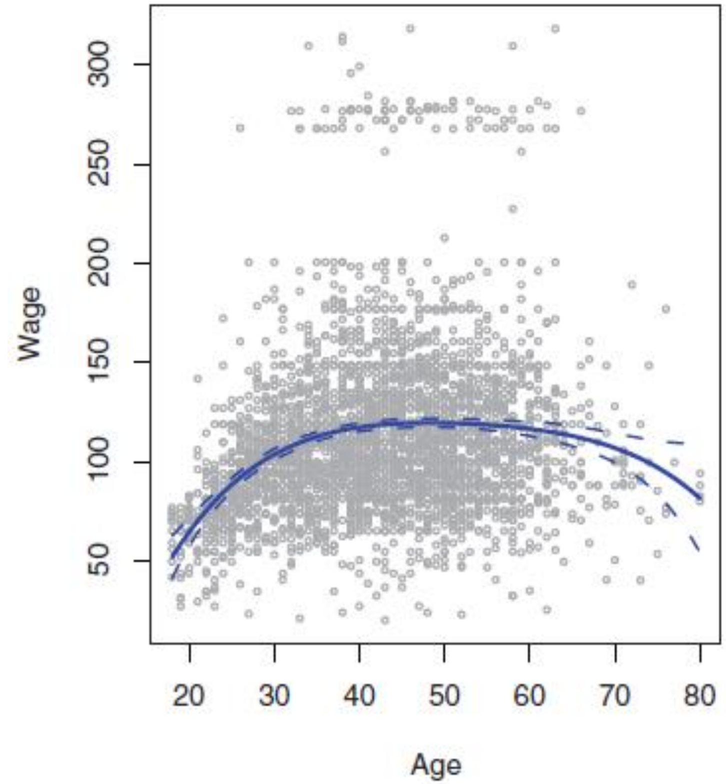
Regresión polinomial

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \ldots + \beta_d x_i^d + \epsilon_i \quad \textbf{(1)}$$

Los coeficientes de (1) son calculados mediante mínimos cuadrados pues es un modelo lineal con p $x_i, x_i^2, x_i^3, \ldots, x_i^d$

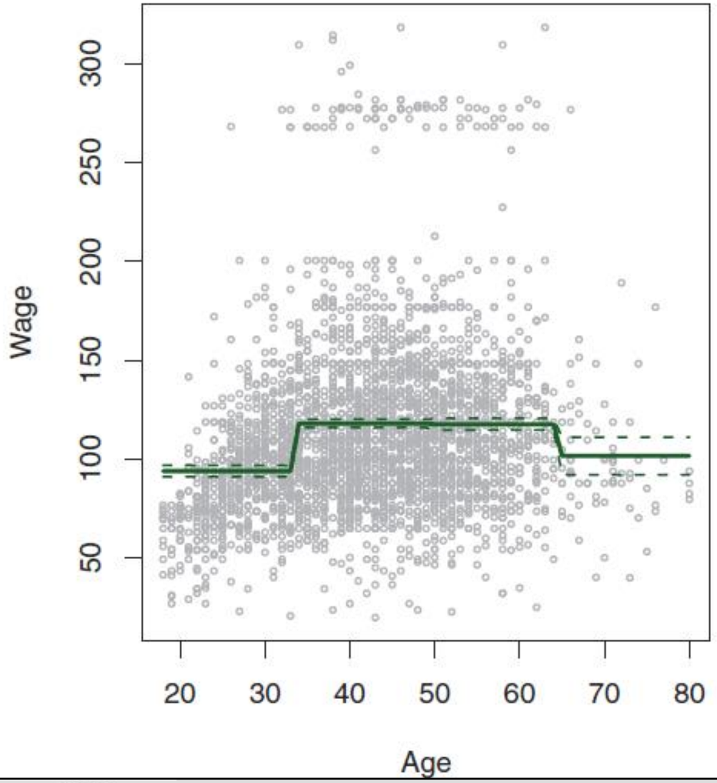
$$\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0 + \hat{\beta}_2 x_0^2 + \hat{\beta}_3 x_0^3 + \hat{\beta}_4 x_0^4$$



Regresión escalonada

Dividimos el rango de X en tramos y ajustamos una nueva constante en cada uno. Asi transformamos una variable continua en una categórica. Elegimos los nodos c_1, c_2, ..., c_k y construimos K+1 nuevas variables

$$
\begin{aligned}
C_0(X) &= I(X < c_1),\\
C_1(X) &= I(c_1 \leq X < c_2),\\
C_2(X) &= I(c_2 \leq X < c_3),\\
&\vdots\\
C_{K-1}(X) &= I(c_{K-1} \leq X < c_K),\\
C_K(X) &= I(c_K \leq X),
\end{aligned}
$$



$$y_i = \beta_0 + \beta_1 C_1(x_i) + \beta_2 C_2(x_i) + \ldots + \beta_K C_K(x_i) + \epsilon_i \quad \textbf{(2)}$$

Regresión por *Splines*

- *Polinomios por tramos*

Ajusto con polinomios cúbicos (grado 3) por tramos

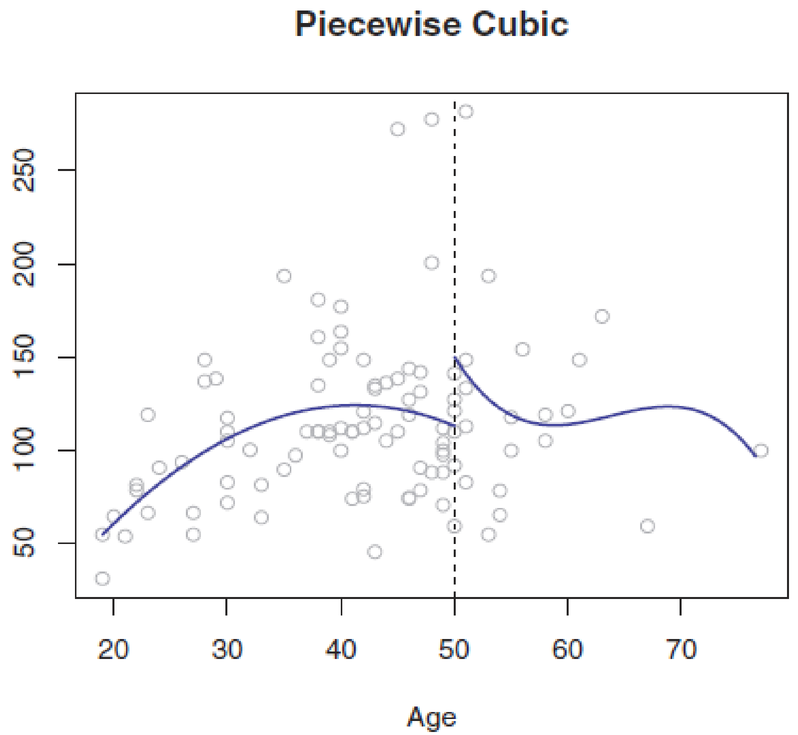
$$y_i = \begin{cases} \beta_{01} + \beta_{11}x_i + \beta_{21}x_i^2 + \beta_{31}x_i^3 + \epsilon_i & \text{if } x_i < c \\ \beta_{02} + \beta_{12}x_i + \beta_{22}x_i^2 + \beta_{32}x_i^3 + \epsilon_i & \text{if } x_i \geq c \end{cases}$$

Los coeficientes a estimar por mínimos cuadrados parta cada polinomio resp. son

$\beta_{01}, \beta_{11}, \beta_{21}, \beta_{31}$

$\beta_{02}, \beta_{12}, \beta_{22}, \beta_{32}$

Veamos un ejemplo

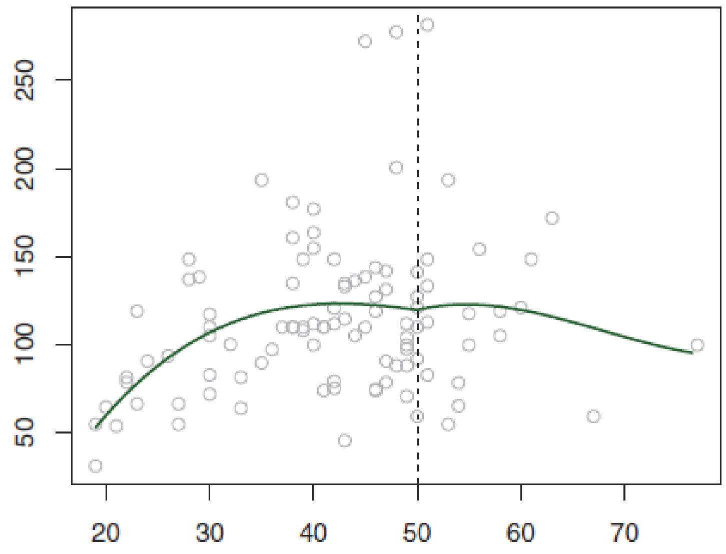


El resultado es muy malo. Tenemos 8 GL

- *Condiciones y Splines*

Vamos a ir quitándole grados de libertad a la curva mediante diferentes restricciones

→ 1. La curva ajustada debe ser continua: igualdad $f(c_i)$ de $P^{(i)}$ y $P^{(i+1)}$

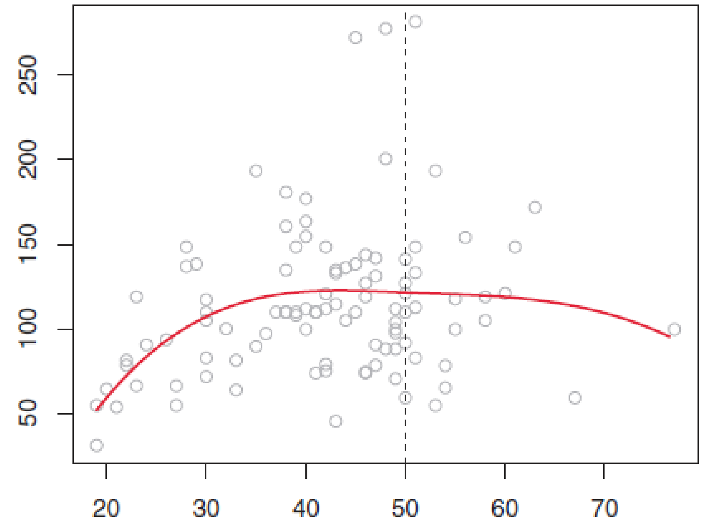


Resultado: en punto c tiene forma de V

→ 2. La curva ajustada debe ser continua y suave los puntos de corte

a) Igualdad de $f'(c_i)$ de $P^{(i)}$ y $P^{(i+1)}$

a) Igualdad de $f''(c_i)$ de $P^{(i)}$ y $P^{(i+1)}$



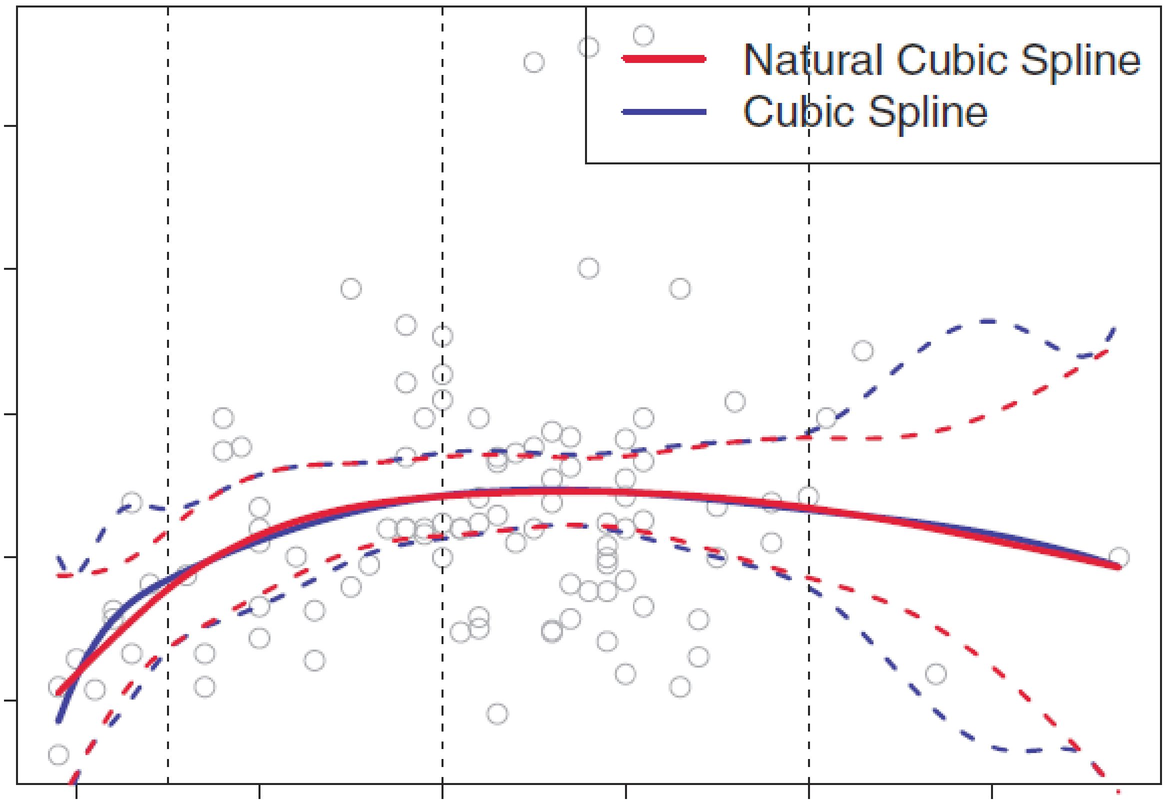
El resultado es mejor Tenemos 5 GL

- *Cantidad y Ubicación de los nodos*

En la práctica se ubican los puntos de corte en forma equidistante. Se especifica la cantidad de GL y el software ubica los nodos. En este ejemplo se especificaron 4 GL y el sistema ubicó los c_i en los cuartiles 1, 2 y 3

- *Splines naturales*

Se incorpora la condición que en los puntos extremos la curva sea una recta.



Splines Suavizados

- *Descripción*

Lo que buscamos al ajustar una curva a los datos es que la suma de los errores al cuadrado sea mínima . Pero si no ponemos ninguna restricción podemos ajustar perfectamente con un polinomio que pase por todos los puntos. Eso es *overfitting.* Lo que queremos es encontrar g(x) que ajuste lo suficiente pero que siga siendo suave.

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

El segundo término es un penalización. La integral de la derivada segunda indica que tan sinuosa e irregular es la curva.

Si $\lambda = 0$, entonces el segundo término no influirá y la curva será sinuosa y tendrá muchos saltos.. Si $\lambda \rightarrow \infty$ el segundo término será tan influyente que hará que la función de error sea una línea recta, que es la curva de error que hace cero la $g''(t)^2$

La función resultante es un spline cúbico natural al que se denomina *spline suavizado*