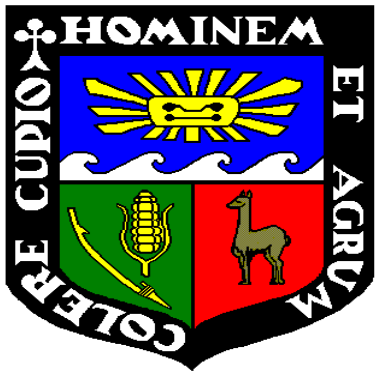
Técnicas Multivariadas Avanzadas

Más alla de la linealidad

Ms Carlos López de Castilla Vásquez

Universidad Nacional Agraria La Molina

2014-2

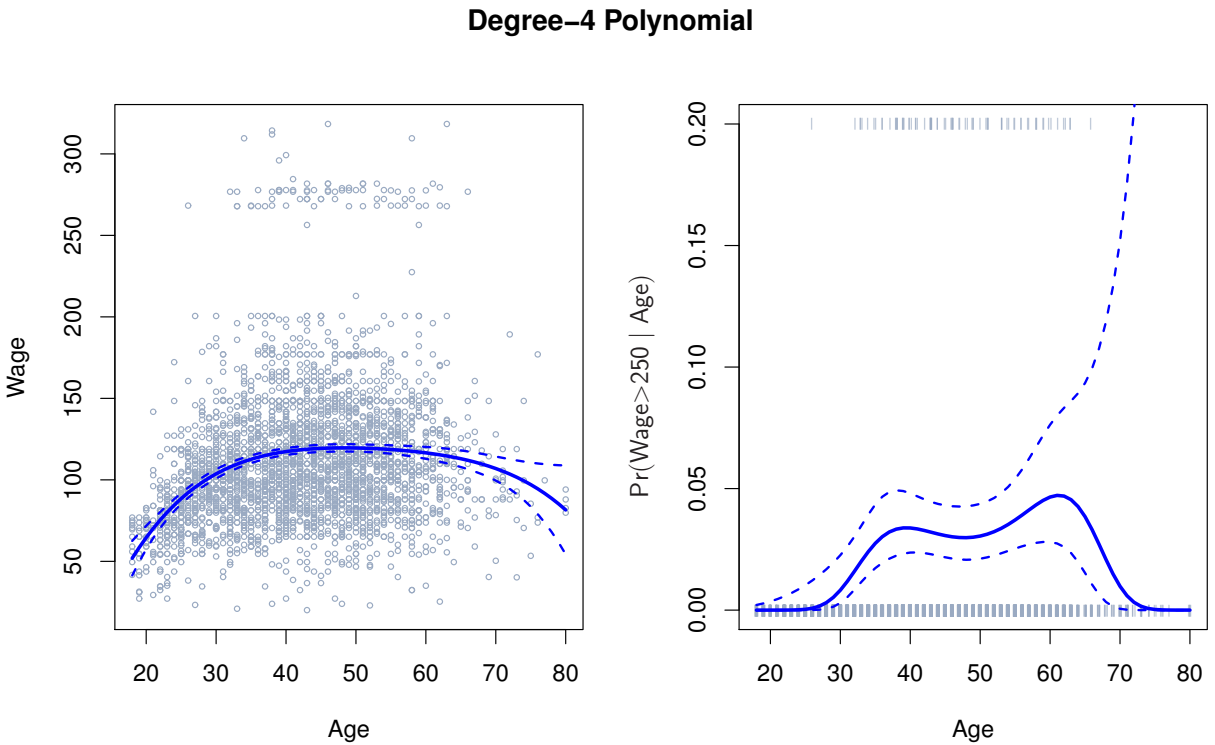


Introducción

- Las relaciones entre variables *casi nunca* son lineales.
- Sin embargo el supuesto de linealidad casi siempre es suficiente.
- Existen algunas alternativas como la regresión polinomial, las funciones paso, splines, regresión local y los modelos aditivos generalizados.
- Estos métodos ofrecen gran flexibilidad sin perder la esencia e interpretabilidad de los modelos lineales.

Regresión Polinomial

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \cdots + \beta_d x_i^d + \epsilon_i$$



Regresión Polinomial

- Se crean nuevas variables: $X_1 = X$, $X_2 = X^2$, etc y luego se estima el modelo de regresión usando MCO.
- El interés principal no esta en los coeficientes sino en los valores estimados para un valor x_0:

$$\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0 + \hat{\beta}_2 x_0^2 + \hat{\beta}_3 x_0^3 + \hat{\beta}_4 x_0^4$$

- Como $\hat{f}(x_0)$ es una función lineal de los $\hat{\beta}_l$ se puede obtener la expresión para la variancia $V(\hat{f}(x_0))$.
- En la gráfica anterior se observan las estimaciones puntuales y los intervalos correspondientes a:

$$\hat{f}(x_0) \pm 2SE(\hat{f}(x_0))$$

Regresión Polinomial

- Para determinar d puede usarse validación cruzada.
- Es posible aplicar una ecuación polinomial a la regresión logística, por ejemplo:

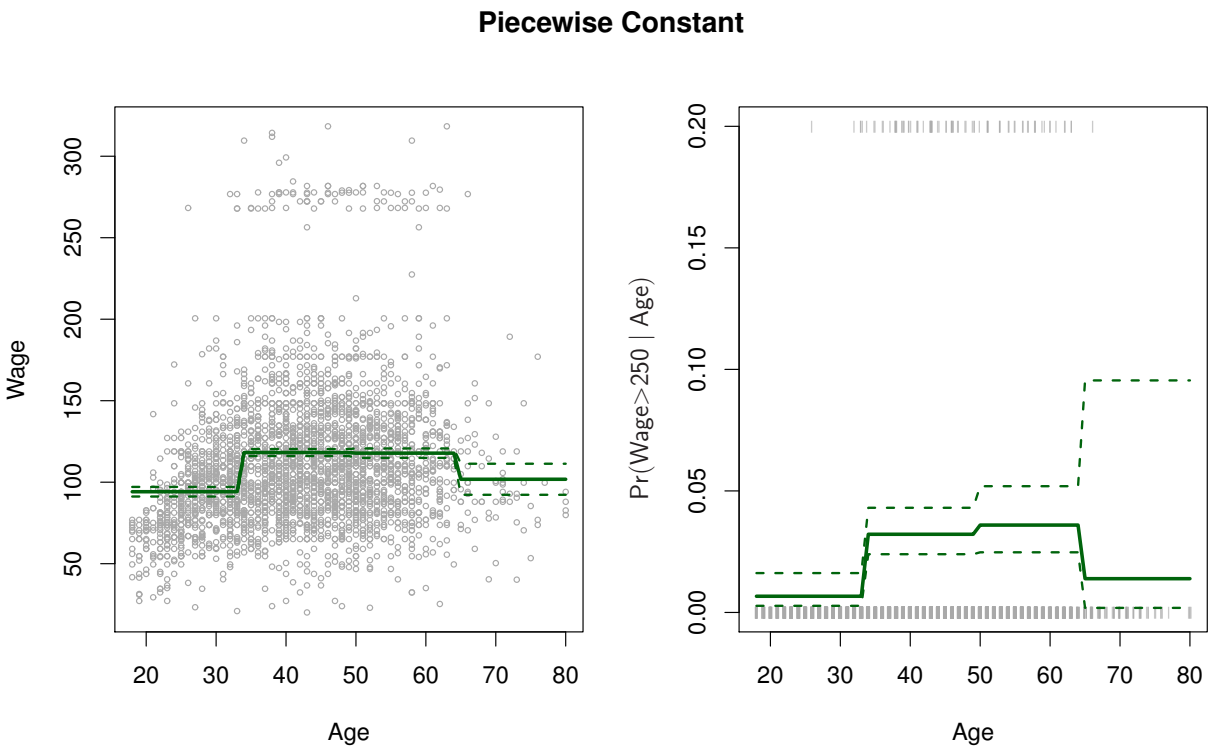
$$\Pr(y_i > 250|x_i) = \frac{\exp\{\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_d x_i^d\}}{1 + \exp\{\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_d x_i^d\}}$$

- Para obtener intervalos de confianza se calculan los limites en la escala logit y luego se convierten en probabilidades.
- *Advertencia*: Los modelos polinomiales son malos para extrapolación.
- En R: `y~poly(x,degree=3)`.

Funciones paso

- Se puede cortar la variables en diferentes regiones, por ejemplo:

$$C_1(X) = I(X < 35) \quad \cdots \quad C_4(X) = I(X \geq 65)$$



Funciones paso

- Se requiere de un conjunto de variables dummy para representar cada grupo.
- Se pueden establecer interacciones que son fáciles de interpretar, por ejemplo:

$$I(\texttt{Year} < 2005) \cdot \texttt{Age} \qquad I(\texttt{Year} \geq 2005) \cdot \texttt{Age}$$

  considera diferentes funciones lineales en cada grupo.
- En R: `I(year<2005)` o `cut(age,c(18,25,40,65))`.
- La elección de los puntos de puntos de corte o *knots* puede ser un problema.

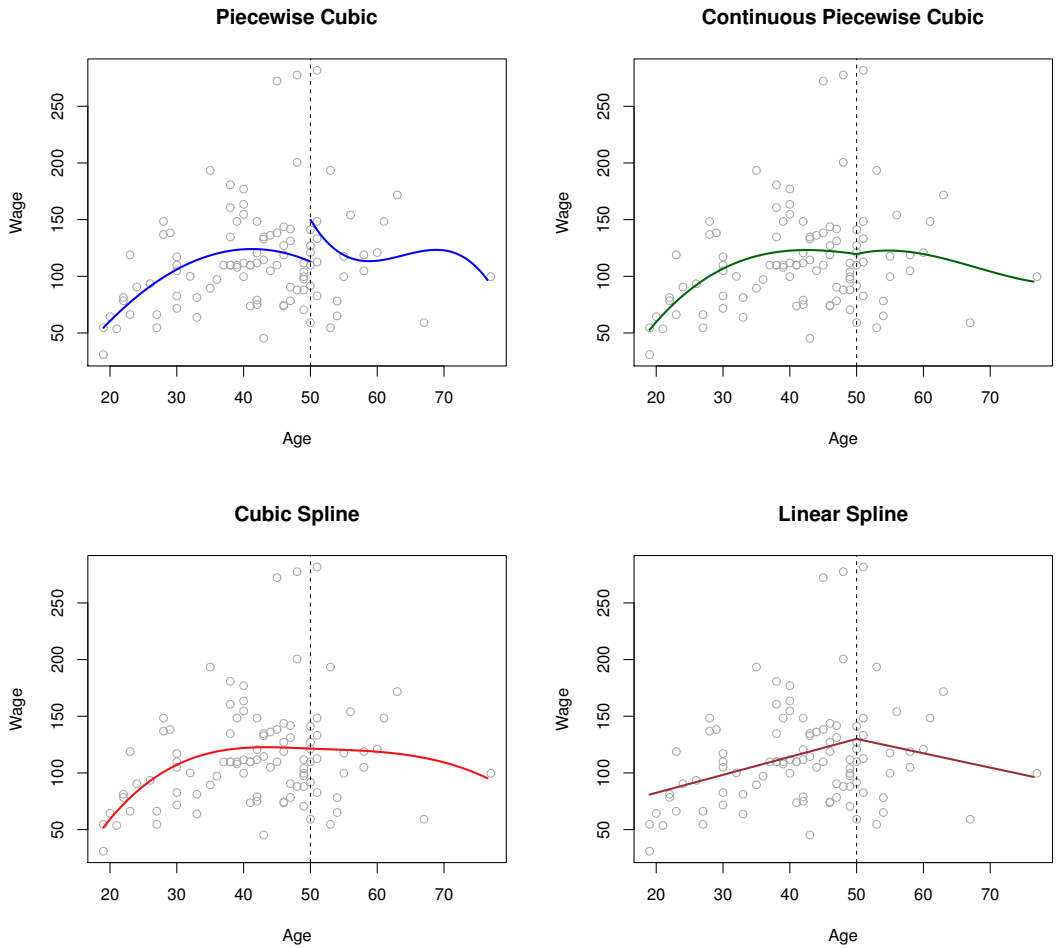
Polinomios por partes

- Se pueden usar polinomios diferentes en las regiones definidas por los knots, por ejemplo:

$$y_i = \begin{cases} \beta_{01} + \beta_{11} x_i + \beta_{21} x_i^2 + \beta_{31} x_i^3 + \epsilon_i & x_i < c_i \\ \beta_{02} + \beta_{12} x_i + \beta_{22} x_i^2 + \beta_{32} x_i^3 + \epsilon_i & x_i \geq c_i \end{cases}$$

- Es necesario agregar restricciones a los polinomios, por ejemplo para lograr continuidad.
- Los *splines* permiten obtener la mayor *cantidad* de continuidad.

Polinomios por partes



Splines lineales

- Un *spline lineal* con knots en ξ_k, $k = 1, \cdots, K$ es un polinomio *lineal* por partes que es continuo en cada knot.
- Se puede representar este modelo como:

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \cdots + \beta_{K+3} b_{K+3}(x_i) + \epsilon_i$$

donde los b_k son *funciones base*:

$$b_1(x_i) = x_i$$
$$b_{k+1}(x_i) = (x_i - \xi_k)_+ \qquad k = 1, \cdots, K+2$$

Splines lineales

- La notación $()_+$ representa la *parte positiva*, es decir:

$$(x_i - \xi_k)_+ = \begin{cases} x_i - \xi_k & x_i > \xi_k \\ 0 & x_i \leq \xi_k \end{cases}$$

Splines cúbicos

- Un *spline cúbico* es un polinomio *cúbico* por partes que tiene derivadas continuas de orden dos o más en cada knot ξ_k para $k = 1, \cdots, K$.
- Se puede representar este modelo como:

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \cdots + \beta_{K+3} b_{K+3}(x_i) + \epsilon_i$$

donde los b_k son *funciones base*:

$$\begin{aligned} b_1(x_i) &= x_i \\ b_2(x_i) &= x_i^2 \\ b_3(x_i) &= x_i^3 \\ b_{k+3}(x_i) &= (x_i - \xi_k)_+^3 \qquad k = 1, \cdots, K \end{aligned}$$

Splines cúbicos

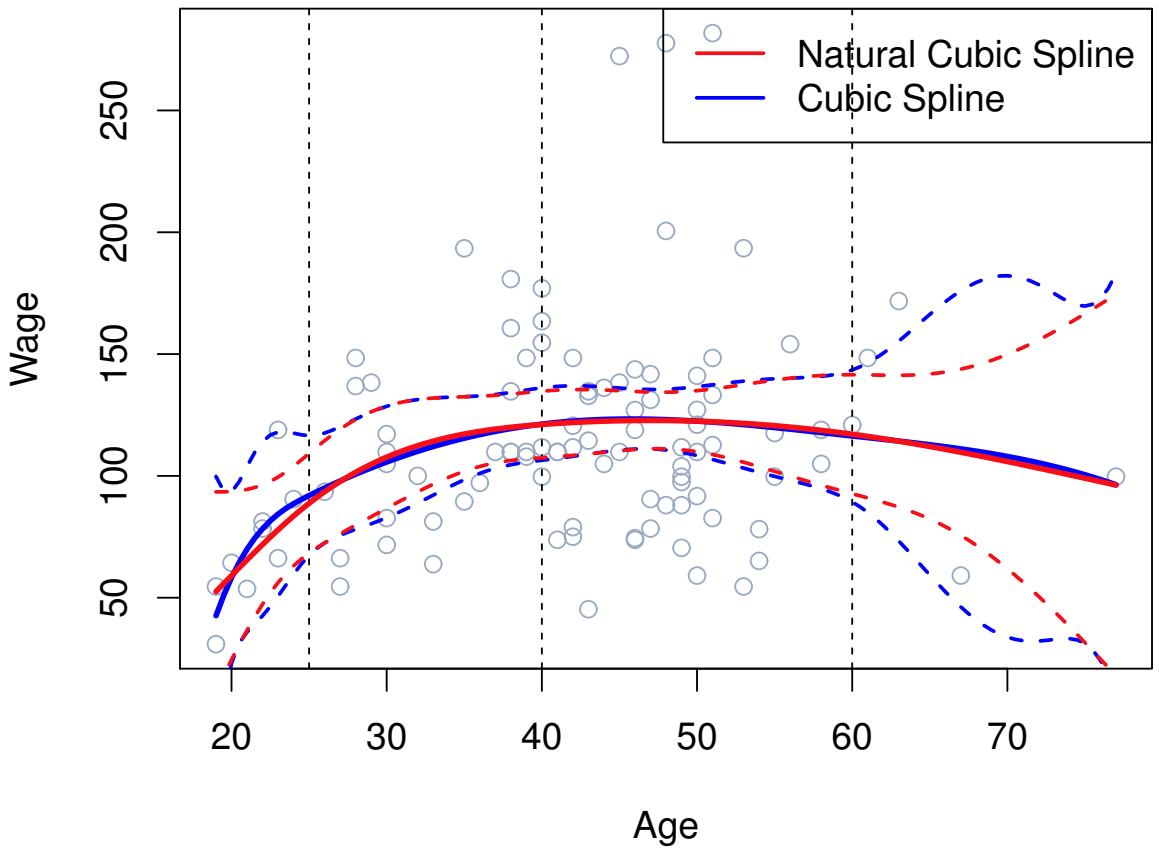
- La notación $()_+$ representa la *parte positiva*, ahora:

$$(x_i - \xi_k)_+^3 = \begin{cases} (x_i - \xi_k)^3 & x_i > \xi_k \\ 0 & x_i \leq \xi_k \end{cases}$$

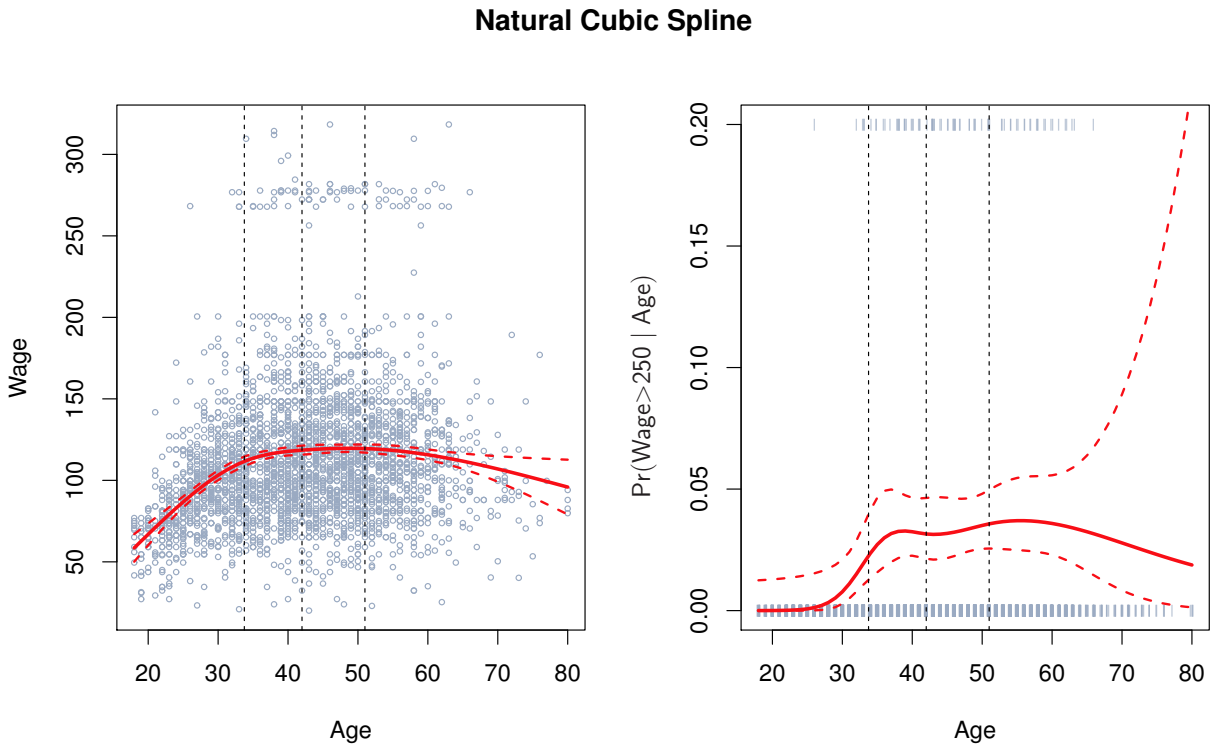
Splines cúbicos naturales

- Los splines tienen mucha varianza fuera del rango de los predictores.
- Un spline cúbico natural considera 4 restricciones adicionales para que la función sea lineal en la *frontera* manteniendo los mismos grados de libertad de un spline cúbico regular.
- Estas restricciones permiten que un spline cúbico natural generalmente produzca estimaciones más estables en la frontera.
- Se pueden estimar splines en R usando `bs(x, ...)` para splines de cualquier grado y `ns(x, ...)` para splines cúbicos naturales dentro de la librería `splines`.

Splines cúbicos naturales



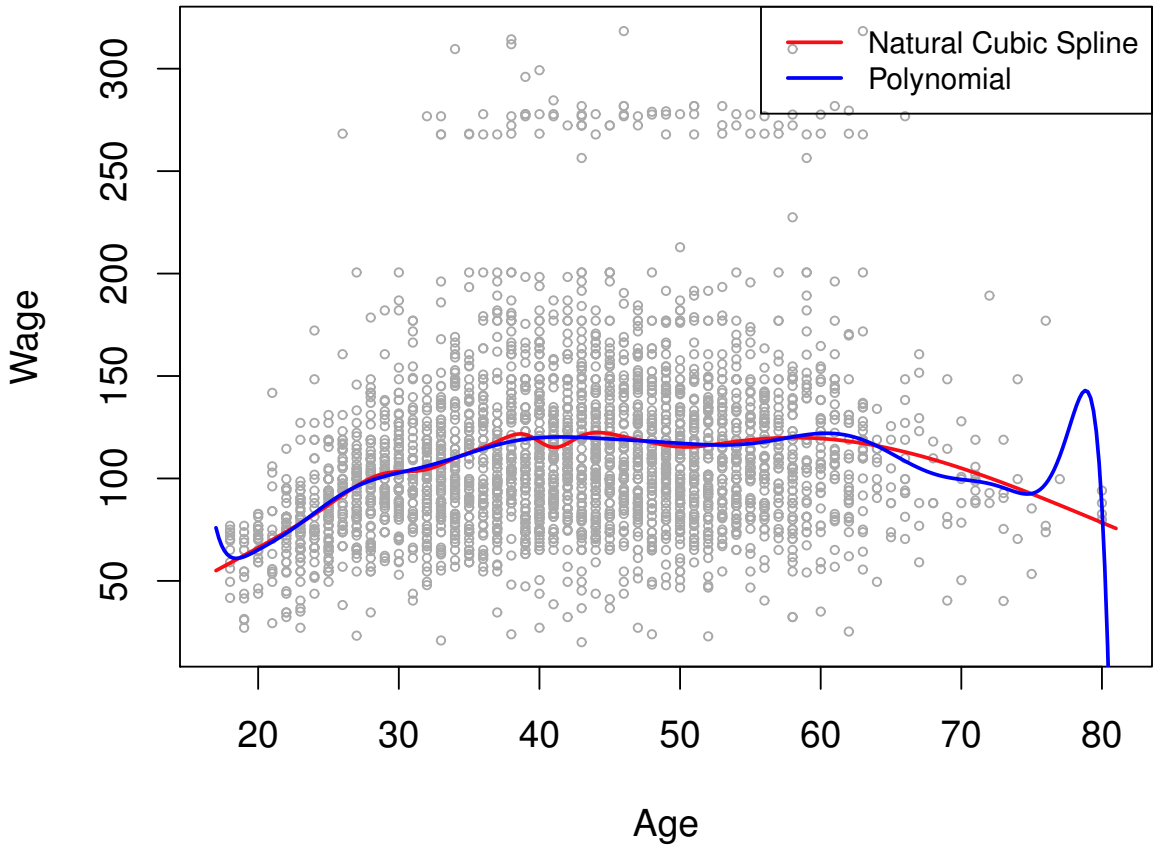
Splines cúbicos naturales



Ubicación de los knots

- Una estrategia es decidir primero el número de knots K y luego usar algunos cuantiles apropiados para los valores observados de X.
- Un spline cúbico con K knots tiene $K + 4$ parámetros o grados de libertad.
- Un spline natural con K knots tiene K grados de libertad.
- En el siguiente gráfico se muestra un polinomio de grado 15 y un spline cúbico natural con 15 grados de libertad usando `ns(age, df=15)` y `poly(age, deg=15)`.

Ubicación de los knots



Suavización por Splines

- Considere el siguiente criterio para estimar una función suavizada $g(x)$ a partir de una data:

$$\min_{g \in \mathcal{S}} \sum_{i=1}^{n} (y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt$$

- El primer término es RSS y busca que $g(x)$ se aproxime a la data en cada x_i.
- El segundo término es una *penalidad por aspereza* y controla que tan rugosa es $g(x)$ y se regula a través de un *parámetro de sintonización* $\lambda \geq 0$.

Suavización por Splines

- Cuanto menor sea el valor de λ más rugosa sera la función y eventualmente interpola cada y_i cuando $\lambda = 0$. Conforme $\lambda \to \infty$ la función $g(x)$ se vuelve lineal.
- La solución al problema anterior es un spline cúbico natural con knots en cada valor único de x_i.
- La penalidad por aspereza se controla usando λ.
- La suavización por splines evita el problema de selección de los knots ya que solo debe elegirse λ.
- Los detalles del algoritmo son algo complejos para ser descritos en esta presentación. En R se usa `smooth.spline()`.

Suavización por Splines

- El vector de n valores estimados puede ser escrito como:

$$\hat{\mathbf{g}}_\lambda = \mathbf{S}_\lambda \mathbf{y}$$

  donde $\mathbf{S}_\lambda$ es una matriz $n \times n$ determinada por x_i y λ.
- Los *grados de libertad efectivos* están dados por:

$$df_\lambda = \sum_{i=1}^{n} \{\mathbf{S}_\lambda\}_{ii}$$

- En R: `smooth.spline(age,df=10)`.

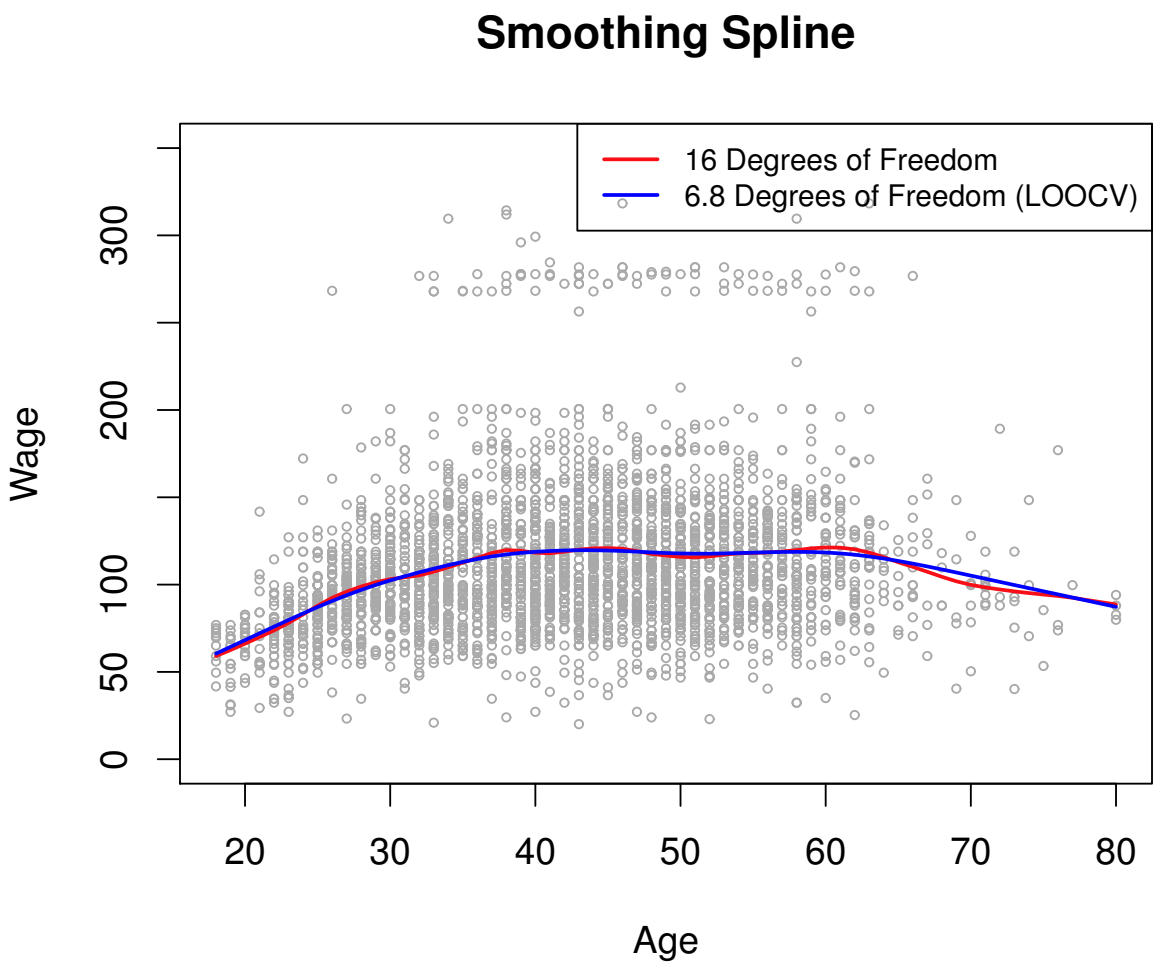
Suavización por Splines

- Para elegir λ se usa el error por validación cruzada dejando uno afuera:

$$\begin{aligned} RSS_{CV}(\lambda) &= \sum_{i=1}^{n}(y_i - \hat{g}_{\lambda}^{(-i)}(x_i))^2 \\ &= \sum_{i=1}^{n}\left[\frac{y_i - \hat{g}_{\lambda}(x_i)}{1 - \{\mathbf{S}_{\lambda}\}_{ii}}\right]^2 \end{aligned}$$

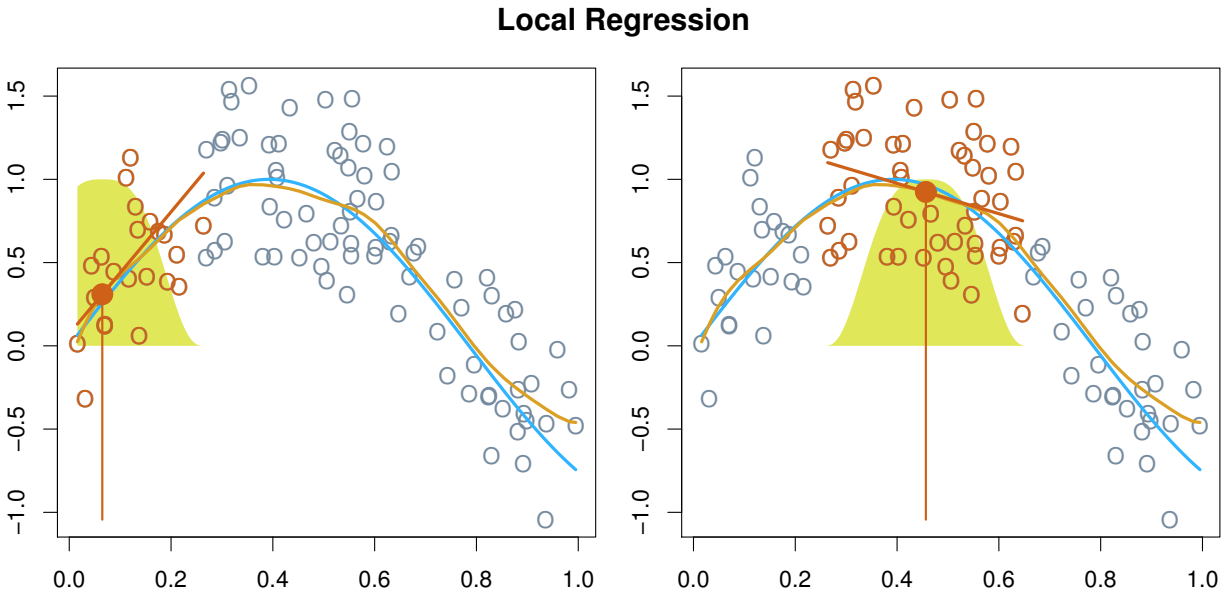
- Se uso `smooth.spline(age,wage)` para obtener el siguiente gráfico.

Suavización por Splines

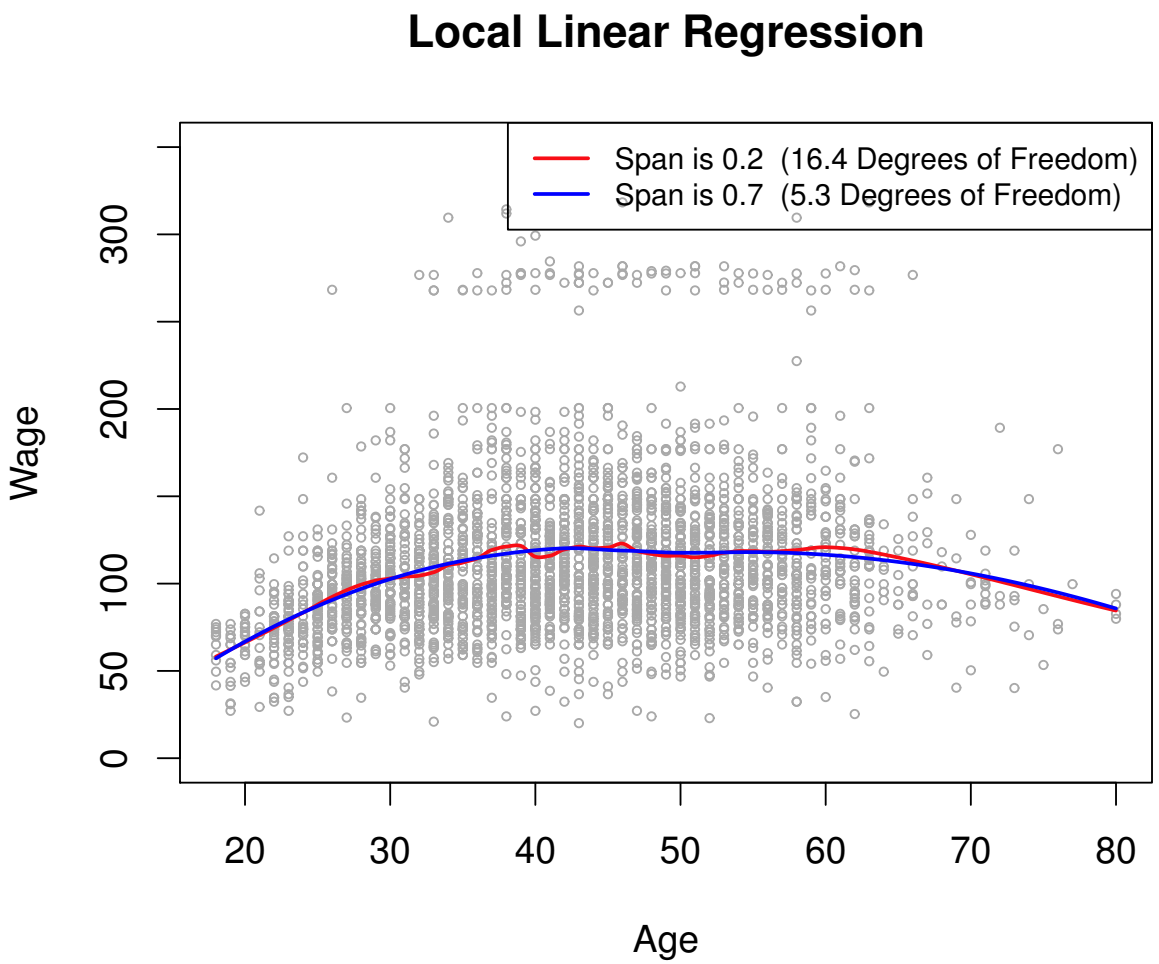


Regresión local

- Usando una función peso de manera *local* se estima una recta por cada región definida en X usando *mínimos cuadrados ponderados*. En R se usa: `loess()`.



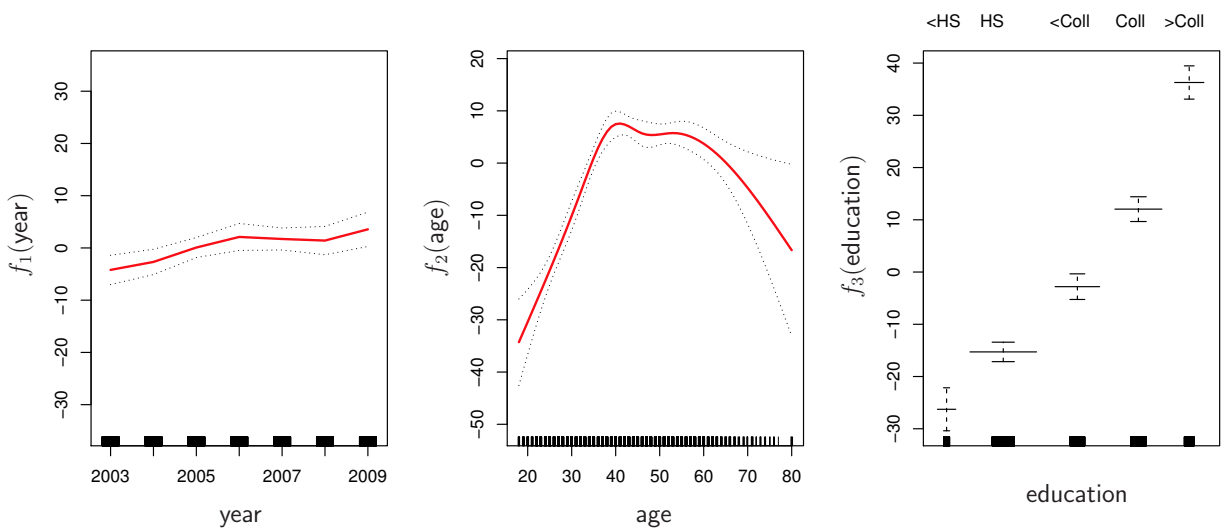
Regresión local



Modelos Aditivos Generalizados

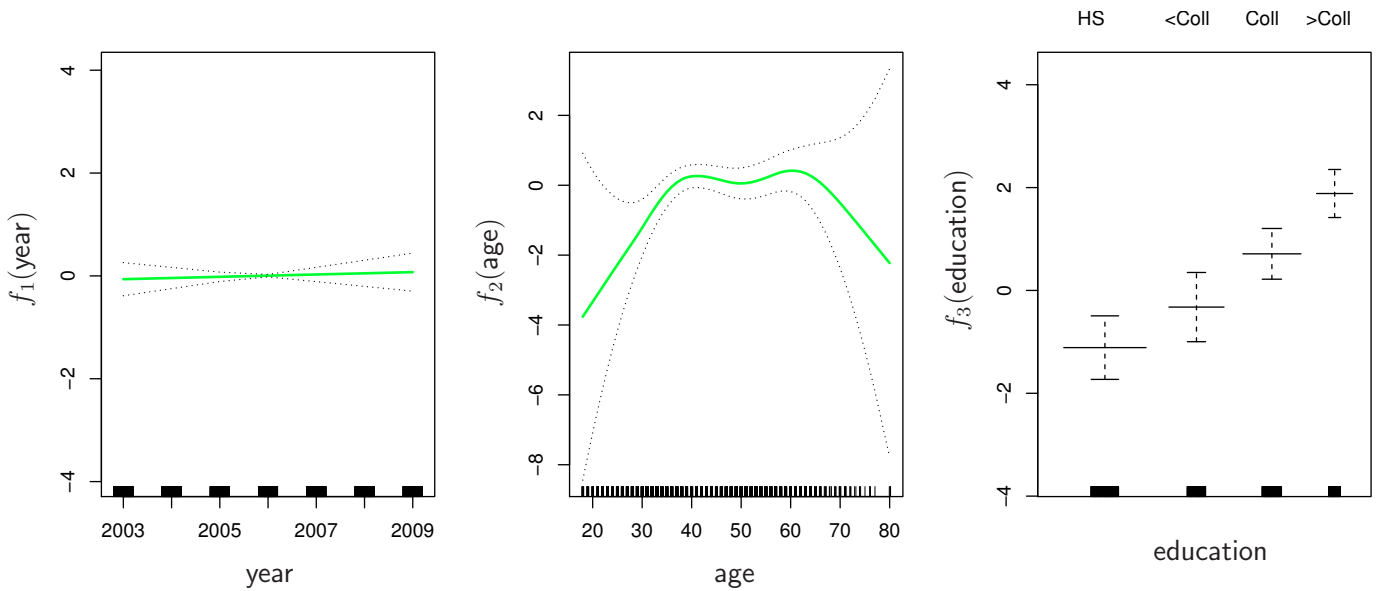
- Los GAM permiten mayor flexibilidad al considerar no linealidad en las variables reteniendo la estructura aditiva de los modelos lineales:

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



Modelos Aditivos Generalizados

- Se puede estimar un GAM usando por ejemplo splines naturales:
  `lm(wage~ns(year.df=5)+ns(age.df=5)+educ)`.
- Los coeficientes no son de mucho interés en cambio la función estimada si.
- Se puede considerar términos mixtos, lineales o no lineales, y usar `anova()` para comparar los modelos resultantes.
- Es posible usar suavización por splines o regresión local:
  `gam(wage~s(year.df=5)+lo(age,span=.5)+educ)`.
- Se pueden incorporar suavizadores bivariados o interacciones de la forma `ns(age,df=5):ns(year,df=5)`.

GAMs para clasificación

$$\log\left(\frac{p(\mathbf{X})}{1-p(\mathbf{X})}\right) = \beta_0 + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p) + \epsilon_i$$