17) Generalized Additive Models (GAMs)

Vitor Kamada

March 2019

Reference

Tables, Graphics, and Figures from

1) James et al. (2017): Ch 7.7, and 7.8.3 2) Hastie et al. (2017): Ch 9.1

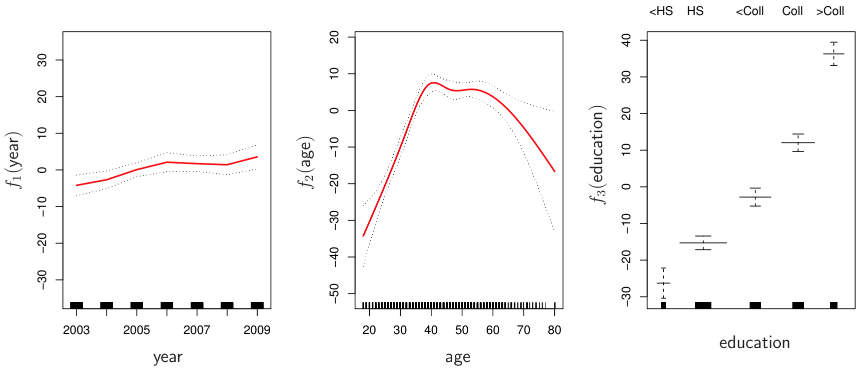
GAMs for Regression Problems

$$y_i = \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i$$

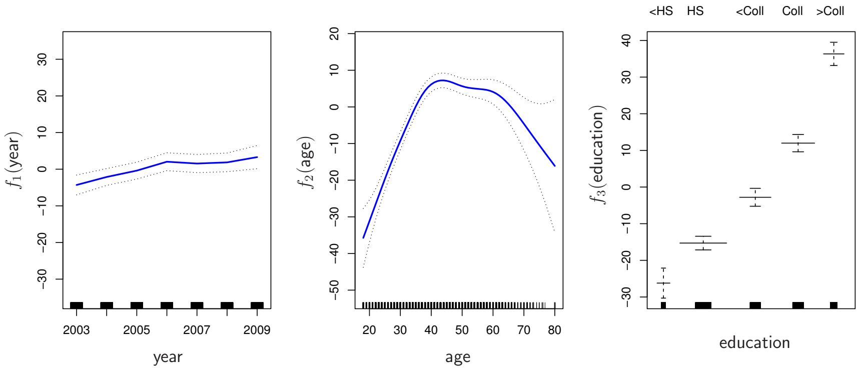
$$= \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \ldots + f_p(x_{ip}) + \epsilon_i$$

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

2 Natural Splines and 1 Step Function



2 Smoothing Splines and 1 Step Function



GAMs for Classification Problems

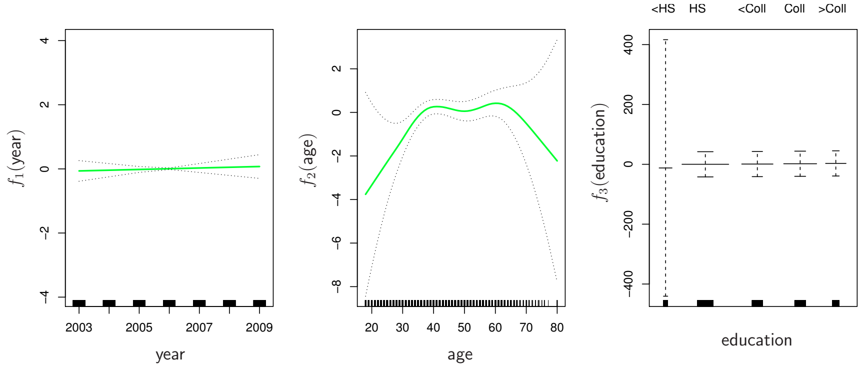
$$log\left[\frac{p(X)}{1-p(X)}\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

$$log\left[\frac{p(X)}{1-p(X)}\right] = \beta_0 + f_1(X_1) + f_2(X_2) + ... + f_p(X_p)$$

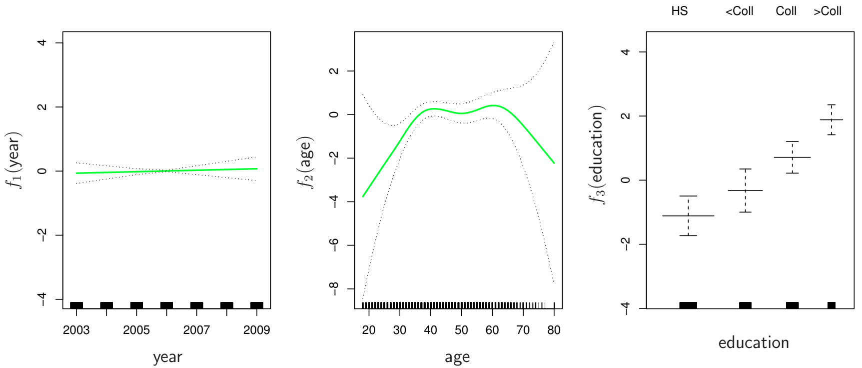
$$log\left[\frac{p(X)}{1-p(X)}\right] = \beta_0 + \beta_1 year + f_2(age) + f_3(educ)$$

$$p(X) = Pr(wage > 250 | year, age, educ)$$

Linear, Smoothing Spline, and Step Function



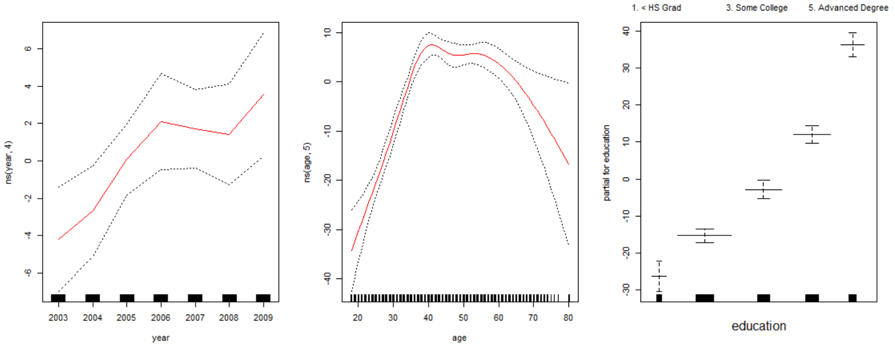
Excluding the Observations for which educ is <HS



library(ISLR); attach(Wage); library(splines); library(gam)

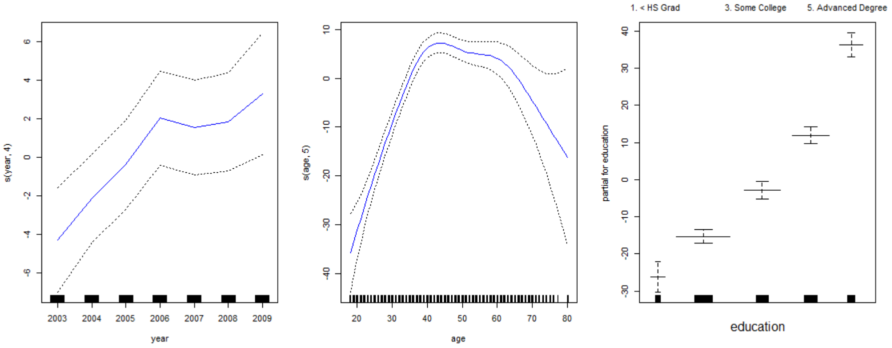
gam1=gam(wage~ns(year,4)+ns(age,5)+education, data=Wage)

par(mfrow=c(1,3)); plot(gam1, se=TRUE,col="red")



gam.m3=gam(wage~s(year,4)+s(age,5)+education, data=Wage)

par(mfrow=c(1,3)); plot(gam.m3, se=TRUE,col="blue")



gam.m1=gam(wage~s(age,5)+education, data=Wage)

gam.m2=gam(wage~year+s(age,5)+education, data=Wage)

anova(gam.m1,gam.m2,gam.m3,test="F")

```
Model 1: wage ~ s(age, 5) + education
Model 2: wage ~ year + s(age, 5) + education
Model 3: wage ~ s(year, 4) + s(age, 5) + education
  Resid. Df Resid. Dev Df Deviance       F    Pr(>F)
1      2990    3711731
2      2989    3693842  1  17889.2 14.4771 0.0001447 ***
3      2986    3689770  3   4071.1  1.0982 0.3485661
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

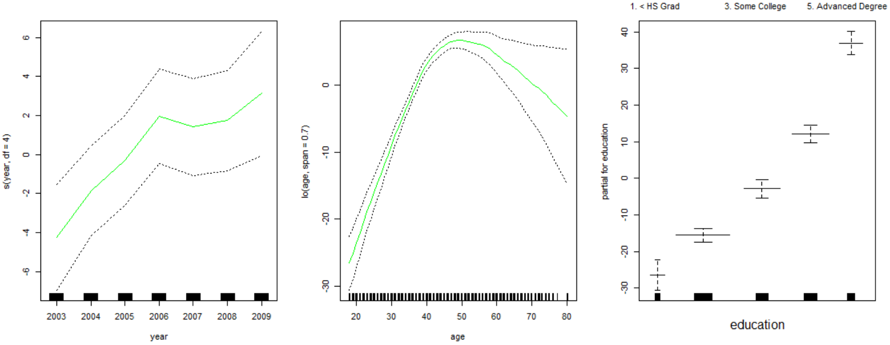
summary(gam.m3)

```
Anova for Parametric Effects
              Df  Sum Sq Mean Sq F value    Pr(>F)
s(year, 4)     1   27162   27162  21.981 2.877e-06 ***
s(age, 5)      1  195338  195338 158.081 < 2.2e-16 ***
education      4 1069726  267432 216.423 < 2.2e-16 ***
Residuals   2986 3689770    1236
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Anova for Nonparametric Effects
            Npar Df Npar F  Pr(F)
(Intercept)
s(year, 4)        3  1.086 0.3537
s(age, 5)         4 32.380 <2e-16 ***
education
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

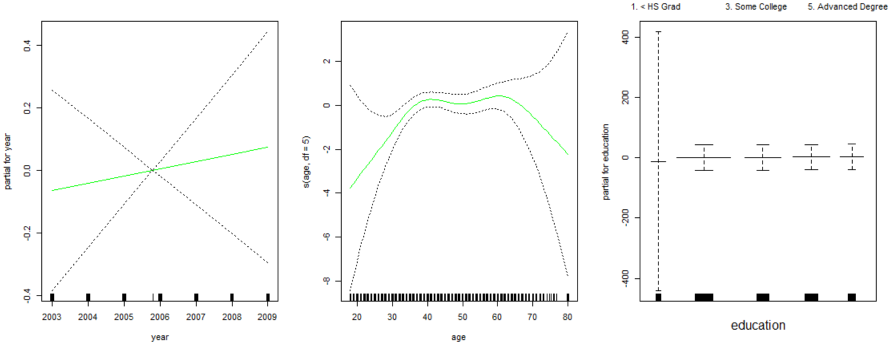
gam.lo=gam(wage~s(year,df=4)+ lo(age,span=0.7)+education, data=Wage)

plot(gam.lo, se=TRUE,col="green")



gam.lr=gam(I(wage>250)~year+s(age,df=5) +education, family=binomial,data=Wage)

par(mfrow=c(1,3)); plot(gam.lr,se=T,col="green")

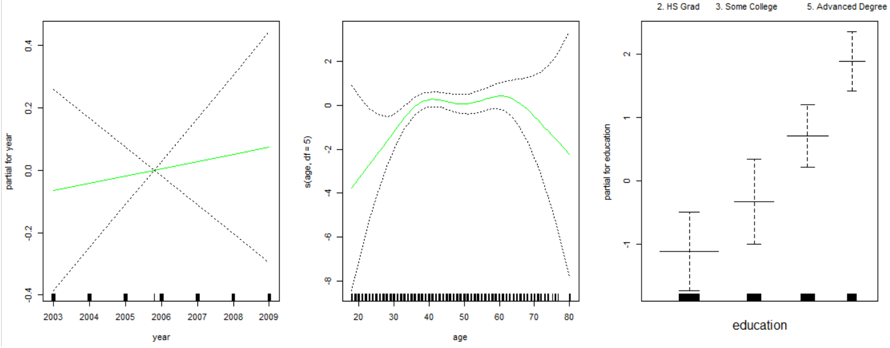


table(education,I(wage>250))

```
education                   FALSE TRUE
  1. < HS Grad                268    0
  2. HS Grad                  966    5
  3. Some College             643    7
  4. College Grad             663   22
  5. Advanced Degree          381   45
```

gam.lr.s=gam(I(wage>250)~year+s(age,df=5) +education, family=binomial, data=Wage, subset=(education!="1. < HS Grad"))

plot(gam.lr.s,se=T,col="green")



Predicting Email Spam

E-mails from Hewlett-Packard laboratories

	Training	Test	Total
E-mail	3065	1536	4601

57 predictors:

% of words: business, free, george

% of characters: ch;, ch!, ch$

CAPAVE: average length of capital letters

CAPMAX: length of the longest capital letters

CAPTOT: sum of the length of capital letters

<table>
<tr><td colspan="2"></td><td colspan="2">True condition</td><td colspan="3"></td></tr>
<tr><td></td><td>Total population</td><td>Condition positive</td><td>Condition negative</td><td>Prevalence $= \frac{\Sigma \text{ Condition positive}}{\Sigma \text{ Total population}}$</td><td colspan="2">Accuracy (ACC) = $\frac{\Sigma \text{ True positive} + \Sigma \text{ True negative}}{\Sigma \text{ Total population}}$</td></tr>
<tr><td rowspan="2">Predicted condition</td><td>Predicted condition positive</td><td>True positive,
Power</td><td>False positive,
Type I error</td><td>Positive predictive value (PPV), Precision = $\frac{\Sigma \text{ True positive}}{\Sigma \text{ Predicted condition positive}}$</td><td colspan="2">False discovery rate (FDR) = $\frac{\Sigma \text{ False positive}}{\Sigma \text{ Predicted condition positive}}$</td></tr>
<tr><td>Predicted condition negative</td><td>False negative,
Type II error</td><td>True negative</td><td>False omission rate (FOR) = $\frac{\Sigma \text{ False negative}}{\Sigma \text{ Predicted condition negative}}$</td><td colspan="2">Negative predictive value (NPV) = $\frac{\Sigma \text{ True negative}}{\Sigma \text{ Predicted condition negative}}$</td></tr>
<tr><td colspan="2" rowspan="2"></td><td>True positive rate (TPR), Recall, Sensitivity, probability of detection $= \frac{\Sigma \text{ True positive}}{\Sigma \text{ Condition positive}}$</td><td>False positive rate (FPR), Fall-out, probability of false alarm $= \frac{\Sigma \text{ False positive}}{\Sigma \text{ Condition negative}}$</td><td>Positive likelihood ratio (LR+) $= \frac{\text{TPR}}{\text{FPR}}$</td><td rowspan="2">Diagnostic odds ratio (DOR) $= \frac{\text{LR+}}{\text{LR−}}$</td><td rowspan="2">$F_1$ score = $2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$</td></tr>
<tr><td>False negative rate (FNR), Miss rate $= \frac{\Sigma \text{ False negative}}{\Sigma \text{ Condition positive}}$</td><td>Specificity (SPC), Selectivity, True negative rate (TNR) $= \frac{\Sigma \text{ True negative}}{\Sigma \text{ Condition negative}}$</td><td>Negative likelihood ratio (LR−) $= \frac{\text{FNR}}{\text{TNR}}$</td></tr>
</table>

GAM: Logistic Cubic Smoothing Spline

Each variable is decomposed into a linear and nonlinear component

Nominal df: $4 = trace[S_j(\lambda_j)] - 1$

True Class	Predicted Class	
	`email` (0)	`spam` (1)
`email` (0)	58.3%	2.5%
`spam` (1)	3.0%	36.3%

	Test Error Rate
Additive Logistic	5.5%
Linear Logistic	7.6%

Additive Logistic - Spam Training Data

Name	Num.	df	Coefficient	Std. Error	Z Score	Nonlinear P-value
Positive effects						
our	5	3.9	0.566	0.114	4.970	0.052
over	6	3.9	0.244	0.195	1.249	0.004
remove	7	4.0	0.949	0.183	5.201	0.093
internet	8	4.0	0.524	0.176	2.974	0.028
free	16	3.9	0.507	0.127	4.010	0.065
business	17	3.8	0.779	0.186	4.179	0.194
hpl	26	3.8	0.045	0.250	0.181	0.002
ch!	52	4.0	0.674	0.128	5.283	0.164
ch$	53	3.9	1.419	0.280	5.062	0.354
CAPMAX	56	3.8	0.247	0.228	1.080	0.000
CAPTOT	57	4.0	0.755	0.165	4.566	0.063
Negative effects						
hp	25	3.9	−1.404	0.224	−6.262	0.140
george	27	3.7	−5.003	0.744	−6.722	0.045
1999	37	3.8	−0.672	0.191	−3.512	0.011
re	45	3.9	−0.620	0.133	−4.649	0.597
edu	46	4.0	−1.183	0.209	−5.647	0.000

Nonlinearity Picks up the Discontinuity at 0