Session 08

GAMs an Introduction

Overview

- Model assumes that the mean response is a sum of terms each depending on (usually) a single predictor:

$$Y = \alpha + \sum_{j=1}^{p} f_j(x_j) + \varepsilon$$

- if the f_j s are linear terms, this is just regression
- if they are step functions – main effect of a factor term
- in general they may be smooth terms, with the degree of smoothness chosen by cross validation

Choices of terms

- In some cases the additive terms may be known
- Smoothing splines
- Local regression
- Splines with fixed degrees of freedom
- Splines with known knots and boundary knot positions
- Harmonic terms, &c

Similarities to, and differences from GLMs

- Additive models are analogous to regression models
- Generalized additive models are akin to the GLMs – they may employ a link function to relate the linear predictor to the mean of the response, they may have a non-normal distribution, &c
- Fitting GAMs is the same process as fitting GLMs (but with one letter different in the function name).
- The fitting process is NOT maximum likelihood if there are any smoother terms present. A likelihood penalized by a roughness term is maximised, with the tuning constant chosen (usually) by cross-validation
- Inference for GAMs is difficult and somewhat contentious. Best regarded as an exploratory technique with standard models to follow (see examples)

Example: the Iowa wheat yield data

- A toy example from Draper N R, and Smith H, *Applied regression analysis*, 2nd Ed., John Wiley & Sons, New York, 1981.
  - Response: Yield of wheat in bushels/acre for the state of Iowa for the years 1930-1962
  - Predictors: Year (as surrogate), Rain0, 1, 2, 3, Temp1, 2, 3, 4
- Problem: Build a predictor for Yield from the predictors available.
  - Note: with only 33 observations and 9 possible predictors some care has to be taken in choosing a model.

An initial linear model

```
iowa.lm1 <- lm(Yield ~ ., Iowa)
iowa.step <- stepAIC(iowa.lm1, scope = list(lower =  ~ Year,
   upper =  ~ .), k = log(nrow(Iowa)), trace = F)
dropterm(iowa.step, test = "F", k = log(nrow(Iowa)),
   sorted = T)

Single term deletions

Model:
Yield ~ Year + Rain0 + Rain2 + Temp4
       Df Sum of Sq      RSS      AIC  F Value      Pr(F)
<none>              1554.605 144.6140
 Temp4  1   187.951 1742.556 144.8838  3.38519 0.07640894
 Rain0  1   196.008 1750.612 145.0361  3.53029 0.07070429
 Rain2  1   240.204 1794.809 145.8589  4.32632 0.04679808
  Year  1  1796.216 3350.821 166.4610 32.35167 0.00000425
```

Initial reflections

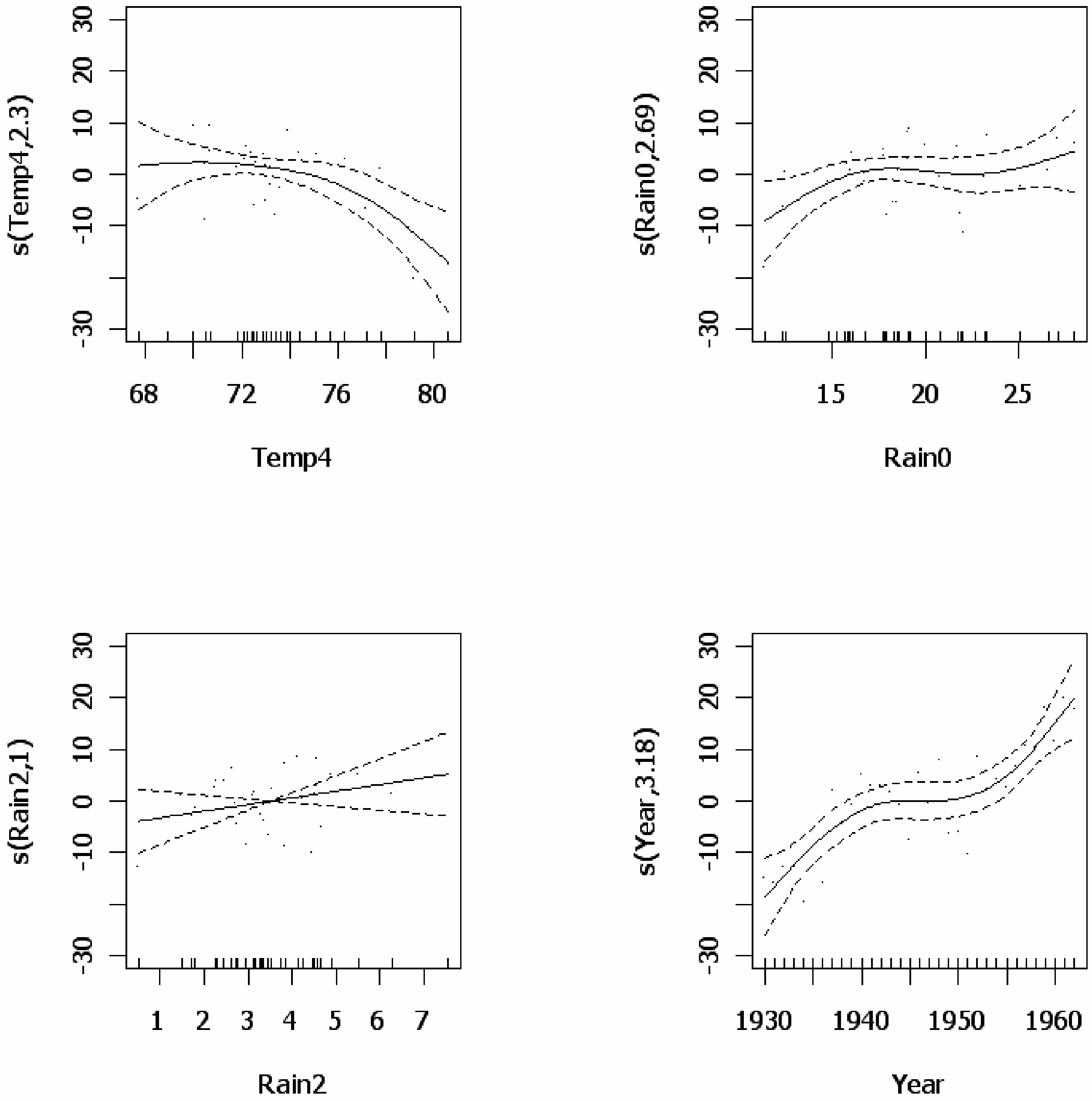
- Even with the more stringent BIC penalty on model complexity, two of the terms found are only borderline significant in the conventional sense – a consequence of the small sample size.
- Nevertheless the terms found are tentatively realistic:
  - `Year`: surrogate for crop improvements
  - `Rain0`: a measure of pre-season sowing conditions
  - `Rain2`: rainfall during the critical growing month
  - `Temp4`: climatic conditions during harvesting
- Are strictly linear terms in these variables reasonable?

Additive models

- Consider a non-parametric smoother in each term:

```
library(mgcv)
iowa.gam <- gam(Yield ~ s(Temp4,k=5) +
  s(Rain0,k=5) + s(Rain2,k=5) + s(Year,k=5),
  data = Iowa, trace=T)

par(mfrow = c(2,2))
plot(iowa.gam, se = T, ylim = c(-30, 30), resid =
  TRUE)
```

- It can be important to keep the y-axes of these plots approximately the same to allow comparisons between terms.


s(Temp4,2.3)
Temp4
68
72
76
80
s(Rain0,2.69)
Rain0
15
20
25
s(Rain2,1)
Rain2
1
2
3
4
5
6
7
s(Year,3.18)
Year
1930
1940
1950
1960
30
20
10
0
-10
-30

Speculative comments

- **`Temp4`**: Two very hot years had crop damage during harvest?
- **`Rain0`**: Wide range where little difference, but very dry years may lead to a reduced yield and very wet years to an enhanced one?
- **`Rain2`**: One very dry growing month led to a reduced yield?
- **`Year`**: Strongest and most consistent predictor by far. Some evidence of a pause in new varieties during the war and immediately post-war period?

Tentative inference

```
> summary(iowa.gam)

Family: gaussian
Link function: identity

Formula:
Yield ~ s(Temp4, k = 5) + s(Rain0, k = 5) + s(Rain2, k = 5) +
    s(Year, k = 5)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   50.000      1.035   48.32   <2e-16

Approximate significance of smooth terms:
           edf Est.rank      F  p-value
s(Temp4) 2.303        4  4.067   0.0124
s(Rain0) 2.686        4  2.343   0.0852
s(Rain2) 1.000        1  1.682   0.2076
s(Year)  3.180        4 14.477 5.02e-06

R-sq.(adj) =  0.797   Deviance explained = 85.5%
GCV score = 51.074   Scale est. = 35.334    n = 33
```

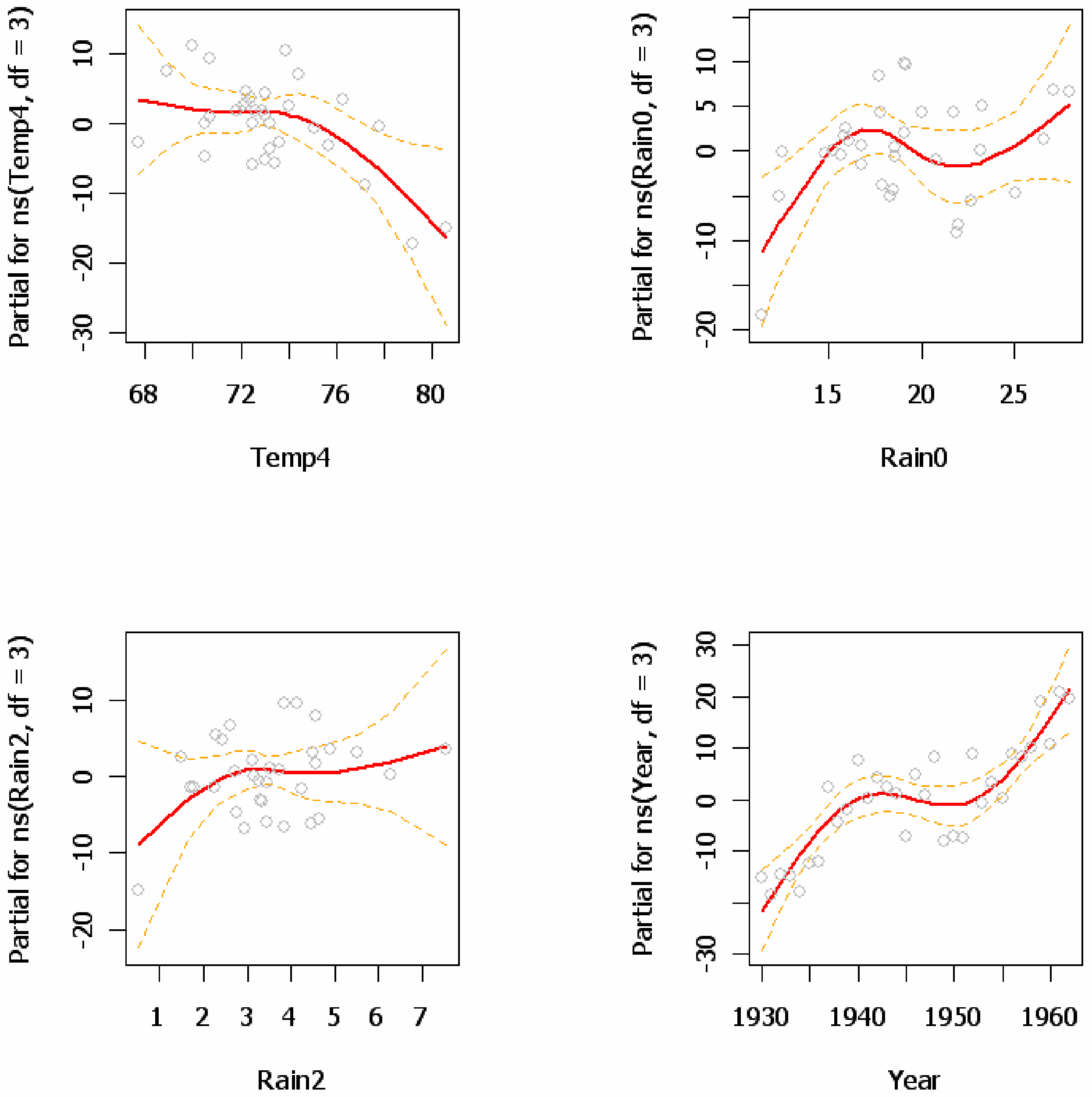
Can we get to the same place with GLMs?

- Spline terms: specified with ns(x, …) or bs(x, …), differ only in behaviour near the end points
- May specify the knot and boundary knot positions (recommended if prediction will be needed) or the equivalent degrees of freedom (OK for exploratory purposes)
- Each spline term is a collection of ordinary linear terms, but the coefficients have no simple meaning and the individual significance tests are meaningless. Best regarded as a single composite term and retained or removed as a block.

```
library(splines)
iowa.ns <- lm(Yield ~ ns(Temp4, df=3) + ns(Rain0, df=3) +
  ns(Rain2, df = 3) + ns(Year, df=3), Iowa)
termplot(iowa.ns, se=T, partial.resid = T)
dropterm(iowa.ns, test = "F", k = log(nrow(Iowa)))
```

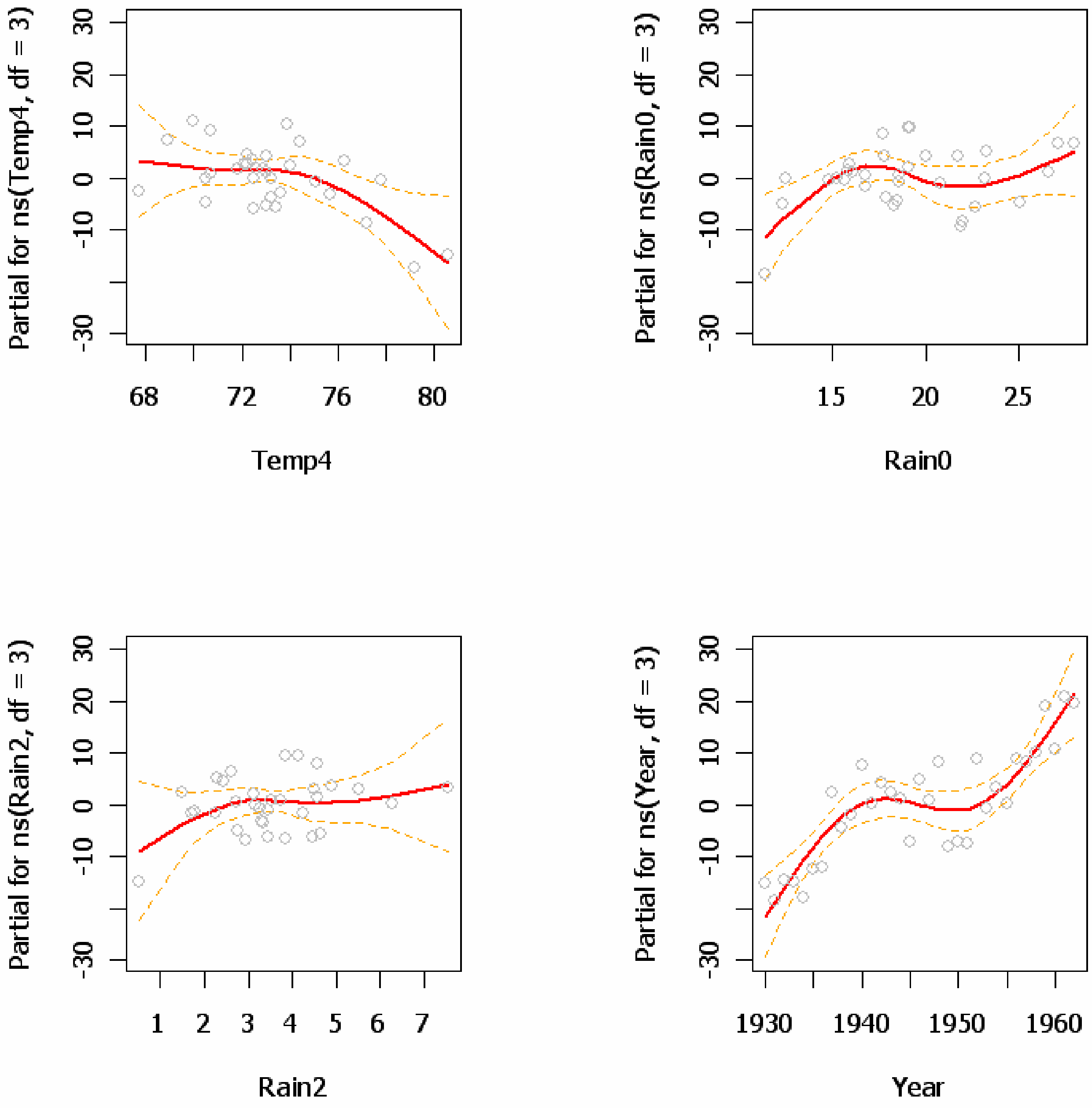
```
Single term deletions

Model:
Yield ~ ns(Temp4, df = 3) + ns(Rain0, df = 3) + ns(Rain2, df = 3)
  +
    ns(Year, df = 3)
                  Df Sum of Sq     RSS     AIC F Value     Pr(F)
<none>                          726.26  147.47
ns(Temp4, df = 3)  3    274.60 1000.86  147.56    2.52   0.08706
ns(Rain0, df = 3)  3    332.31 1058.57  149.41    3.05   0.05231
ns(Rain2, df = 3)  3     70.61  796.87  140.04    0.65   0.59327
ns(Year, df = 3)   3   2022.93 2749.19  180.91   18.57 5.339e-06
```

CSIRO



Partial for ns(Temp4, df = 3)
10
0
-10
-20
-30
68
72
76
80
Temp4
Partial for ns(Rain0, df = 3)
10
5
0
-10
-20
15
20
25
Rain0
Partial for ns(Rain2, df = 3)
10
0
-10
-20
1
2
3
4
5
6
7
Rain2
Partial for ns(Year, df = 3)
30
20
10
0
-10
-30
1930
1940
1950
1960
Year

CSIRO



Partial for ns(Temp4, df = 3)
Temp4
68
72
76
80
Partial for ns(Rain0, df = 3)
Rain0
15
20
25
Partial for ns(Rain2, df = 3)
Rain2
1
2
3
4
5
6
7
Partial for ns(Year, df = 3)
Year
1930
1940
1950
1960
-30
-10
0
10
20
30

Final remarks

- Very similar pattern to the components as for the additive model
- Now clear that the term in `Rain2` is not useful and `Temp4` and `Rain0` terms will need to be re-assessed.
- The term in Year stands out as dominant with a clear pattern in the response curve and the partial residuals following it closely
- *Small data sets like this can be very misleading!* Extreme caution is needed.

Second example: Rock data (V&R p. 233 ff)

- Response: permeability
- Predictors: area, perimeter and shape
- Problem: build a predictor for log(perm) using the available predictors

```
rock.lm <- lm(log(perm) ~ area + peri + shape, data = rock)
summary(rock.lm)

Coefficients:
              Value Std. Error t value Pr(>|t|)
(Intercept)  5.3331  0.5487     9.7200  0.0000
       area  0.0005  0.0001     5.6021  0.0000
       peri -0.0015  0.0002    -8.6228  0.0000
      shape  1.7565  1.7559     1.0003  0.3226
```

Strategy

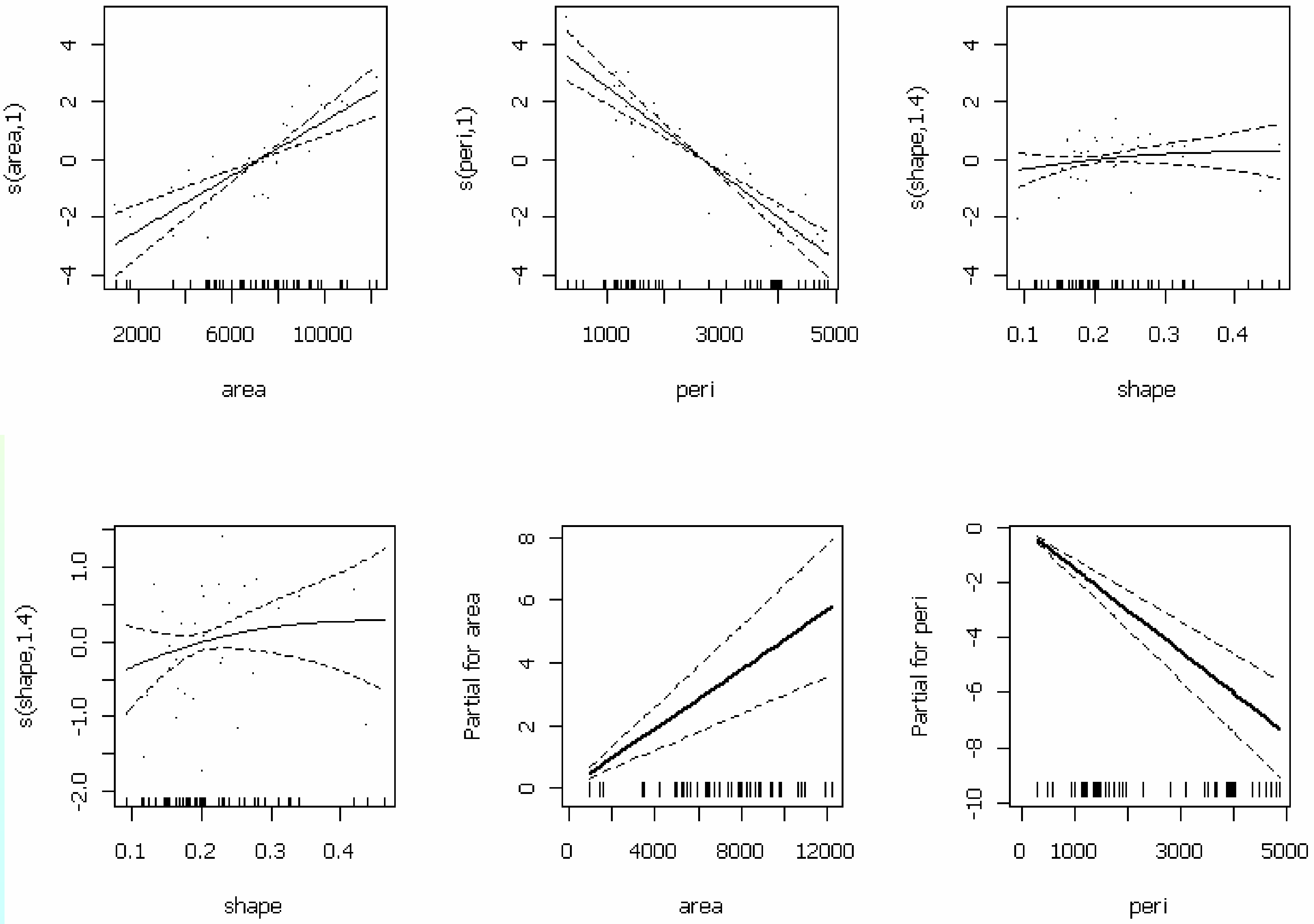
```
rock.gam <- gam(log(perm) ~ s(area) + s(peri) +
   s(shape),
  control = gam.control(maxit = 50), data = rock)
summary(rock.gam)
anova(rock.lm, rock.gam) # shows no improvement

par(mfrow = c(2, 3), pty = "s")
plot(rock.gam, se = T)
rock.gam1 <- gam(log(perm) ~ area + peri +
   s(shape), data = rock)
plot(rock.gam1, se = T)

anova(rock.lm, rock.gam1, rock.gam)
```

```
> summary(rock.gam)

Family: gaussian
Link function: identity

Formula:
log(perm) ~ s(area) + s(peri) + s(shape)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   5.1075     0.1222   41.81   <2e-16

Approximate significance of smooth terms:
          edf Est.rank      F  p-value
s(area)  1.000        1 29.878 2.09e-06
s(peri)  1.000        1 72.664 7.77e-11
s(shape) 1.402        3  1.324    0.279

R-sq.(adj) =  0.735   Deviance explained = 75.4%
GCV score = 0.78865   Scale est. = 0.71631   n = 48
```

Testing lm within a gam model

```
> anova(rock.lm, rock.gam)
Analysis of Variance Table

Model 1: log(perm) ~ area + peri + shape
Model 2: log(perm) ~ s(area) + s(peri) + s(shape)
    Res.Df    RSS       Df Sum of Sq      F Pr(>F)
1 44.00000 31.949
2 43.59763 31.230  0.40237     0.719 2.4951 0.1250
```

```
> summary(rock.gam1)

Family: gaussian
Link function: identity

Formula:
log(perm) ~ area + peri + s(shape)

Parametric coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.747e+00  3.615e-01  15.896  < 2e-16
area         4.727e-04  8.648e-05   5.466 2.09e-06
peri        -1.505e-03  1.766e-04  -8.524 7.77e-11

Approximate significance of smooth terms:
           edf Est.rank     F p-value
s(shape) 1.402        3 1.324   0.279

R-sq.(adj) =  0.735   Deviance explained = 75.4%
GCV score = 0.78865   Scale est. = 0.71631   n = 48
```



s(area,1)
area
2000
6000
10000
s(peri,1)
peri
1000
3000
5000
s(shape,1.4)
shape
0.1
0.2
0.3
0.4
s(shape,1.4)
shape
Partial for area
area
0
4000
8000
12000
Partial for peri
peri
0
1000
3000
5000

Comparing 3 models

```
> anova(rock.lm, rock.gam1, rock.gam)
Analysis of Variance Table

Model 1: log(perm) ~ area + peri + shape
Model 2: log(perm) ~ area + peri + s(shape)
Model 3: log(perm) ~ s(area) + s(peri) + s(shape)
       Res.Df    RSS          Df  Sum of Sq      F    Pr(>F)
1  4.4000e+01 31.949
2  4.3598e+01 31.230  4.0237e-01      0.719 2.4951    0.1250
3  4.3598e+01 31.230 -3.5094e-06 -5.028e-06 2.0001 2.107e-05
>
```

Lessons

- Although suggestive, the curve in shape is not particularly convincing.
- In this case, **`bruto`** also suggests essentially linear terms, at most, in all three variables (V&R p 235)