RLab 8

Jon Griffith

2025-05-01

```
library(ISLR2)
```

```
## Warning: package 'ISLR2' was built under R version 4.4.3
```

```
library(fields)
```

```
## Loading required package: spam
```

```
## Spam version 2.10-0 (2023-10-23) is loaded.
## Type 'help( Spam)' or 'demo( spam)' for a short introduction
## and overview of this package.
## Help for individual functions is also obtained by adding the
## suffix '.spam' to the function name, e.g. 'help( chol.spam)'.
```

```
##
## Attaching package: 'spam'
```

```
## The following objects are masked from 'package:base':
##
##     backsolve, forwardsolve
```

```
## Loading required package: viridisLite
```

```
##
## Try help(fields) to get started.
```

```
library(splines)
library(gam)
```

```
## Loading required package: foreach
```

```
## Loaded gam 1.22-5
```

```
library(akima)
```

```
## Warning: package 'akima' was built under R version 4.4.3
```

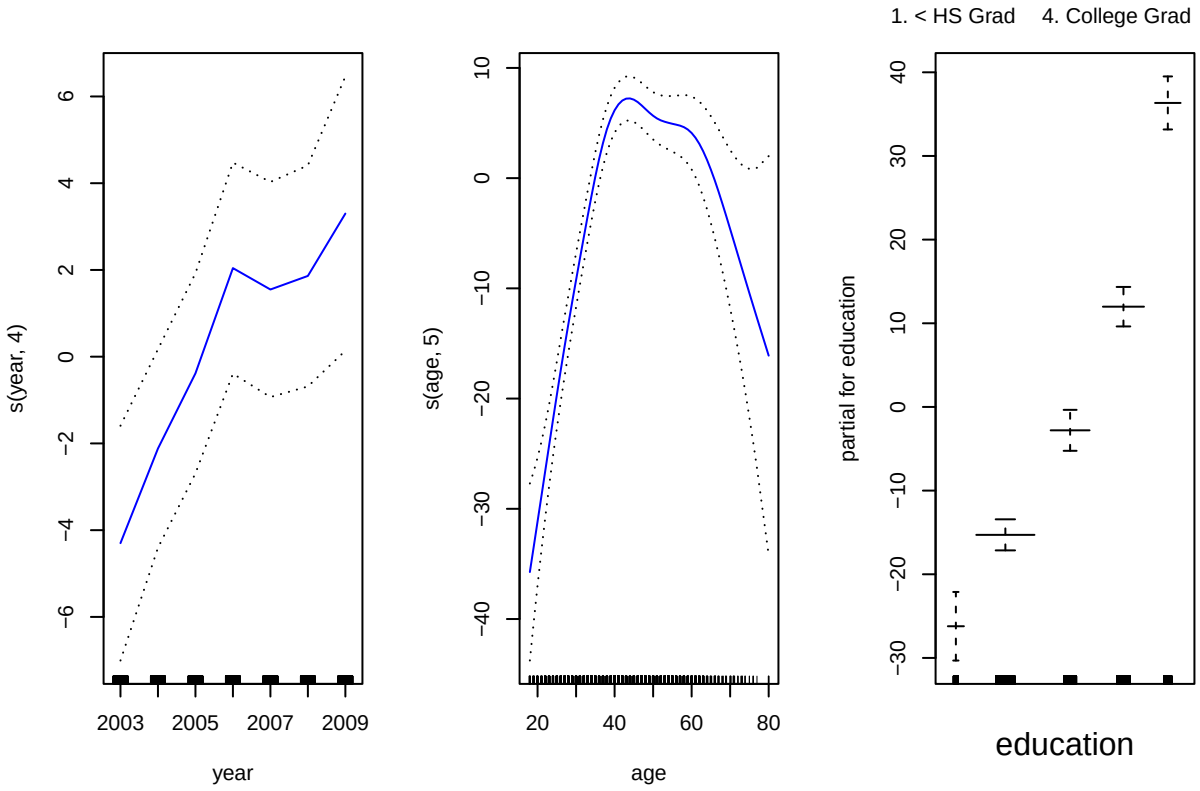
7.8.3 GAMs

We begin by using the same dataset, 'Wages', as we did in the last lab. We first fit a GAM to predict wage using natural spline functions of year and age, and we'll treat education as a qualitative predictor. We first fit the model using the lm() function and the ns() function inside to fit the natural splines.

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

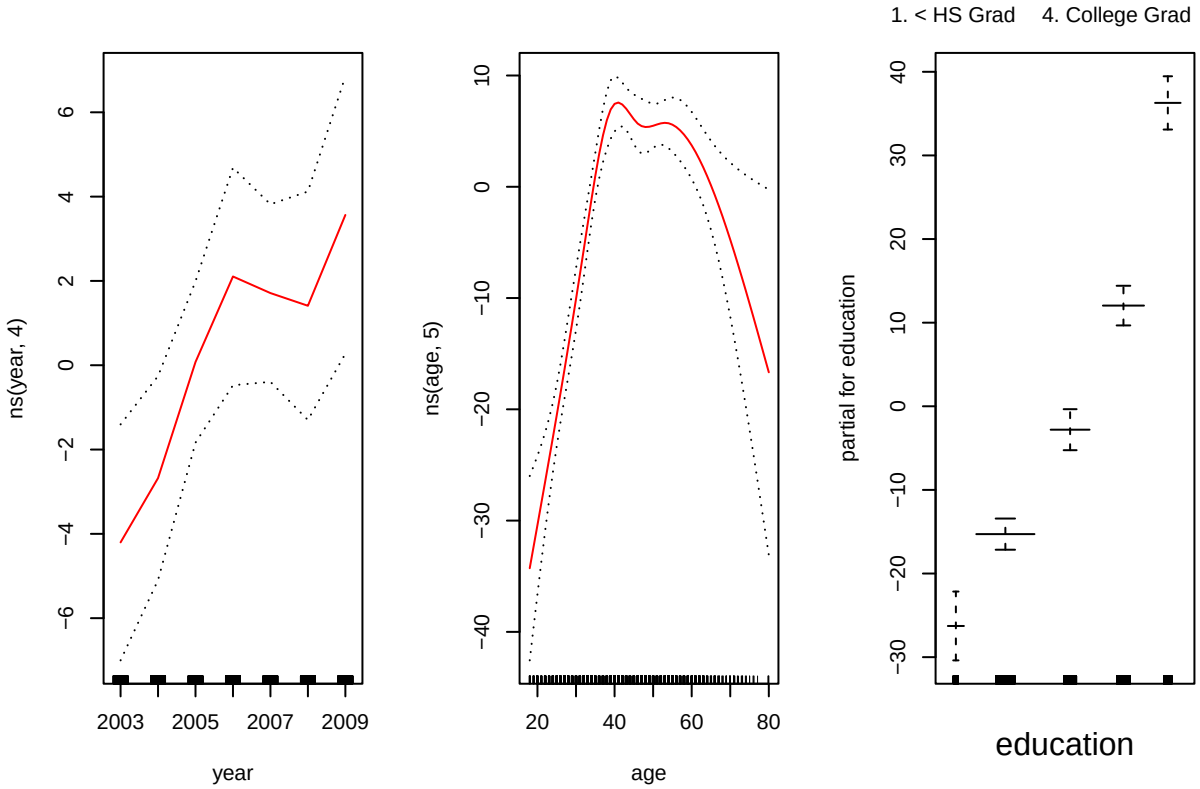
Now we'll fit a GAM using smoothing splines, but this time using the gam() function from the gam library. We specify that age will have 5 DF, year will have 4 DF, and we leave education as is since it is qualitative. We then reproduce Figure 7.12.

```
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')
```



And we'll plot our first GAM model to reproduce Figure 7.11 using plot.Gam().

```
par(mfrow=c(1,3))
plot.Gam(gam1, se = TRUE, col='red')
```



We can see from above that in both models, year appears to be linear. We'll conduct ANOVA tests to determine which model is best from one that excludes year, one that uses a linear function of year, or one that uses a spline function of year.

```
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')
```

```
## Analysis of Deviance Table
##
## 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
```

Based on the results above, we can see that the optimal model out of these three is the model that uses a linear function of year. Lets look at the summary for our third model that uses splines for year and notice that we can see the p-value for the parametric and nonparametric effects of each variable. For age, we see that the nonlinear version is statistically significant, confirming our visual belief from the plots. We also see that year is not significant for this panel and therefore, the linear transformation is valid based on the results in the parametric effects table.

```
summary(gam.m3)
```

```
##
```

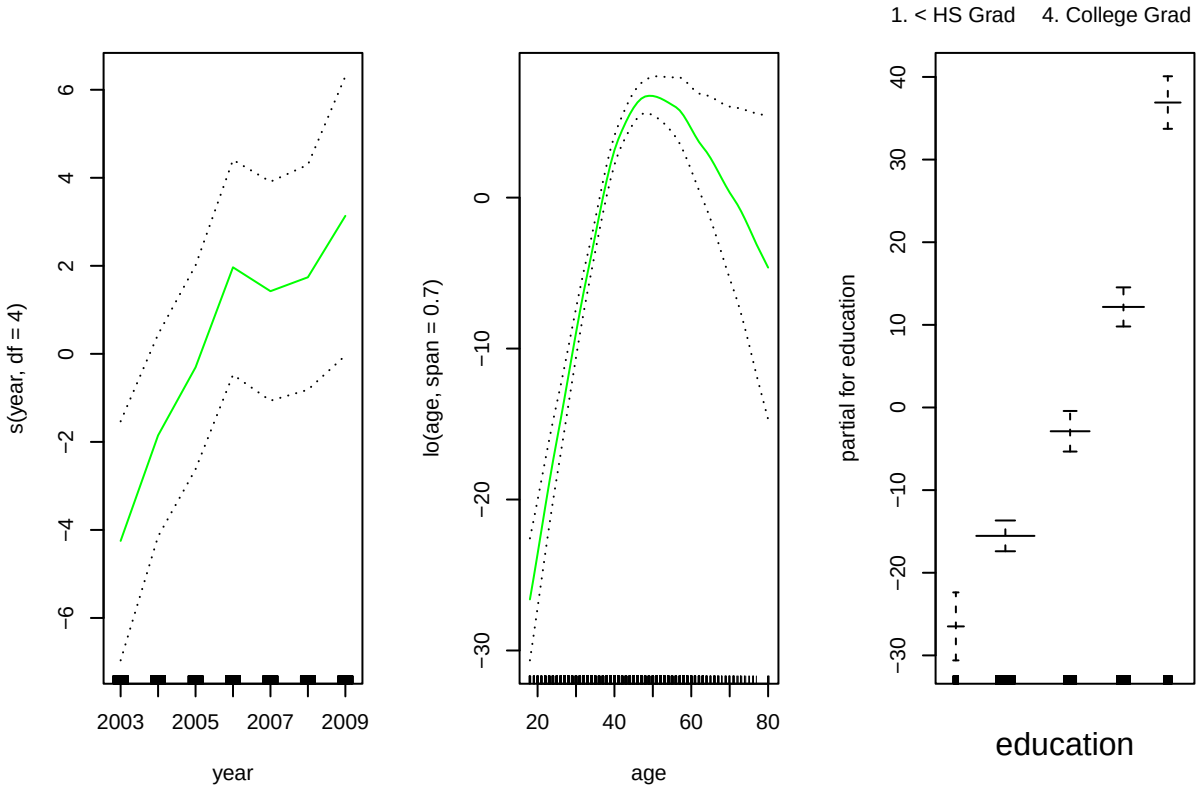
```
## Call: gam(formula = wage ~ s(year, 4) + s(age, 5) + education, data = Wage)
## Deviance Residuals:
##     Min      1Q  Median      3Q     Max
## -119.43  -19.70   -3.33   14.17  213.48
##
## (Dispersion Parameter for gaussian family taken to be 1235.69)
##
##     Null Deviance: 5222086 on 2999 degrees of freedom
## Residual Deviance: 3689770 on 2986 degrees of freedom
## AIC: 29887.75
##
## Number of Local Scoring Iterations: NA
##
## 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
```

We'll make predictions using our second model.

```
preds <- predict(gam.m2, newdata=Wage)
```

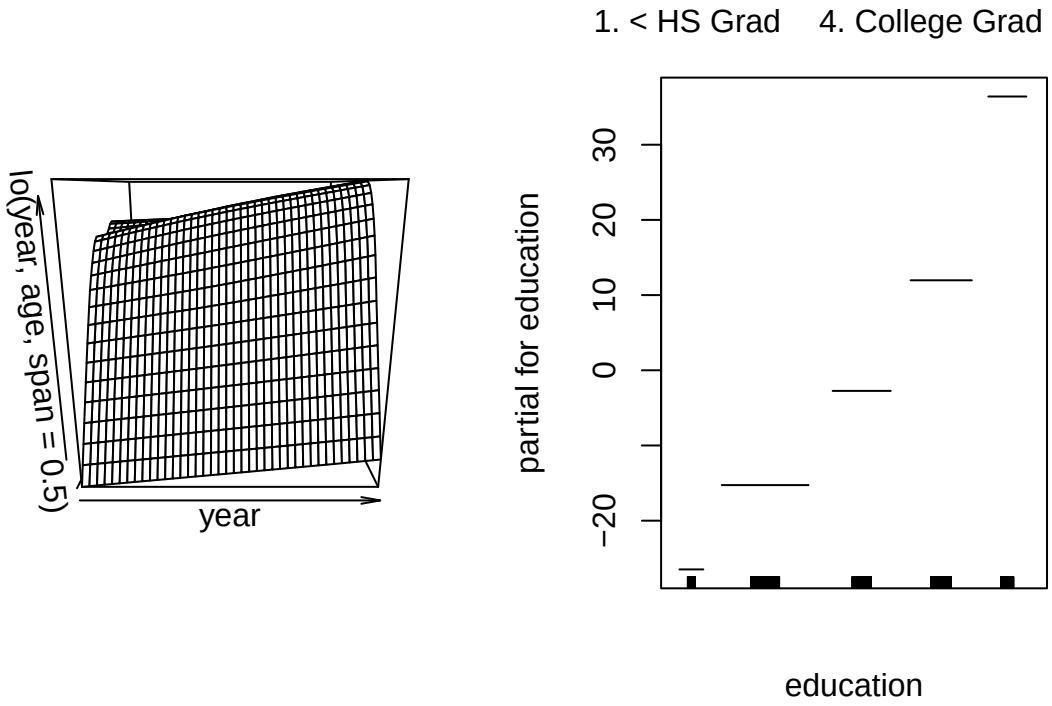
Now we'll create a new GAM that uses local regression fits as the building blocks. This is done using the lo() function where we can fit age and specify the span.

```
par(mfrow=c(1,3))
gam.lo <- gam(wage ~ s(year, df=4) + lo(age, span=0.7) + education, data=Wage)
plot(gam.lo, se=TRUE, col='green')
```



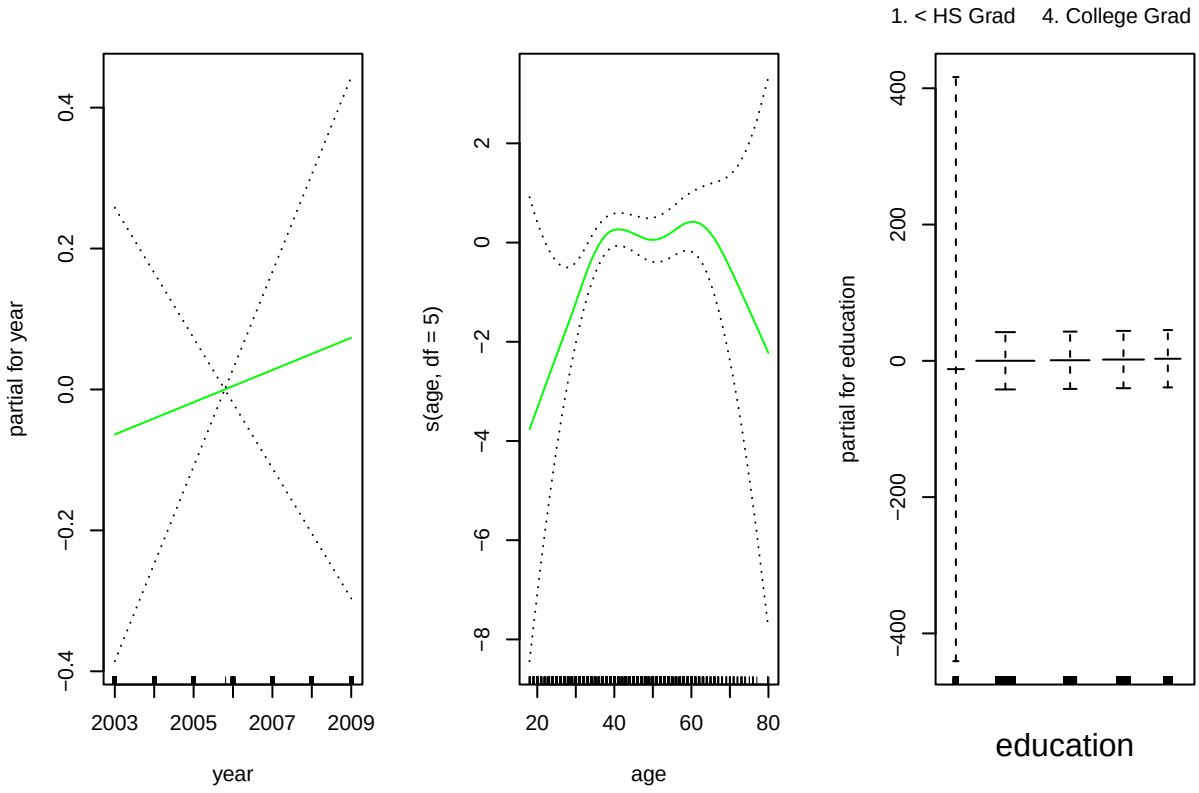
The lo() function can also be used to create interactions before calling gam(). We use a local regression surface between year and age and visualize that in the plot below.

```
gam.lo.i <- gam(wage ~ lo(year, age, span=0.5) + education, data=Wage)

par(mfrow=c(1,2))
plot(gam.lo.i)
```



Now we'll fit a logistic regression GAM using the I() function.

```
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')
```



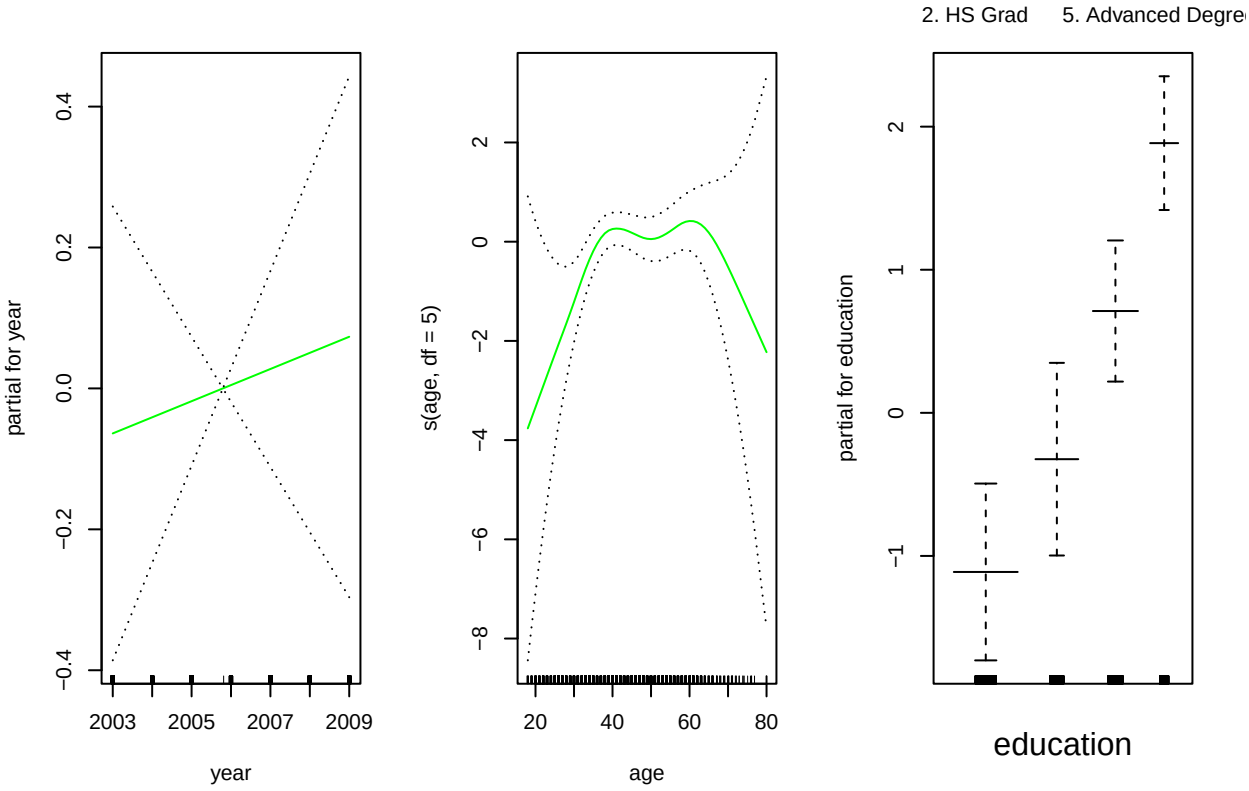
We can see that there aren't any high earners in the <HS category.

```
table(Wage$education, I(Wage$wage > 250))
```

```
##
##                      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
```

Since we can see that there are indeed 0, we will refit the model without this category.

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

par(mfrow=c(1,3))
plot(gam.lr.s, se=T, col='green')
```


partial for year
0.4
0.2
0.0
−0.2
−0.4
2003
2005
2007
2009
year
s(age, df = 5)
2
0
−2
−4
−6
−8
20
40
60
80
age
2. HS Grad
5. Advanced Degre
partial for education
2
1
0
−1
education