ORIGIN ≡ 0

General Additive Models using Regression Splines

In constructing a statistical model for non-linear relationships between a dependent variable and multiple independent variables, common strategies include transforming variables (such as taking the square root or log) or employing higher-order terms (involving squares, cubes) in a linear model. These strategies often work quite well. Such linear models and estimated coefficients, transformed back into the equations in units of the original variables, often form highly useful interpretations. (A good example of this is the so-called "allometric" equation in assessing measured shape relationships in organisms, derived from analysis of logged linear models.) However, when one wishes to describe a complex curvilinear relationship, but no *a priori* rules exist for higher-order terms, curvilinear curve fitting may be used instead as a heuristic. General Additive Models (GAM) form a flexible methodology for doing this. Employing cross validation, an automated methodology exists to more-or-less objectively decide upon the degree of curvature required for a specific curvilinear fit. In addition, evaluation of properly constructed GAMs, assessed with AIC or by formal Full Model versus Reduced model test, is substantially analogous to that utilized in linear models.

Shown here are are Additive Models (GAM with identity link) utilizing using the gam() function in the {mgcv} package in R (part of the base installation). The example is derived from Chapter 3 in Zuur et al. 2009, *Mixed Effects Models and Extensions in Ecology with R.* This book is highly recommended as readable introduction. However, changes in {mgcv} since publication of the book require reading of errata and use of modified R scripts found on their website: http://www.highstat.com

Example in R:

```
#LOADING ZUUR ET AL'S ISIT EXAMPLE DATA
#EXTRACTING DATA FOR STATIONS 8 & 13 ONLY
#IF PREFERRED THIS MIGHT BE DONE INSTEAD IN MS EXCEL
setwd("c:/DATA/Models")
ISIT=read.table("ISIT.txt",header=T)
ISIT
S8=ISIT$Sources[ISIT$Station==8]
D8=ISIT$SampleDepth[ISIT$Station==8]
S13=ISIT$Sources[ISIT$Station==13]
D13=ISIT$SampleDepth[ISIT$Station==13]
So=c(S8, S13);
De=c(D8, D13)
ID=rep(c(8, 13), c(length(S8), length(S13)))
mi=max(min(D8), min(D13))
ma=min(max(D8), max(D13))
I1=De > mi & De < ma
ISIT813=data.frame(So=So[I1],De=De[I1],ID=ID[I1])
ISIT813
#NEW DATASET ISIT813 WRITTEN TO DISK
write.table(ISIT813,file="ISIT813.txt")
```

Zuur et al. provide a highly useful set of instructions for extracting and working with a subsample from a larger dataset, shown here.

Their original dataset is available as a textfile in a zip folder, or in the R package {AED}. The R package must be downloaded from their website and installed using "Install package(s) from local zip files" option on the console main menu.

Using their data set ISIT.txt, information for observation stations 8 & 13 are extracted and then subsampled for the same range of values of De shared by the two sites. They then proceed to work with this subsample directly.

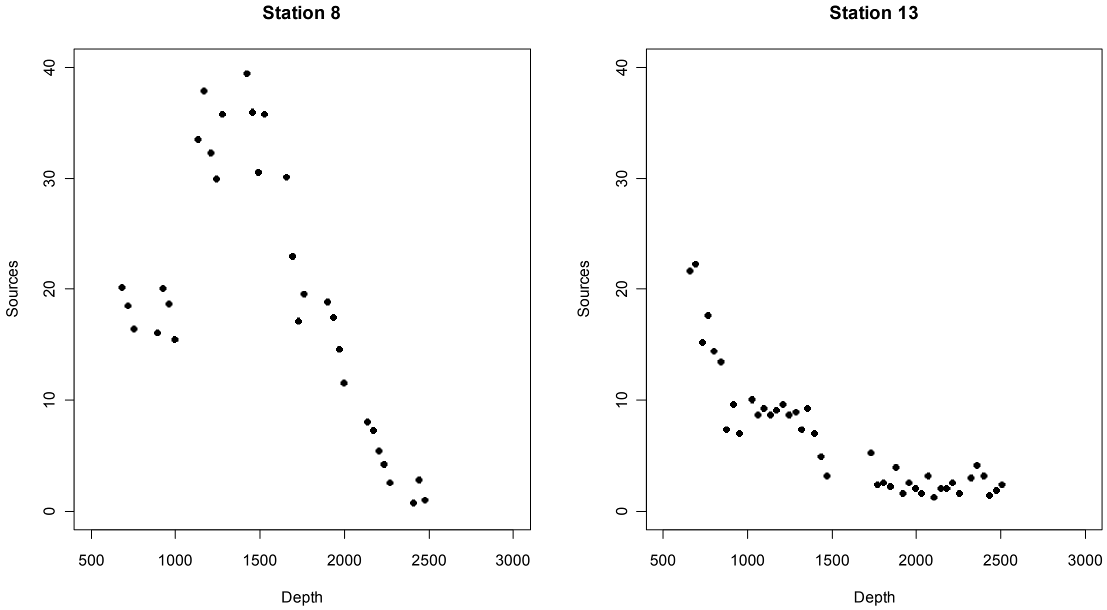
In what follows below, the subsample is written as the new dataset ISIT813.txt, and then treated as if the data were instead modified externally using, for instance, MS Excel. This approach separates the problem of using gam() from the problem of data manipulation in R.

```
##IMPORTANT - RESTART R AT THIS POINT
#THIS IS AS IF LOADING THE REVISED DATASET ISIT813 FROM SCRATCH

#READING ISIT813 SUBSAMPLE DATA TABLE FROM DISK
setwd("c:/DATA/Models")
I=read.table("ISIT813.txt",header=T)
I
```

Note: Restart R if proceeding with the accompaning script.

Start work fresh at this point.

```
#PLOTS FROM ISIT813 DATASET
op=par(mfrow = c(1, 2))
plot(I$De[I$ID==8], I$So[I$ID==8], pch = 16, xlab = "Depth",
   ylab = "Sources", col = 1, main = "Station 8",
   xlim = c(500, 3000), ylim = c(0, 40))

plot(I$De[I$ID==13], I$So[I$ID==13], pch = 16, xlab = "Depth",
   ylab = "Sources", col = 1, main = "Station 13",
   xlim = c(500, 3000), ylim = c(0, 40))
par(op)
```



```
> I
      So   De ID
1  39.43 1417  8
2  37.91 1167  8
3  35.93 1452  8
4  35.78 1276  8
5  35.78 1521  8
6  33.49 1132  8
7  32.28 1203  8
8  30.60 1487  8
9  30.14 1658  8
10 29.99 1239  8
11 22.99 1692  8
12 20.25  682  8
...
61  2.67 1805 13
62  2.67 1953 13
63  2.67 2213 13
64  2.48 1768 13
65  2.48 2503 13
66  2.29 1844 13
67  2.10 1991 13
68  2.10 2140 13
69  2.10 2175 13
70  1.90 2468 13
71  1.71 1918 13
72  1.71 2029 13
73  1.71 2250 13
74  1.52 2431 13
75  1.33 2102 13
```

^ Plots of original data

```
#GAM MODEL FOR ISIT813 SINGLE CURVE WITH OFFSET ONLY
library(mgcv)
attach(I)
fID=factor(ID)
M4=gam(So~s(De)+fID)
M4
anova(M4)
summary(M4)
```

The "Formula" follows normal R syntax showing So as the dependent variable and fID as a standard "parametric" factor. The s(De) term indicates a "smoother" (i.e., non-parametric curve fit) for numeric variable De. The P-value for parametric fID tests the null hypothesis H_0: no difference between stations 8 & 13 mean response.

Under description of the smooth term:
edf = "effective degrees of freedom":
edf = 1 linear relationship is sufficient.
edf > 1 higher numbers indicate greater amounts of curvature are required.

```
> anova(M4)

Family: gaussian
Link function: identity

Formula:
So ~ s(De) + fID

Parametric Terms:
    df     F  p-value
fID  1 77.46 7.59e-13

Approximate significance of smooth terms:
        edf Ref.df     F p-value
s(De) 4.849  5.904 14.77 1.2e-10
```

Test Statistic F and associated P-values are not adequately defined in Zuur et al., but according to the R documentation for summary.gam(), this tests the whether the smoother term s() should be in the model or not. Like the general F-test of nested models, the Null hypothesis H_0 (Reduced Model) excludes the s() term whereas H_1 includes it. P-values are only approximate and must be used with great caution when near 0.05.

> summary(M4)

```
Family: gaussian
Link function: identity

Formula:
So ~ s(De) + fID

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   19.198      1.054  18.207  < 2e-16 ***
fID13        -12.296      1.397  -8.801 7.59e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
        edf Ref.df     F p-value
s(De) 4.849  5.904 14.77 1.2e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.695   Deviance explained = 71.9%
GCV score = 38.802  Scale est. = 35.259    n = 75
```

Parametric coefficients for factors may be interpreted in the same way as in standard ANOVA. Because "treatment" contrasts is default in R, the estimate of (Intercept) gives mean value for station 8 and FID13 provides the difference between stations 13 and 8.

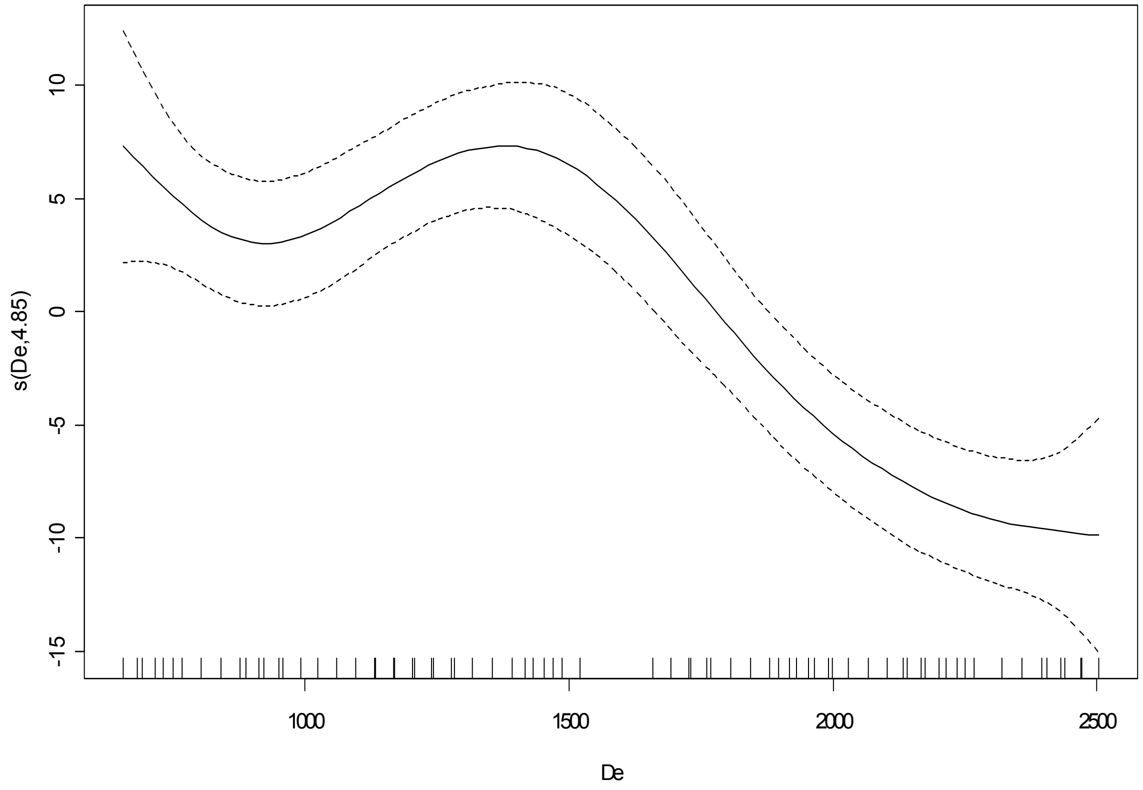
Scale est. = variance of the residuals.

GCV score is a measure that may be used like AIC in judging the relative merits of competing models. Smaller is always better.

plot(M4)

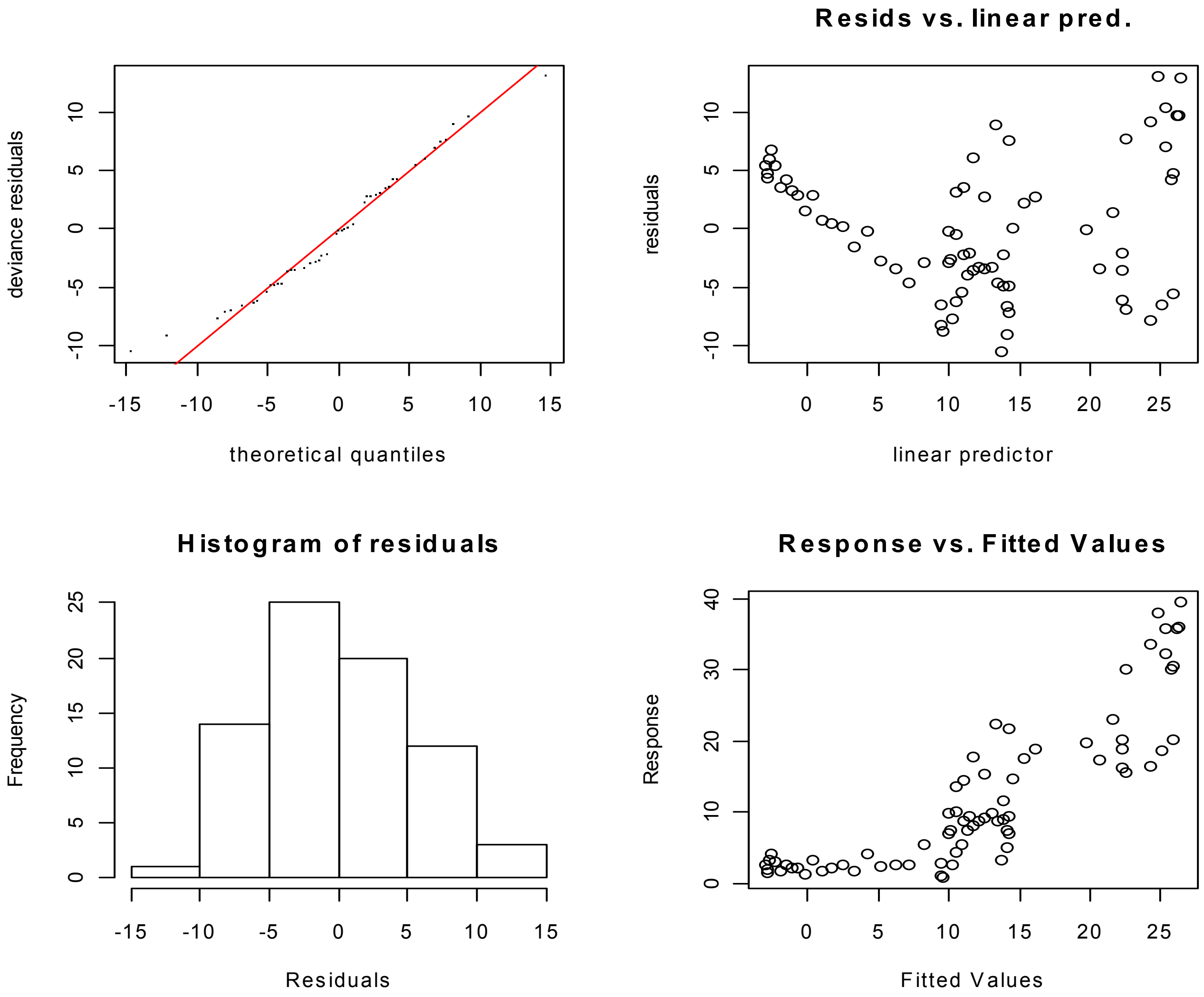
Plot of smoother s(De):

Predicted values of smoother s(De) values (Y-axis) are plotted against De (X-axis). Dashed lines indicate 95% confidence bands. Small hatching along X-axis indicate observed values of De in the dataset.



```
gam.check(M4)
```

Diagnostic plots for validation of the model. Zuur et al. would have us note that the two right-hand panels show patterns, and thus indicate the model needs further work by inclusion of additional independent variables or other corrections.



```
#GAM MODEL FOR ISIT813 WITH INTERACTION
M5=gam(So~s(De)+s(De,by=as.numeric(ID==13)))
#NOTE: CODE HERE IS MODIFIED ACCORDING TO ERRATA AT www.highstat.com
#NUMERICAL VALUES REPORTED CORRESPOND TO R OUTPUT FROM THE ONLINE
#R SCRIPT, BUT NOT TO THE VALUES REPORTED ON P.60
M5
anova(M5)
```

First of three models described as including "interaction" between variables De and fID.

```
> anova(M5)

Family: gaussian
Link function: identity

Formula:
So ~ s(De) + s(De, by = as.numeric(ID == 13))

Approximate significance of smooth terms:
                             edf Ref.df     F p-value
s(De)                      8.073  8.608 101.7  <2e-16
s(De):as.numeric(ID == 13) 7.696  8.663 113.7  <2e-16
```

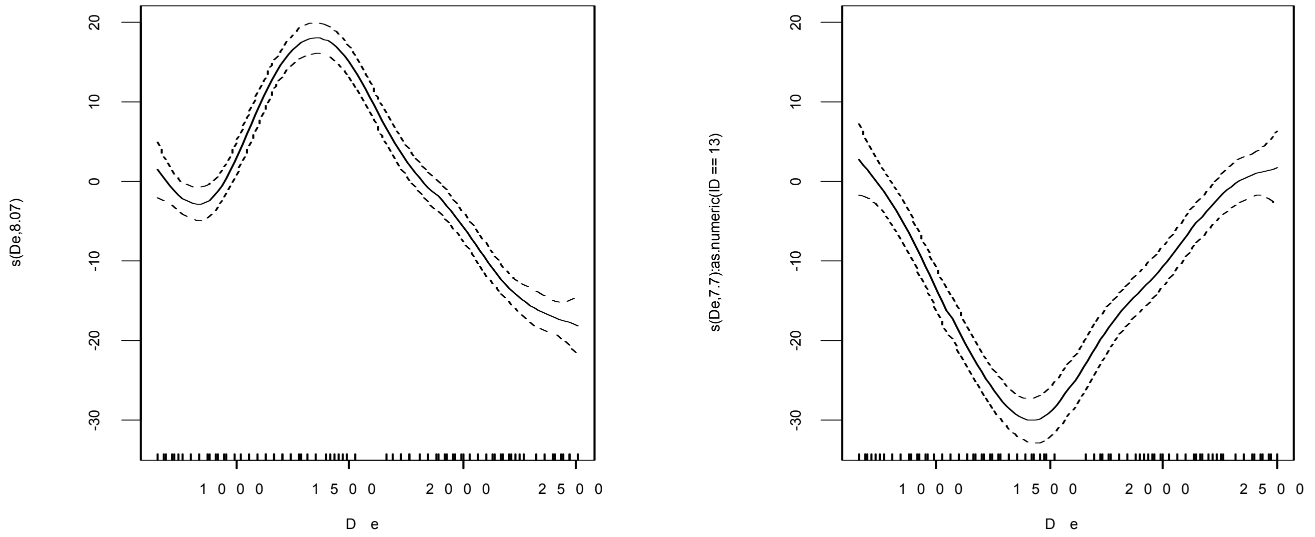
```
summary(M5)
```

```
> summary(M5)

Family: gaussian
Link function: identity

Formula:
So ~ s(De) + s(De, by = as.numeric(ID == 13))

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  18.8552     0.3879    48.6   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
                                edf Ref.df     F p-value
s(De)                         8.073  8.608 101.7  <2e-16 ***
s(De):as.numeric(ID == 13) 7.696  8.663 113.7  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =  0.959   Deviance explained = 96.8%
GCV score =  6.037  Scale est. = 4.6872    n = 75
```

Note: edf value for the second smoother is incorrect in Zuur et al.'s text, but the value here corresponds to R output running their modified R script found online.

```
op=par(mfrow = c(1, 2))
plot(M5)
par(op)
```

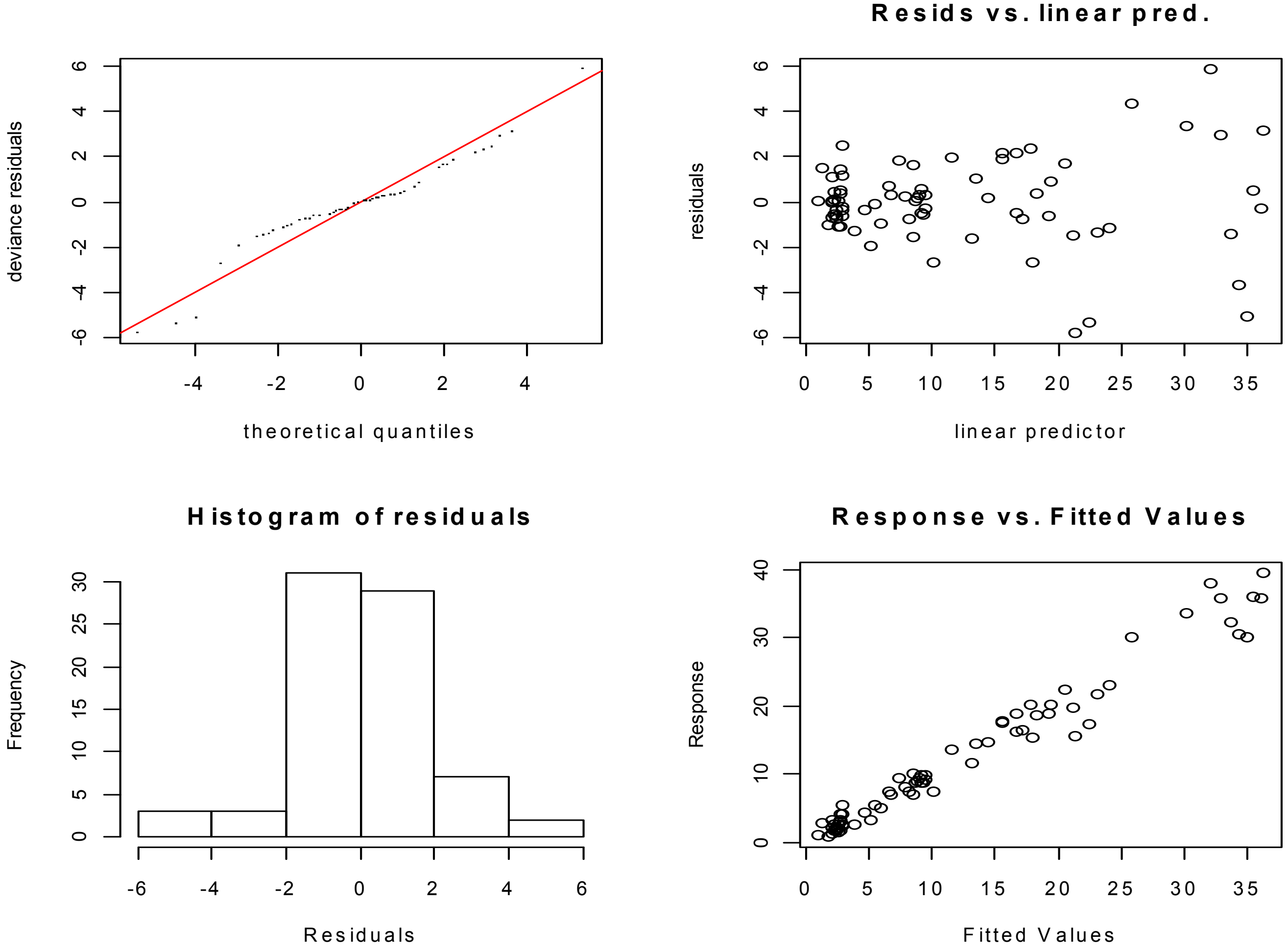
Left graph below shows the smoother for the combined data. Right graph below shows deviation from the overall pattern for data from station 13.



Model M4 has only one smoother and therefore implies only a single kind of non-linear response to De, with Stations 8 & 13 only allowed to differ by an offset (or difference) in their respective mean levels. By contrast, Model M5 contains a general smoother s(De) plus an offset smoother s(De, by ID=13) permitting data from station 13 to deviate from the general pattern in a specific way. This is one of three possible "interaction" models offered by Zuur et al. The others are models M6a & M6b given below.

```
gam.check(M5)
```

Zuur et al. will have us note much improved apparent randomness in the two right-hand panels.



Resids vs. linear pred.

Histogram of residuals

Response vs. Fitted Values

Testing Nested Models:

```
#F-TEST COMPARING NESTED MODELS
anova(M4,M5,test='F')
AIC(M4)
AIC(M5)
```

The Analysis of Deviance Table is treated analogously to a regular ANOVA table. Here the F-test uses as Null hypothesis the simpler model (M4) = Reduced Model (RM), with the more specified model M5 as the the Full Model (FM). This procedure is strictly analogous to the regular F-test of nested models.

```
> anova(M4,M5,test='F')
Analysis of Deviance Table

Model 1: So ~ s(De) + fID
Model 2: So ~ s(De) + s(De, by = as.numeric(ID == 13))
  Resid. Df Resid. Dev     Df Deviance      F    Pr(>F)
1    68.151    2402.90
2    58.231     272.94 9.9198     2130 45.809 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> AIC(M4)
[1] 488.5602
> AIC(M5)
[1] 345.2614
```

Both F-test and relative values of AIC (smaller is better) calculated for both models strongly indiate preference for M5 over M4.

Two other kinds of "interactions" models:

```
#TWO OTHER INTERACTION MODELS
M6a=gam(So~s(De,by = as.numeric(ID== 8))+s(De,by = as.numeric(ID== 13))-1)
summary(M6a)
anova(M6a)

M6b=gam(So~s(De,by = ID) + factor(ID))
summary(M6b)
anova(M6b)
```

```
> anova(M6b)
Family: gaussian
Link function: identity

Formula:
So ~ s(De, by = ID) + factor(ID)

Parametric Terms:
          df     F p-value
factor(ID)  1 121.4  <2e-16

Approximate significance of smooth terms:
          edf Ref.df     F p-value
s(De):ID 5.057  5.991 42.94  <2e-16
```

```
> summary(M6b)
Family: gaussian
Link function: identity

Formula:
So ~ s(De, by = ID) + factor(ID)

Parametric coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    11.089      1.006   11.03   <2e-16 ***
factor(ID)13  -17.678      1.604  -11.02   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1

Approximate significance of smooth terms:
          edf Ref.df     F p-value
s(De):ID 5.057  5.991 42.94  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1

R-sq.(adj) =  0.596   Deviance explained = 62.4%
GCV score = 50.736  Scale est. = 46.631    n = 75
```

```
> anova(M6a)
Family: gaussian
Link function: identity

Formula:
So ~ s(De, by = as.numeric(ID == 8)) + s(De, by =
as.numeric(ID ==
    13)) - 1

Approximate significance of smooth terms:
                             edf Ref.df     F p-value
s(De):as.numeric(ID == 8)  8.581  9.408 374.8  <2e-16
s(De):as.numeric(ID == 13) 7.520  8.647  77.7  <2e-16
```

```
> summary(M6a)
Family: gaussian
Link function: identity

Formula:
So ~ s(De, by = as.numeric(ID == 8)) + s(De, by =
as.numeric(ID ==
    13)) - 1

Approximate significance of smooth terms:
                             edf Ref.df     F p-value

s(De):as.numeric(ID == 8)  8.581  9.408 374.8  <2e-16
***
s(De):as.numeric(ID == 13) 7.520  8.647  77.7  <2e-16
***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
0.1 ' ' 1

R-sq.(adj) =   0.96   Deviance explained = 98.6%
GCV score = 5.8509  Scale est. = 4.5948     n = 75
```

```
anova(M5,M6a,M6b)
anova(M5,M6a)
anova(M5,M6b)
anova(M6a,M6b)
AIC(M5)
AIC(M6a)
AIC(M6b)
```

```
> AIC(M5)
[1] 345.2614
> AIC(M6a)
[1] 343.2881
>  AIC(M6b)
[1] 508.8192
```

M5 = So ~ s(De) + s(De, by ID13)
M6a = So ~ s(De,by ID8) + s(De,by ID13) - 1
M6b = So ~ s(De,by ID) + fID

The anova() functions provide deviance calculation but no test statistics, implying that the models do not nest. Relative values of AIC suggest slight preference for M6a over M5.

Interpretation of the three different "interactions" models needs confirmation from another sourse, such as Wood 2006, *Generalized Additive Models: An Introduction with R.*

However, by analogy with linear models, they appear readable:

M5 - models a general pattern plus offset for ID13.
M6a - models separate smoothers for ID8 & ID13 without intercept (Cell Means model analog).
M6b - models separate smoothers for ID8 & ID13 (perhaps as offsets from intercept) plus intercept.