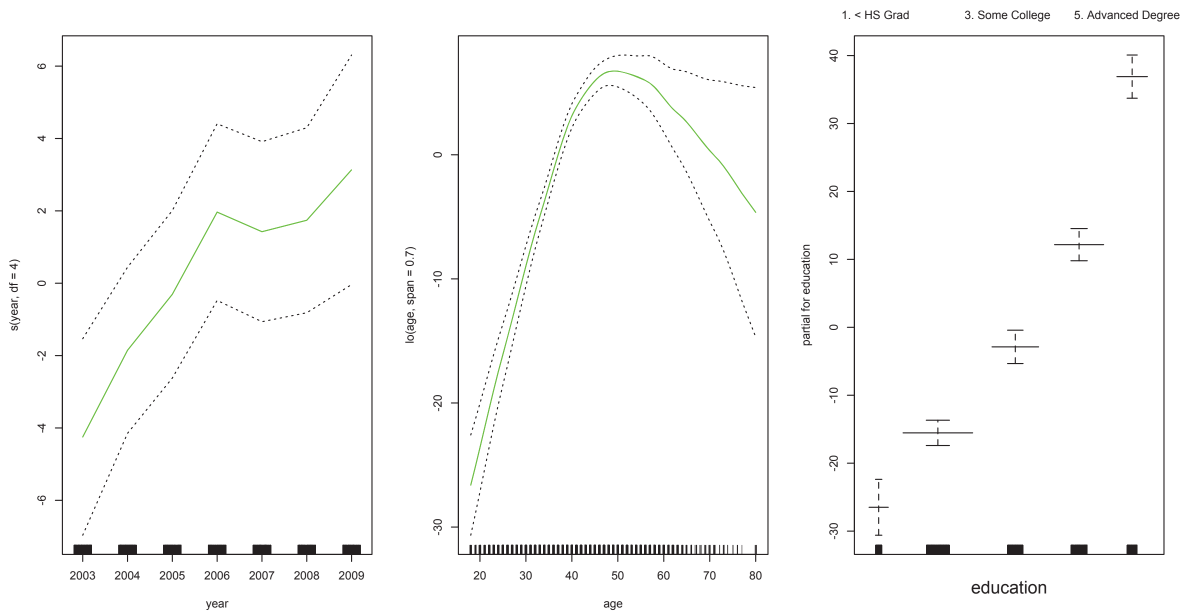
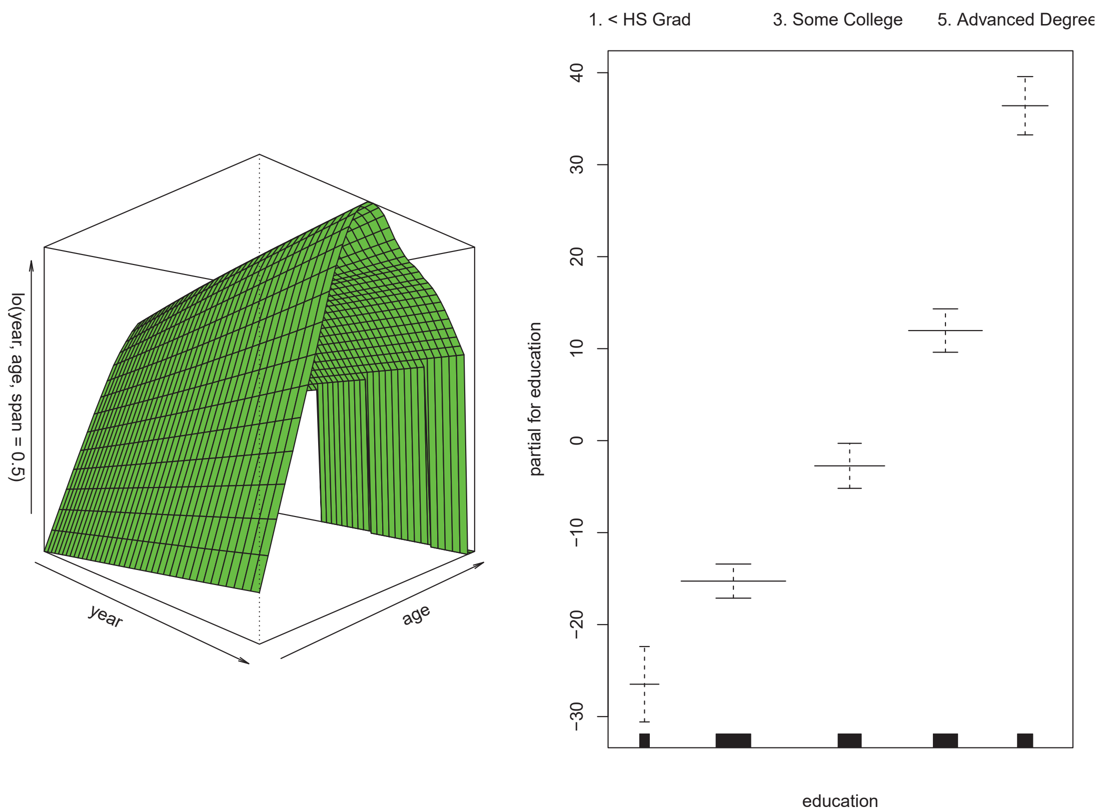


Here we have used local regression for the age term, with a span of 0.7. We can also use the lo() function to create interactions before calling the gam() function. For example,

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

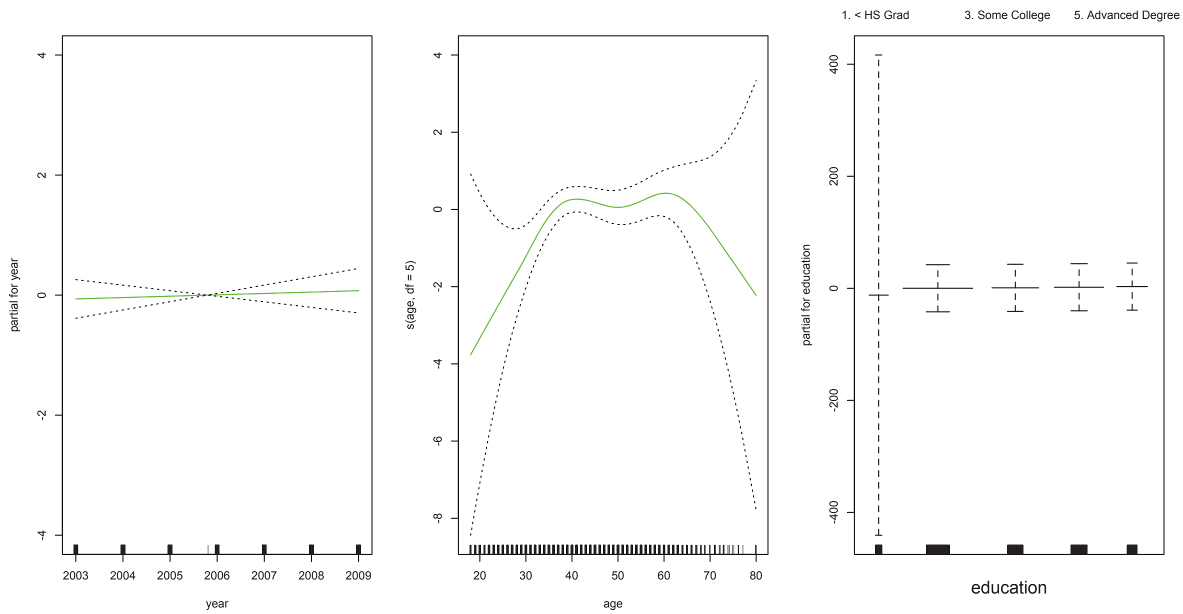
fits a two-term model, in which the first term is an interaction between year and age, fit by a local regression surface. We can plot the resulting two-dimensional surface if we first install the akima package.

```
> library(akima)
> par(mfrow = c(1, 2))
> plot(gam_lo_i, theta = 45, phi =0 ,se=T, col="green")
```



In order to fit a logistic regression GAM, we once again use the I() function in constructing the binary response variable, and set family=binomial.

```
> 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, ylim = c(-4,4), col = "green")
```



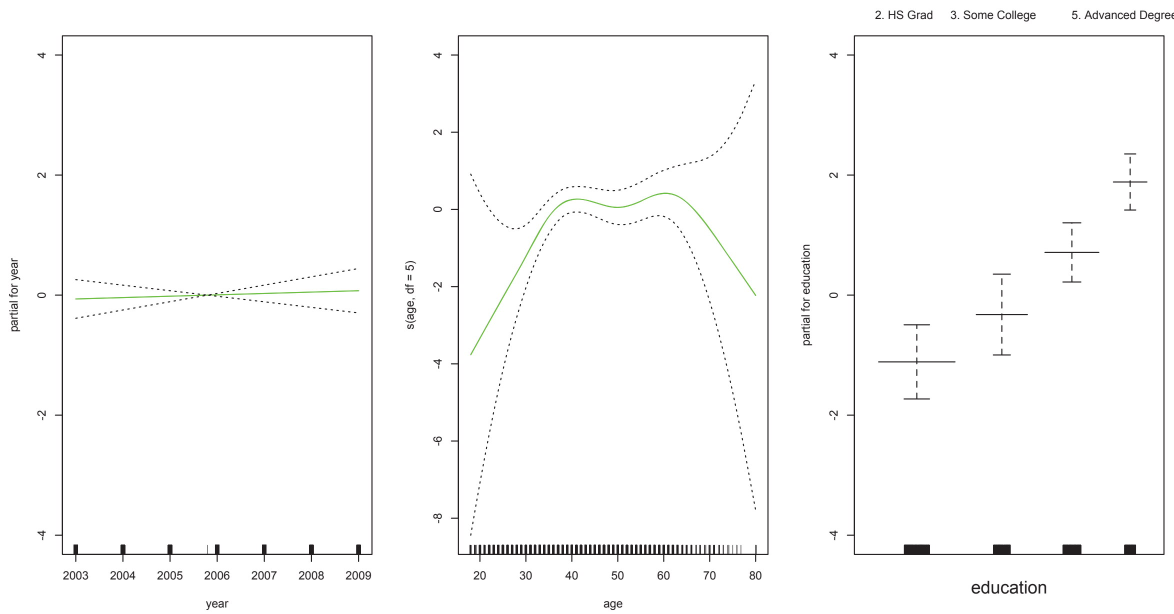
It is easy to see that there are no high earners in the < HS category:

```
> 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
```

Hence, we fit a logistic regression GAM using all but this category. This provides more sensible results.

```
> 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, ylim = c(-4,4), col = "green")
```


2. HS Grad
3. Some College
5. Advanced Degree
partial for year
year
2003
2004
2005
2006
2007
2008
2009
s(age, df = 5)
age
20
30
40
50
60
70
80
partial for education
education
4
2
0
-2
-4
-6
-8

Tree-Based Methods

- In this chapter, we describe tree-based methods for regression and classification. These involve stratifying or segmenting the predictor space into a number of simple regions.
- In order to make a prediction for a given observation, we typically use the mean or the mode response value for the training observations in the region to which it belongs.
- Since the set of splitting rules used to segment the predictor space can be summarized in a tree, these types of approaches are known as decision tree methods.

The Basics of Decision Trees

Decision trees can be applied to both regression and classification problems.

Regression Trees

In order to motivate regression trees, we begin with a simple example.

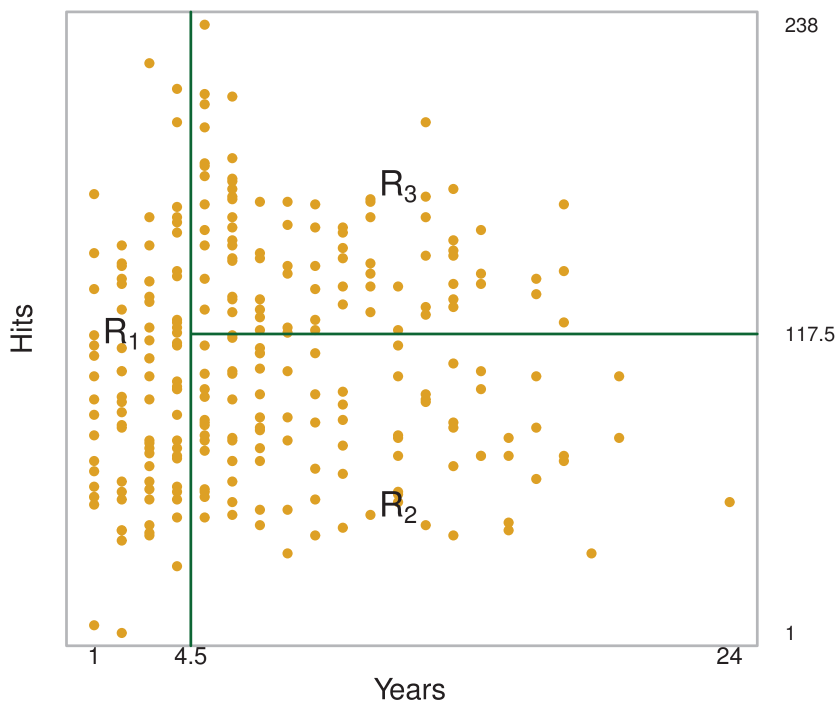
For the Hitters data, a regression tree for predicting the log salary of a baseball player, based on the number of years that he has played in the major leagues and the number of hits that he made in the previous year. At a given internal node, the label (of the form $X_j < t_k$) indicates the left-hand branch emanating from that split, and the right-hand branch corresponds to $X_j \geqslant t_k$. The tree has two internal nodes and three terminal nodes, or leaves. The number in each leaf is the mean of the response for the observations that fall there.

The predicted salary for the players in the left branch is given by the mean response value for the players in the data set with Years < 4.5. For such players, the mean log salary is 5.11, and so we make a prediction of $e^{5.11}$ thousands of dollars, that is \$165,174, for these players.

The tree stratifies or segments the players into three regions of predictor space:

$$R_1 = \{X | \text{Years} < 4.5\},$$
$$R_2 = \{X | \text{Years} \geqslant 4.5, \text{Hits} < 117.5\},$$
$$R_3 = \{X | \text{Years} \geqslant 4.5, \text{Hits} \geqslant 117.5\}$$

The regions R_1, R_2, and R_3 are known as terminal nodes or leaves of the tree.



The three-region partition for the Hitters data set from the regression tree illustrated in the previous figure.

The predicted salaries for these three groups are:

R_1: $\$1{,}000 \times e^{5.11} = \$165{,}174$

R_2: $\$1{,}000 \times e^{6.0} = \$402{,}834$

R_3: $\$1{,}000 \times e^{6.74} = \$845{,}346$

We might interpret the last regression tree as follows:

- Years is the most important factor in determining Salary, and players with less experience earn lower salaries than more experienced players.
- Given that a player is less experienced, the number of hits that he made in the previous year seems to play little role in his salary.
- But among players who have been in the major leagues for five or more years, the number of hits made in the previous year does affect salary, and players who made more hits last year tend to have higher salaries.

Prediction via Stratification of the Feature Space

We now discuss the process of building a regression tree. Roughly speaking, there are two steps.

1. We divide the predictor space — that is, the set of possible values for $X_1, X_2, \ldots, X_p$ — into J distinct and non-overlapping regions, R_1, R_2, ..., R_J.
2. For every observation that falls into the region R_j, we make the same prediction, which is simply the mean of the response values for the training observations in R_j.