Statistics 452: Statistical Learning and Prediction

Chapter 7, Part 4: Generalized Additive Models

Brad McNeney

2018-10-26

Generalized Additive Models (GAMs)

- We now consider extending the linear model when we have p explanatory variables, $X = (X_1, \ldots, X_p)$.
- In linear regression, the function $f(X)$ is of the form

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$$

- In a GAM we use (up to) p smooth functions

$$f(X) = \beta_0 + f_1(X_1) + f_2(X_2) + \ldots + f_p(X_p)$$

- The component functions $f_j(\cdot)$ can be any of the smoothers discussed in Sections 7.1-7.6 (e.g., polynomial, spline or local regression; smoothing spline)

Example: Wage Data (Again)

- Fit a model for `wage` of the form

$$wage = \beta_0 + f_1(year) + f_2(age) + f_3(education) + \epsilon$$

- Recall that `education` is categorical, so f_3 is an expansion into dummy variables.

```r
library(ISLR)
data(Wage)
table(Wage$education)
```

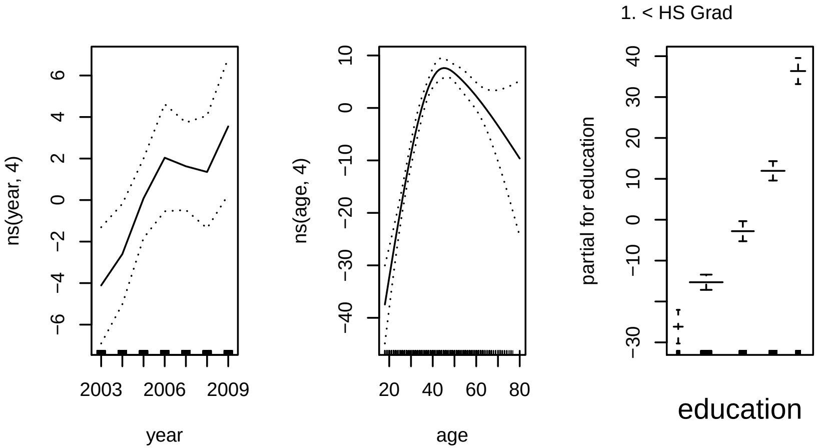
```
##
##       1. < HS Grad         2. HS Grad    3. Some College
##                268                971                650
##    4. College Grad 5. Advanced Degree
##                685                426
```

- We can use natural cubic splines with 4 df in `year` and `age`.

```r
library(splines)
gfit <- lm(wage ~ ns(year,4) + ns(age,4) + education,data=Wage)
```

- To plot we can use a plotting function from the `gam` package.

```
library(gam)
par(mfrow=c(1,3))
plot.Gam(gfit, se=TRUE)
```



Model Selection

- We could do CV-based model selection on the df for the two splines, but this would now be a search over a 2-d grid of df's.
- However, we notice that a linear fit in `year` looks plausible, and we can use an ANOVA F-test to test the null hypothesis of linearity.
  - Importantly, the model that is linear in `year` is a sub-model of the natural cubic spline model (`ns(age,df=1)` is linear in age)

```
gfit2 <- lm(wage ~ year+ns(age,4)+education,Wage)
anova(gfit2,gfit)
```

```
## Analysis of Variance Table
##
## Model 1: wage ~ year + ns(age, 4) + education
## Model 2: wage ~ ns(year, 4) + ns(age, 4) + education
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1   2990 3700055
## 2   2987 3697241  3    2813.8 0.7577 0.5178
```

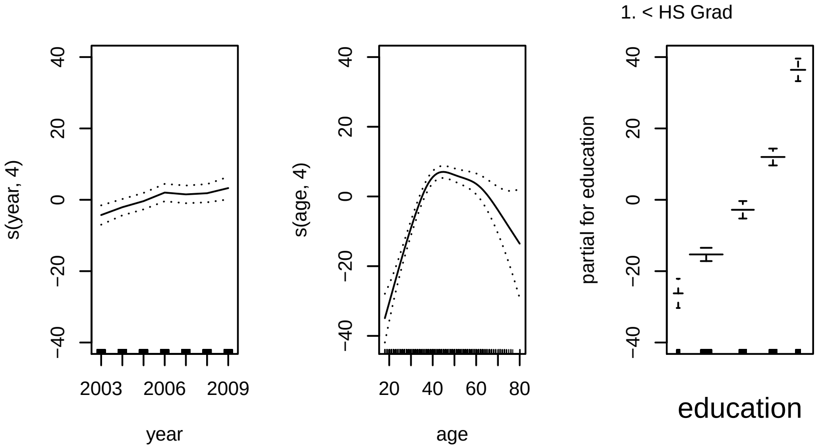
- We retain the hypothesis that f is linear in `year`.

GAM with Smoothing Splines

- Smoothing splines shrinkage estimators and so are not fit simply by least squares.
- Use the `gam` package.
  - Written by Hastie and Tibshirani (also authors of a book on the subject)

```
library(gam)
gfit3 <- gam(wage ~  s(year,4) + s(age,4) + education,data=Wage)
```

```r
par(mfrow=c(1,3))
plot(gfit3, se=TRUE,ylim=c(-40,40))
```



GAM Intepretation

- Each smooth is the estimated effect of changing one variable holding the others fixed.
- For example, holding age and education fixed, wage increases slightly, and approximately linearly by year.
- Holding year and education fixed, wage increases until about 40, then is levels out, and the drops after 60.
- Holding year and age fixed, wage increases with education level.

Model Reduction

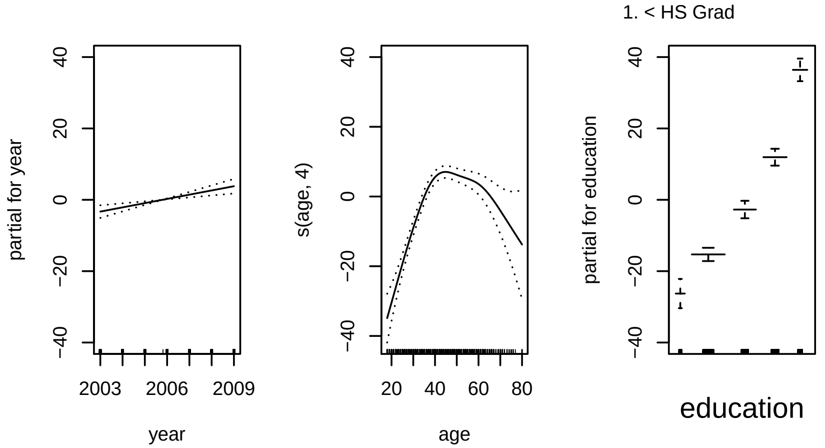
- A smoothing spline with 2df in `year` is linear, so we can use the F-test to compare models.

```
gfit4 <- gam(wage ~  year + s(age,4) + education,data=Wage)
anova(gfit4,gfit3)
```

```
## Analysis of Deviance Table
##
## Model 1: wage ~ year + s(age, 4) + education
## Model 2: wage ~ s(year, 4) + s(age, 4) + education
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      2990    3696846
## 2      2987    3692824  3   4021.7   0.3542
```

- We retain the hypothesis of linear in `year`.

```r
par(mfrow=c(1,3))
plot(gfit4,se=TRUE,ylim=c(-40,40))
```



Model Summary

```r
summary(gfit4)
```

```
##
## Call: gam(formula = wage ~ year + s(age, 4) + education, data = Wage)
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -119.463  -19.649   -3.284   13.928  213.522
##
## (Dispersion Parameter for gaussian family taken to be 1236.403)
##
##     Null Deviance: 5222086 on 2999 degrees of freedom
## Residual Deviance: 3696846 on 2990 degrees of freedom
## AIC: 29885.5
##
## Number of Local Scoring Iterations: 2
##
## Anova for Parametric Effects
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## year         1   26745   26745  21.631 3.447e-06 ***
## s(age, 4)    1  194578  194578 157.374 < 2.2e-16 ***
## education    4 1072774  268194 216.914 < 2.2e-16 ***
## Residuals 2990 3696846    1236
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Anova for Nonparametric Effects
##             Npar Df Npar F     Pr(F)
## (Intercept)
## year
## s(age, 4)         3 42.444 < 2.2e-16 ***
## education
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

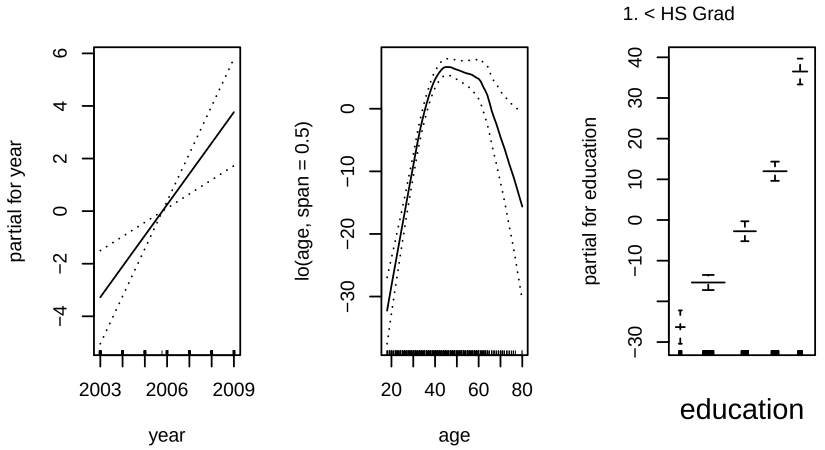
GAM Predictions

```r
newdat <- expand.grid(
  year=2003:2009,
  age=c(20,30,40,50,60,70,80),
  education=levels(Wage$education)) # head(newdat)
preds <- predict(gfit4,newdata=newdat)
preds[,,5] # Advanced degree
```

```
##              age
## year           age=20   age=30   age=40   age=50   age=60   age=70   age=80
##   year=2003 114.3619 135.7681 150.3541 151.0054 148.3792 140.7746 131.0273
##   year=2004 115.5477 136.9539 151.5399 152.1912 149.5650 141.9603 132.2131
##   year=2005 116.7335 138.1396 152.7257 153.3770 150.7508 143.1461 133.3989
##   year=2006 117.9192 139.3254 153.9115 154.5628 151.9366 144.3319 134.5847
##   year=2007 119.1050 140.5112 155.0972 155.7485 153.1223 145.5177 135.7705
##   year=2008 120.2908 141.6970 156.2830 156.9343 154.3081 146.7035 136.9563
##   year=2009 121.4766 142.8828 157.4688 158.1201 155.4939 147.8893 138.1421
```

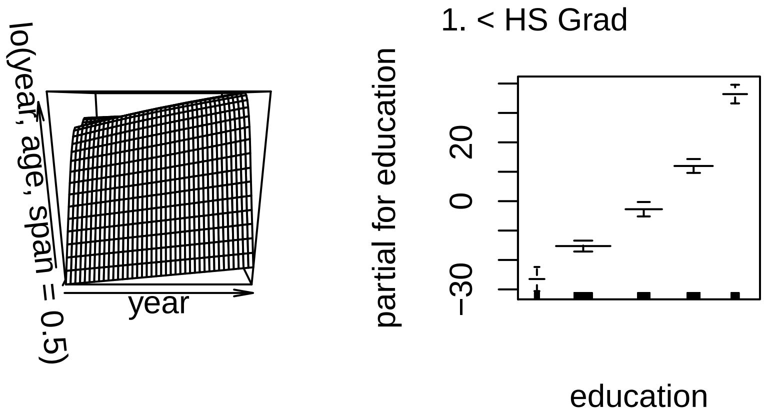
GAM with Local Regression

```
gfit5 <- gam(wage ~ year + lo(age,span=0.5)+education,data=Wage)
par(mfrow=c(1,3))
plot(gfit5,se=TRUE)
```



GAM with Multiple Local Regression ("Interaction")

```r
gfit6 <- gam(wage ~ lo(year,age,span=0.5)+education,data=Wage)
par(mfrow=c(1,2))
library(akima)
plot(gfit6,se=TRUE)
```



Advantages and Limitations of GAMs

- Advantages
  - Allows non-linear relationships that simple linear regression might miss, or might take a lot of work to discover (think age effect).
  - Can interpret components of the GAM (holding other variables fixed)
  - Smoothness of the component functions can be controlled by their df.
- Disadvantages
  - Restricted to additive models, though can fit interactions with local regression.

GAMs for Classification

- The “G” in GAM also stands for a generalization beyond gaussian linear models.
- Recall the logistic regression model formulation for modelling $p(X) = P(Y = 1|X)$:

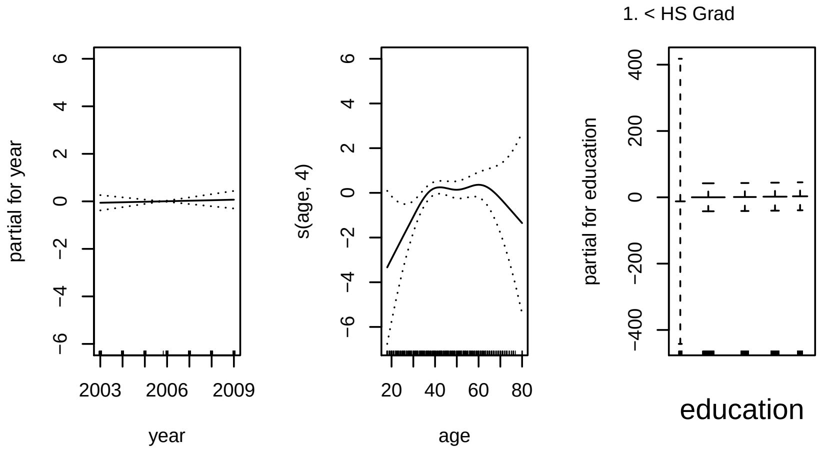
$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + X_1\beta_1 + \ldots + X_p\beta_p.$$

- Generalize to

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + f_1(X_1) + \ldots + f_p(X_p).$$

Example Logistic GAM

```
gfit7 <- gam(I(wage>250) ~ year + s(age,4) + education,
             data=Wage,family=binomial)
par(mfrow=c(1,3))
plot(gfit7,se=TRUE,ylim=c(-6,6))
```

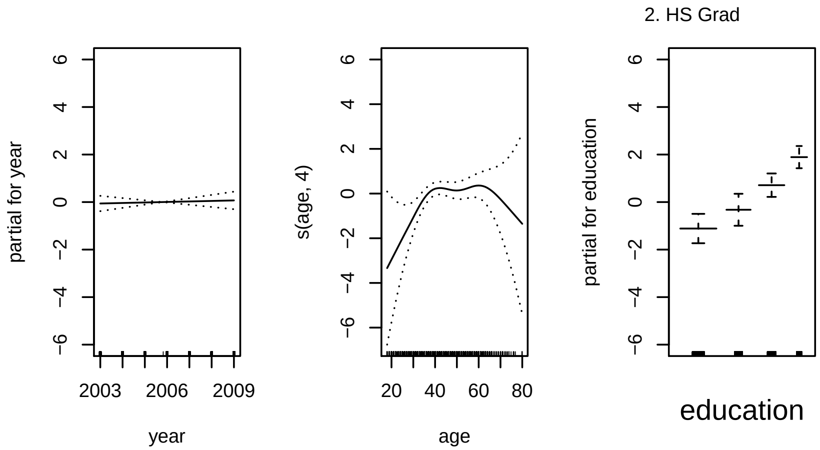


Removing < `HS Grad`

- There are no high-income earners with < `HS Grad` education, so our esimate of the education effect in this category is essentially $-\infty$.
  - Remove this category and re-fit

```
gfit7.s <- gam(I(wage>250) ~ year + s(age,4) + education,
               data=Wage,family=binomial,subset=(education!="1. < HS Grad"))
```

```r
par(mfrow=c(1,3))
plot(gfit7.s,se=TRUE,ylim=c(-6,6))
```



Remove Year

```r
gfit8.s <- gam(I(wage>250) ~  s(age,4) + education,
              data=Wage,family=binomial,subset=(education!="1. < HS Grad"))
anova(gfit8.s,gfit7.s)
```

```
## Analysis of Deviance Table
##
## Model 1: I(wage > 250) ~ s(age, 4) + education
## Model 2: I(wage > 250) ~ year + s(age, 4) + education
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      2724     603.94
## 2      2723     603.78  1  0.16156   0.6877
```

Alternative Implementation of GAMs

- `mgcv` is another well-developed R package that fits GAMS.
- The focus in `mgcv` is on penalized regression splines.
  - Penalty term may be selected by CV or other estimate of test set error.
- Allows interactions through "thin plate" regression splines