# 6.1 Generalised Additive Models

## Generalised Additive Models

This section discusses *additive models* and explains how they extend to *generalised additive models*. Examples show how to fit both additive models and generalised models to data and how to interpret the obtained R output.

### Additive models

A natural way to extend the multiple linear regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

in order to allow for non-linear relationships between each feature and the response is to replace each linear component $\beta_j x_{ij}$ with a smooth non-linear function $f_j(x_{ij})$.

We can now write the model as

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

This is an example of a generalised additive model (GAM). It is called **additive** model.

### Example in R: Natural Splines

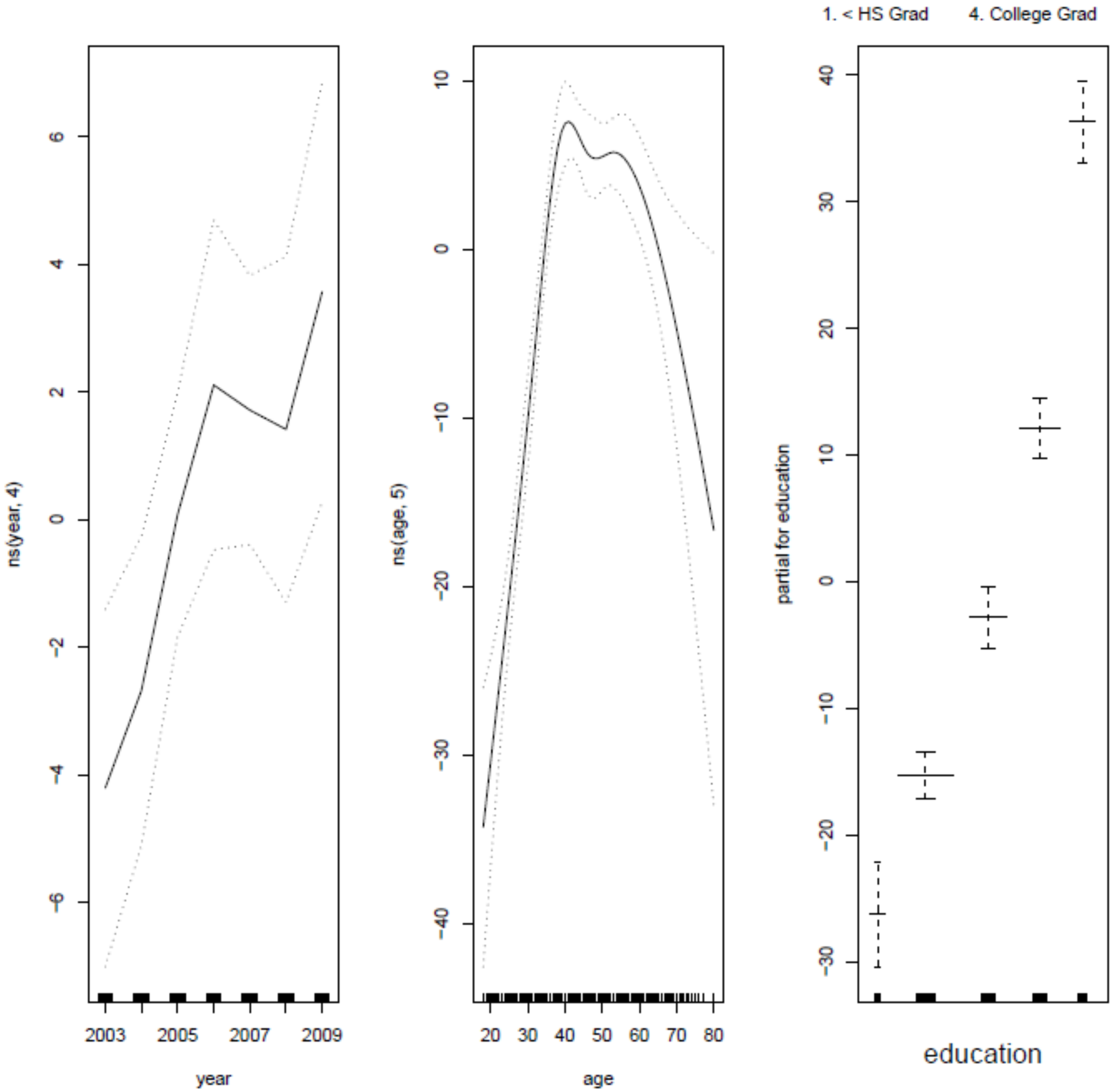
We consider the **Wage** data and fit the model

$$\mathtt{wage} = \beta_0 + \mathtt{f}_1(\mathtt{year}) + \mathtt{f}_2(\mathtt{age}) + \mathtt{education} + \varepsilon\mathtt{.}$$

Recall that natural splines can be constructed using an appropriately chosen set of basis functions. Therefore, the entire problem is just a big regression onto spline basis variables and dummy variables, all packed into one big regression (design) matrix.

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

data("Wage")
attach(Wage)

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

```
par(mfrow=c(1,3))

plot.Gam(gam1, se=TRUE)
```




We can see from the figure on the left that keeping **age** and **education** fixed, **wage** tends to increase slightly with **year**. The centre figure shows that holding **education** and **year** fixed, **wage** tends to be highest for intermediate values of **age**. Moreover, as indicated on the right hand side figure, keeping **year** and **age** fixed, **wage** tends to increase with **education**. All these findings are quite intuitive.

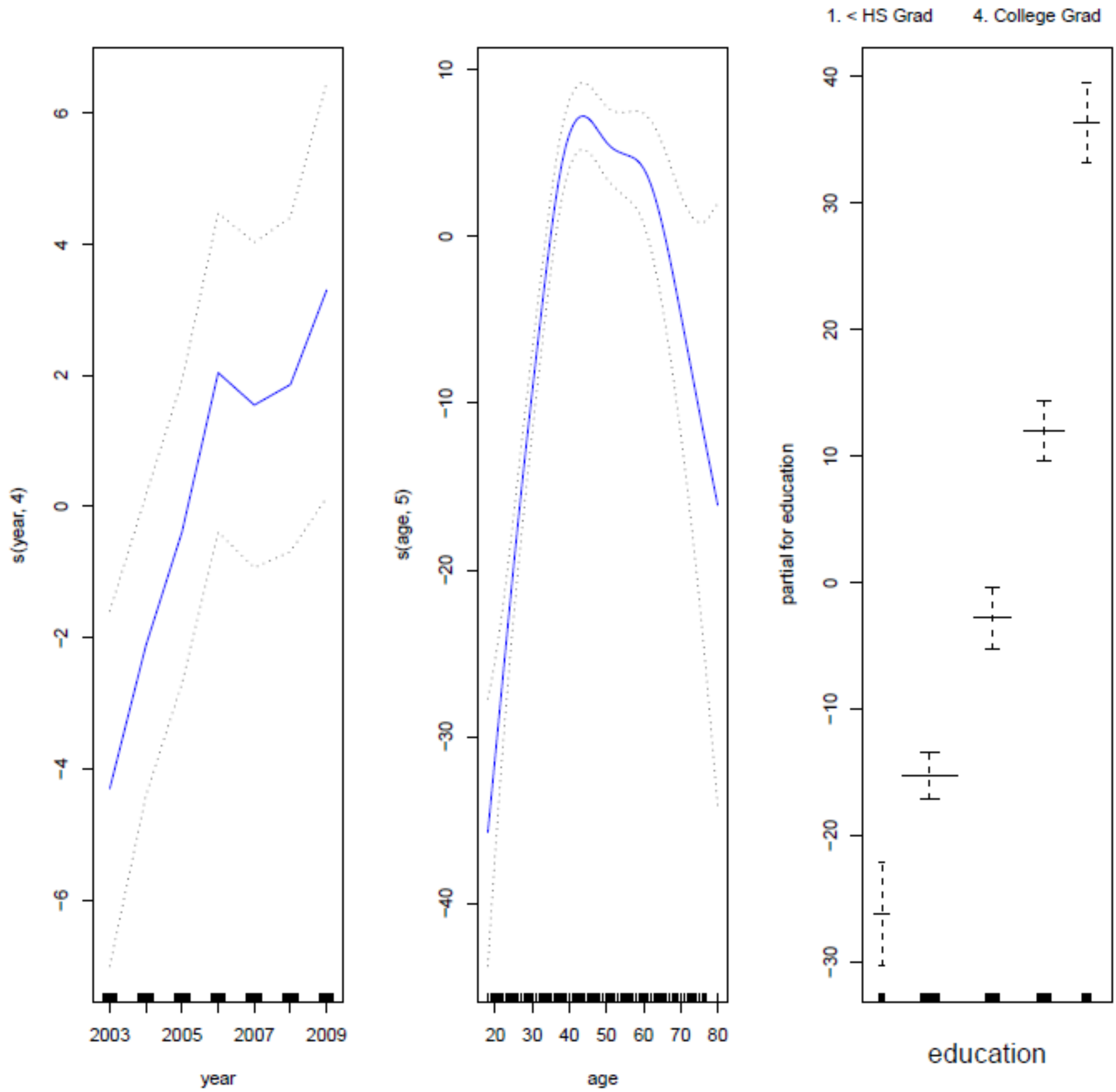
## Example in R: Smoothing Splines

We will now use smoothing splines in the above example instead of a natural spline. Fitting GAM with smoothing spline is not as simple as fitting GAM with natural splines. In the case of smoothing splines least squares cannot be used.

Why is that the case? The penalised sum of squares for this problem can be minimised by an additive cubic spline model where each of the functions $f_j$ is a cubic spline in the component $X_j$ , with knots at each of the unique values of $x_{ij}, i = 1, \ldots, N$. Without further restrictions on the model, the solution is not unique. The constant $\lambda$ is not identifiable, since we can add or subtract any constants to each of the functions $f_j$ , and adjust $\lambda$ accordingly.

Luckily, the $\mathbf{gam}()$ function in $\mathbf{R}$ can fit this model via an approach known as **backfitting**, which we will discuss later on.

```
library(ISLR)
library(gam)

data("Wage")
attach(Wage)

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")
```




The fitted functions look somewhat similar to the previous example. **In most situations, the difference in the GAMs obtained using smoothing splines versus natural splines are small.**

Note that here we used $\mathbf{s}()$ to tell $\mathbf{gam}()$ to fit a smoothing spline of degrees of freedom indicated by the second parameter of this wrapper.

## Example: `gam()` output interpretation and prediction

Let us now inspect the summary of the `gam()` function corresponding to $M_3$:

```
library(ISLR)
library(gam)
```

```
data("Wage")
attach(Wage)

gam.m3=gam(wage~s(year,4)+s(age,5)+education, data=Wage)
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: 2

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
```

The way the output of this approach to fitting GAMs is structured is to group the linear parts of the smoothers in with the other parametric terms. Notice `education` has an entry in the first table but its entry is empty in the second. This is because `education` is a strictly parametric term; it is a factor variable and hence is associated with an estimated parameter which represents the effect of `education`. The reason the smooth terms are separated into two types of effect is that this output allows you to decide if a smooth term has

1. *a nonlinear effect*: we use the nonparametric table and assess significance. If significance, we leave it as a smooth nonlinear effect. If insignificant, consider the linear effect (2. below)
2. *a linear effect*: look at the parametric table and assess the significance of the linear effect. If significant you can turn the term into a smooth `s(x)=x` in the formula describing the model. If insignificant you might consider dropping the term from the model entirely (note that this

amounts to a strong statement that the true effect is equal to $0$).

The p-values in the nonparametric part of the output corresponding to a null hypothesis of a linear relationship versus the alternative of a non-linear relationship. The large p-value for **year** supports our output from the ANOVA test that a linear function is adequate for this term (confirmed by the small pvalue in the ANOVA for Parametric Effects table). However, there is apparent evidence that a non-linear term is required for **age**.

We can also make predictions from **gam** objects, just like from **lm** objects, using the **predict**() function:

```
library(ISLR)
library(gam)

data("Wage")
attach(Wage)

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

We can also use local regression fits as building blocks in GAM:

```
library(ISLR)
library(gam)

data("Wage")
attach(Wage)

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

We can also use local regression to create interactions before calling the **gam**() function:

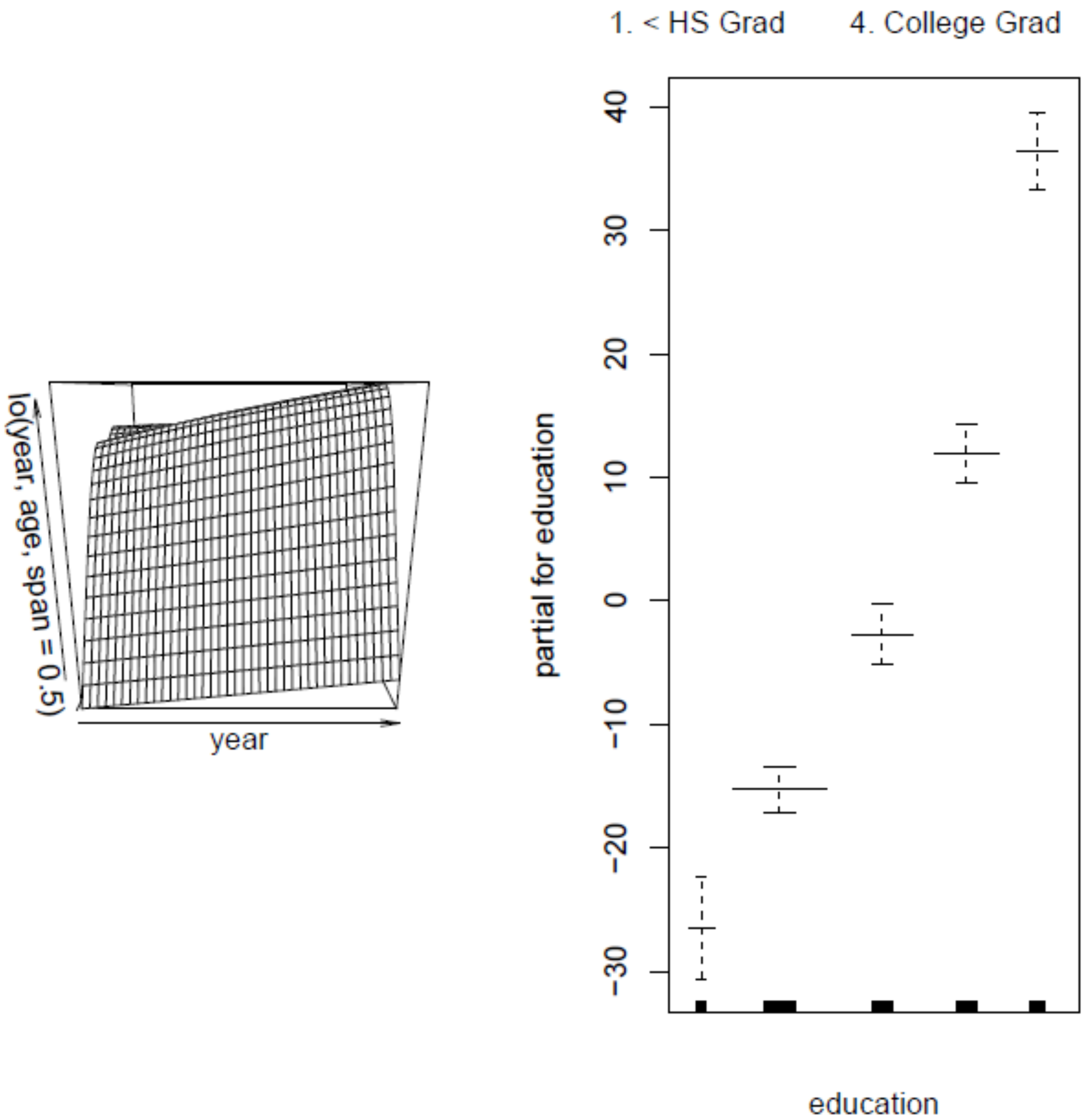
```
library(ISLR)
library(gam)

data("Wage")
attach(Wage)

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

We can plot the resulting surface using package **akima**:

```
library(ISLR)
library(gam)

data("Wage")
attach(Wage)

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

```
library(akima)
par(mfrow=c(1,2))
plot(gam.lo.i,se=TRUE)
```

```
```




## Generalised additive models

In general, the conditional mean $\mu$ of a response $Y$ is related to an additive function of the predictors via a **link** function $g$:

$$g(\mu) = \alpha + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p).$$

Recall the classical examples of link functions:

(a) $g(\mu) = \mu$ is the identity link, used for linear and additive models for Gaussian response data;

(b) $g(\mu) = \mathrm{logit}(\mu)$ for modeling binomial probabilities;

(c) $g(\mu) = \log(\mu)$ for log-linear or log-additive models for Poisson count data.

All three of these arise from exponential family sampling models. These families generate the previously mentioned generalised linear models, which are all extended in the same way to generalised additive models.

As we observed before not all $f_j$ functions need to be non-linear. We can mix in linear and other parametric forms with the nonlinear terms, a necessity when some of the inputs are qualitative variables. The nonlinear terms are not restricted to main effects either: we can have nonlinear components in two or more variables, or separate curves in $X_j$ for each level of the factor $X_k$.

Examples:

(a) $g(\mu) = X^\top \beta + \alpha_k + f(Z)$ - a **semiparametric** model, where $X$ is a vector of predictors to be modelled linearly, $\alpha_k$ the effect for the $k$th level of a qualitative input $V$, and the effect of predictor $Z$ is modelled nonparametrically.

(b) $g(\mu) = f(X) + g_k(Z)$ - $k$ indexes the levels of a qualitative input $V$, and creates an interaction term $g(V, Z) = g_k(Z)$ for the effects of $V$ and $Z$.

(c) $g(\mu) = f(X) + g(Z, W)$ where $g$ is a nonparametric function in two variables.

For fully general models, we have to look for even more flexible approaches as random forests and boosting. GAMs provide a useful compromise between linear and fully nonparametric models.

## Example: GAM logistic regression

In order to fit logistic regression GAM we execute:

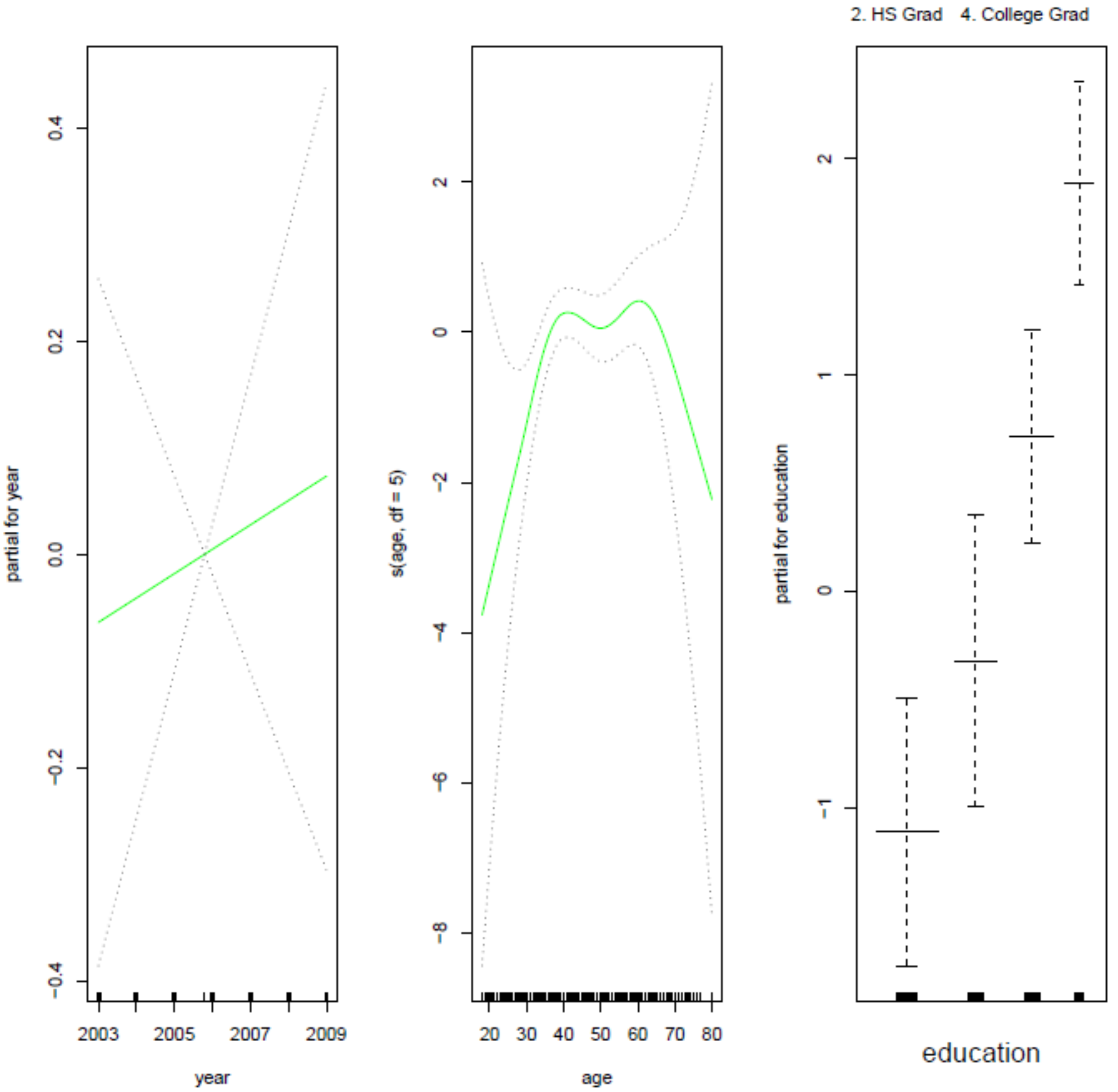
```
library(ISLR)
library(gam)

data("Wage")
attach(Wage)

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

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

```
library(ISLR)
library(gam)

data("Wage")
attach(Wage)

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.

```
library(ISLR)
library(gam)

data("Wage")
attach(Wage)

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")
```




All three panels have the same vertical scale which allows to visually assess the relative contributions of each of the variables. Here we observe that `age` and `education` have a much larger effect than `year` on the probability of being a high earner.

## Activity in R: Model Comparison

In the pervious examples the function of **`year`** looks rather linear. Perform a series of ANOVA tests in order to determine which of these three models is best:

(a) $M_1$ a GAM that excludes **`year`**;

(b) $M_2$ a GAM that uses a linear function of **`year`**;

(c) $M_3$ a GAM that uses spline function of **`year`**.

Note that each model should also include the **`age`** and **`education`** variables as previously defined.

# Activity in R: Generalised Additive Models

This question relates to the `College` data set form the library `ISLR`.

(a) Split the data into a training set and a test set. Using out-of-state tuition as the response and the other variables as the predictors, perform forward stepwise selection on the training set in order to identify a satisfactory model that uses just a subset of the predictors.

(b) Fit a GAM on the training data, using out-of-state tuition as the response and the features selected in the previous step as the predictors. Plot the results, and explain your findings.

(c) Evaluate the model obtained on the test set and explain the results obtained.

(d) For which variables, if any, is there evidence of a non-linear relationship with the response?

## Additional Activity

**Question 1** *Submitted Mar 17th 2023 at 1:03:42 am*

Additive models arise by

- ( ) introducing a link funtion in the multiple linear regression model;
- ( ) allowing non-linear releationships between the predictors and the response variable in the logistic regression model;
- (•) extending the multiple regression model to allow for non-linear relationships between the predictors and the response variable.

**Question 2** *Submitted Mar 17th 2023 at 1:03:49 am*

Which of the following model fits in R can be used when fitting an additive model with smoothing splines, where $\mathtt{y}$ is the response variable and $\mathtt{x1}, \mathtt{x2}$ are predictors:

- [ ] $\mathtt{lm}(\mathtt{y}\tilde{}\,\mathtt{s}(\mathtt{x1}, 4) + \mathtt{s}(\mathtt{x2}, 4))$
- [x] $\mathtt{gam}(\mathtt{y}\tilde{}\,\mathtt{s}(\mathtt{x1}, 4) + \mathtt{s}(\mathtt{x2}, 4))$
- [x] $\mathtt{gam}(\mathtt{y}\tilde{}\,\mathtt{s}(\mathtt{x1}, 4) + \mathtt{s}(\mathtt{x2}, 4), \mathtt{family} = \mathtt{binomial})$

**Question 3** *Submitted Mar 17th 2023 at 1:03:59 am*

Which of the following statements are true about the generalised additive models:

- [ ] the generalised additive model is a special case of an additive model;
- [x] the generalised additive model extends the additive model by introducing a link function to the model;

- [x] an example of a generalised additive model is

$$log(\mu) = X^\top \beta + \alpha_k + f(Z)$$

where $X$ is a vector of predictors to be modelled linearly, $\alpha_k$ the effect for the $k$th level of a qualitative input $V$, and the effect for predictor $Z$ is modelled nonparametrically.