

UNIVERSITY OF WOLLONGONG
School of Mathematics and Applied Statistics
STAT902. Advanced Data Analysis
ASSIGNMENT 12

The submission of this assignment is optional. If you do not submit this assignment then the 30% contribution to your mark from assignments will be based solely on your scores for Assignments 1-11. However, if you do submit this assignment then I am happy to grade it and include it among your best 11 assignment scores (i.e. if your score in this assignment is better than at least one of your scores on Assignments 1–11.). If you do decide to submit this assignment then the due time is: 5:00pm Thursday, 5th February 2009; and it please should be put under the door of my office (Room 139 of the Austin Keane Building).

Preliminary: R Package Installation

An important part of advanced data analysis with R is the installation of specialist packages. These reside on the **Comprehensive R Archive Network (CRAN)** and give the user the of using thousands of additional R functions written by analysts around the World. This assignment requires the package **gam**, written by Trevor Hastie (Stanford University, USA), for Generalised Additive Model fitting and inference. There are a few ways by which packages can be installed – although some of them depend on the set-up of the computer. The following instructions are, hopefully, foolproof, and specific to **Windows** (similar commands apply to other operating systems):

1. Go to the CRAN web-site: <http://cran.r-project.org/>
2. Click on **Packages** in the menu on the left.
3. This results in a long list of R packages. Scroll down to the one named **gam** and click on its link.
4. Download the file with name of the form **gam_xxxx.zip** to your computer. Here **xxxx** corresponds to the version number. In late May, 2008, the version number was 0.98.
5. Start an R session and under the **Packages** menu click on **Install package(s) from local zip files...**
6. Select the file that you just downloaded. This should install the package **gam** on your computer.
7. If successful then the command `library(gam)` should make functions in this package available to the current R session.

The Hand-in Part of the Assignment

1. The R package **gam** supports a stepwise procedure for choosing among generalised additive models; known as `step.gam()`. Its use is illustrated for the Intensive Care Unit (ICU) data in the script `ICUStepGAM.Rs`. In this particular script the three continuous predictors `age`, `SBP` and `heartRate` are permitted to enter the model either linearly, or as smooth functions with either 3, 6 or 9 degrees of freedom. For example, `s(SBP,6)` is

the way that `gam()` and `step.gam()` specify a smooth function of SBP with 6 degrees of freedom.

- (a) Download `ICUStepGAM.Rs` from the `Computer Code and Data` page on the course web-site. Make sure that the data file `ICU.txt` is available to the current session type `source("ICUStepGAM.Rs")` to run an example of `step.gam()`.
- (b) Download the data `trade.union.txt` from the `Computer Code and Data` page on the course web-site. The trade union data consists of:

<code>years.educ</code>	number of years of education.
<code>south</code>	indicator of living in southern region of U.S.A.
<code>female</code>	gender indicator: 0=male,1=female.
<code>years.experience</code>	number of years of work experience
<code>union.member</code>	indicator of trade union membership: 0=non-member, 1=member.
<code>wage</code>	wages in dollars per hour.
<code>age</code>	age in years.
<code>race</code>	1=black, 2=Hispanic, 3=white.
<code>occupation</code>	1=management, 2=sales, 3=clerical, 4=service, 5=professional, 6=other.
<code>sector</code>	0=other, 1=manufacturing, 2=construction.
<code>married</code>	indicator of being married: 0=unmarried, 1=married.

- (c) In R create a binary indicator of a worker being white via:

```
white <- as.numeric(trade.union$race==3)
```

- (d) Starting with the predictors specified in the call:

```
gam.object <- gam(union.member~years.educ+south+female
                 +years.experience+wage+age+white+married,
                 data=trade.union,family=binomial)
```

use `step.gam()` to select a logistic additive model for $P(\text{union.member} = 1)$. As in the ICU example, you should allow each continuous variable to have either 3, 6 or 9 degrees of freedom. Summarise the final model through a table and plot.

2. There is another function in R for fitting generalised additive models, also named `gam()`, and part of the package `mgcv` by Simon N. Wood (University of Bath, United Kingdom). Note that `mgcv` is included in the 'base' version of R so, unlike the `gam` package, does not need to be downloaded from CRAN and installed. An advantage of the `gam()` function in `mgcv` is that the degrees of freedom of smooth functions are chosen via generalised cross validation (GCV).

- (a) Start a new R session and, to avoid confusion with the two versions of `gam()`, do not issue the command `library(gam)`.
- (b) Make sure that the file `trade.union.txt`, corresponding to the previous question of this assignment, is available. Issue the commands:

```
library(mgcv)
trade.union <- read.table("trade.union.txt",header=TRUE)
white <- as.numeric(trade.union$race==3)
```

- (c) Issue the commands:

```
fit1 <- gam(union.member~s(age)+s(years.educ)+s(wage),
           family=binomial,data=trade.union)
print(summary(fit1))
par(mfrow=c(2,2))
plot(fit1,shade=TRUE,shade.col="tan2")
```

3. Download the data set `fossil.txt` and the scripts `fossilBayesModel.txt` and `fossilBayes.Rs` from the **Computer Code and Data** page of the course web-site.

- (a) Start an R session and enter the commands:

```
fossil <- read.table("fossil.txt",header=TRUE)
plot(fossil$age,fossil$strontium.ratio)
```

The scatterplot corresponds to data on ratios of strontium isotopes found in fossil shells and their age.

- (b) Type `source("fossilBayes.Rs")`. This invokes **BRugs** and **WinBUGS** to fit the Bayesian penalised spline model

$$y_i = \beta_0 + \beta_1 x_i + \sum_{k=1}^{35} u_k (x_i - \kappa_k)_+ + \varepsilon_i$$

$$[u_1, \dots, u_{35} | \sigma_u^2] \sim N(0, \sigma_u^2)$$

$$\beta_0, \beta_1 \stackrel{\text{ind.}}{\sim} N(0, 10^8) \quad \sigma_u^2, \sigma_\varepsilon^2 \stackrel{\text{ind.}}{\sim} \text{IG}(0.01, 0.01)$$

where (x_i, y_i) are standardised versions of age and strontium ratio and $\kappa_1, \dots, \kappa_{35}$ is a set of knots. The script produces a plot of the fit, with pointwise 95% credible intervals for $E(y|x)$, and the estimated posterior degrees of freedom.

Important note: Spline-based models have many more parameters than those previously considered in this course. Consequently, MCMC is quite a bit slower. You should expect the fitting to take minutes rather than seconds. (A good excuse for a stretch, and checking out the latest news on Britney Spears.)

- (c) Fit the above model to the LIDAR data from Assignment 11 and produce similar plots to those produced by `fossilBayes.Rs`.

Hint: Given the slowness of MCMC for spline models, it is strongly recommended that the code be developed with much lower numbers of burn-in and iteration values (e.g. 100). Once the code seems to be running alright then these numbers can be re-set to their usual values.

While this question deals with the simplest version of Bayesian semiparametric regression, it illustrates how MCMC might be used to fit more complicated semi-parametric regression models – such as those involving generalised responses, several predictors and repeated measures.

4. The R script `SmithWand08.Rs` on the **Computer Code and Data** page of the course web-site corresponds to the Appendix of the 2008 *Statistics in Medicine* paper by A.D.A.C. Smith and M.P. Wand. The script data first generates simulated according to the *additive mixed model*

$$y_{ij} = \beta x_{ij} + U_i + f(s_{ij}) + \varepsilon_{ij}$$

$$U_i \stackrel{\text{ind.}}{\sim} N(0, \sigma_U^2), \quad \varepsilon_{ij} \stackrel{\text{ind.}}{\sim} N(0, \sigma_\varepsilon^2)$$

where the x variable is binary, the s variable is continuous and $f(x) = -\sin(2\pi x)$. It then fits the model to the data using maximum likelihood via `lme()` with f modelled as a penalised spline (with a *radial cubic* basis).

- (a) Download the script as a text file and run it. Write down the resulting approximate 95% confidence interval for β .
- (b) In the script change the seed (currently specified by the command `set.seed(39402)`) to be your student number. Also change f to be the function

$$f(x) = 4 + 2\Phi(8x - 4)$$

where $\Phi(x) = P(Z \leq x)$, $Z \sim N(0, 1)$ is the cumulative distribution function of the standard normal distribution. (Note that Φ in **R** is called `pnorm()`.) Obtain a plot of the fit to these new data, and report 95% approximate confidence intervals for β , σ_U and σ_ε .