Clase 9 - Módulo 2: Introducción a la analítica

Mauricio Alejandro Mazo Lopera

Universidad Nacional de Colombia
Facultad de Ciencias
Escuela de Estadística
Medellín

UNIVERSIDAD
NACIONAL
DE COLOMBIA

Pasemos de:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ip} + \epsilon_i$$

Al modelo GAM (Generalized Additive Model):

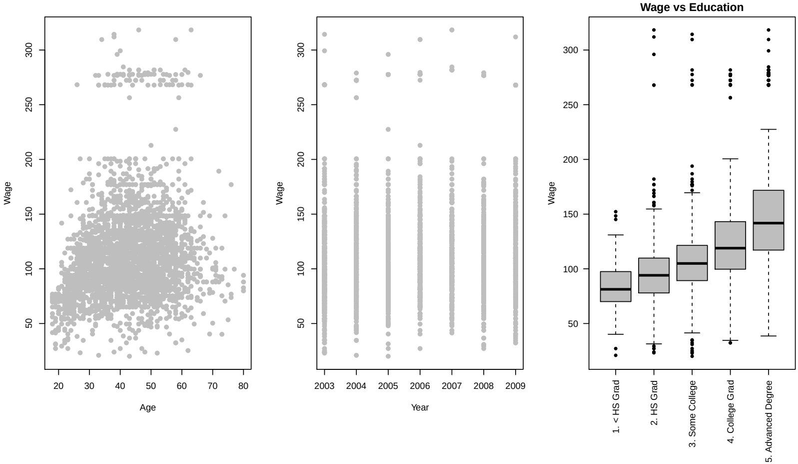
$$Y_i = \beta_0 + f_1(X_{i1}) + f_2(X_{i2}) + \cdots + f_p(X_{ip}) + \epsilon_i$$

¿Quiénes son las funciones $f_1, f_2, \ldots, f_p$?

$f_1, f_2, \ldots, f_p$ pueden ser funciones:

- **Paramétricas:** Polinomios, splines cúbicos, etc.
- **No paramétricas:** Smoothing splines, k-means, etc.
- **Semi-paramétricas:** Combinaciones de las dos anteriores.

Modelos aditivos generalizados



En cada caso, ¿cuál función seleccionaría para modelar **wage**?

Modelos aditivos generalizados

```r
require(splines)
mod1 <- lm(wage~ns(year,4)+ns(age,5)+education,data=Wage)
summary(mod1)
```

```
##
## Call:
## lm(formula = wage ~ ns(year, 4) + ns(age, 5) + education, data = Wage)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -120.513  -19.608   -3.583   14.112  214.535
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   46.949      4.704   9.980  < 2e-16 ***
## ns(year, 4)1                   8.625      3.466   2.488  0.01289 *
## ns(year, 4)2                   3.762      2.959   1.271  0.20369
## ns(year, 4)3                   8.127      4.211   1.930  0.05375 .
## ns(year, 4)4                   6.806      2.397   2.840  0.00455 **
## ns(age, 5)1                   45.170      4.193  10.771  < 2e-16 ***
## ns(age, 5)2                   38.450      5.076   7.575 4.78e-14 ***
## ns(age, 5)3                   34.239      4.383   7.813 7.69e-15 ***
## ns(age, 5)4                   48.678     10.572   4.605 4.31e-06 ***
## ns(age, 5)5                    6.557      8.367   0.784  0.43328
## education2. HS Grad           10.983      2.430   4.520 6.43e-06 ***
## education3. Some College      23.473      2.562   9.163  < 2e-16 ***
## education4. College Grad      38.314      2.547  15.042  < 2e-16 ***
## education5. Advanced Degree   62.554      2.761  22.654  < 2e-16 ***
## ---
```

Para ver los nodos seleccionados por la función **ns()**:

```
attr(ns(Wage$year,4),"knots")
```

```
##  25%  50%  75%
## 2004 2006 2008
```

```
attr(ns(Wage$age,5),"knots")
```

```
## 20% 40% 60% 80%
##  32  39  46  53
```

Las funciones serían:

$$f_1(year) = \textbf{ns(year,4)}$$
$$f_2(age) = \textbf{ns(age,5)}$$

$$\begin{aligned} f_3(education) &= \boldsymbol{I(< HS\ Grad) + I(HS\ Grad) +} \\ &\quad \boldsymbol{I(Some\ College) + I(College\ Grad) +} \\ &\quad \boldsymbol{I(Advanced\ Degree)} \end{aligned}$$

donde

$$I(a) = \begin{cases} 1, & \text{si } education = a \\ 0, & \text{e.o.c} \end{cases}$$

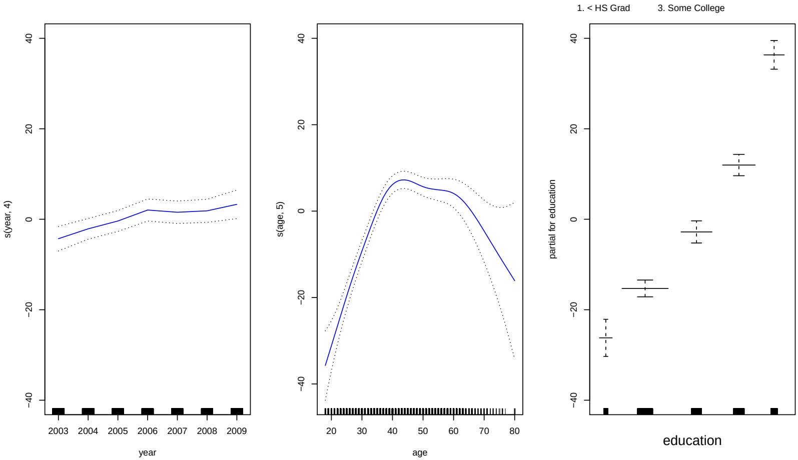
El paquete **gam** permite ajustar un modelo GAM y a diferencia del método anterior, donde usamos la función **lm**, esta permite usar smoothing splines.

```
require(gam)
mod.gam1<-gam(wage~s(year,4)+s(age,5)
              +education,data=Wage)
```

La función **s()** permite ajustar smothing spline. Para ver la ayuda, poner **?s**. Otra función de este paquete es **lo** que permite ajustar un LOESS.

Modelos aditivos generalizados

```r
par(mfrow =c(1,3))
plot(mod.gam1, se=TRUE ,col ="blue",ylim=c(-40,40))
```



```r
summary(mod.gam1)
```

```
##
## 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
```

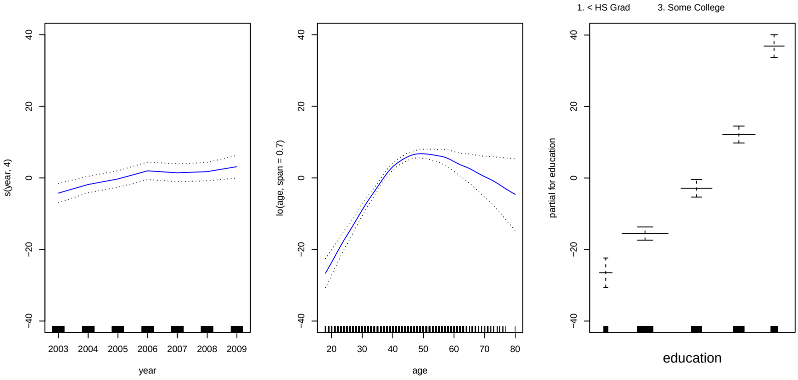
Comparando modelos:

```r
mod.gam2<-gam(wage~s(age,5)+education,data=Wage)
mod.gam3<-gam(wage~year+s(age,5)+education,data=Wage)
mod.gam4<-gam(wage~s(year,4)+s(age,5)+education,data=Wage)
anova(mod.gam2,mod.gam3,mod.gam4,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 ' '
```

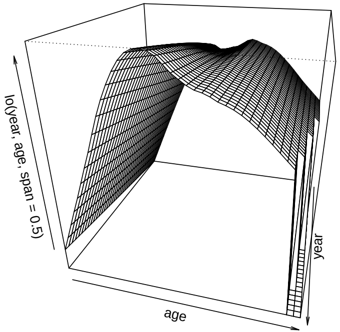
Ajustando otro modelo con LOESS:

```
mod.gam5<-gam(wage~s(year,4)+lo(age,span=0.7)
              +education,data=Wage)
par(mfrow =c(1,3))
plot(mod.gam5, se=TRUE ,col ="blue",ylim=c(-40,40))
```



Ajustando otro modelo LOESS con interacción:

```
require(akima)
mod.gam6<-gam(wage~lo(year ,age ,span =0.5)
              +education,data=Wage)
plot(mod.gam6,phi = 25, theta=105)
```

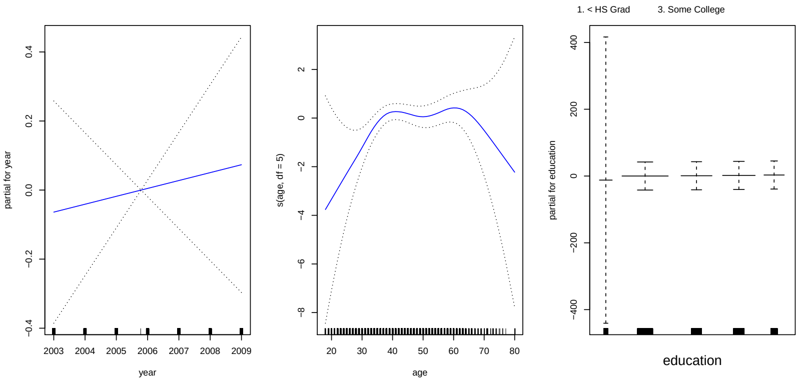


El modelo GAM se puede plantear de manera más general:

$$g[E(Y_i)] = \beta_0 + f_1(X_{i1}) + f_2(X_{i2}) + \cdots + f_p(X_{ip})$$

donde $g(\cdot)$ es conocida como la función link.

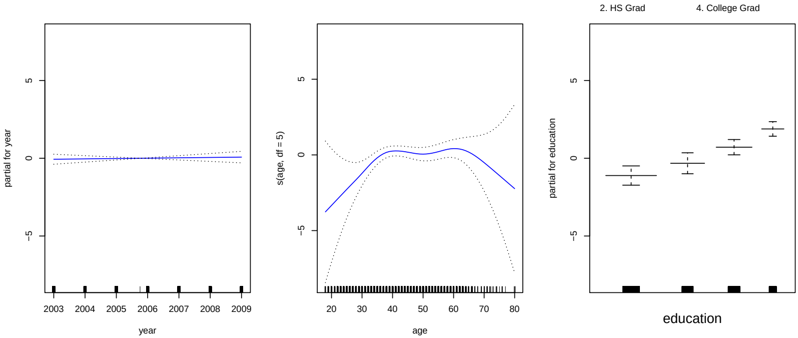
```
mod.gam7<-gam(I(wage>250)~year+s(age,df=5)+education,
             family=binomial,data=Wage)
par(mfrow =c(1,3))
plot(mod.gam7,se=T,col="blue")
```



```
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
```

```r
mod.gam8<-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(mod.gam8,se=T,col="blue",ylim=c(-8,8))
```



Actividad para realizar en clase:

Considere la base de datos **Credit** del paquete **ISLR**. Suponiendo que nuestra variable de interés es el **Balance**, realice la siguiente actividad:

a. Realice gráficos que permitan ver las relaciones existentes entre todas las variables de la base de datos.
b. Seleccione un conjunto de variables que considere útiles para modelar el **Balance**.
c. Por medio un análisis de varianza de modelos anidados, plantee al menos 4 modelos GAM.
d. Usando CV, seleccione el mejor modelo entre los cuatro modelos planteados en el item anterior.