STAT40810 — Stochastic Models

Brendan Murphy

Week 5

Generalized Additive Models

Multivariate Smoothing

- In today's class we will look at how we can go multivariate with the spline modelling idea.
- This will lead to the ideas of *generalized additive models* and *projection-pursuit* regression.
- First, we will look at the concept of "backfitting"; this is a general technique that arises many model fitting contexts.

Linear Models

- Let's return to multiple linear regression for the moment.
- Suppose we have a dataset with K predictor variables $(X_1, X_2, \ldots, X_K)$ and a response variable Y.
- We want to fit the model,

$$Y_i = b_0 + \sum_{k=1}^{K} b_k X_{ik} + \epsilon_i,$$

  where $\epsilon_i \sim N(0, \sigma^2)$ independently.
- We typically do this by minimising,

$$\sum_{i=1}^{n} \left[y_i - b_0 - \sum_{k=1}^{K} b_k X_{ik} \right]^2.$$

Backfitting

- We can fit the multiple linear regression model by fitting a series of univariate simple linear regression models.
- The following steps could be used:
  0. Let $b_0 = \overline{y}$ and $b_1, b_2, \ldots, b_K = 0$. Let $\ell = 1$.
  1. To estimate b_1, fit a linear regression (with zero intercept) with response
  $$y_i - b_0 - \sum_{\substack{k=1 \\ k \neq \ell}}^{K} b_k X_{ik}$$
  and predictor $X_{i\ell}$.
  2. Increment ℓ by 1. If $\ell \leq K$ return to step 1. Otherwise, if $\ell > K$ check if estimates have converged.
  If not, let $\ell = 1$, $b_0 = \sum_{i=1}^{N}(y_i - \sum_{k=1}^{K} b_k X_{ik})/n$ and return to step 1.

Backfitting: Alternative Explanation

- We are minimising,

$$\sum_{i=1}^{n}\left[y_i - b_0 - \sum_{k=1}^{K} b_k X_{ik}\right]^2,$$

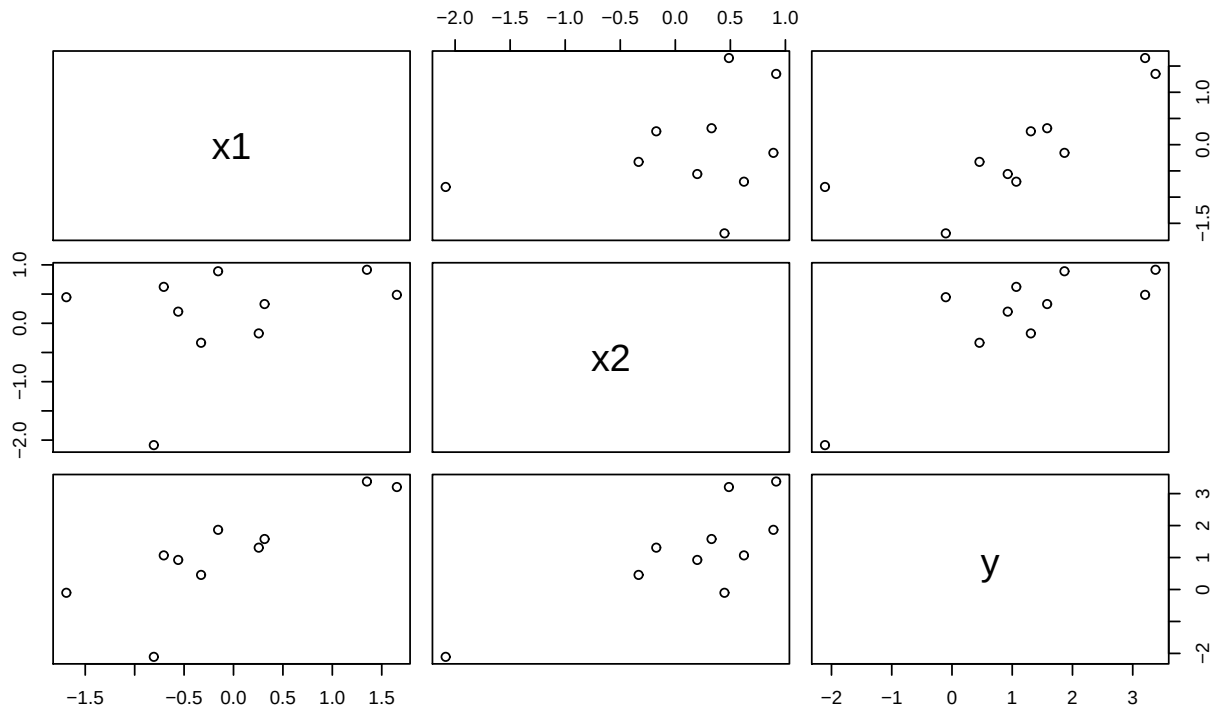
  with respect to $b_0, b_1, \ldots, b_K$.
- Backfitting is simply reducing this to a series of univariate minimizations:
  - For $k = 0, 1, \ldots, K$ minimize

$$\sum_{i=1}^{n}\left[y_i - b_0 - \sum_{k=1}^{K} b_k X_{ik}\right]^2,$$

    with respect to b_k.
  - Repeat until convergence.

Example: Bivariate Regression

- Scatterplots of a dataset with one response (y) and two predictors (x1,x2) are shown.



Example: Backfitting Results

- The parameter estimates for the first 9 iterations are:

```
b0 b1 b2
[1,] 1.157511 1.2682460 0.9935901
[2,] 1.113851 1.0000186 1.0940428
[3,] 1.082707 0.9709355 1.1097723
[4,] 1.078701 0.9664456 1.1120537
[5,] 1.078101 0.9657930 1.1123885
[6,] 1.078014 0.9656973 1.1124375
[7,] 1.078001 0.9656833 1.1124447
[8,] 1.077999 0.9656812 1.1124457
[9,] 1.077999 0.9656809 1.1124459
```

- The final estimates are the same as the multiple regression estimates.

Example: Generalized Additive Models

- The generalized additive model is a multivariate equivalent of a spline model.
- Whereas in spline modelling, we modelled the data as,

$$Y_i = \alpha + f(X_i) + \epsilon_i, \text{ where } f(\cdot) \text{ is a smooth function.}$$

- In generalized additive models, we model the data using the model,

$$Y_i = \alpha + \sum_{k=1}^{K} f_k(X_{ik}) + \epsilon_i, \text{ where the } f_k(\cdot) \text{ are smooth functions.}$$

- This compares to multiple regression where the model is,

$$Y_i = \alpha + \sum_{k=1}^{K} \beta_k X_{ik} + \epsilon_i.$$

Fitting Generalized Additive Models

- Generalized Additive Models are fitted using backfitting.
- The general idea of the fitting algorithm is that it iterates $\ell = 1, 2, \ldots, K$ the process of fitting a smoothing spline, with response variable

$$y_i - \alpha - \sum_{\substack{k \neq \ell \\ k=1}}^{K} f_k(X_{ik})$$

  and predictor $X_{i\ell}$.
- Also, α is estimated by the mean of the values,

$$y_i - \sum_{k=1}^{K} f_k(X_{ik}).$$

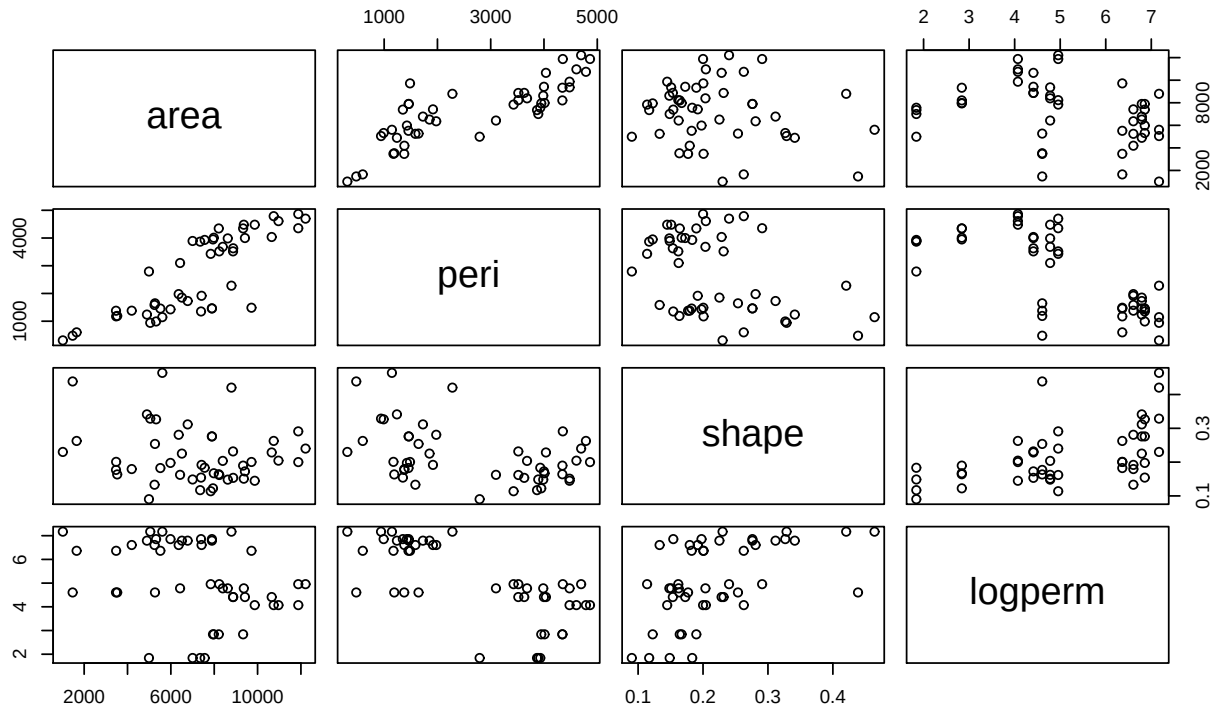
- This iteratively minimizes the least squares fitting criterion.

Example: Rock Permeabilities

- Data were collected on the permeability (`perm`) of 12 oil-bearing rock core samples from petroleum resevoirs.
- Four cross-sections were taken from each core sample.
- Each core sample was measured for permeability, and each cross-section has total area of pores, total perimeter of pores, and shape.
- Due to the values occurring the log(`perm`) was used as a response variable.

Example: Scatter Plot

- Scatter plots of the data were produced



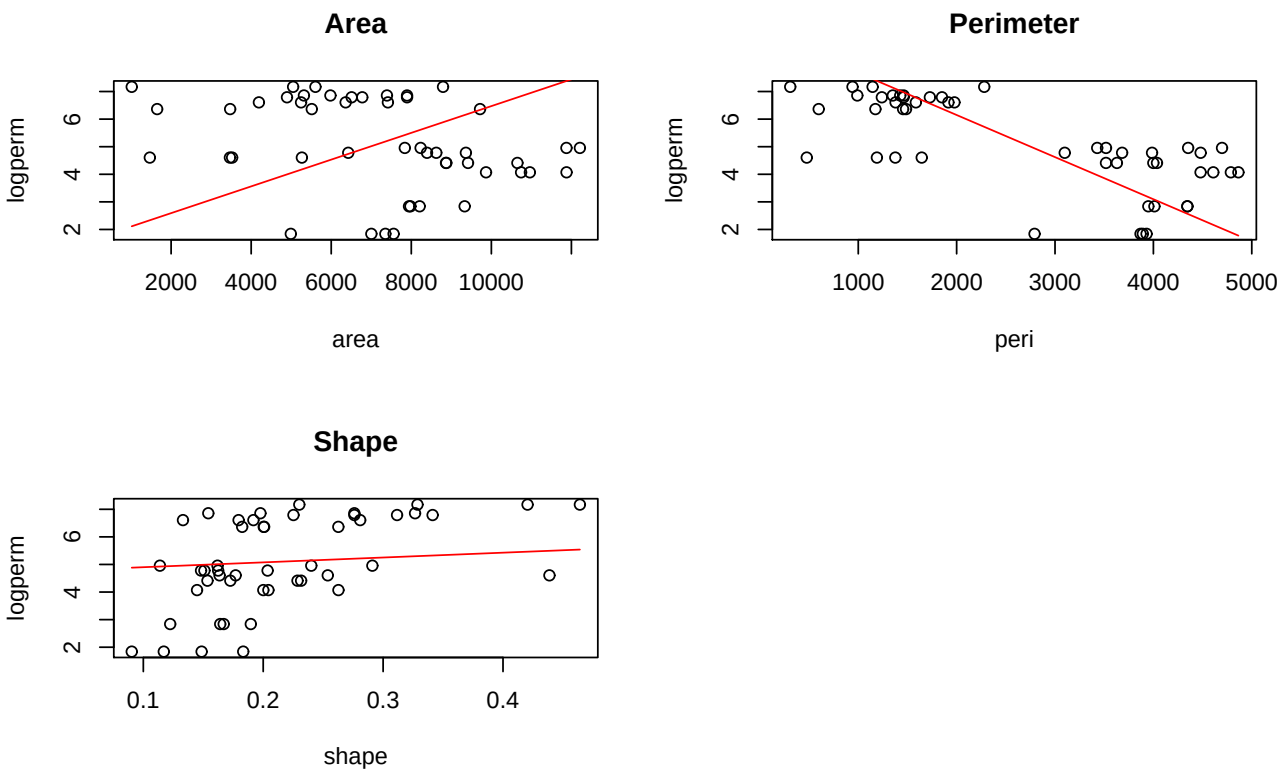
R Code

```
# Load the rock data
data(rock)

# Produce a pairs plot of the data
pairs(rock)

# Add a column with log(perm)
rock$lperm <- log(rock$perm)

# Fit a multiple linear regression model
fit0 <- lm(lperm~area+peri+shape,data=rock)

# Fit a generalized additive model
library(mgcv)
fit <- gam(lperm~s(area)+s(peri)+s(shape),data=rock)

par(mfrow=c(2,2))
plot.gam(fit,scale=0,se=FALSE)
par(mfrow=c(1,1))

# Re-fit the generalized additive model with
# area and peri being linear and only shape
# being transformed
fit1 <- gam (lperm~area+peri+s(shape),data=rock)

plot.gam(fit1)
```

Example: Multiple Regression

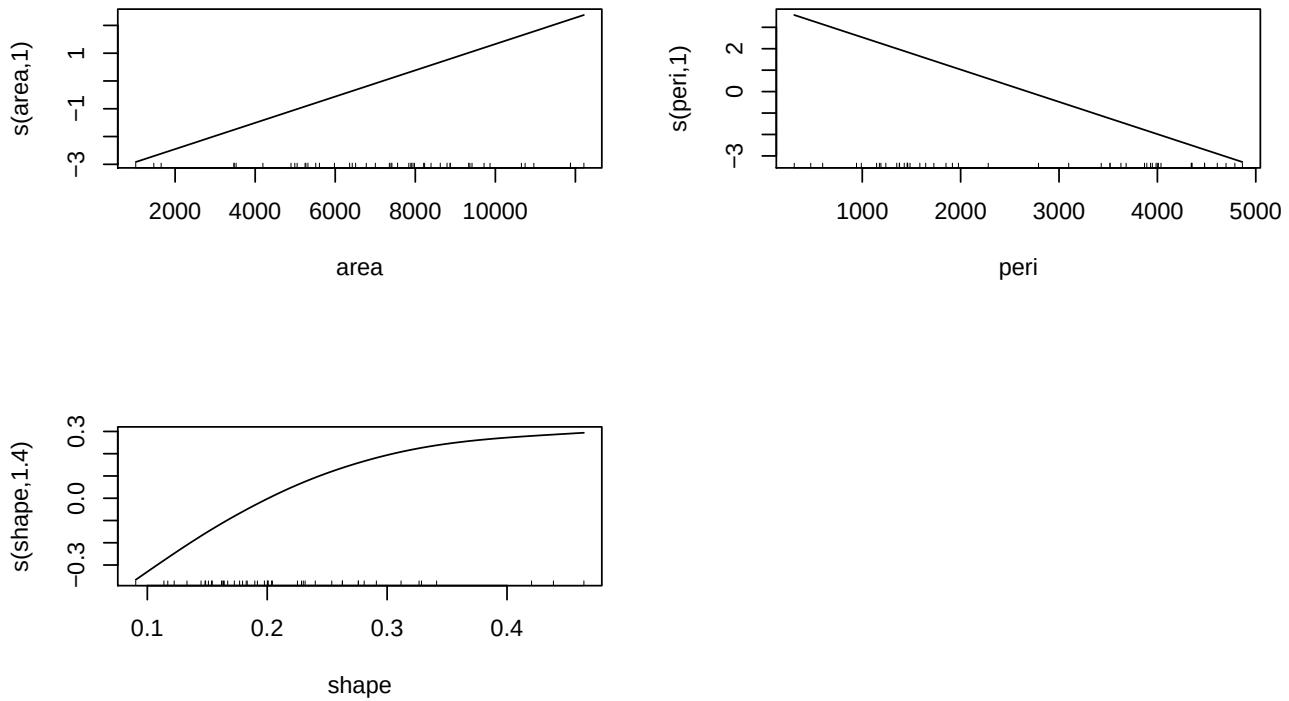
- A multiple regression gives the following fitted model,

$$\begin{aligned}\log(Perm) &= 5.333146 + 0.000485 Area \\ &\quad -0.001527 Perimeter + 1.756519 Shape\end{aligned}$$



Smooth Transformations

- The transformations found by the generalized additive model fit are shown below.



- Note that only the `shape` variable was transformed.