NRES_798_12_201501

Generalized additive models

GAM

GAMs

Statistical power (prediction) vs. ecological interpretation

ELSEVIER

Ecological Modelling 157 (2002) 89–100

ECOLOGICAL MODELLING

www.elsevier.com/locate/ecolmodel

Generalized linear and generalized additive models in studies of species distributions: setting the scene

Antoine Guisan [a,b,*], Thomas C. Edwards, Jr [c], Trevor Hastie [d]

[a] *Swiss Center for Faunal Cartography (CSCF), Terreaux 14, CH-2000 Neuchâtel, Switzerland*
[b] *Institute of Ecology, University of Lausanne, BB, CH-1015 Lausanne, Switzerland*
[c] *USGS Biological Resources, Utah Cooperative Fish and Wildlife Research Unit, Utah State University, Logan, UT 84322-5210, USA*
[d] *Statistics Department, Stanford University, Sequoia Hall, Stanford, CA 94305, USA*

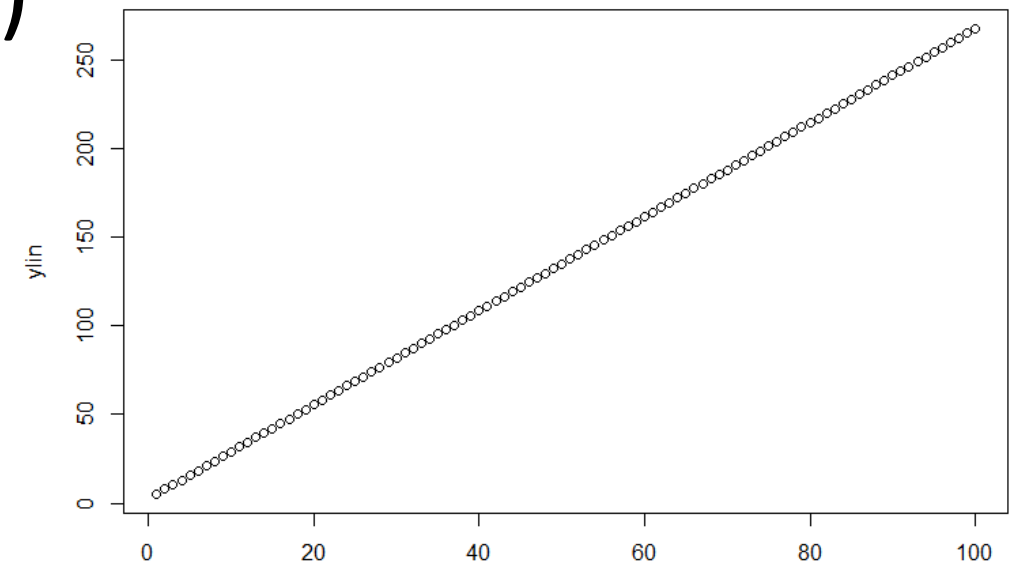
Generalized additive models

- Linear model (LM)
  - Models response variable using explanatory variables that are associated to the response variable using **specific parametric functions**
    - Linear, log, quadratic

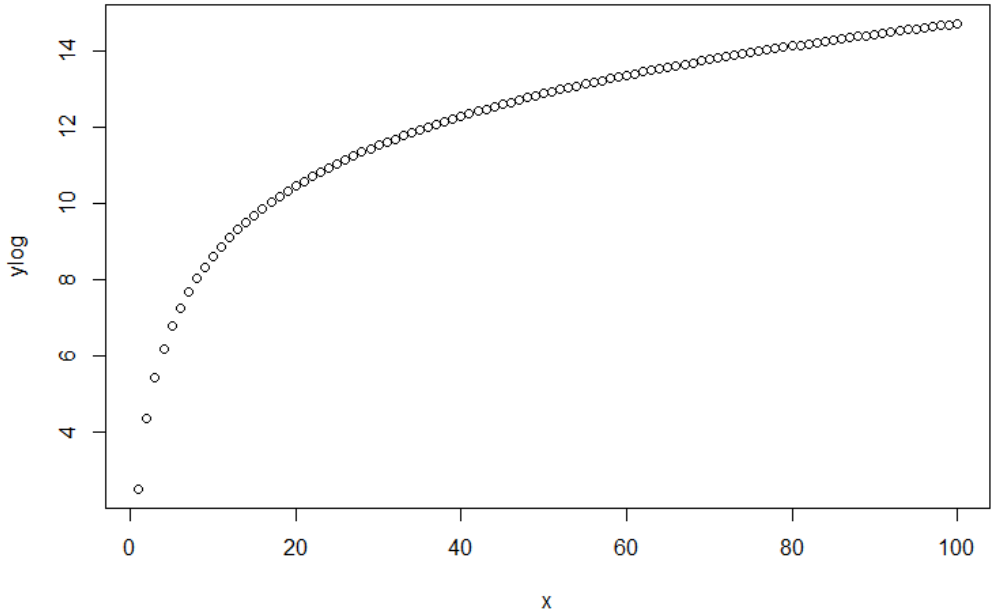
- Generalized additive models (GAM)
  - Models response variable using explanatory variables that are fitted as arbitrary smoothed functions using **non-parametric smoothers**

Linear models (lm)

```
# linear parametric functions
x <- seq(1,100,1)

ylin   <- 2.5 + 2.65*x
plot(x,ylin)
```

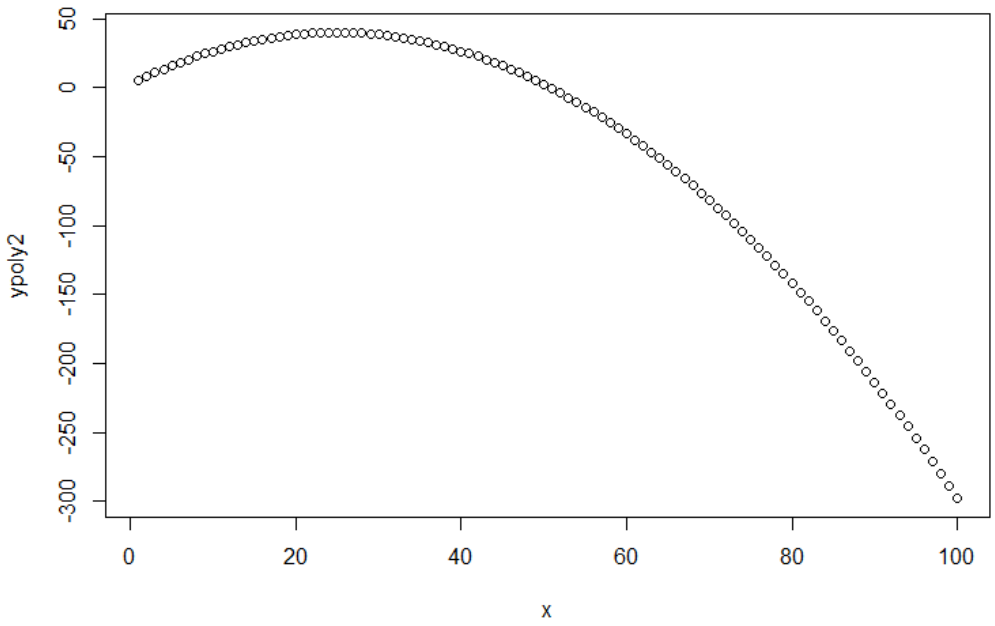



```
ylog   <- 2.5 + 2.65*log(x)
plot(x,ylog)
```

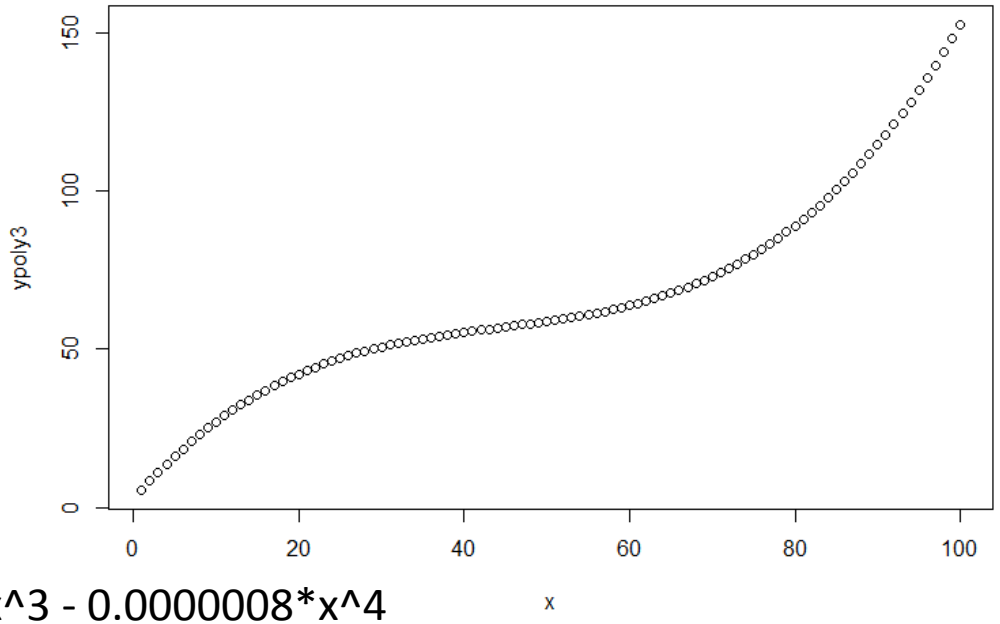


Parametric response functions

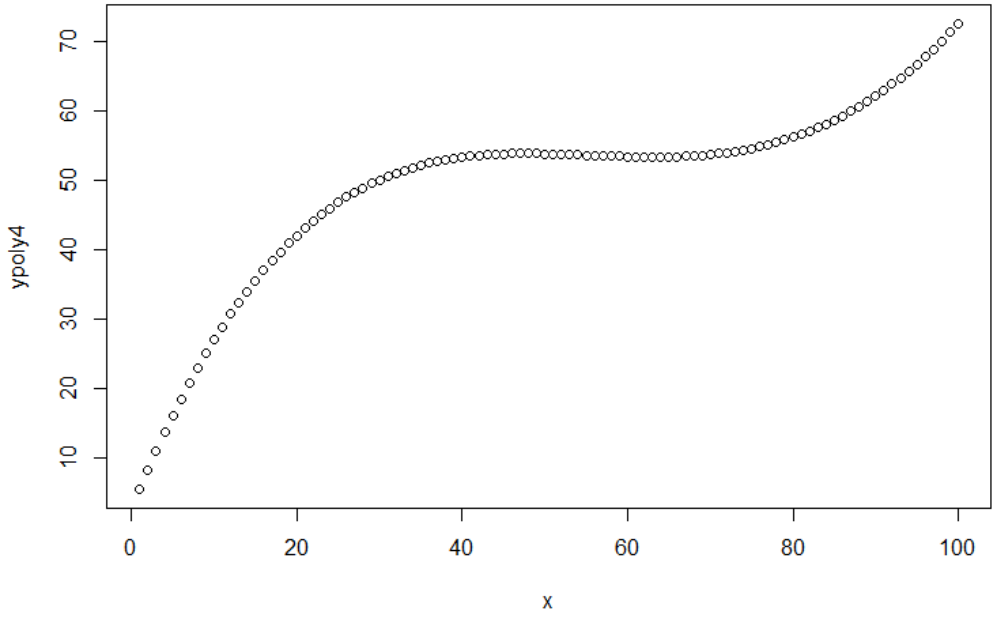
```
# linear parametric functions
x <- seq(1,100,1)

# Quadratic, second order polynomial

ypoly2 <- 2.5 + 3*x - 0.06*x^2
plot(x,ypoly2)
```



```
ypoly3 <- 2.5 + 3*x - 0.06*x^2 + 0.00045*x^3
plot(x,ypoly3)   # cubic, third order polynomial
```




```
ypoly4 <- 2.5 + 3*x - 0.06*x^2 + 0.00045*x^3 - 0.0000008*x^4
plot(x,ypoly4)
```

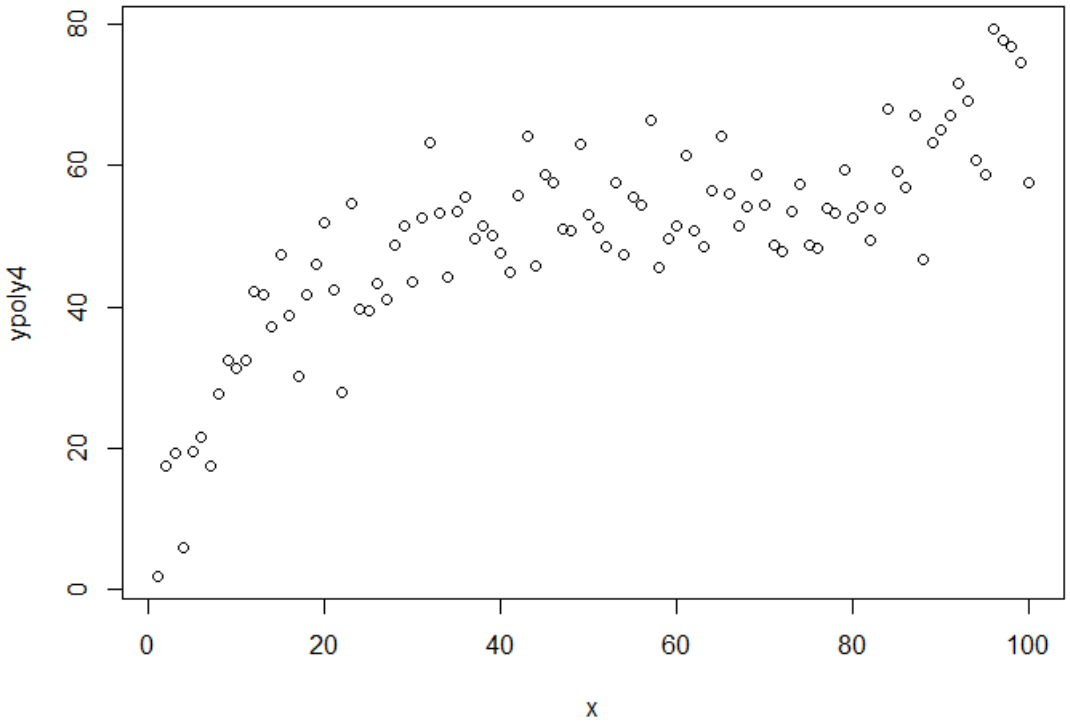


No a-priori relationship between x and y

Non-parametric smoothed functions

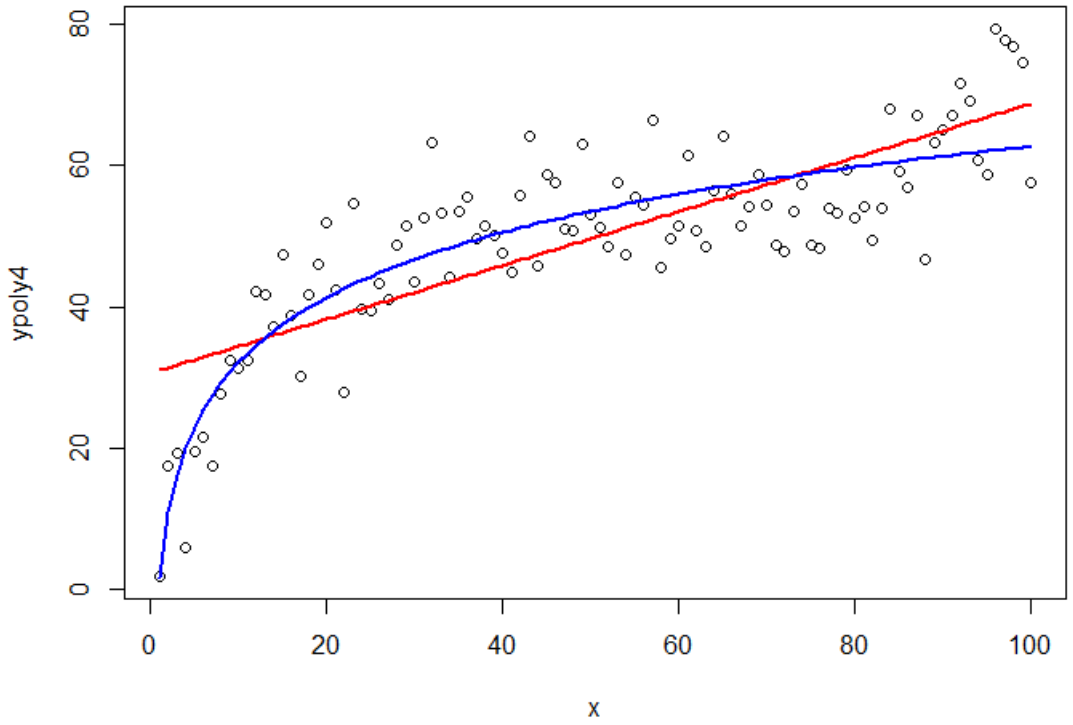
```
ypoly4 <- 2.5 + 3*x - 0.06*x^2 + 0.00045*x^3 - 0.0000008*x^4 + rnorm(length(x),0,6)
plot(x,ypoly4)
```

Let the data determine the shape of the response cure



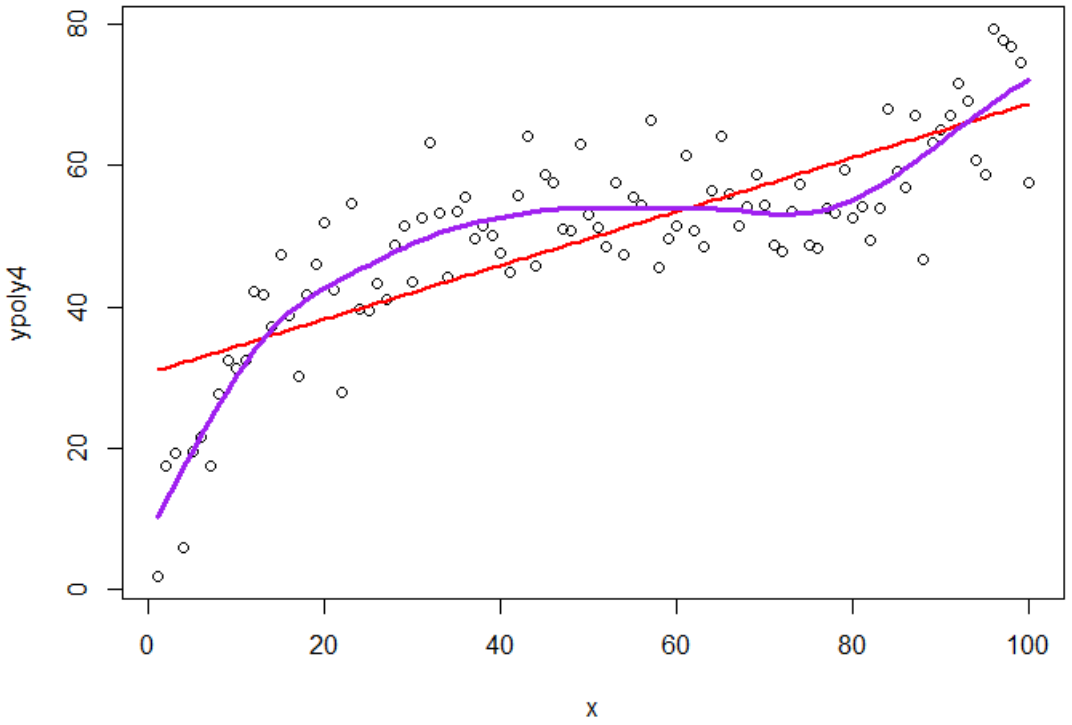
Non-parametric smoothed functions

```
ypoly4 <- 2.5 + 3*x - 0.06*x^2 + 0.00045*x^3 - 0.0000008*x^4 + rnorm(length(x),0,6)

ylm1 <- lm(ypoly4 ~ x)
ylm2 <- lm(ypoly4 ~ log(x))

plot(x,ypoly4)
lines(x,ylm1$fitted.values,col="red",lwd=2)
lines(x,ylm2$fitted.values,col="blue",lwd=2)
```



Non-parametric smoothed functions

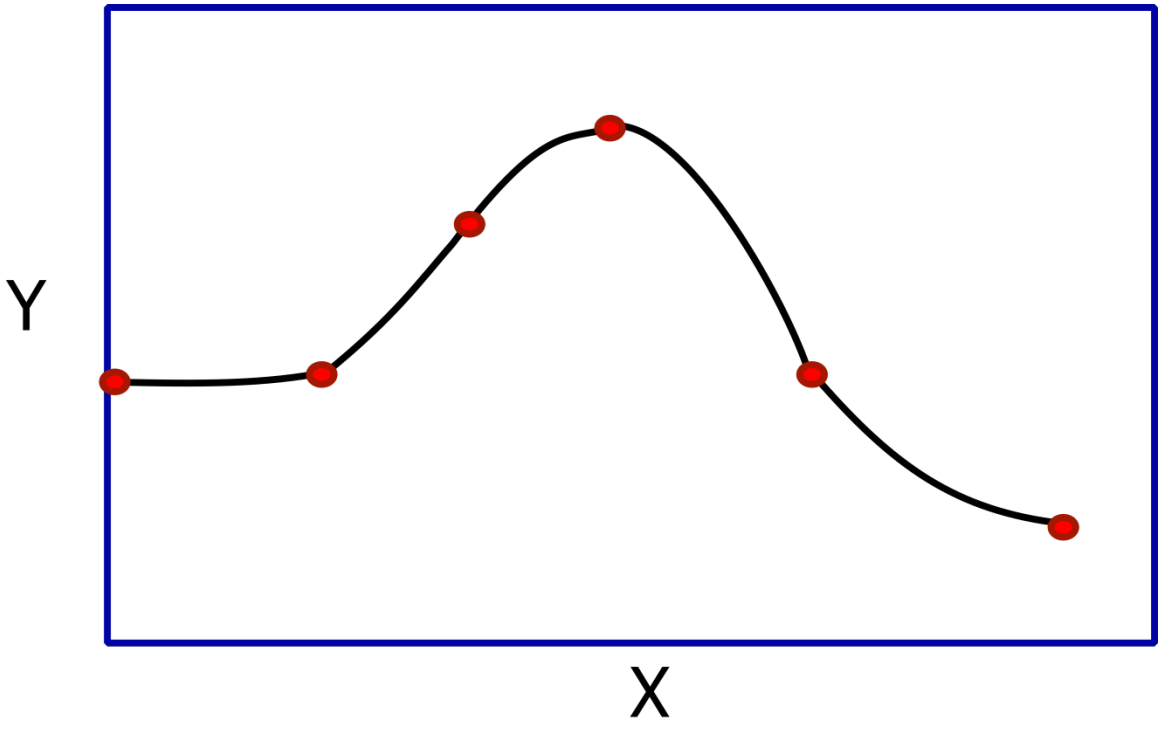
```
ypoly4 <- 2.5 + 3*x - 0.06*x^2 + 0.00045*x^3 - 0.0000008*x^4 + rnorm(length(x),0,6)

ylm1 <- lm(ypoly4 ~ x)
spline1 <- smooth.spline(x,ypoly4)

plot(x,ypoly4)
lines(x,ylm1$fitted.values,col="red",lwd=2)
lines(predict(spline1,x) ,col="purple",lwd=3)
```



Non-parametric smoothed functions

- Spline
  - A numeric function that is **piecewise-defined** by **polynomial functions**, and which possesses a sufficiently high degree of **smoothness** at the places where the polynomial pieces connect (knots).

Spline



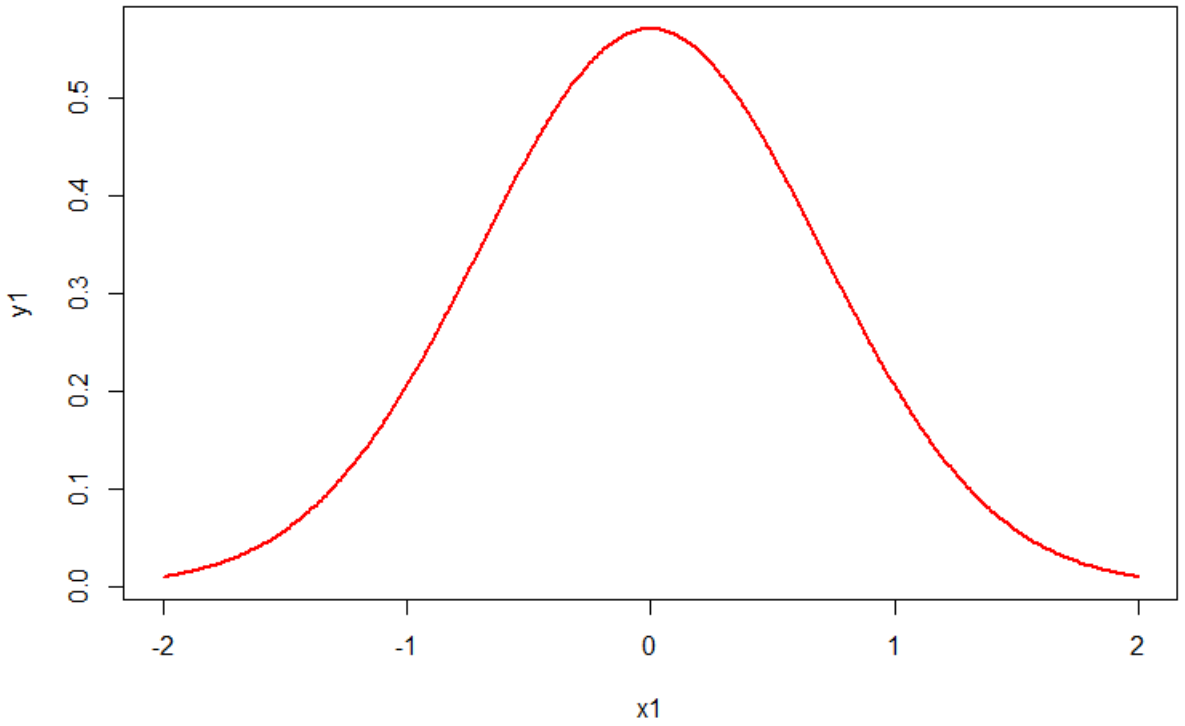
- Knots

polynomial function

cubic splines: third order polynomials

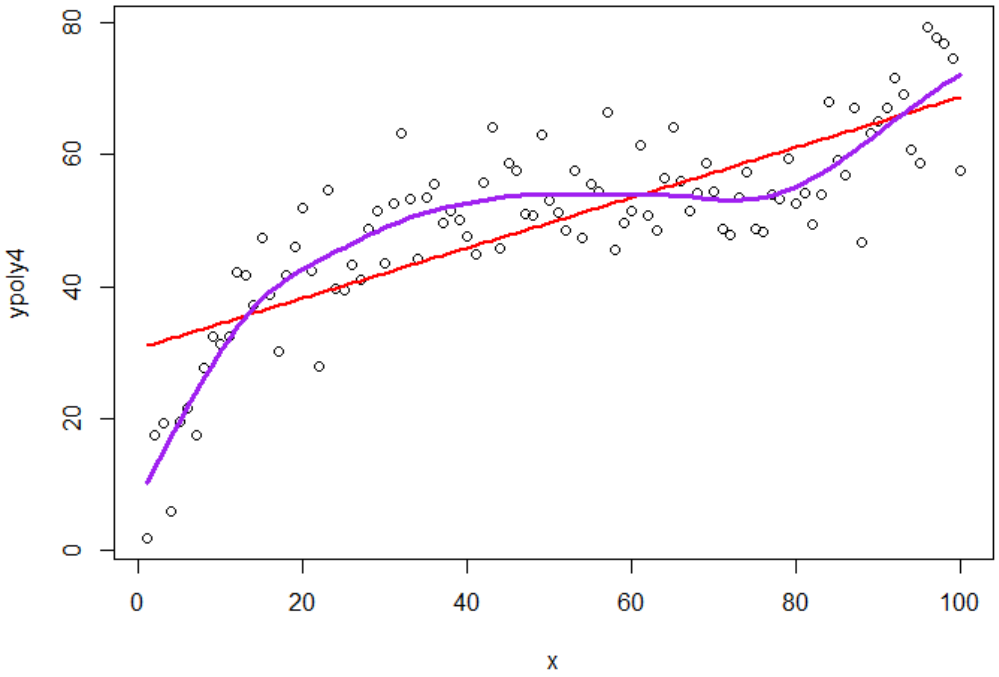
Splines

$$f_X(x) = \begin{cases} \frac{1}{4}(x+2)^3 & -2 \leq x \leq -1 \\ \frac{1}{4}\left(3|x|^3 - 6x^2 + 4\right) & -1 \leq x \leq 1 \\ \frac{1}{4}(2-x)^3 & 1 \leq x \leq 2 \end{cases}$$



Overfitting (noise vs. process)?

- Penalizing badness of fit?
- Penalizing wiggleness?



Generalized additive models

- Linear model:
  - Linear combination of parametric terms predict response variable
- Generalized linear model:
  - Parametric predictor terms related to a response variable through a linker function
- Generalized additive model:
  - A GAM is a generalized linear model in which the linear predictor is given by a user specified sum of smooth functions of the covariates plus a conventional parametric component of the linear predictor

lm, glm, gam

$$y_i = B_0 + B_1 x_i + \cdots + B_p X_p + \varepsilon_i$$

$$g(y_i) = n_i = B_0 + B_1 x_i + \cdots + B_p X_p + \varepsilon_i$$

$$g(y_i) = n_i = B_0 + f_1(x_{1i}) + \cdots + f_p(x_{1p}) + \varepsilon_i$$

GAM is a GLM where the linear predictor depends on smooth functions of covariates

GAM specification

```
install.packages("mgcv")
library(mgcv)
```

$y \sim s(x)$

Single non-parametrically smoothed function

$y \sim s(x) + s(w) + s(z)$

Multiple smoothed functions

$y \sim s + s(w)$

Mixture of parametric parameters and smoothed function

$y \sim s(x) + s(z) + s(x, z)$

Inclusion of nested smoothed terms (two-dimensional)

GAM specification

```
gm <- gam(growth ~ s(temp) + s(precip), family=gaussian())
```

Non-parametric smoothed function

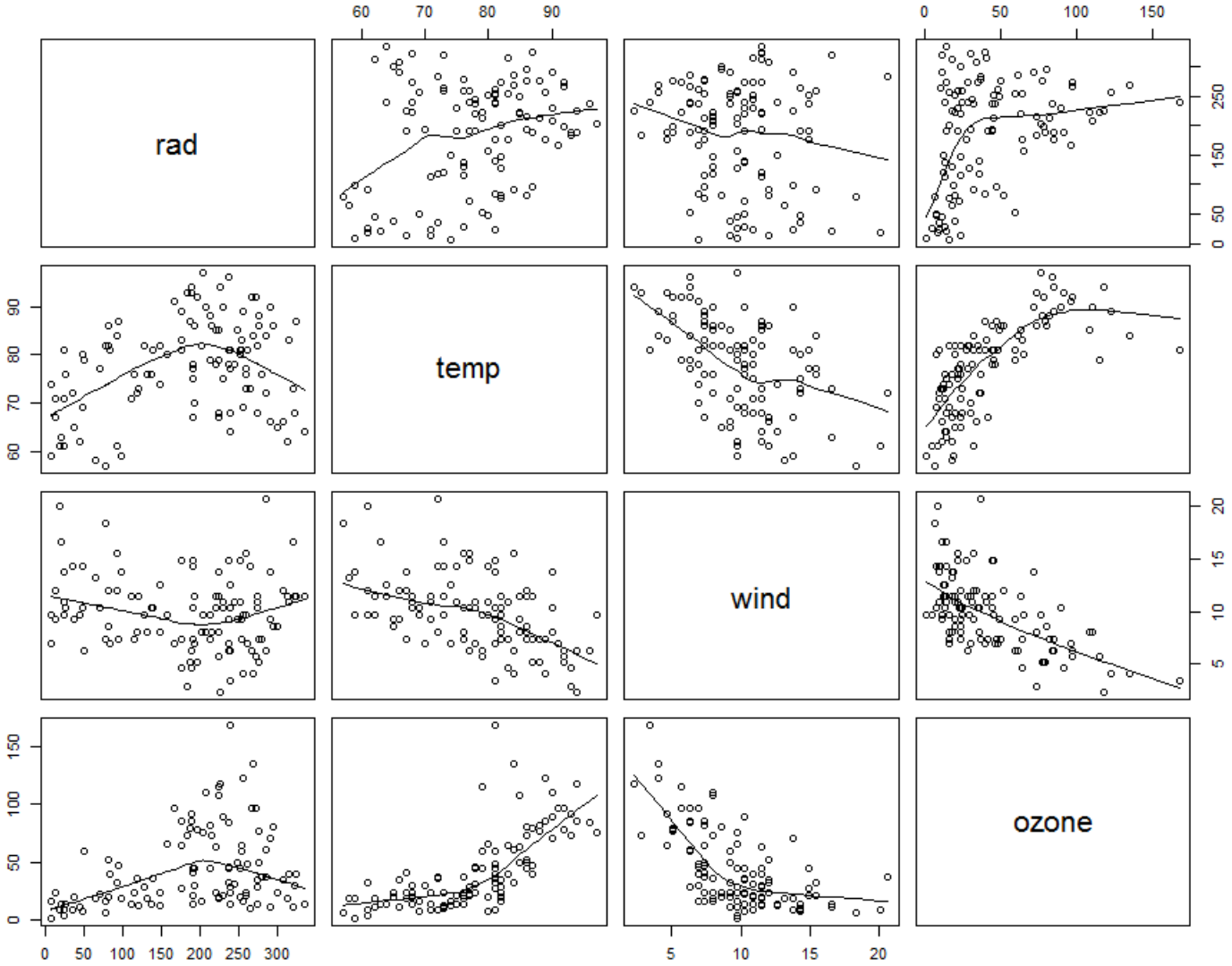
Linker function

GAM example 1

- Local ozone concentration

rad	temp	wind	ozone
Min. : 7.0	Min. :57.00	Min. : 2.300	Min. : 1.0
1st Qu.:113.5	1st Qu.:71.00	1st Qu.: 7.400	1st Qu.: 18.0
Median :207.0	Median :79.00	Median : 9.700	Median : 31.0
Mean :184.8	Mean :77.79	Mean : 9.939	Mean : 42.1
3rd Qu.:255.5	3rd Qu.:84.50	3rd Qu.:11.500	3rd Qu.: 62.0
Max. :334.0	Max. :97.00	Max. :20.700	Max. :168.0

Local ozone concentration



Local ozone concentration

```
model <- gam(ozone ~ s(rad) + s(temp) + s(wind))
summary(model)
```

Family: gaussian
Link function: identity

Formula:
ozone ~ s(rad) + s(temp) + s(wind)

Parametric coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	42.10	1.66	25.36	<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(rad)	2.763	3.451	3.823	0.00931 **
s(temp)	3.841	4.762	11.767	7.18e-09 ***
s(wind)	2.918	3.666	13.770	1.46e-08 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

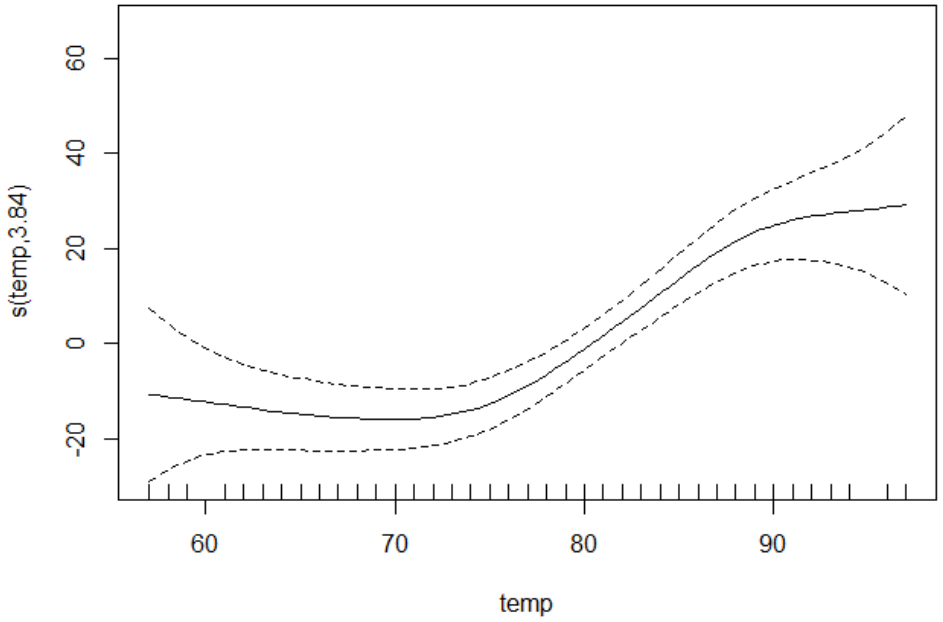
R-sq.(adj) = 0.724 Deviance explained = 74.8%
GCV = 338 Scale est. = 305.96 n = 111

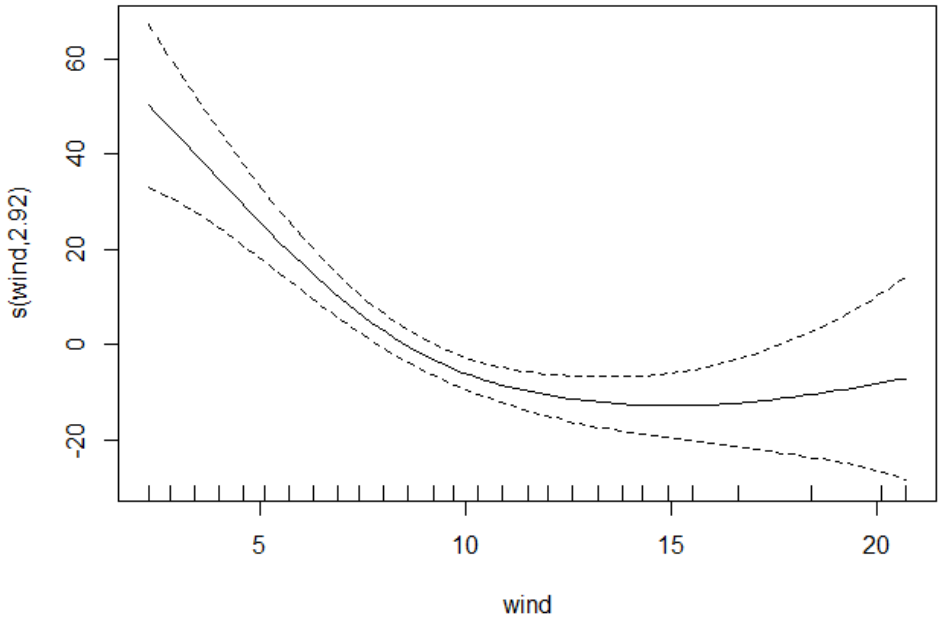
Local ozone concentration

plot(model)

Temp	wind
Min. :57.00	Min. : 2.300
1st Qu.:71.00	1st Qu.: 7.400
Median :79.00	Median : 9.700
Mean :77.79	Mean : 9.939
3rd Qu.:84.50	3rd Qu.:11.500
Max. :97.00	Max. :20.700

ozone
Min. : 1.0
1st Qu.: 18.0
Median : 31.0
Mean : 42.1
3rd Qu.: 62.0
Max. :168.0





GAM example 2 (binary data)

Presence
Absence

incidence
Min. :0.00
1st Qu.:0.00
Median :1.00
Mean :0.58
3rd Qu.:1.00
Max. :1.00

area
Min. :0.153
1st Qu.:2.248
Median :4.170
Mean :4.319
3rd Qu.:6.431
Max. :9.269

isolation
Min. :2.023
1st Qu.:4.823
Median :5.801
Mean :5.856
3rd Qu.:7.191
Max. :9.577

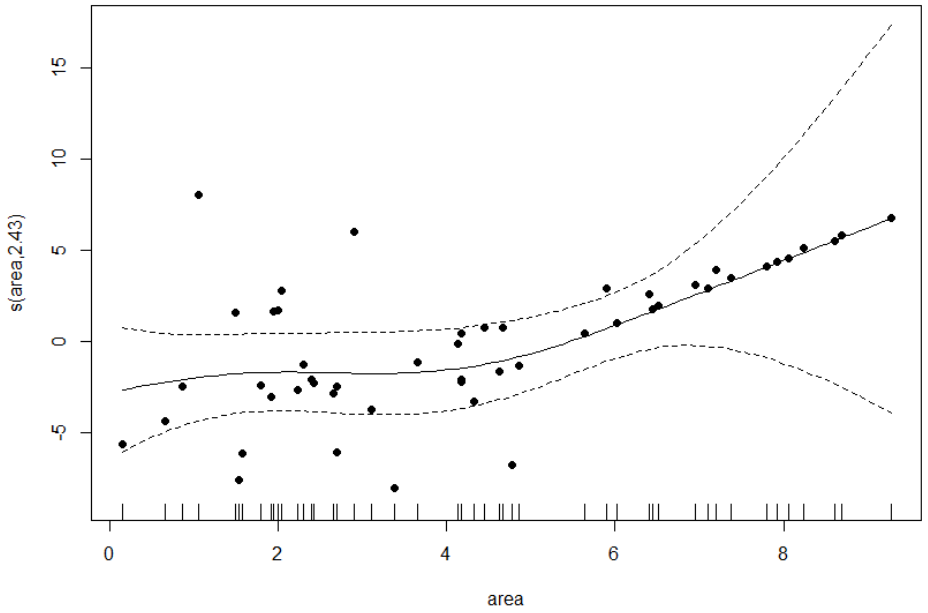
Species occurrence

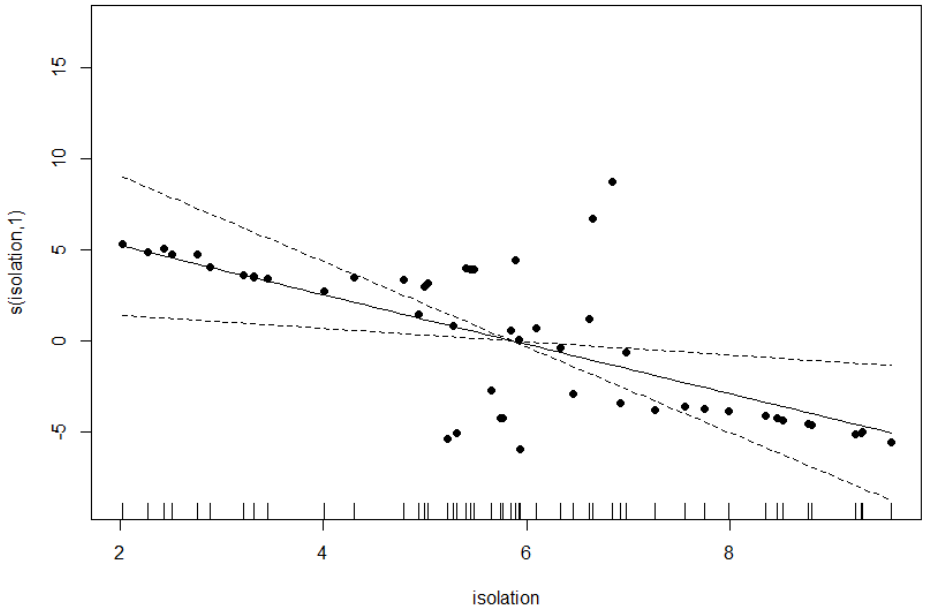
```
model <- gam(incidence ~ s(area) + s(isolation), family = binomial)
summary(model)
```

```
Family: binomial
Link function: logit

Formula:
incidence ~ s(area) + s(isolation)

Parametric coefficients:
             Estimate   Std. Error   z value   Pr(>|z|)
(Intercept)  1.6371     0.9898       1.654     0.0981 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
             edf        Ref.df       Chi.sq    p-value
s(area)      2.429      3.066        3.455     0.33606
s(isolation) 1.000      1.000        7.480     0.00624 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

R-sq.(adj) =   0.63   Deviance explained = 63.1%
UBRE = -0.32096  Scale est. = 1        n = 50
```

Species occurrence

plot.gam(model,residuals=T,pch=16)

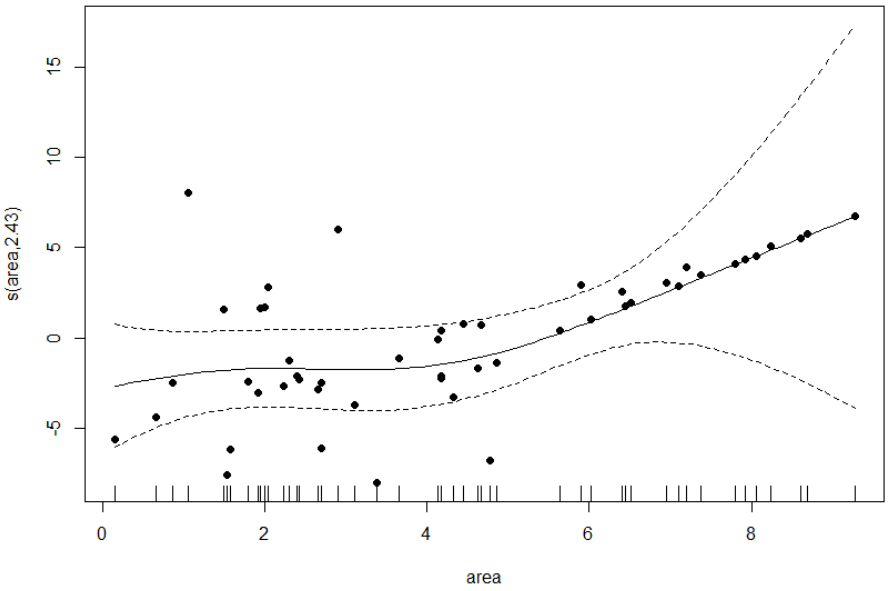




Species occurrence

```
model1 <- gam(incidence~s(area) + isolation,family=binomial)
summary(model1)
```

```
Family: binomial
Link function: logit

Formula:
incidence ~ s(area) + isolation

Parametric coefficients:
             Estimate   Std. Error  z value    Pr(>|z|)
(Intercept)  9.5755     3.1859      3.006      0.00265 **
isolation    -1.3555    0.4956      -2.735     0.00624 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Approximate significance of smooth terms:
            edf     Ref.df  Chi.sq  p-value
s(area)     2.429   3.066   3.555   0.323

R-sq.(adj) =  0.63   Deviance explained = 63.1%
UBRE = -0.32096  Scale est. = 1       n = 50
```



GAM vs. GLM

- GAM
  - Good
    - Extremely flexible models for fitting smooth curves
    - No a-priori assumptions of response curve shape
    - With ecological data the fit of GAM is often better than GLM or LM
  - Bad
    - Hard to evaluate (except graphically)
      - Can't provide an equation for a paper
    - Hard (impossible) to interpret and understand
      - Biological and physical processes underlying response curve
    - Computationally intensive (not an issue normally)
  - Statistical power vs. ecological interpretability

GAM details

Wood S.N. (2006)
Generalized Additive Models: An Introduction with R.
Chapman and Hall/CRC Press