

22) Generalized Additive Models (GAMs)

Vitor Kamada

March 2018

Tables, Graphics, and Figures from
An Introduction to Statistical Learning

James et al. (2017): Chapters: 7.7, and 7.8.3

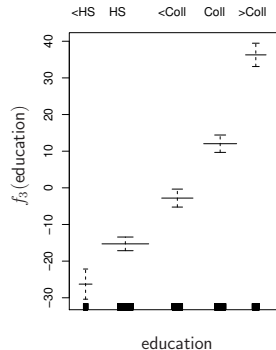
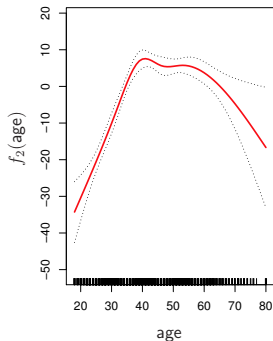
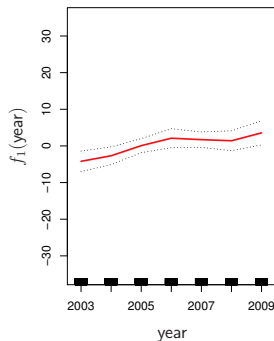
GAMs for Regression Problems

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

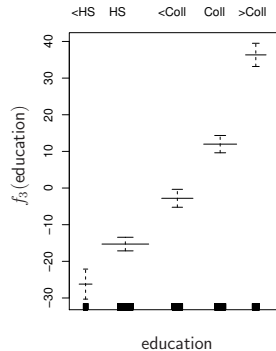
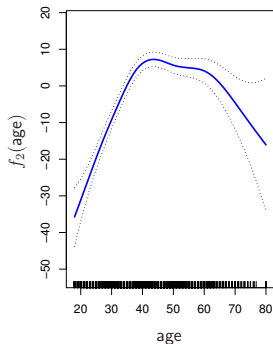
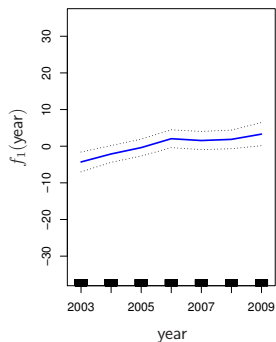
$$= \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \dots + f_p(x_{ip}) + \epsilon_i$$

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

2 Natural Splines and 1 Step Function



2 Smoothing Splines and 1 Step Function



GAMs for Classification Problems

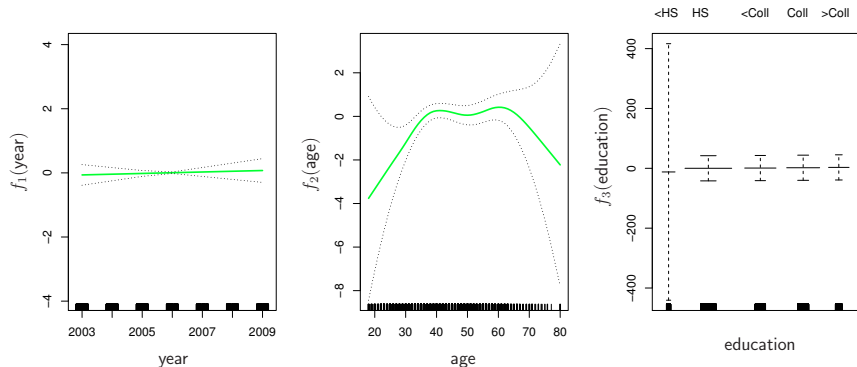
$$\log \left[\frac{p(X)}{1-p(X)} \right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p$$

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

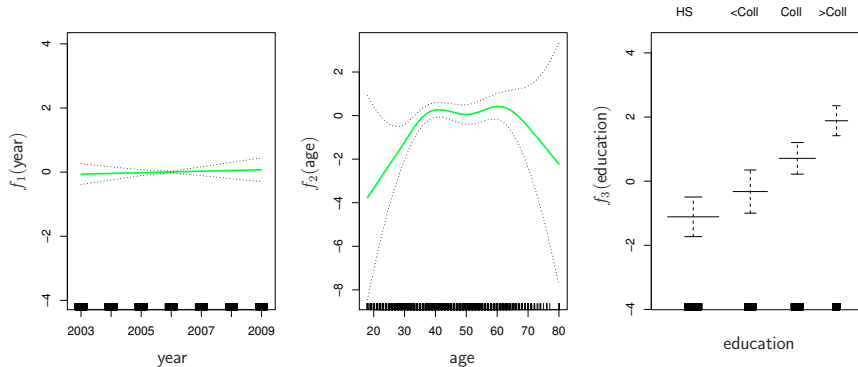
$$\log \left[\frac{p(X)}{1-p(X)} \right] = \beta_0 + \beta_1 \textit{year} + f_2(\textit{age}) + f_3(\textit{educ})$$

$$p(X) = \textit{Pr}(\textit{wage} > 250 | \textit{year}, \textit{age}, \textit{educ})$$

Linear, Smoothing Spline, and Step Function



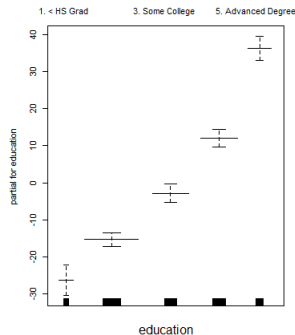
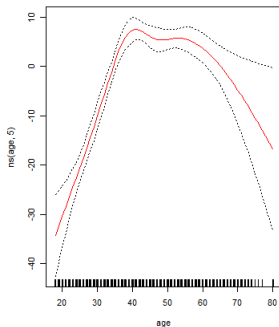
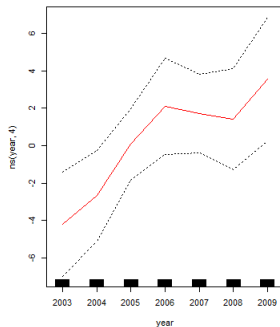
Excluding the Observations for which educ is < HS




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

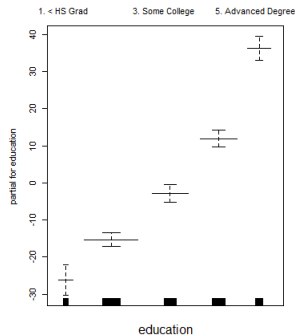
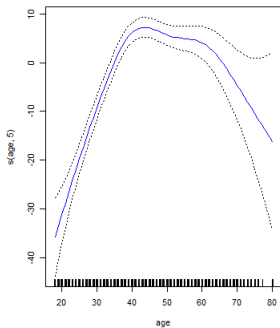
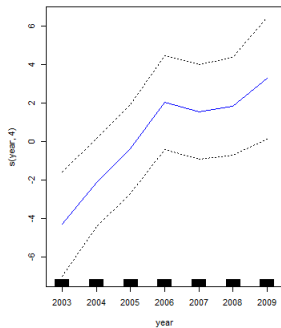
```
gam1=gam(wage~ns(year,4)+ns(age,5)+education,  
data=Wage)
```

```
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

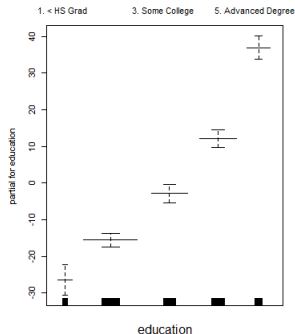
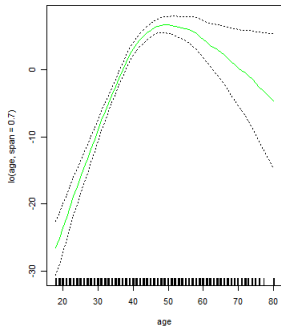
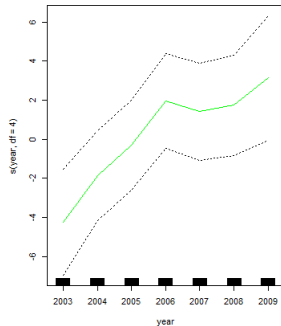
Anova for Nonparametric Effects

	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

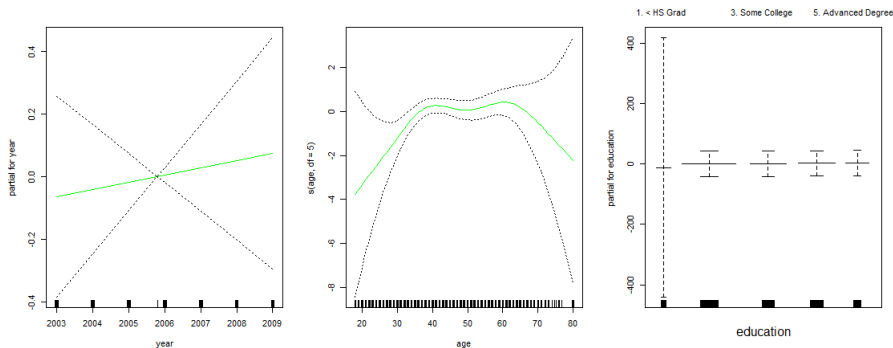
```
gam.lo=gam(wage~s(year,df=4)+  
lo(age,span=0.7)+education, data=Wage)
```

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



```
gam.lr=gam(l(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,l(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(l(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")
```

