Stat 8053, Fall 2013: Generalized Additive Models

For generalized additive models, we have a linear predictor,

$$\eta(x) = \beta_0 + \sum_{j=1}^{p} s_j(x)$$
$$= \beta_0 + \sum_{j=1}^{p} \sum_{j=1}^{d_j} \beta_{jk}\phi_{jk}(x)$$

Assuming the ϕs and d_j are known, by selecting a link function and an appropriate error distribution we could fit a generalized linear model. For a gam, we maximize the penalized likelihood function,

$$\ell_p(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - \frac{1}{2}\sum_j \lambda_j \boldsymbol{\beta}_j' B_j \boldsymbol{\beta}_j$$

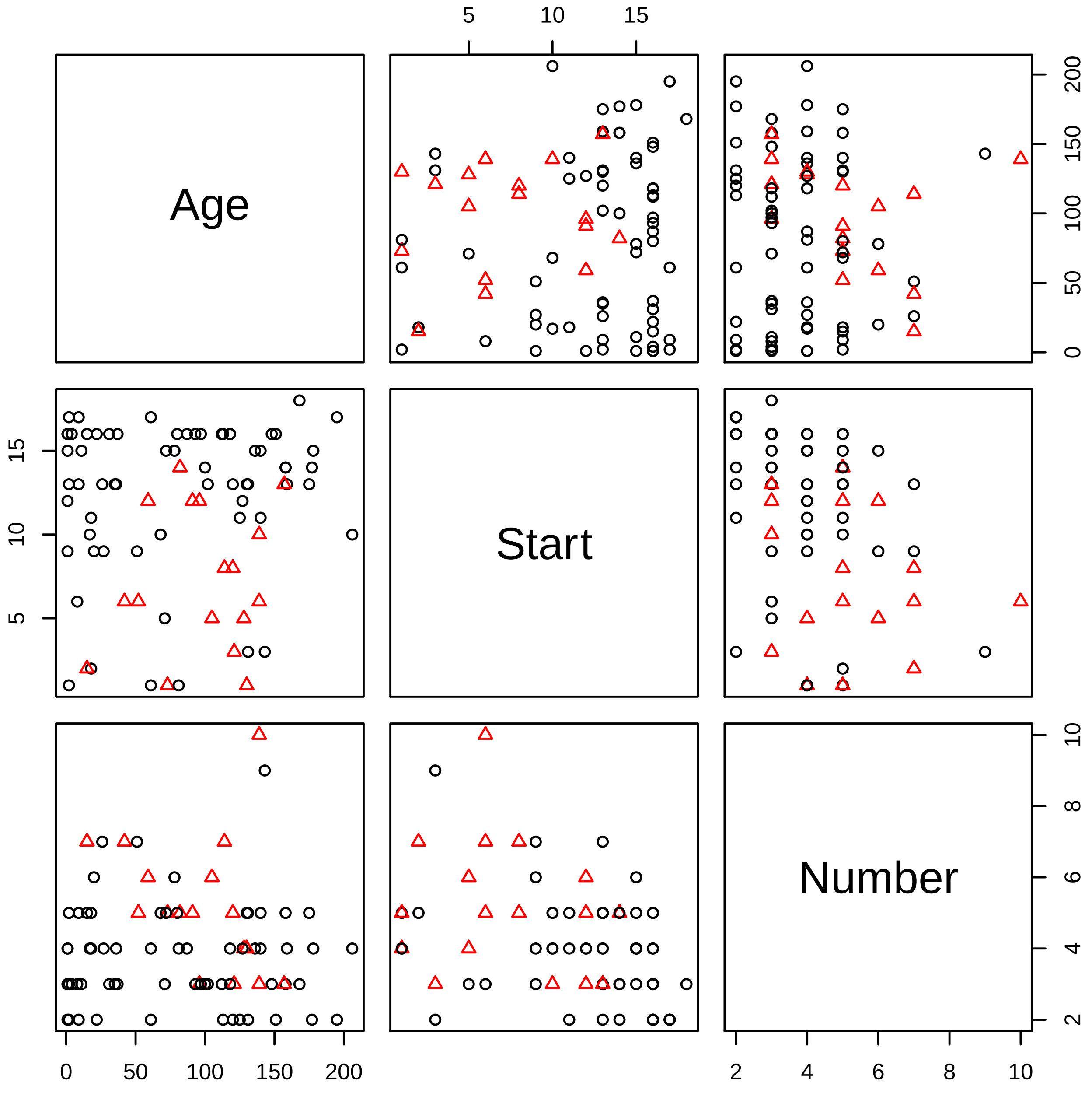
where $\ell_p(\boldsymbol{\beta})$ is the log-likelihood for the generalized linear model, B_j is a known matrix, λ_j is the smoothing parameter for the j-th smooth, the penalty has a negative sign because the log-likelihood is to be maximized rather than minimized as for least squares. The fraction 1/2 is unimportant but it makes the log-likelihood match the least square objective function for normal data.

```
data(kyphosis, package="gam")
str(kyphosis)
```

```
'data.frame':        81 obs. of  4 variables:
 $ Kyphosis: Factor w/ 2 levels "absent","present": 1 1 2 1 1 1 1 1 1 2 ...
 $ Age     : int  71 158 128 2 1 1 61 37 113 59 ...
 $ Number  : int  3 3 4 5 4 2 2 3 2 6 ...
 $ Start   : int  5 14 5 1 15 16 17 16 16 12 ...
```

These data are on the results of a spinal "laminectomy" on children to correct a condition called *kyphosis*, curvature of the spine. The response is presence/absence of kyphosis after surgery. Predictors are `Age` if the child, the `Start`ing vertebrae number, and the `Number of vertebra` effected.

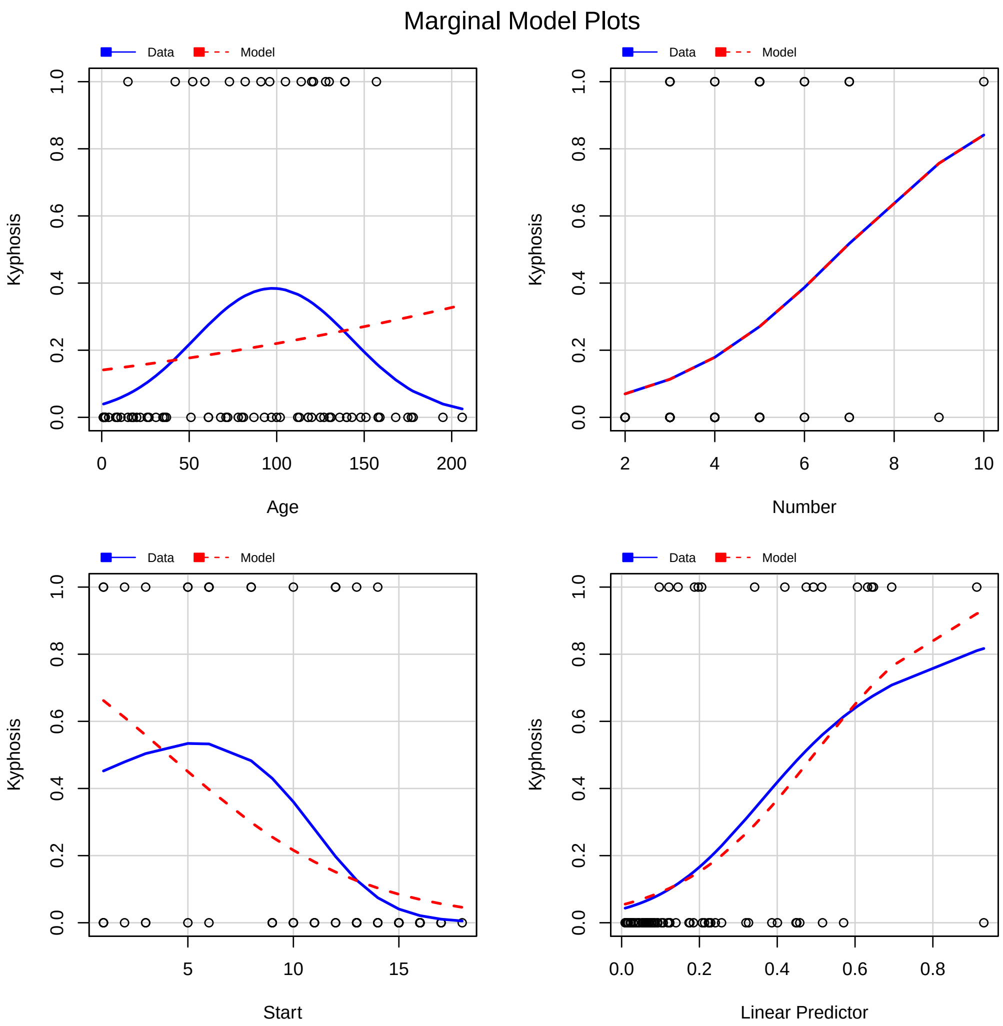
```
pairs(~ Age + Start + Number, kyphosis, col=as.numeric(kyphosis$Kyphosis),
    pch=as.numeric(kyphosis$Kyphosis))
```


Age
Start
Number
5
10
15
0
50
100
150
200
2
4
6
8
10

```
library(car)
summary(m0 <- glm(Kyphosis ~ Age + Number + Start, data=kyphosis, family=binomial))
```

```
Call:
glm(formula = Kyphosis ~ Age + Number + Start, family = binomial,
    data = kyphosis)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.312  -0.548  -0.363  -0.166   2.161

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.03693    1.44957   -1.41   0.1600
Age          0.01093    0.00645    1.70   0.0900
Number       0.41060    0.22486    1.83   0.0678
Start       -0.20651    0.06770   -3.05   0.0023

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 83.234  on 80  degrees of freedom
Residual deviance: 61.380  on 77  degrees of freedom
AIC: 69.38

Number of Fisher Scoring iterations: 5
```

```
mmps(m0)
```



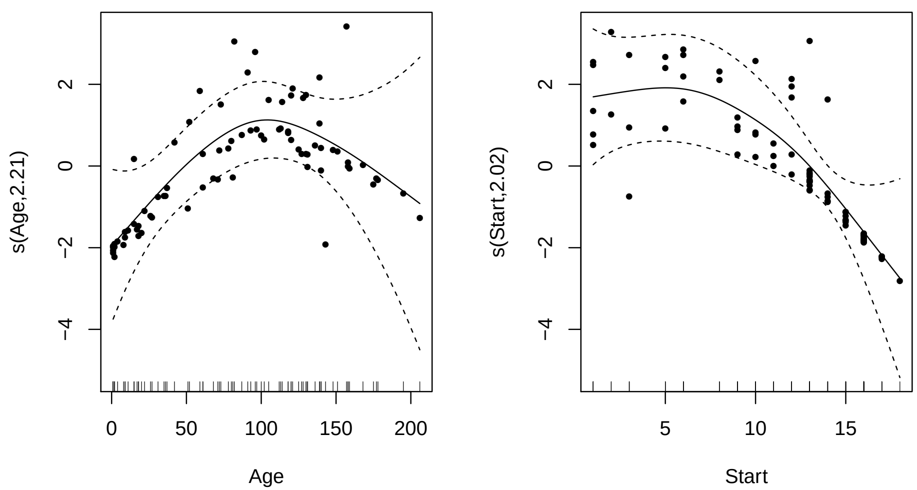
There appears to be an obvious problem with `Age`, and possible `Start`.

```
library(mgcv)
m1 <- gam(Kyphosis ~ s(Age) + s(Start) + Number, data=kyphosis, family=binomial)
summary(m1)
```

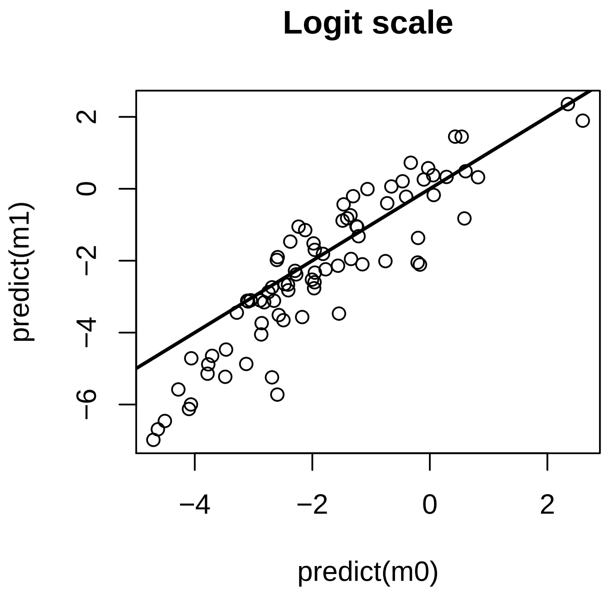
```
Family: binomial
Link function: logit
```

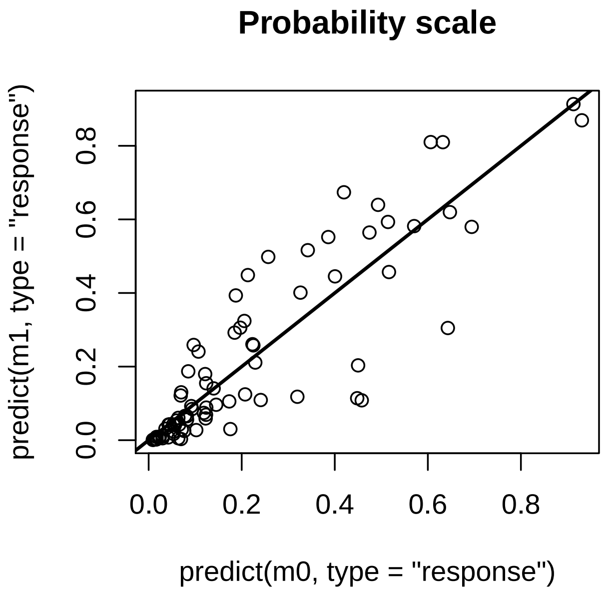
```
Formula:
Kyphosis ~ s(Age) + s(Start) + Number

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.593      1.146   -3.13   0.0017
Number         0.333      0.232    1.43   0.1515

Approximate significance of smooth terms:
          edf Ref.df Chi.sq p-value
s(Age)   2.21   2.79   6.30   0.084
s(Start) 2.02   2.52   9.76   0.014

R-sq.(adj) =  0.355   Deviance explained = 39.4%
UBRE score = -0.22384  Scale est. = 1         n = 81
```

```
plot(m1, residuals=TRUE, pch=16, cex=.7, pages=1)
```



```
par(mfrow=c(1, 2))
plot(predict(m1) ~ predict(m0), main="Logit scale")
abline(0, 1, lwd=2)
plot(predict(m1, type="response") ~ predict(m0, type="response"), main="Probability scale")
abline(0, 1, lwd=2)
```






```
m2 <- update(m1, ~ . - s(Start) + Start)
anova(m2, m1, test="Chisq")
```

```
Analysis of Deviance Table

Model 1: Kyphosis ~ s(Age) + Number + Start
Model 2: Kyphosis ~ s(Age) + s(Start) + Number
  Resid. Df Resid. Dev  Df Deviance Pr(>Chi)
1      75.9       55.1
2      74.8       50.4 1.1     4.64    0.036
```

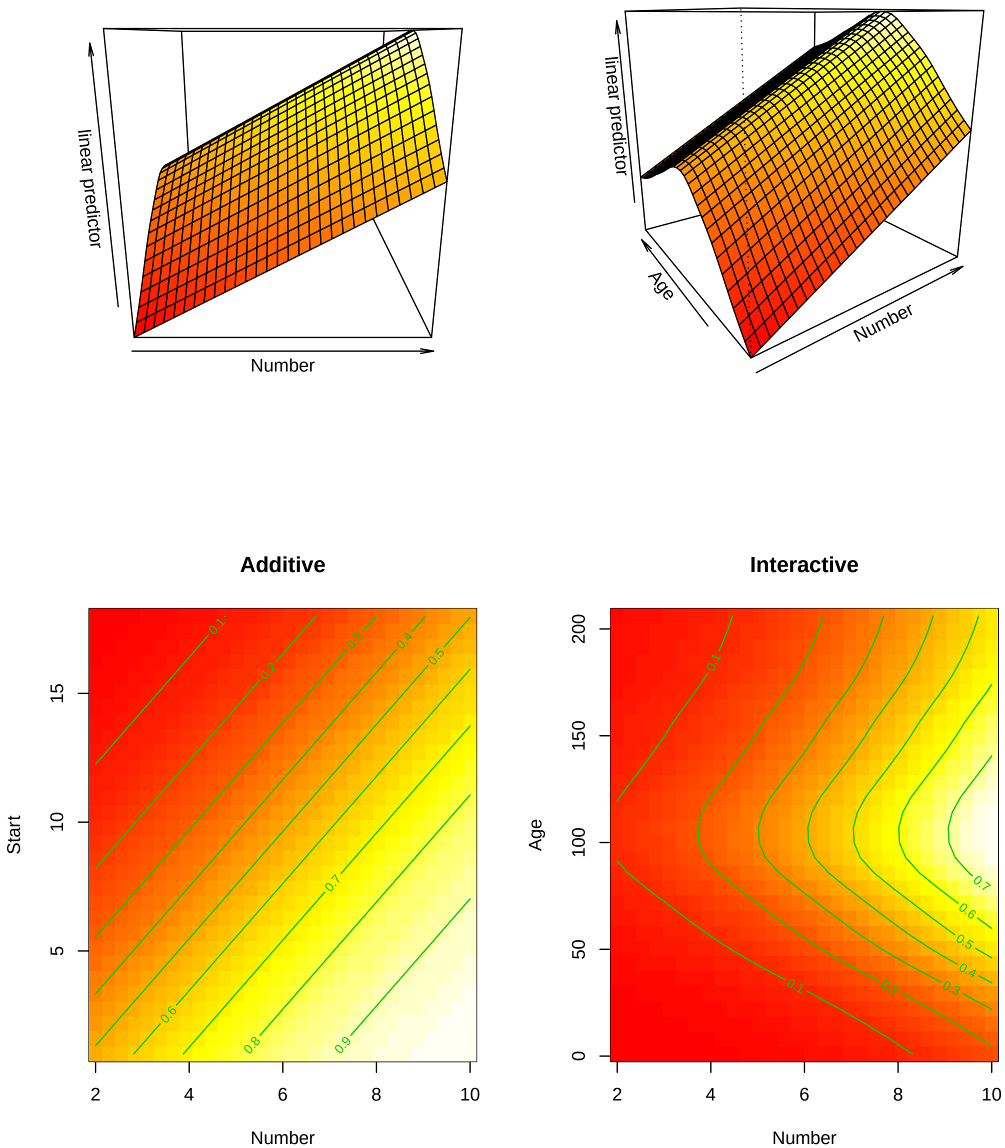
. . . and then with an interaction:

```
summary(m3 <- update(m1, ~ s(Age, Start) + Number))
```

```
Family: binomial
Link function: logit

Formula:
Kyphosis ~ s(Age, Start) + Number

Parametric coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.668      1.134   -3.23   0.0012
Number         0.418      0.231    1.81   0.0701

Approximate significance of smooth terms:
              edf Ref.df Chi.sq p-value
s(Age,Start) 3.53   4.48   12.7   0.019

R-sq.(adj) =  0.316   Deviance explained = 33.9%
UBRE score = -0.18431  Scale est. = 1         n = 81
```

UBRE stands for *unbiased risk estimator*, Wood, p. 172, and is similar to an AIC statistic.

```
par(mfrow=c(2, 2))
vis.gam(m3)
vis.gam(m3, theta=-35)
vis.gam(m2, plot.type="contour", type="response", main="Additive")
vis.gam(m3, plot.type="contour", type="response", main="Interactive")
```


linear predictor
Number
linear predictor
Age
Number
Additive
Interactive
Start
Age
Number
Number