verify the expression given for $\mathsf{var}\{\}\, \hat{f}(y)$ above by employing a change of variable and a Taylor series expansion.

(ii) Use the Taylor expansions discussed above to show that MISE can be expressed as

$$\text{MISE}(\hat{f}) \approx \frac{1}{4} h^4 \sigma_w^4 \int f''(y)^2 dy + \frac{1}{nh} \alpha(w).$$

(iii) Now show that the value of h which minimizes MISE in an asymptotic sense is

$$h_{\text{opt}} = \left\{ \frac{\gamma(w)}{\beta(f) n} \right\}^{1/5},$$

where $\gamma(w) = \alpha(w)/\sigma_w^4$, and $\beta(f) = \int f''(y)^2 dy$.

(iv) Show that when f is a normal density this yields the simple formula

$$h = \left(\frac{4}{3n} \right)^{1/5} \sigma,$$

where σ denotes the standard deviation of the distribution.

(v) Considering $\hat{f}(y)$ as a density function in y, what is the mean and variance of the distribution which this density function represents? Consider what this says about the operation of smoothing.

4 Deciding how much to smooth

The theory sketched in the exercise above shows that an optimal smoothing parameter can be defined as the value which minimises MISE. This has the form

$$h_{\text{opt}} = \left\{ \frac{\gamma(w)}{\beta(f)n} \right\}^{1/5}.$$

Of course, this is of rather limited use because it is a function of the unknown density. However, there are two practical approaches which can be taken to deciding on a suitable smoothing parameter to use. One is to construct an estimate of MISE and minimise this. Another is to estimate the optimal smoothing parameter. These two approaches are outlined below.

Cross-validation

The integrated squared error (ISE) of a density estimate is

$$\int \{\hat{f}(y) - f(y)\}^2 dy = \int \hat{f}(y)^2 dy - 2\int f(y)\hat{f}(y)dy + \int f(y)^2 dy.$$

Only the first two of these terms involve h and these terms can be estimated by

$$\frac{1}{n}\sum_{i=1}^{n} \int \hat{f}_{-i}^2(y)dy - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{-i}(y_i),$$

where $\hat{f}_{-i}(y)$ denotes the estimator constructed from the data without the observation y_i. The value of h which minimises this expression is known as the *cross-validatory* smoothing parameter.

Plug-in methods

By inserting suitable estimates of the unknown quantities in the formula for the optimal smoothing parameter, a *plug-in* choice can be constructed. The difficult part is the estimation of $\beta(f)$ as this involves the second derivative of the density function. Sheather & Jones (JRSSB 53, 683–90) came up with a good, stable way of doing this. The Sheather-Jones remains one of the most effective strategies for choosing the smoothing parameter.

A very simple plug-in approach is to use the normal density function in the expression for the optimal smoothing parameter. An earlier exercise showed that this produces the simple formula $\left(\frac{4}{3n}\right)^{1/5}\sigma$, where the standard deviation σ can be estimated from the data. This approach is surprisingly effective, in large part because it is very stable.

Exercise: Smoothing parameter selection in practice

The R expression `sm.density(y, panel = TRUE)` allows interactive experimentation in the construction of density estimates from a data vector `y`. Try this function out with data of your own choice and look at the comparative performances of the methods of smoothing parameter section outlined here.

Try this out with two-dimensional data, using the aircraft data or some simple randomly generated data, for example by `y <- cbind(rnorm(50), rnorm(50))`. The `sm.density` function operates similarly, although the Sheather-Jones plug-in method is not available because it is trickier to implement beyond the one-dimensional setting.

5 Some simple inferential tools

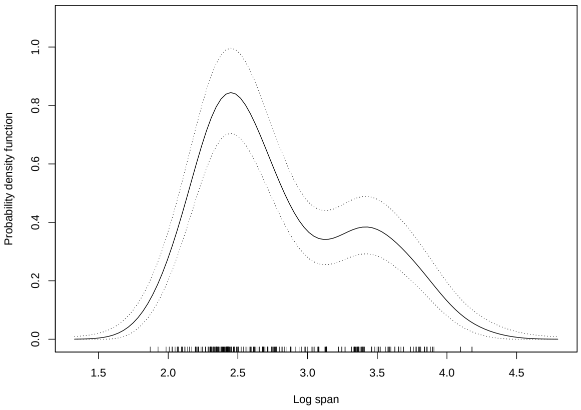
Once an estimate has been constructed, a natural next step is to find its standard error. The earlier result on the variance of $\hat{f}$ is a natural starting point, but this expression involves the unknown density. A helpful route is to consider a 'variance stabilising' transformation. For any transformation $t(\cdot)$, a Taylor series argument shows that

$$\mathsf{var}\{\}\, t(\hat{f}(y)) \approx \mathsf{var}\{\}\, \hat{f}(y) \left[t'\left(\mathbb{E}\{\}\, \hat{f}(y)\right)\right]^2 .$$

When $t(\cdot)$ is the square root transformation, the principal term of this expression becomes

$$\mathsf{var}\{\}\, \sqrt{\hat{f}(y)} \approx \frac{1}{4}\frac{1}{nh}\,\alpha(w),$$

which does not depend on the unknown density f. This forms the basis of a useful *variability band*. We cannot easily produce proper confidence intervals because of the bias present in the estimate. However, if the standard error is constructed and the intervals corresponding to two s.e.'s on the square root scale are transformed back to the origin scale, then a very useful indication of the variability of the density estimate can be produced. This is shown below for the aircraft span data from period 3.



A useful variation on this arises when the true density function is assumed to be normal with mean μ and variance σ^2, and the kernel function w is also normal. If the standard normal density function is denoted by ϕ, then the mean and variance of the density estimate at the point y are then

$$\begin{aligned}
\mathbb{E}\{\}\, \hat{f}(y) &= \phi\left(y-\mu; \sqrt{h^2+\sigma^2}\right) \\
\mathsf{var}\{\}\, \hat{f}(y) &= \frac{1}{n}\phi\left(0; \sqrt{2}\,h\right)\, \phi\left(y-\mu; \sqrt{\sigma^2+\frac{1}{2}h^2}\right) \\
&\quad -\frac{1}{n}\phi\left(y-\mu; \sqrt{\sigma^2+h^2}\right)^2
\end{aligned}$$

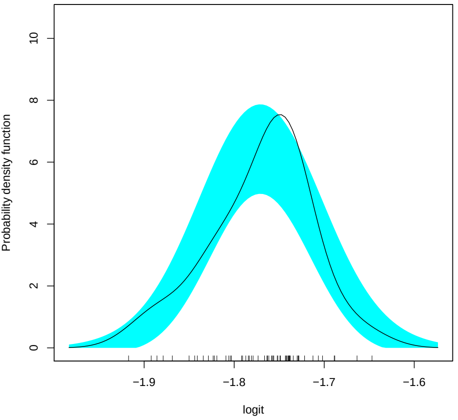
These expressions allow the likely range of values of the density estimate to be calculated, under the assumption that the data are normally distributed. This can be expressed graphically through a *reference band*.

Exercise: Icelandic tephra layer

Data on the percentages of aluminium oxide found in samples from a tephra layer resulting from a volcanic eruption in Iceland around 3500 years ago are available in the `tephra` dataset in the `sm` package. TO deal with the percentage scale, apply the logit transformation

```
logit <- log(tephra$Al2O3/(100-tephra$Al2O3)).
```

Now use the `sm.density` function with the interactive controls provided by `panel = TRUE` to add standard errors and a reference band for a normal model. Is there convincing evidence that the tephra data exhibit non-normal features?



Exercise: Means and variance again

Verify the results stated above on the approximate variance of $\sqrt{\hat{f}(y)}$ and the exact mean and variance of $\hat{f}(y)$ when the underlying distribution is normal with mean μ and variance σ^2.

The bootstrap is sometimes a useful way of gaining information about the properties and sampling variation of quantities of interest. Here is a simple version of the bootstrap for density estimation.

1. Construct a density estimate $\hat{f}$ from the observed data $\{y_1, \ldots, y_n\}$.
2. Resample the data with replacement to produce a bootstrap sample $\{y_1^*, \ldots, y_n^*\}$.
3. Construct a bootstrap density estimate $\hat{f}^*$ from the bootstrap data $\{y_1^*, \ldots, y_n^*\}$.
4. Repeat steps 2 and 3 a large number of times to create a collection of bootstrap density estimates $\{\hat{f}_1^*, \ldots, \hat{f}_B^*\}$.
5. Use the empirical distribution of $\hat{f}^*$ about $\hat{f}$ to mimic the distribution of $\hat{f}$ about f.

However, we need to be careful about the interpretation of this. Since the distribution of y_i^* is uniform over $\{y_1, \ldots, y_n\}$, it follows that

$$\mathbb{E}\{*\}\, \hat{f}^*(y) = \mathbb{E}\{*\}\, w(y - y_i^*; h) = \hat{f}(y)$$

and so the bias which we know is present in the distribution of $\hat{f}$ is absent in the bootstrap version. However, the bootstrap does usefully mimic the variance of $\hat{f}$.

Exercise: Bootstrapped variability bands

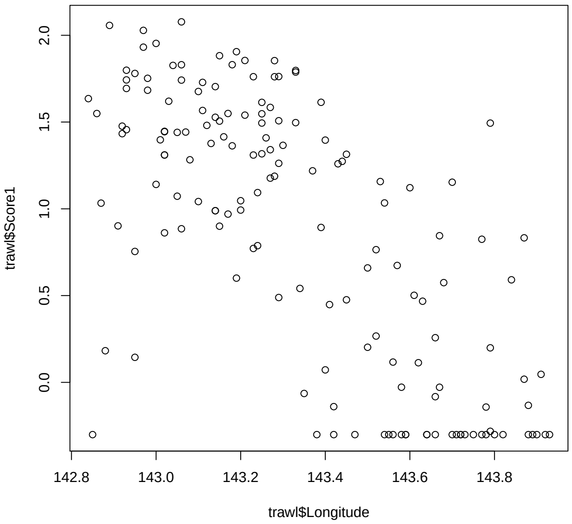
Write a few lines of R code which will construct and plot bootstrap density estimates of the aircraft log span data. Compare the results with the variability band shown near the start of Section 5.

6 A quick look at flexible regression

The emphasis of the preliminary material has been density estimation. However, here are two problems which might whet your appetite for regression, which will be the main focus of the APTS course.

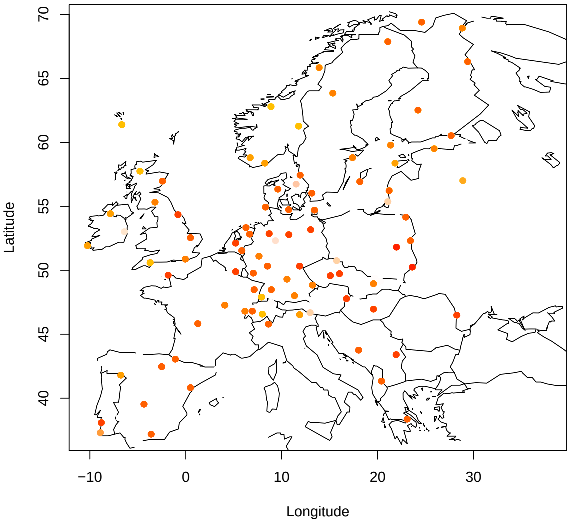
Exercise: Great Barrier Reef

In a study of the effects of trawling on the Great Barrier Reef, measurements of the quantity of marine life sampled from the sea bed were taken at a variety of locations. These data are available in the data frame `trawl` in the `sm` package. The plot on the right shows the response variable `Score1` against `Longitude`, which here is an approximate proxy for distance offshore. Try `sm.regression(trawl$Latitude, trawl$Score1, panel = TRUE)` to see whether you gain any further insight into the relationship between these two variables.



Exercise: SO_2 over Europe

Europe has an extensive system of air monitoring stations at which levels of SO_2 are regularly measured. The dataset `SO2` in the `rpanel` package for R contains measurements from these stations over several years. We should expect spatial effects, seasonal effects and time effects. How might we construct a flexible regression model which could capture and explore the nature of these effects? Feel free to experiment with the dataset, if you have some ideas.



7 Further reading

A classic text on density estimation:

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis.* Chapman & Hall, London.

A variety of texts on flexible regression:

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models.* Chapman & Hall: London.

Bowman, A.W. & Azzalini, A. (1996). *Applied Smoothing Techniques for Data Analysis.* OUP: Oxford.

Ruppert, D., Wand, M.P. & Carroll, R.J. (2003). *Semiparametric Regression.* CUP: Cambridge.

Wood, S. (2006). *Generalized additive models: an introduction with* R. Chapman & Hall, London.