Lecture 15. Nonparametric GLMs (cont.)

Nan Ye

School of Mathematics and Physics
University of Queensland

Recall: Generalized additive model (GAM)

- Recall: A GLM has the following structure

$$\text{(systematic)} \quad \mathbb{E}(Y \mid \mathbf{x}) = h(\beta^\top \mathbf{x}),$$
$$\text{(random)} \quad Y \mid \mathbf{x} \text{ follows an exponential family distribution.}$$

- A *generalized additive model* has the following structure

$$\text{(systematic)} \quad \mathbb{E}(Y \mid \mathbf{x}) = \beta_0 + \sum_i h_i(x_i)$$
$$\text{(random)} \quad Y \mid \mathbf{x} \text{ follows an exponential family distribution.}$$

This defines a conditional probability model

$$p(y \mid \mathbf{x}, \beta_0, h_1, \ldots, h_d)$$

Reall: Roughness penalty approach for GAM

- We want to choose β_0, $h_1, \ldots, h_d$ to maximize

$$\sum_i \ln p(y_i \mid \mathbf{x}_i, \beta_0, h_1, \ldots, h_d) - \sum_j \lambda_j \int h_j''(x_j)^2 dx_j.$$

- Again, if each $\lambda_j > 0$, then each h_j must be a natural cubic spline with knots at the unique values of x_j.
- This reduces the problem to a parametric regression problem.

Recall: Remarks

- Higher order derivatives may be used in the regularizer (smoothness penalty).
- We can also use regression splines instead of smoothing splines to represent h_i's.
- h_i's may use a mix of different representations.

  e.g. $h_1(x_1) = x_1$, $h_2(x_2)$ *a regression spline,* $h_3(x_3)$ *a smoothing spline...*

This Lecture

GAM using mgcv

- Model options
- Model checking

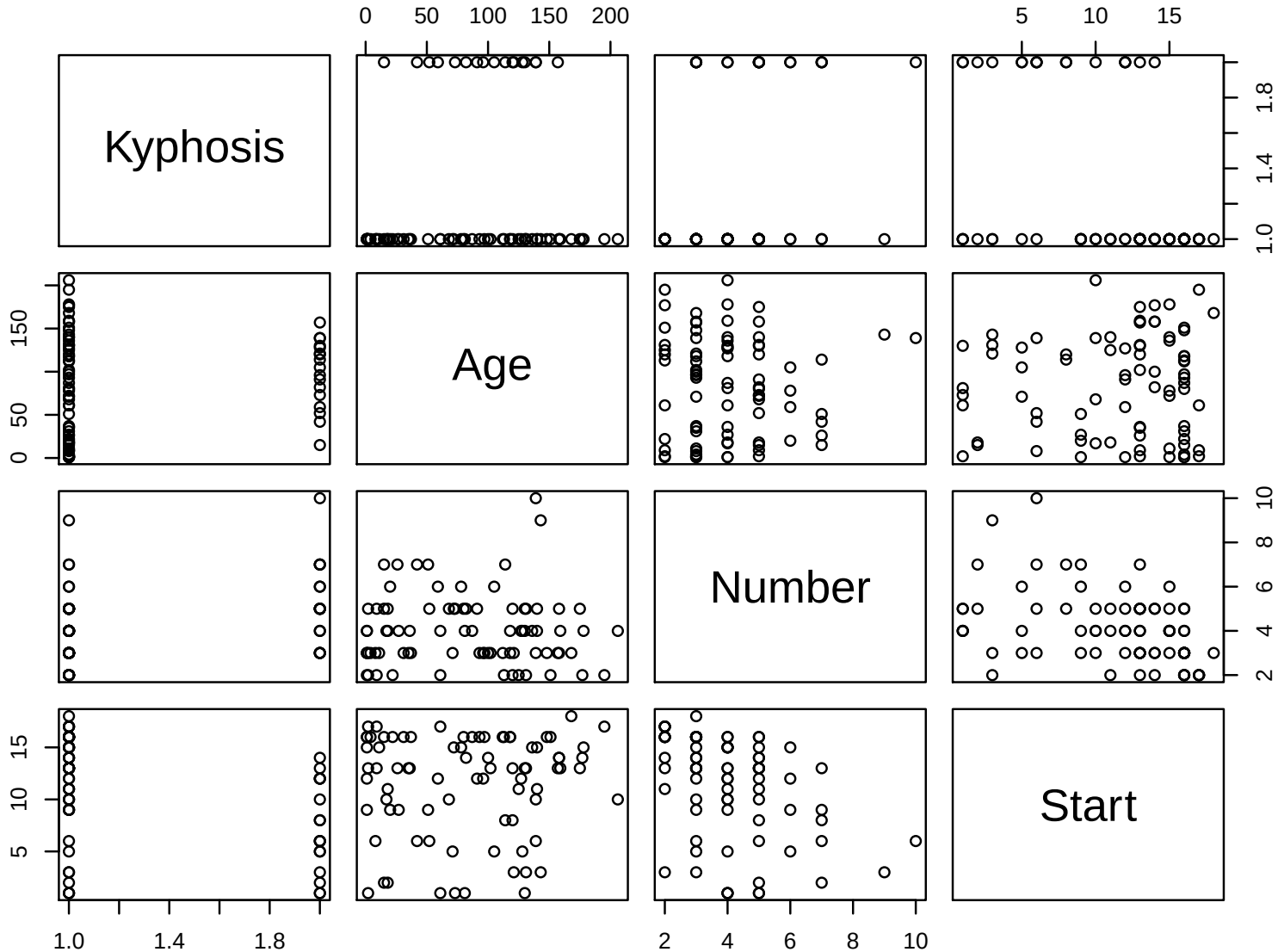
Data

```
> library(rpart) # contains the kyphosis dataset
> dim(kyphosis)
[1] 81  4
> head(kyphosis)
  Kyphosis Age Number Start
1   absent  71      3     5
2   absent 158      3    14
3  present 128      4     5
4   absent   2      5     1
5   absent   1      4    15
6   absent   1      2    16
```

Data on children who have had corrective spinal surgery.

- Kyphosis: if a kyphosis is present after the surgery.
- Age: age in month.
- Number: number of veterbrae involved.
- Start: the number of the topmost vertebra operated on.

Pairwise scatterplots



- None of Age, Number, and Start is a good predictor on Kyphosis alone.
- Age, Number and Start show little correlation between each other.

mgcv overview

- R package for estimating penalized Generalized Linear models including Generalized Additive Models and Generalized Additive Mixed Models.
- The `gam` function is for fitting penalized regression splines with automatic smoothness estimation (documentation at goo.gl/TmaoFW).
- Smoothness selection in `gam` is by GCV, AIC/Mallows' Cp, GACV, REML or ML (see the method argument in the documentation).

GLM via gam

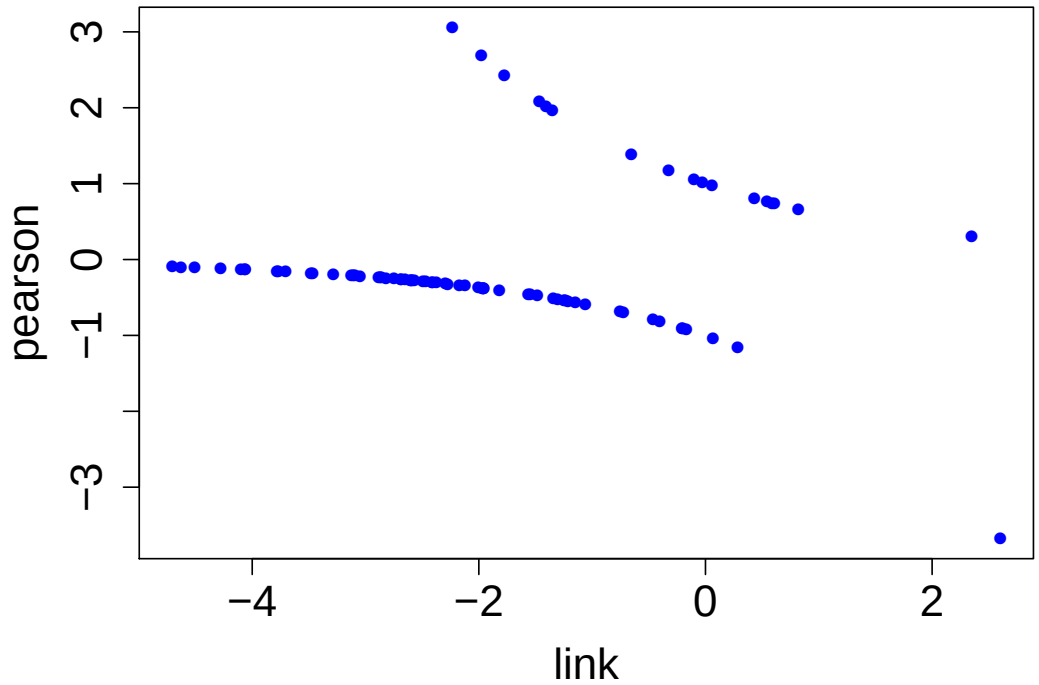
```
> library(mgcv)
> fit.gam.glm = gam(Kyphosis ~ Age + Number + Start, data=kyphosis,
    family=binomial)
> summary(fit.gam.glm)
Parametric coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.036934   1.449622  -1.405  0.15998
Age          0.010930   0.006447   1.696  0.08997 .
Number       0.410601   0.224870   1.826  0.06786 .
Start       -0.206510   0.067700  -3.050  0.00229 **
```

- The `gam` function has essentially the same syntax as the `glm` function when fitting a GLM.

```
> fit.glm = glm(Kyphosis ~ Age + Number + Start, data=kyphosis,
    family=binomial)
> summary(fit.glm)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.036934   1.449575  -1.405  0.15996
Age          0.010930   0.006446   1.696  0.08996 .
Number       0.410601   0.224861   1.826  0.06785 .
Start       -0.206510   0.067699  -3.050  0.00229 **
```

- The two logistic models fitted using `gam` and `glm` are essentially the same.

Residual plot for the fitted logistic model



Is this plot useful?

Characteristics of residual plots for a logistic model

- Recall that $\mathbf{x}^\top \beta$ is the value of the link function $g(\mu)$
- The Pearson residual for a logistic model is $r_P = \frac{y-\mu}{\sqrt{\mu(1-\mu)}}$.
- For each link value $g(\mu)$, there is a negative point and a positive point

$$(g, -\mu/\sqrt{\mu(1-\mu)}), \text{ and } (g, 1-\mu/\sqrt{\mu(1-\mu)}),$$

  where the two residual values have product -1.
- As g increases, the positive point is roughly $(g, \sqrt{1-\mu})$.
- As g decreases, the negative point is roughly $(g, -\sqrt{\mu})$.

Residual plot is not quite informative for binary data...

- If the model is true...
  - as g increases, the negative point becomes less common.
  - as g decreases, the positive point becomes less common.
- Unless one observes quite a few positive points for small g, or quite a few negative points for large g, then nothing is obviously wrong.
  - Such abnormality is not observed for the fitted model.

Fitting a GAM

```
> fit.gam = gam(Kyphosis ~ s(Age) + s(Number,k=8) + s(Start,k=16),
    data=kyphosis, family=binomial)
```

- Syntax pretty much the same as `glm`.
- However, we can now specify a smoothing term using `s`.
- This fits a nonparametric logistic model of the form

$$\begin{aligned}&\text{logit}P(Kyphosis \mid Age, Number, Start)\\ &= \beta_0 + h_1(Age) + h_2(Number) + h_3(Start).\end{aligned}$$

Specifying smooth terms using `s`

- The k parameter specifies the number of basis functions to use.
- $k = 10$ by default, but need to be at most the number of unique values (8 for Number, and 16 for Start).
- The default basis functions are a class of thin plate regression splines.
- `bs='cr'`: the basis functions are cubic regression splines with knots spread evenly through the covariate values.

  A broad class of alternative smooth terms and basis functions are available: https://goo.gl/AJ8qgP.

Inspecting the GAM

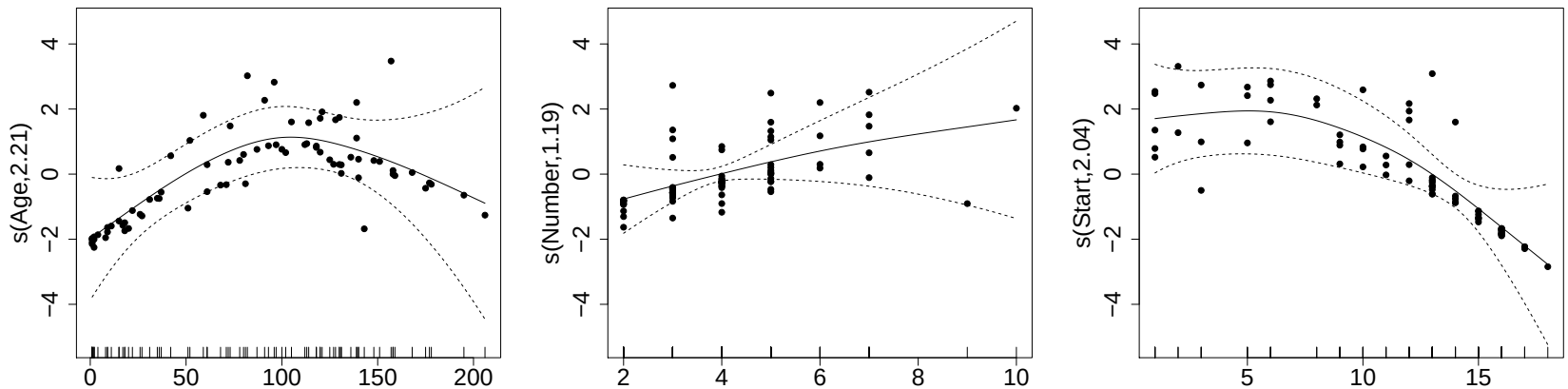
```
> summary(fit.gam)
Parametric coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.2706     0.5015  -4.528 5.96e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Approximate significance of smooth terms:
                 edf Ref.df Chi.sq p-value
s(Age)     2.212  2.791  6.367  0.0768 .
s(Number) 1.193  1.358  2.577  0.1959
s(Start)  2.035  2.542  9.814  0.0144 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

- The EDF (estimated degree of freedom) for Age is 2.212, so $\hat{h}_1$ has a complexity between a quadratic and a cubic polynomial.
- Similarly, $\hat{h}_2$ is like a linear function, and $\hat{h}_3$ is like a quadratic polynomial.

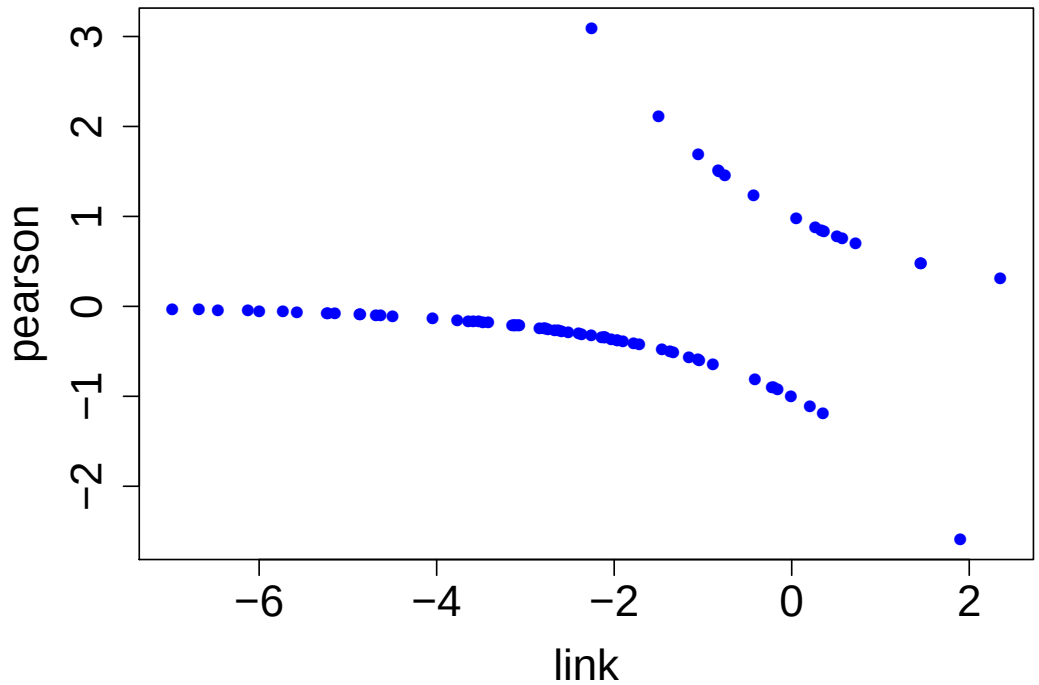
Checking fitted smooth terms

- `plot(fit.gam, residuals=TRUE, pch=19, pages=1)`



- Dotted lines represent 95% Bayesian confidence intervals.
- Black dots are obtained by adding partial residuals to each fitted $\hat{h}_i$. Systematic departure from $\hat{h}_i$ indicates a problem.

Checking GAM residuals



Almost the same as the residual plot for logistic regression. Not really useful.

Checking basis dimension k

```
> summary(gam(Kyphosis ~ s(Age, k=15) + s(Number,k=8) +
  s(Start,k=10), data=kyphosis, family=binomial))
Parametric coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.2704     0.5014  -4.528 5.96e-06 ***
---
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(Age)     2.217  2.802  6.370  0.0774 .
s(Number) 1.192  1.356  2.575  0.1957
s(Start)  2.031  2.533  9.707  0.0143 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

- Already using maximum k for Number and Start.
- Increasing k for Age does not have much effect.

Checking training set accuracy

```
> y = kyphosis$Kyphosis == 'present'
> max(sum(y)/length(y),  1 - sum(y)/length(y))
[1] 0.7901235
> pred.glm = predict(fit.glm, type='response') > 0.5
> sum(y == pred.glm)/length(y)
[1] 0.8395062
> pred.gam = predict(fit.gam, type='response') > 0.5
> sum(y == pred.gam)/length(y)
[1] 0.8765432
```

- Always predicting 'absent' has an accuracy of 0.79.
- Nonparametric logistic model has best training set accuracy.

What You Need to Know

GAM using mgcv

- Various model options in gam
  - type of basis functions, number of basis functions, method of estimating smoothing parameter...
- Model checking
  - Residual plot not useful for binary data.
  - Check things like fitted smooth terms, basis dimension, training set accuracy to see whether something is obviously wrong/inadequate.