비선형회귀모형들 - Part II

이재용,임요한

서 울 대 학 교
통 계 학 과

2017년 8월

일반화가법모형(generalized additive models) I

모형

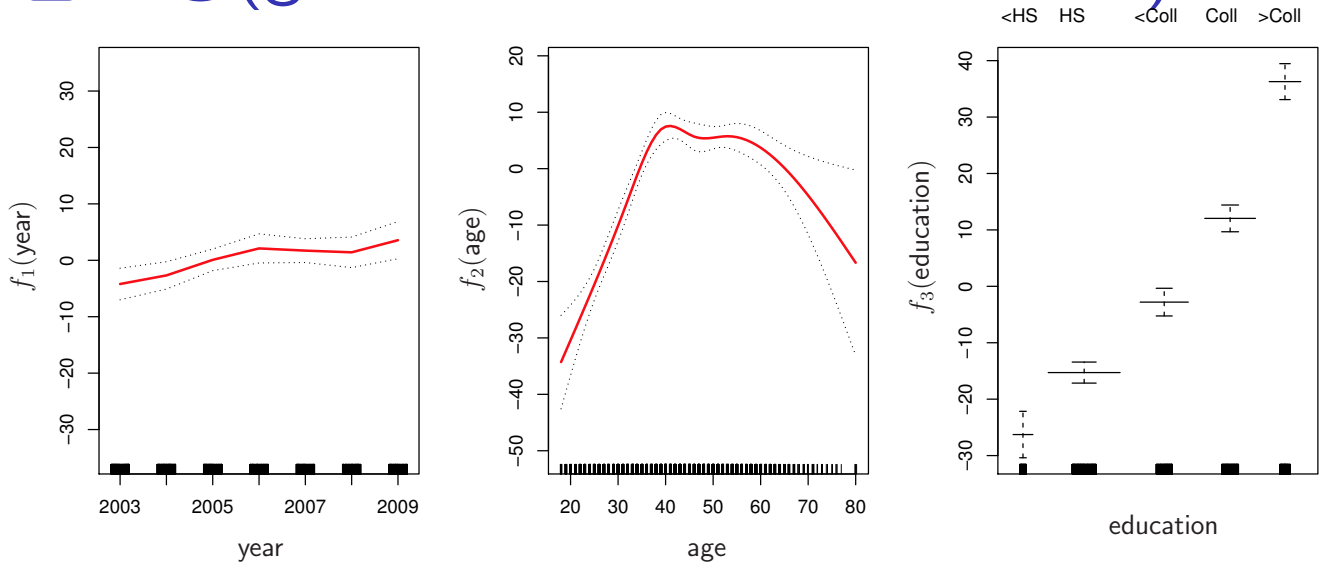
여러 개의 설명변수가 있을 때, 각 변수에 1변수 회귀함수를 적용하는 것을 말한다.

$$y_i = \beta_0 + f_1(x_{i1}) + \ldots + f_p(x_{ip}) + \epsilon_i, \;\; i = 1, 2, \ldots, n.$$

연봉자료의 예

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

일반화가법모형(generalized additive models) II



역적합(backfitting)

설명변수 하나씩 돌아가면서 나머지 설명변수와 그의 적합된 함수를 고정한 후, 한 개의 변수에 관해서만 회귀함수를 구하는 방법

노트. 참고.

평활스플라인의 경우 최소제곱법을 사용할 수 없기 때문에,
일반화가법모형 전체에 최소제곱법을 적용하는 것이 어렵다.

일반화가법모형(generalized additive models) III

일반화가법모형의 장점과 단점

1. 각 설명변수에 비선형 함수 f_j를 적용하기 때문에, 설명변수를 변환할 필요가 없다.
2. 설명변수의 비선형 함수는 예측 성능을 향상시킬 수 있다.
3. 모형이 가법이기 때문에, 다른 설명변수들을 고정시켰을 때, 한 설명변수의 효과가 f_j이다.
4. f_j의 유연성은 자유도로 요약된다.
5. (단점) 가법모형이기 때문에 교호작용을 표현하지 못한다. 하지만, 교호작용을 표현하고 싶으면 $f(x_i, x_j)$를 적합하면 된다.

이산형 반응변수와 가법모형

반응변수가 이산형일 때도 가법모형을 적용할 수 있다.

$$\log\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + f_1(x_1) + \ldots + f_p(x_p)$$

일반화 가법모형 R 코드 I

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

예측변수 모두가 자연스플라인 기저로 구성된 가법모형을 적용하려면 lm을 이용하면 된다. 가법모형이 큰 선형모형일 뿐이기 때문이다.

```
library(gam)

## Loading required package: foreach
## Loaded gam 1.12

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

하나 이상의 예측변수에 평활스플라인을 적용하려면 gam 패키지의 함수 gam이 필요하다. 함수 gam 안의 함수 s는 평활스플라인의 기저를 생성하는 함수로 gam 패키지 안에 있는 함수이다.

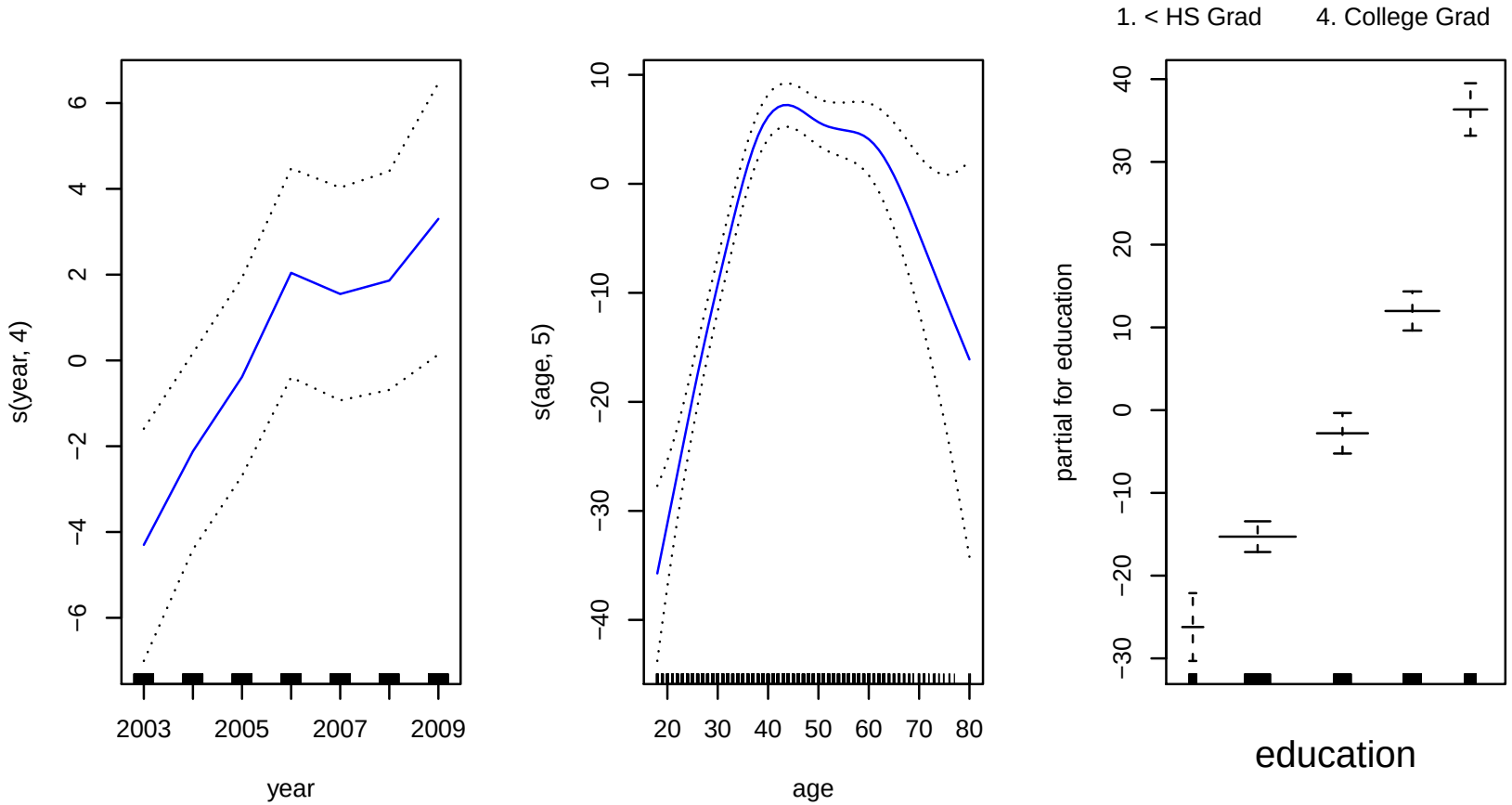
일반화 가법모형 R 코드 II

```
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


preds=predict(gam.m3,newdata=Wage)
```

일반화 가법모형 R 코드 III

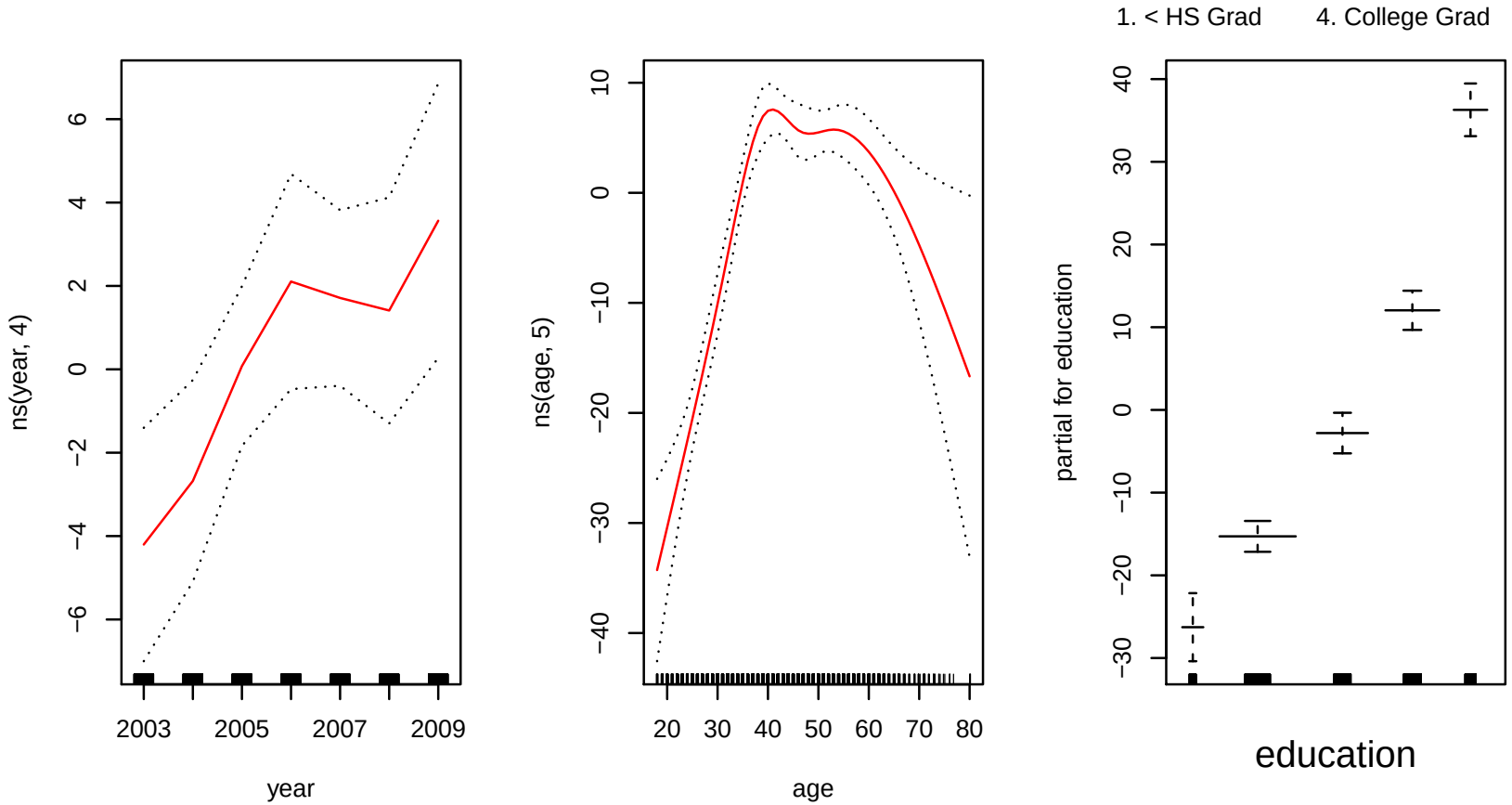
summary 함수와 predict 함수를 쓸 수 있다.

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



일반화 가법모형 R 코드 IV

```
plot.gam(gam1, se=TRUE, col="red")
```



gam 객체에 plot을 적용하면 세 개의 그림을 준다. gam1은 gam 객체가 아니지만 plot.gam 함수를 이용해서 동일한 그림을 그렸다.

일반화 가법모형 R 코드 V

```
gam.m3=gam(wage~s(year,1)+s(age,5)+education,data=Wage
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")
```

nested 모형 구조만 가능

```
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")
```

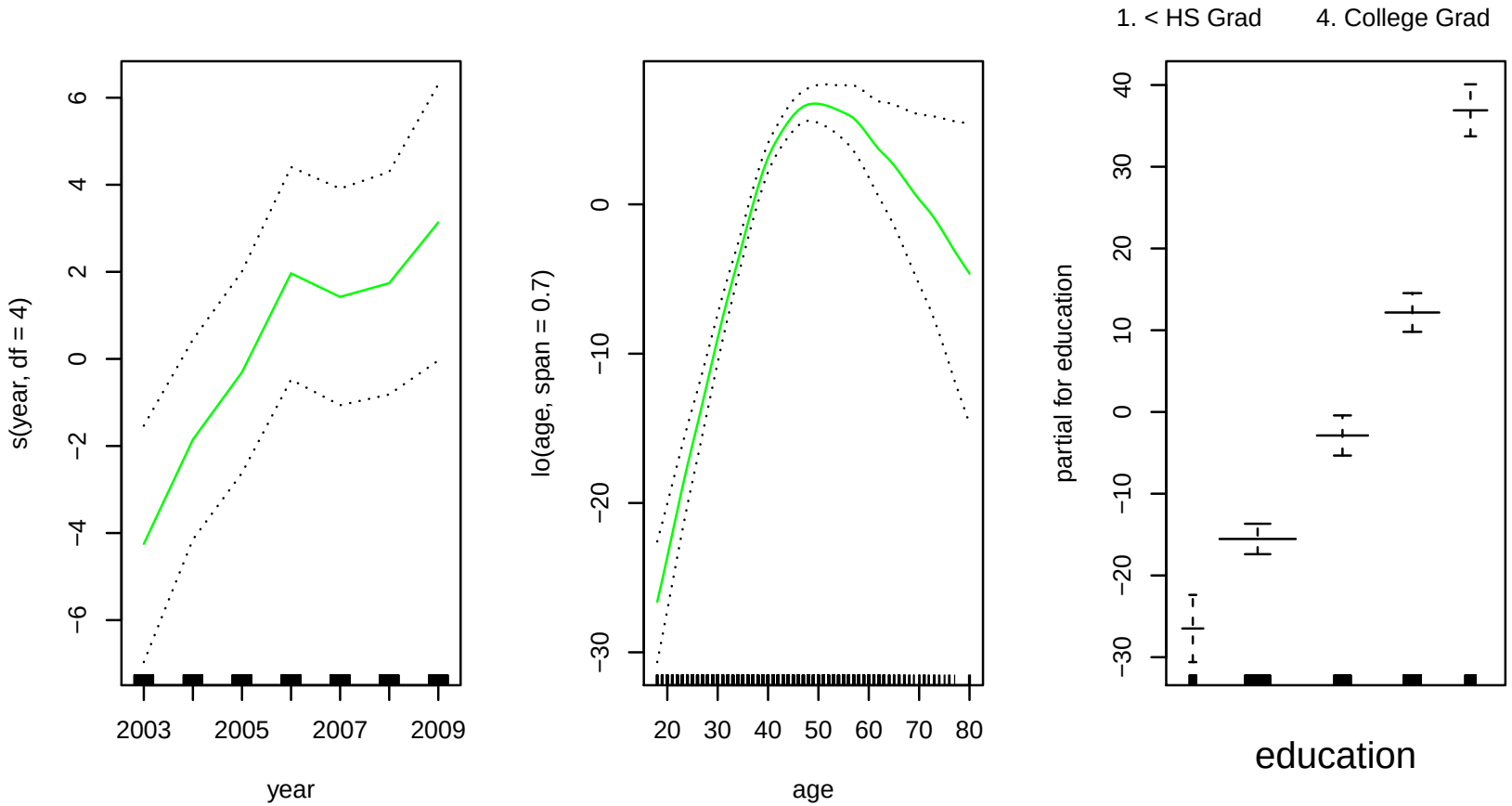
```
par(mfrow=c(1,3))
plot(gam.m3)

summary(gam.m2)
summary(gam.m3)
```

```
## 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
```

세 개의 모형을 anova 함수를 이용해서 비교하였다.

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

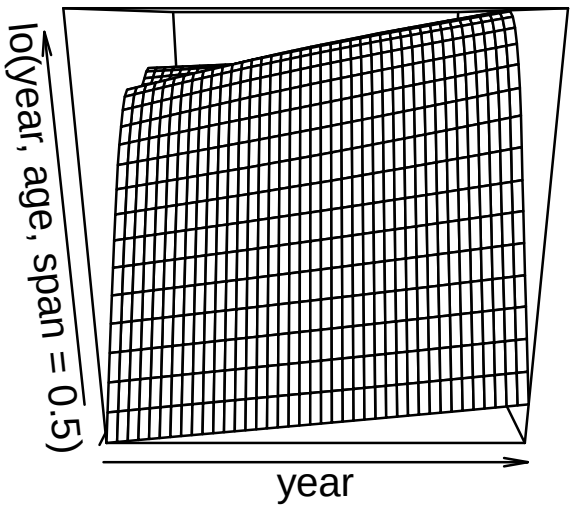
일반화 가법모형 R 코드 VI

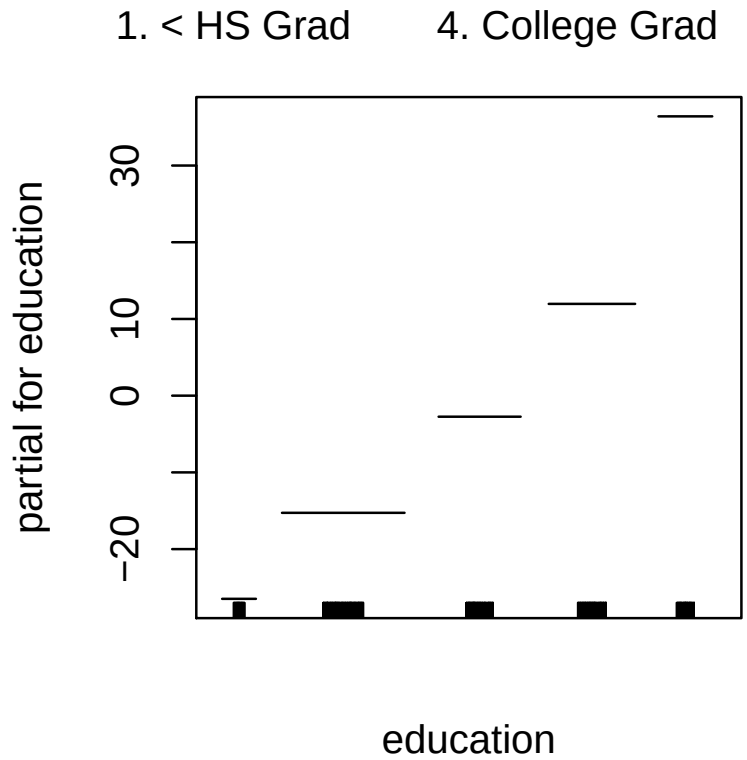


gam 함수안에 국소회귀를 쓸 수도 있다. lo는 국소회귀를 나타낸다.

```
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)
```

일반화 가법모형 R 코드 VII

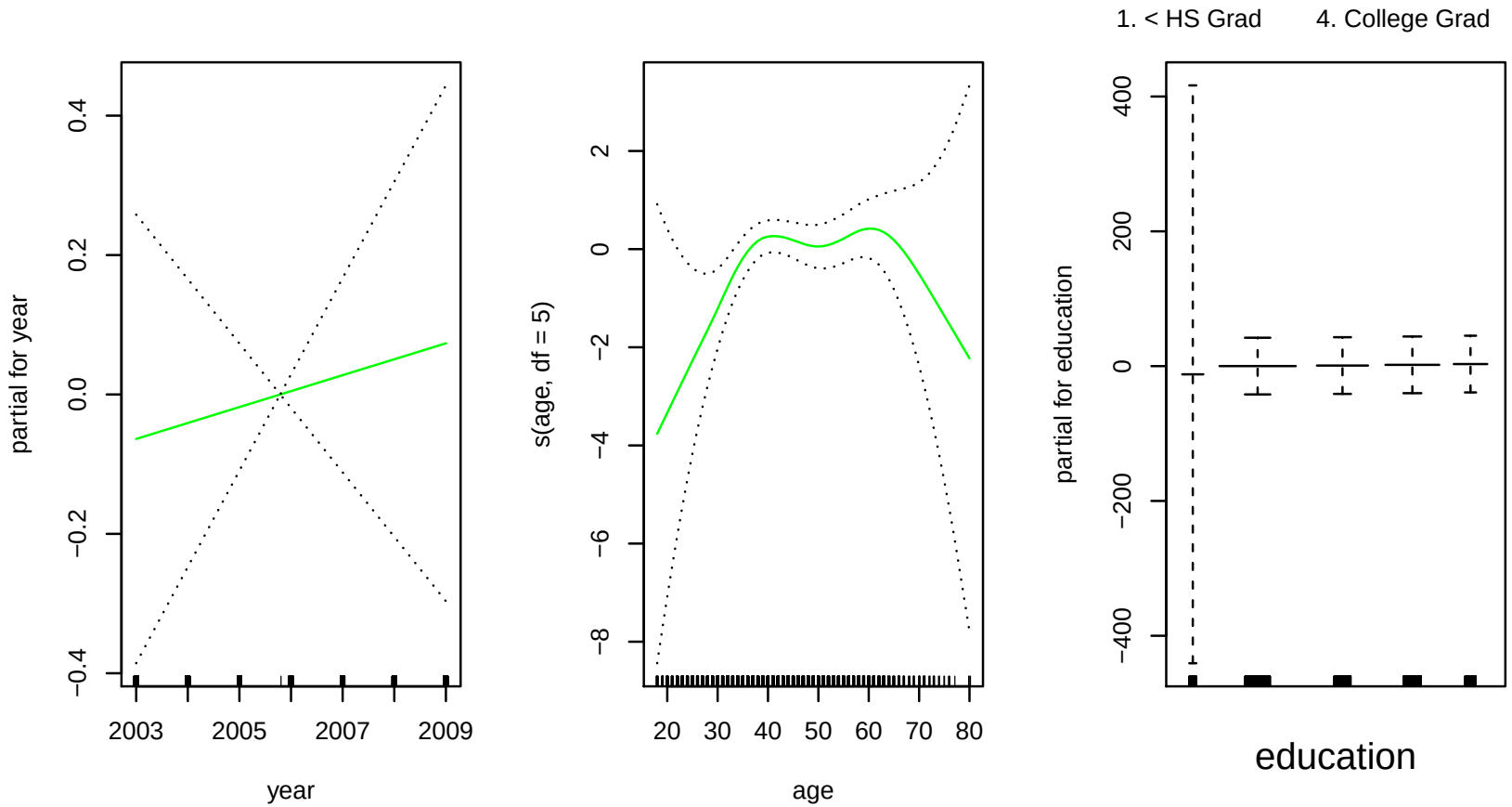




국소회귀는 두 개의 변수에 적용할 수 있다. 두 변수의 교호작용의 그림은 akima 패키지를 이용해 그릴 수 있다.

일반화 가법모형 R 코드 VIII

```
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")
```



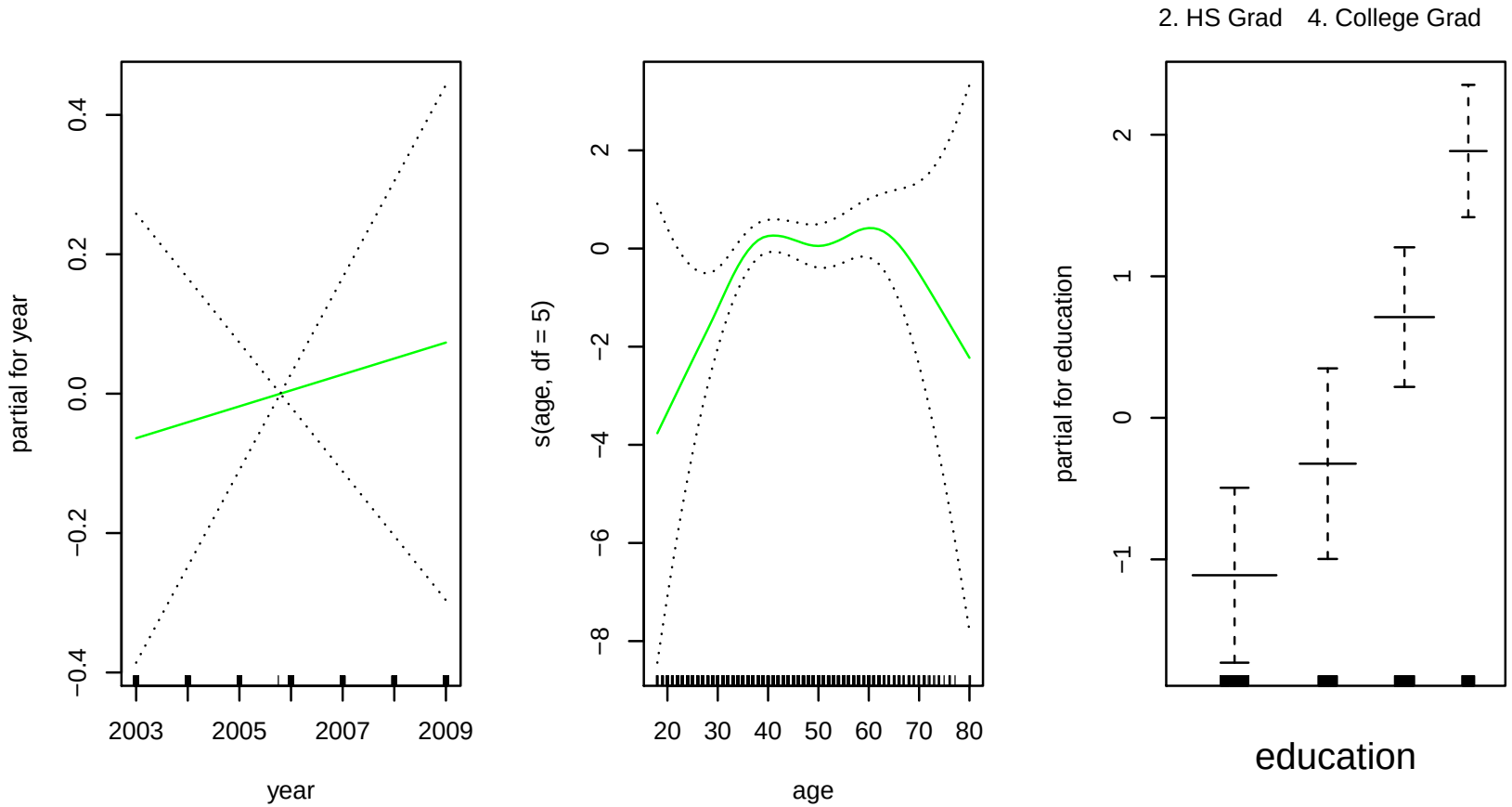
gam 함수를 이용해서 로지스틱모형을 적합할 수도 있다.

일반화 가법모형 R 코드 IX

```
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

par(mfrow=c(1,3))
gam.lr.s=gam(I(wage>250)~year+s(age,df=5)+education,family=binomial,data=Wage,subset=(educa
plot(gam.lr.s,se=T,col="green")
```

subset=(education!="1. < HS Grad"))

일반화 가법모형 R 코드 X



고등학교 졸업 미만의 교육을 받은 사람들 중에는 소득이 25만불 이상되는 고소득자가 없다. 고등학교 졸업 미만의 교육을 받은 사람들을 빼고 다시 적합했다.

참고문헌

아래의 책에서 제공하는 그림들을 사용하였다.

1. Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*. Springer, 2013.