Extensiones del modelo lineal: modelos aditivos

Contents

1 Modelo

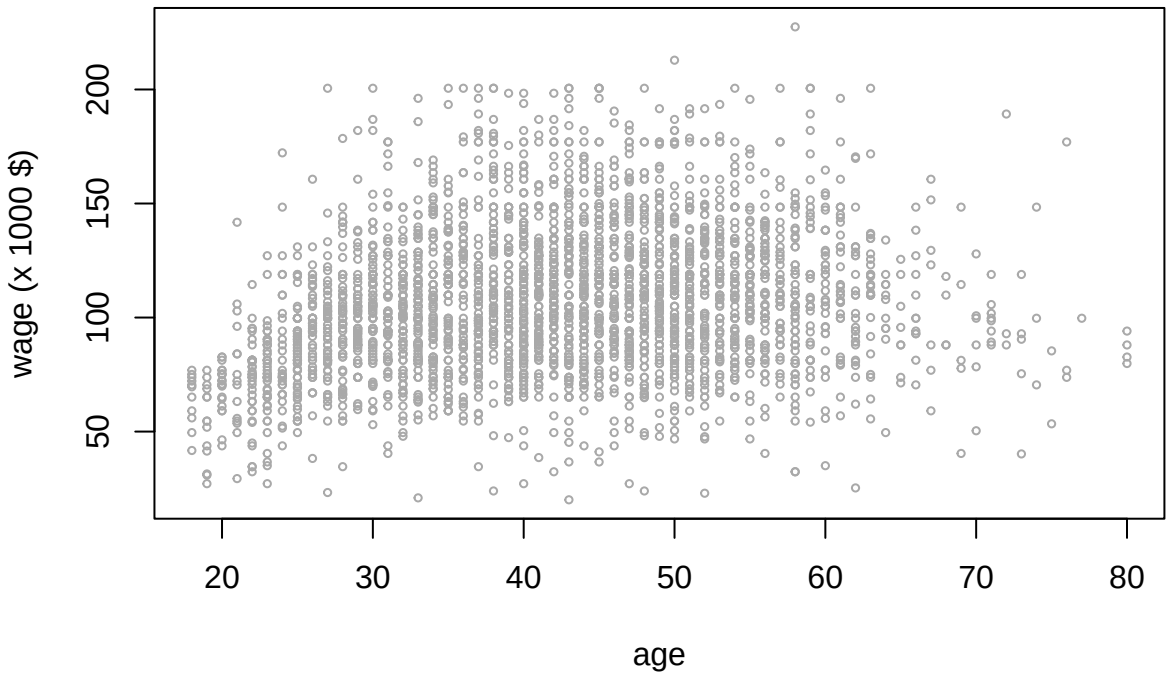
- Los extensiones del modelo lineal estudiados hasta ahora utilizan solo un regresor.
- Los Modelos Aditivos constituyen la manera natural de extender el modelo:
- Datos: $\{y_i, x_{1i}, x_{2i}, \ldots, x_{pi}\},\ i = 1, \ldots, n$

$$y_i = \beta_0 + f(x_{1i}) + f(x_{2i}) + \cdots + f(x_{pi}) + u_i$$

donde cada $f(x_i)$ puede ser un polinomio, un spline, un término $\beta_k x_{ki}$, una interacción,...

2 Ejemplo

```
d = read.csv("datos/Wage.csv")
d = d[d$wage<250,]
plot(d$age,d$wage, cex = 0.5, col = "darkgrey", ylab = "wage (x 1000 $)", xlab = "age")
```



Queremos trabajar con el modelo:

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

- Cuando las f_i están datas en términos de **funciones base**, se puede estimar el modelo utilizando mínimos cuadrados:

```
m1 = lm(wage ~ poly(age, degree = 3) + education, data = d)
summary(m1)
```

```
##
## Call:
## lm(formula = wage ~ poly(age, degree = 3) + education, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -100.235  -16.770   -0.634   16.194   90.930
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    85.371      1.622  52.621  < 2e-16 ***
## poly(age, degree = 3)1        307.473     26.666  11.530  < 2e-16 ***
## poly(age, degree = 3)2       -346.911     26.641 -13.022  < 2e-16 ***
## poly(age, degree = 3)3         92.928     26.569   3.498 0.000476 ***
## education2. HS Grad             9.971      1.832   5.442 5.71e-08 ***
## education3. Some College       21.329      1.930  11.049  < 2e-16 ***
## education4. College Grad       33.318      1.925  17.311  < 2e-16 ***
## education5. Advanced Degree    48.243      2.125  22.705  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.53 on 2913 degrees of freedom
## Multiple R-squared:  0.3119, Adjusted R-squared:  0.3102
## F-statistic: 188.6 on 7 and 2913 DF,  p-value: < 2.2e-16
```

```
m2 = lm(wage ~ year + poly(age, degree = 3) + education, data = d)
summary(m2)
```

```
##
## Call:
## lm(formula = wage ~ year + poly(age, degree = 3) + education,
##     data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -103.84  -16.77   -0.81   15.76   92.14
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -2175.3992   485.2418  -4.483 7.64e-06 ***
## year                           1.1271     0.2419   4.659 3.32e-06 ***
## poly(age, degree = 3)1       302.7188    26.5913  11.384  < 2e-16 ***
## poly(age, degree = 3)2      -350.2735    26.5566 -13.190  < 2e-16 ***
## poly(age, degree = 3)3        96.7958    26.4880   3.654 0.000262 ***
## education2. HS Grad            9.9366     1.8257   5.443 5.69e-08 ***
## education3. Some College      21.3851     1.9237  11.117  < 2e-16 ***
## education4. College Grad      33.2497     1.9179  17.336  < 2e-16 ***
```

```
## education5. Advanced Degree    48.1041     2.1174  22.718  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.43 on 2912 degrees of freedom
## Multiple R-squared:  0.317,  Adjusted R-squared:  0.3151
## F-statistic: 168.9 on 8 and 2912 DF,  p-value: < 2.2e-16
```

```
library(splines)
m3 = lm(wage ~ ns(year, df = 4) + poly(age, degree = 3) + education, data = d)
summary(m3)
```

```
##
## Call:
## lm(formula = wage ~ ns(year, df = 4) + poly(age, degree = 3) +
##     education, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -103.784  -16.784   -0.805   15.869   92.258
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    81.415      1.909  42.648  < 2e-16 ***
## ns(year, df = 4)1               7.482      2.642   2.832 0.004663 **
## ns(year, df = 4)2               3.338      2.254   1.481 0.138761
## ns(year, df = 4)3               8.609      3.204   2.686 0.007262 **
## ns(year, df = 4)4               6.178      1.826   3.383 0.000726 ***
## poly(age, degree = 3)1        303.571     26.597  11.414  < 2e-16 ***
## poly(age, degree = 3)2       -349.741     26.587 -13.155  < 2e-16 ***
## poly(age, degree = 3)3         97.189     26.499   3.668 0.000249 ***
## education2. HS Grad             9.858      1.827   5.397 7.33e-08 ***
## education3. Some College       21.240      1.926  11.030  < 2e-16 ***
## education4. College Grad       33.223      1.919  17.310  < 2e-16 ***
## education5. Advanced Degree    48.033      2.118  22.680  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.43 on 2909 degrees of freedom
## Multiple R-squared:  0.3177, Adjusted R-squared:  0.3152
## F-statistic: 123.2 on 11 and 2909 DF,  p-value: < 2.2e-16
```

- Comparamos los tres modelos con el contraste de la F:

```
anova(m1,m2)
```

```
## Analysis of Variance Table
##
## Model 1: wage ~ poly(age, degree = 3) + education
## Model 2: wage ~ year + poly(age, degree = 3) + education
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1   2913 2049905
## 2   2912 2034737  1     15168 21.707 3.319e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Los modelos no son equivalentes, y como en m2 ha salido significativo el coeficiente de year, nos quedamos con m2.

```
anova(m2,m3)
```

```
## Analysis of Variance Table
##
## Model 1: wage ~ year + poly(age, degree = 3) + education
## Model 2: wage ~ ns(year, df = 4) + poly(age, degree = 3) + education
##   Res.Df     RSS Df Sum of Sq     F Pr(>F)
## 1   2912 2034737
## 2   2909 2032438  3    2299.3 1.097 0.3491
```

Los modelos son equivalentes, nos quedamos con m2 ya que tiene menos parámetros.