Machine Learning

Hernán Aguirre

Universidad San Franciso de Quito

Universidad de Shinshu, Japón

Contenido del Curso

1. **Introducción**
2. **Aprendizaje estadístico**
3. **Regresión lineal**
4. **Clasificación**
5. **Métodos de remuestreo**
6. **Selección y regularización de modelos lineales**
7. **Más allá de la linealidad**
8. **Métodos basados en árboles**
9. **Máquinas de vectores de soporte**
10. Aprendizaje profundo
11. Análisis de supervivencia y datos censurados
12. Aprendizaje sin supervisión
13. Pruebas múltiples

Más allá de la Linealidad

1. Regresión Polinomial
2. Funciones de Paso
3. Funciones Base
4. Splines de Regresión
5. Splines de Suavizado
6. Regresión Local
7. Modelos Aditivos Generalizados (GAMs)
8. Lab: Modelado no Lineal
9. Ejercicios

Hasta Ahora

- Nos hemos centrado principalmente en modelos lineales
  - Relativamente sencillos de describir e implementar y tienen ventajas sobre otros enfoques en términos de interpretación e inferencia
- Sin embargo, la regresión lineal estándar puede tener limitaciones significativas en términos de poder predictivo
  - El supuesto de linealidad es casi siempre una aproximación y, a veces, deficiente

➡ Relajamos el supuesto de linealidad y al mismo tiempo intentamos mantener la mayor interpretabilidad posible

Regresión Polinomial

- Regresión Lineal estándar

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- Regresión Polinomial

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

- Para un grado d suficientemente grande → curva extremadamente no lineal
- Es inusual usar d mayor que 3 o 4
  - $d > 4$, la curva se vuelve demasiado flexible y adopta formas extrañas, particularmente cerca del límite de la variable X
- $\beta_0, \beta_1, \cdots, \beta_d$ se estiman fácilmente utilizando la regresión lineal de mínimos cuadrados, i.e. es un modelo lineal estándar con predictores $x_i, x_i^2, \cdots, x_i^d$

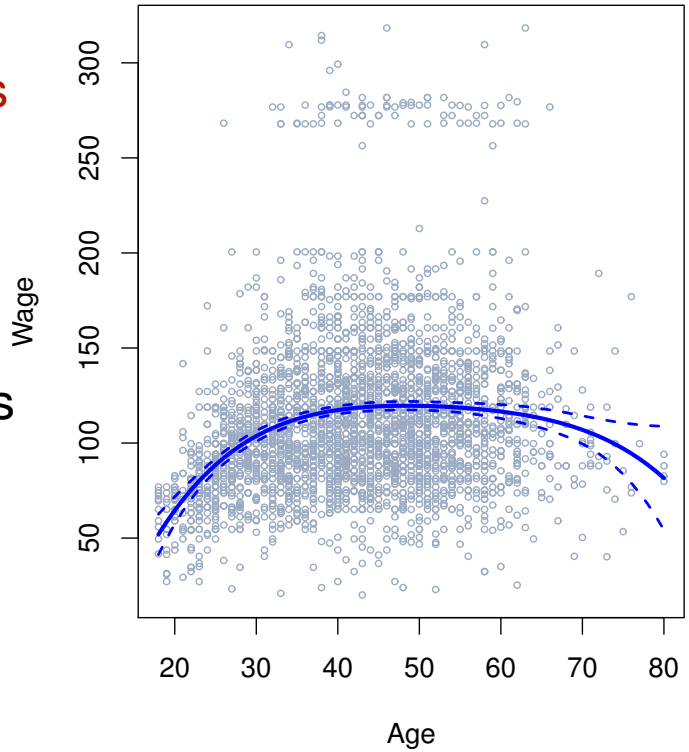
Datos de Salario

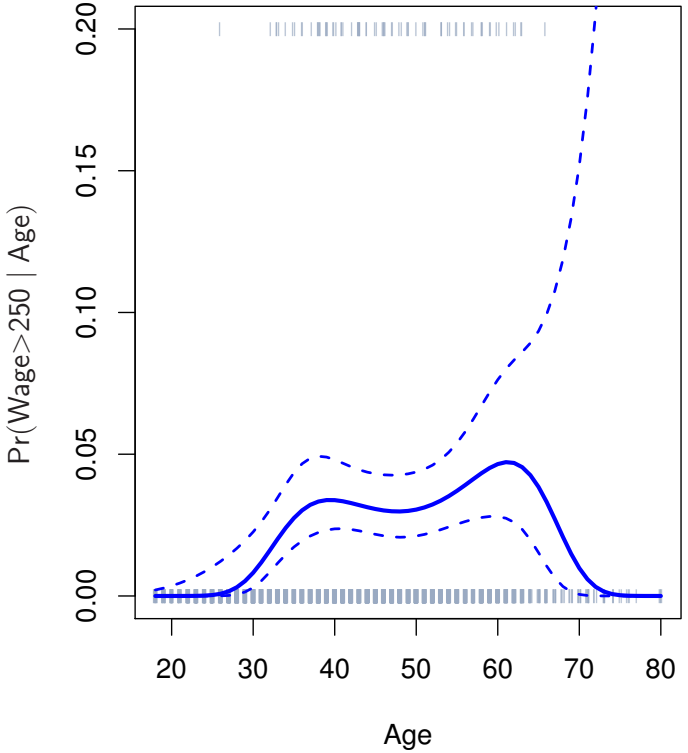
$$Wage = \beta_0 + \beta_1 Age + \cdots + \beta_4 Age^4$$

$$p(Age) = \frac{e^{\beta_0+\beta_1 Age+\cdots+\beta_4 Age^4}}{1 + e^{\beta_0+\beta_1 Age+\cdots+\beta_4 Age^4}}$$

ingresos altos

ingresos bajos





Datos de 3000 personas, solo 79 tienen salarios altos → Varianza alta

Regresión del salario en función de la edad usando un polinomio de grado 4 con un intervalo de confianza estimado del 95 %

Regresión logística, con un polinomio de grado 4, para el evento binario salario>250. La probabilidad posterior ajustada de que salario>250 con un intervalo de confianza estimado del 95%

Funciones de Paso

- El uso de funciones polinomiales de los predictores en un modelo lineal impone una estructura global a la función no lineal de X
- Es posible utilizar funciones escalonadas para evitar imponer una estructura global de este tipo
- Se divide el rango de X en contenedores y ajustamos una constante diferente en cada contenedor
- Equivale a convertir una variable continua en una variable categórica ordenada

Funciones Escalonadas

$$y_i = \beta_0 + \beta_1 C_1(x_i) + \beta_2 C_2(x_i) + \cdots + \beta_K C_K(x_i) + \epsilon_i$$

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

- $I(\cdot)$ es una función indicadora que devuelve un 1 si la condición es verdadera y un 0 en caso contrario.

Funciones Escalonadas (2)

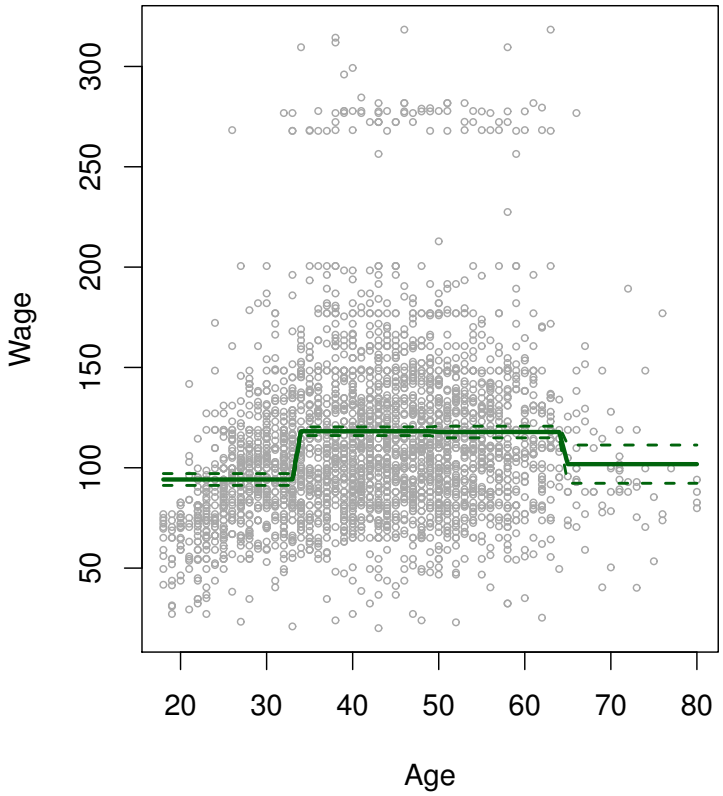
$$y_i = \beta_0 + \beta_1 C_1(x_i) + \beta_2 C_2(x_i) + \cdots + \beta_K C_K(x_i) + \epsilon_i$$

- Para un valor dado de X, como máximo uno de $C_1, C_2, \cdots, C_K$ puede ser distinto de cero
- Cuando $X < c_1$, todos los predictores son cero → β_0 se interpreta como el valor medio de Y para $X < c_1$
- Para $c_j \leq X < c_{j+1}$, el modelo predice una respuesta $\beta_0 + \beta_j$ → β_j representa el aumento promedio en la respuesta cuando $c_j \leq X < c_{j+1}$ en relación con $X < c_1$

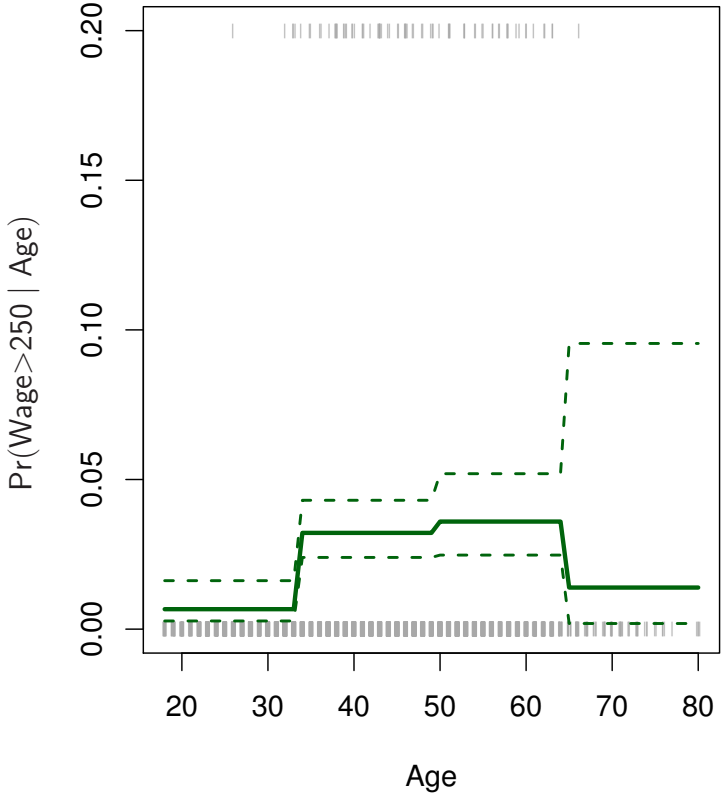
Ajuste de Funciones Escalonadas Datos de Salario

$$y_i = \beta_0 + \beta_1 C_1(x_i) + \cdots + \beta_K C_K(x_i)$$

$$Pr(y_i > 250 \mid x_i) = \frac{e^{\beta_0+\beta_1 C_1(x_i)+\cdots+\beta_K C_K(x_i)}}{1 + e^{\beta_0+\beta_1 C_1(x_i)+\cdots+\beta_K C_K(x_i)}}$$



Regresión Lineal del salario en función de la edad



Regresión Logística para predecir la probabilidad de que un individuo tenga altos ingresos en función de su edad

Observación

- A menos que existan puntos de interrupción naturales en los predictores, las funciones constantes por partes pueden perder las transiciones en los datos

- Por ejemplo, el primer grupo claramente pasa por alto la tendencia creciente del salario con la edad.

- Sin embargo, los enfoques de función escalonada son muy populares en bioestadística y epidemiología, entre otras disciplinas. Por ejemplo, a menudo se utilizan grupos de edad de 5 años para definir las ubicaciones.

Funciones Base

- La idea es tener a la mano una familia de funciones o transformaciones que se puedan aplicar a una variable X: $b_1(X), b_2(X), \cdots, b_K(X)$

- En lugar del modelo lineal en X, ajustamos el modelo

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \cdots + \beta_K b_K(x_i) + \epsilon_i$$

- Las funciones base $b_1(\,\cdot\,), b_2(\,\cdot\,), \cdots, b_K(\,\cdot\,)$ son fijas y conocidas, i.e. elegimos las funciones con anticipación

Casos Especiales de Funciones Base

- Los modelos de *regresión polinomial* y *funciones constante por partes* son casos especiales del enfoque que aplica funciones base

- Para la *regresión polinomial,* las funciones base son

  - $b_j(x_i) = x_i^j,$

- Para *funciones constantes por partes* son

  - $b_j(x_i) = I(c_j \leq x_i < c_{j+1})$

Funciones Base Alternativas

- Funciones polinómicas
- Funciones constantes por partes
- Funciones base contruidas con a partir de
  - Wavelets
  - Series de Fourier
- Splines de regresión
- …

Modelo Lineal Stándar: Funciones Base como Predictores

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \beta_2 b_2(x_i) + \cdots + \beta_K b_K(x_i) + \epsilon_i$$

- *Modelo lineal estándar* con predictores

$$b_1(x_i), b_2(x_i), \cdots, b_K(x_i)$$

➡ Mínimos cuadrados para estimar $\beta_0, \beta_1, \beta_2, \cdots, \beta_K$

➡ Todas las herramientas de inferencia para modelos lineales están disponibles en este entorno

- errores estándar para las estimaciones de coeficientes
- estadísticas F para la significancia general del modelo, etc.

7.4

Splines de Regresión

- Polinomios por partes
- Restricciones y splines
- La representación de la base spline
- Elección del número y ubicación de los nudos
- Comparación con la regresión polinómica

¿Que es un Spline de Regresión?

- Una clase flexible de funciones bases que extiende los enfoques de
  - regresión polinomial y
  - regresión constante por partes

7.4.1

Polinomios por Partes

- En lugar de ajustar un polinomio de alto grado en todo el rango de X,
  - La *Regresión Polinomial por Partes* implica ajustar polinomios separados de bajo grado en diferentes regiones de X
- Por ejemplo, un *Polinomio Cúbico por Partes* funciona ajustando un modelo de regresión cúbica de la forma

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

- donde *los coeficientes $\beta_0, \beta_1, \beta_2, \beta_3$ cambian en diferentes partes* del rango de X
- Los puntos donde cambian los coeficientes se llaman *nudos*

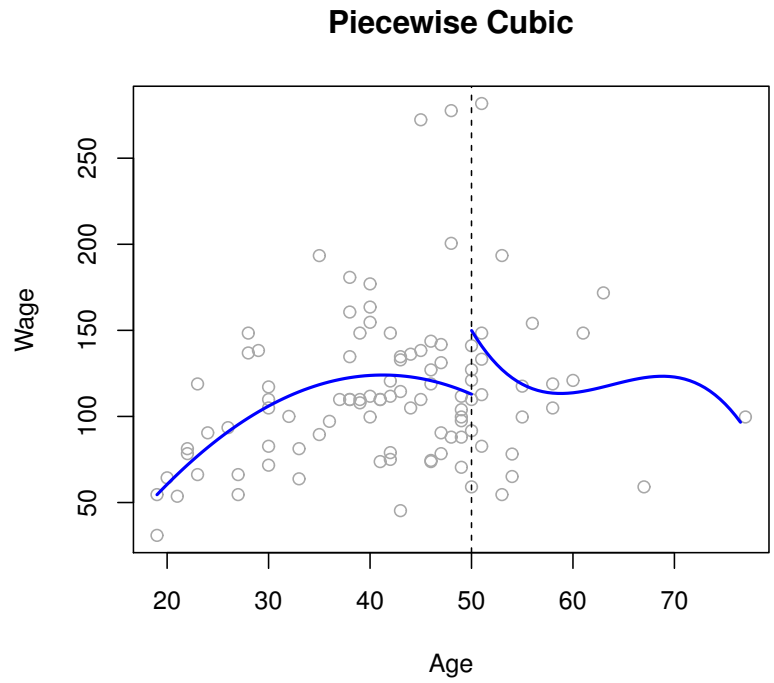
Ejemplo
Polinomio por Partes

- Un *Polinomio Cúbico por Partes* con un solo *nudo* en un punto c toma la forma

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

- Ajustamos dos funciones polinomiales diferentes a los datos,
  - una en el subconjunto de observaciones con $x_i < c$ y
    - $\beta_{01}, \beta_{11}, \beta_{21}, \beta_{31}$
  - otra en el subconjunto de observaciones con $x_i \geq c$
    - $\beta_{01}, \beta_{11}, \beta_{21}, \beta_{31}$

Polinomio Cúbico por Partes Datos de Salario



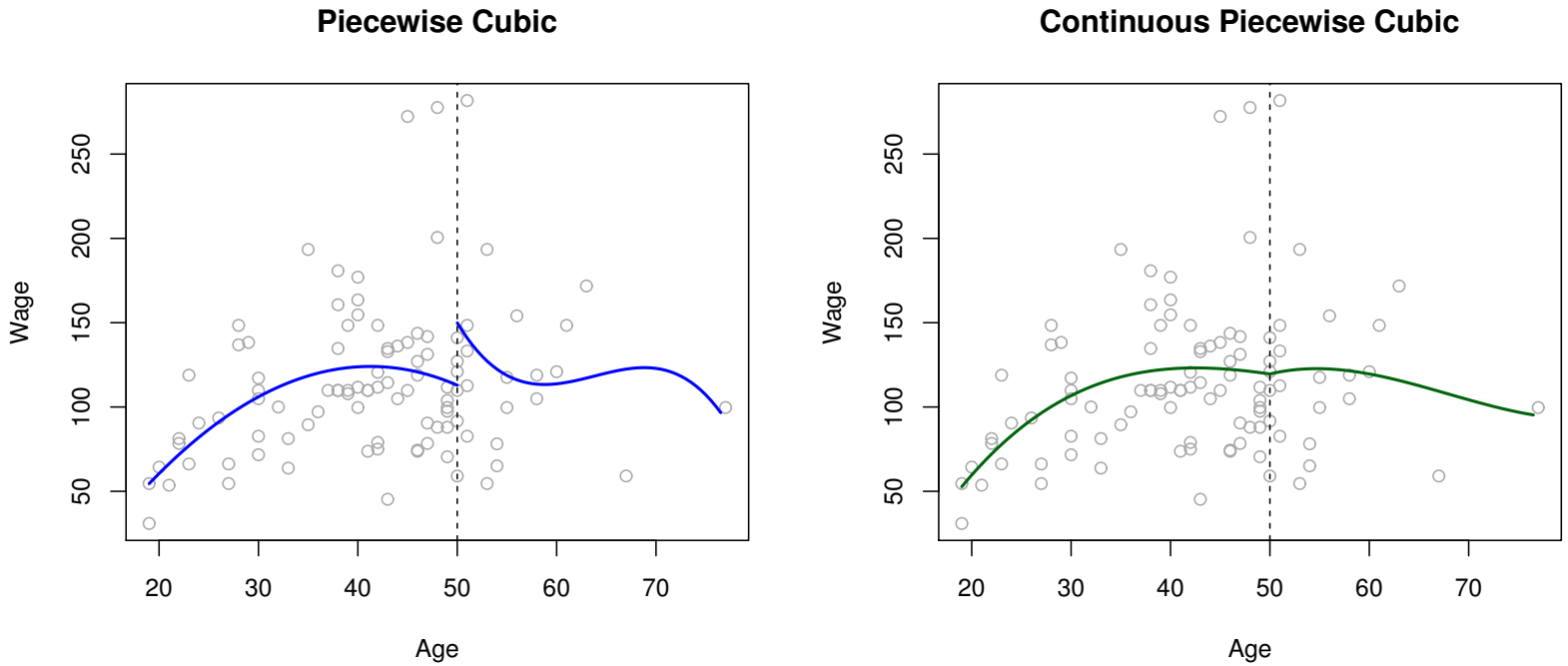
- La figura muestra un ajuste Polinomial Cúbico por Partes a un subconjunto de datos de Salarios, con un solo nudo a la *edad=50*
- *De inmediato* se *ve un problema:* ¡la función es discontinua y luce extraña!

Observaciones

- Dado que cada polinomio cúbico tiene *cuatro parámetros*, utilizamos un total de *ocho grados de libertad* para ajustar este modelo polinomial por partes
- Usar más nudos conduce a un polinomio por partes más flexible
- En general,
  - Si colocamos K *nudos* diferentes en todo el rango de X, terminaremos ajustando $K + 1$ polinomios diferentes
  - *Grados de libertad* = $(K + 1) \times (d + 1)$ coeficientes a estimar
  - donde d es el grado del polinomio

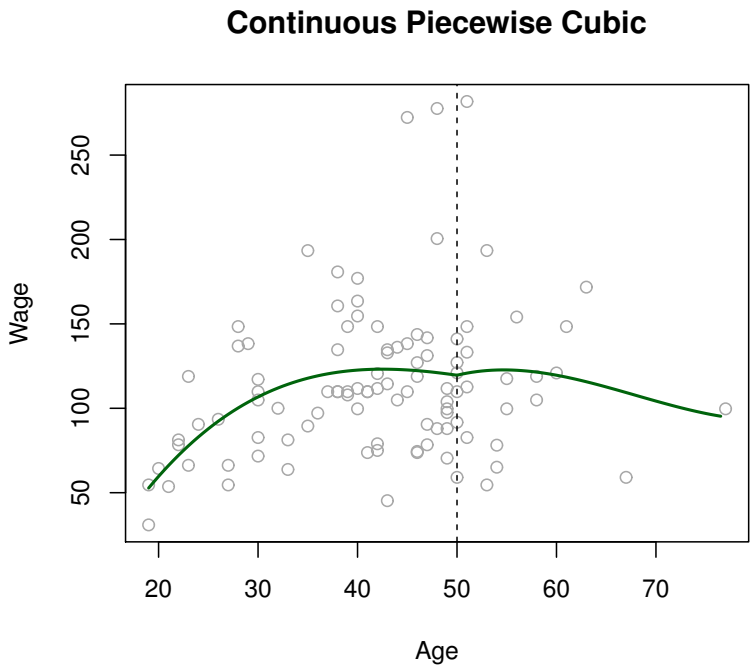
7.4.2

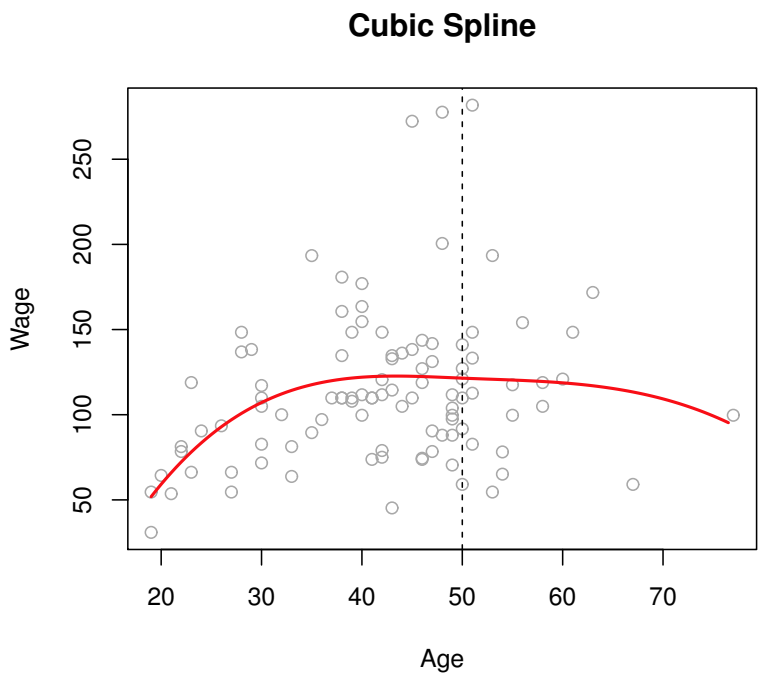
Restricciones y Splines



- Se ajusta un polinomio por partes bajo la restricción de que *la curva ajustada debe ser continua*
  - En otras palabras, no puede haber un salto cuando la edad = 50 años
- El gráfico derecho muestra el ajuste resultante. Se ve mejor que el gráfico izquierdo, pero la unión en forma de V no parece natural

Primera y Segunda Derivada Iguales en los Nodos





- En el gráfico derecho hemos agregado dos restricciones adicionales: ahora *tanto la primera como la segunda derivada de los polinomios por partes son continuas a edad = 50 (en el nodo)*.
- En otras palabras, requerimos que el polinomio por partes no sólo sea *continuo* cuando edad = 50 años, sino que *también sea muy suave*.

Spline Cúbica

- Imponiendo tres restricciones al *Polinomio Cúbico por Partes*
  - continuidad
  - continuidad de la primera derivada
  - continuidad de la segunda derivada

➡ *Spline Cúbica*

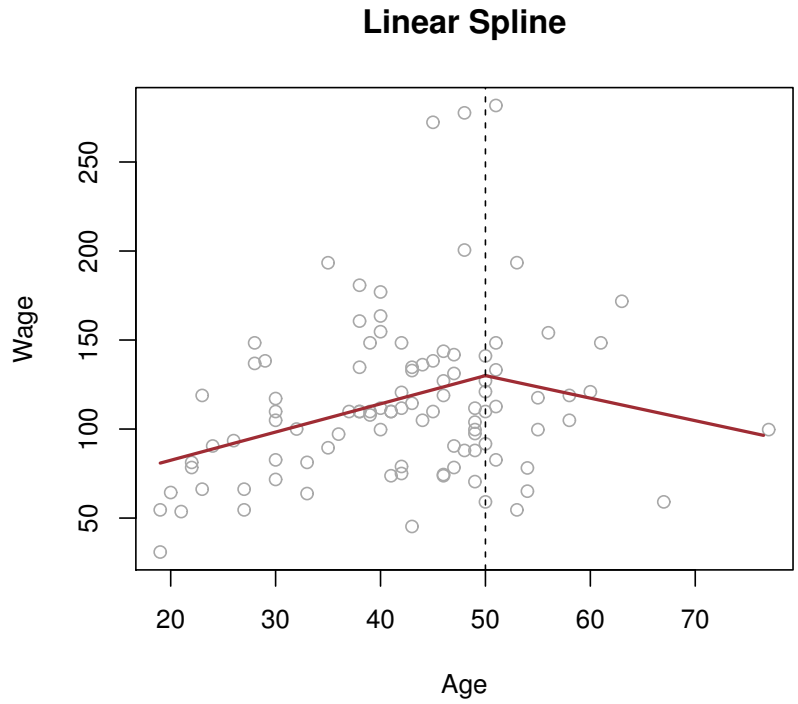
Grados de Libertad y Número de Nodos

- Cada restricción que imponemos a los polinomios cúbicos por partes libera efectivamente un grado de libertad → reduce la complejidad del ajuste del polinomio por partes resultante
- Inicialmente usamos ocho grados de libertad
- En la Spline Cúbica nos quedan cinco grados de libertad
- En general, una Spline Cúbica con K nudos utiliza un total de 4 + K grados de libertad.

Spline de Grado d

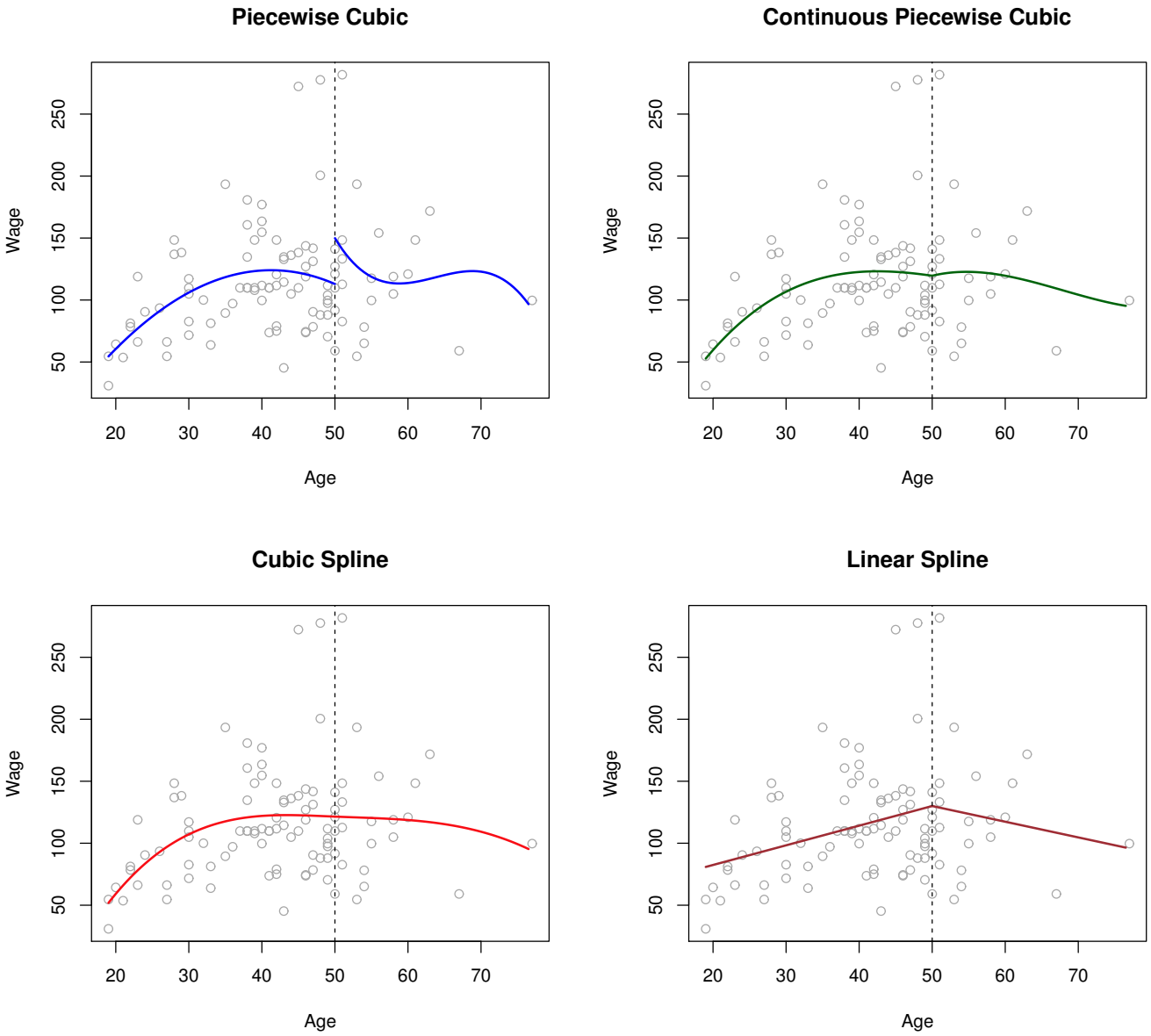
- La definición general de un spline de grado d es que es un polinomio de grado d por partes, con continuidad en las derivadas hasta el grado $d-1$ en cada nudo

Spline Lineal



- Spline lineal, continua a la edad = 50 años.
- Se obtiene un spline lineal ajustando una línea en cada región del espacio predictor definido por los nudos, lo que requiere continuidad en cada nudo.

Varios Polinomios por Partes



7.4.3

Representación de Spline con Funciones Bases

- Se puede usar un modelo de funciones base para representar una Spline de Regresión
- Una Spline Cúbico con K nudos se puede modelar como

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

- para una elección adecuada de las funciones base $b_1, b_2, \cdots, b_{K+3}$.
- El modelo puede ajustarse usando mínimos cuadrados

Función de Base de Potencia Truncada

- Hay muchas formas equivalentes de representar *Splines Cúbicos* usando diferentes opciones de funciones base
- La forma más directa es comenzar con una base para un polinomio cúbico (es decir, x, x^2 y x^3) y luego agregar una *función de base de potencia truncada* por nudo
- *Función de base de potencia truncada*

$$h(x, \xi) = (x - \xi)_+^3 = \begin{cases} (x - \xi)^3 & if\ x > \xi \\ 0 & otherwise \end{cases}$$

- donde ξ es el nodo

Regresión de Mínimos Cuadrados

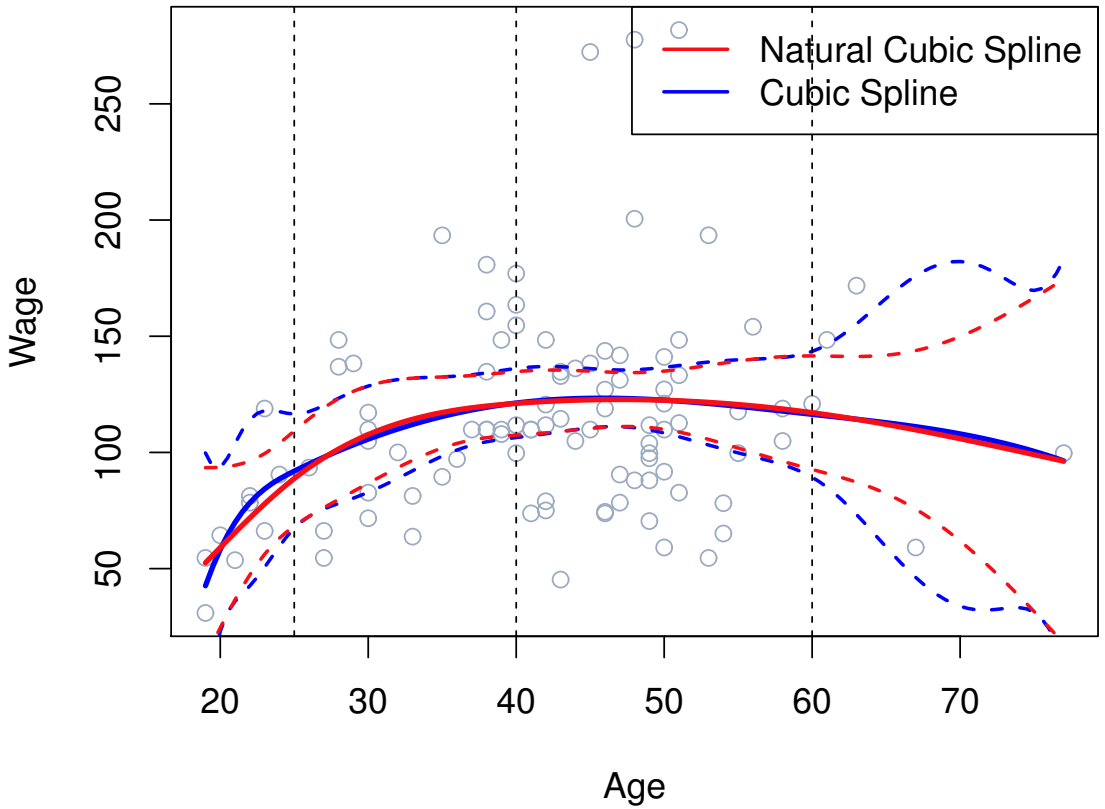
- Para ajustar un Spline Cúbico a un conjunto de datos con K nudos, realizamos una regresión de mínimos cuadrados con una intersección y $3 + K$ predictores,

$$y_i = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 h(X, \xi_1) + \cdots + \beta_{3+K} h(X, \xi_K)$$

- donde $\xi_1, \cdots, \xi_K$ son los nudos.

- Esto equivale a estimar un total de $K + 4$ coeficientes de regresión (grados de libertad)

Splines Ajustados a Datos de Salario



- Los splines pueden tener una gran variación en el rango exterior de los predictores (valor muy pequeño o muy grande de X)
- Los intervalos de confianza en la región límite varian grandemente

Spline Natural

- Un spline natural es un spline de regresión con restricciones de límite adicionales:
  - *se requiere que la función sea lineal en el límite* (en la región donde X es menor que el nudo más pequeño o mayor que el nudo más grande).
  - Esta restricción adicional significa que los splines naturales generalmente producen estimaciones más estables en los límites.
- Un Spline Cúbico natural también se muestra en la figura como una línea roja.
  - Los intervalos de confianza correspondientes son más estrechos

7.4.4

Elegir el Número y la Ubicación de los Nudos

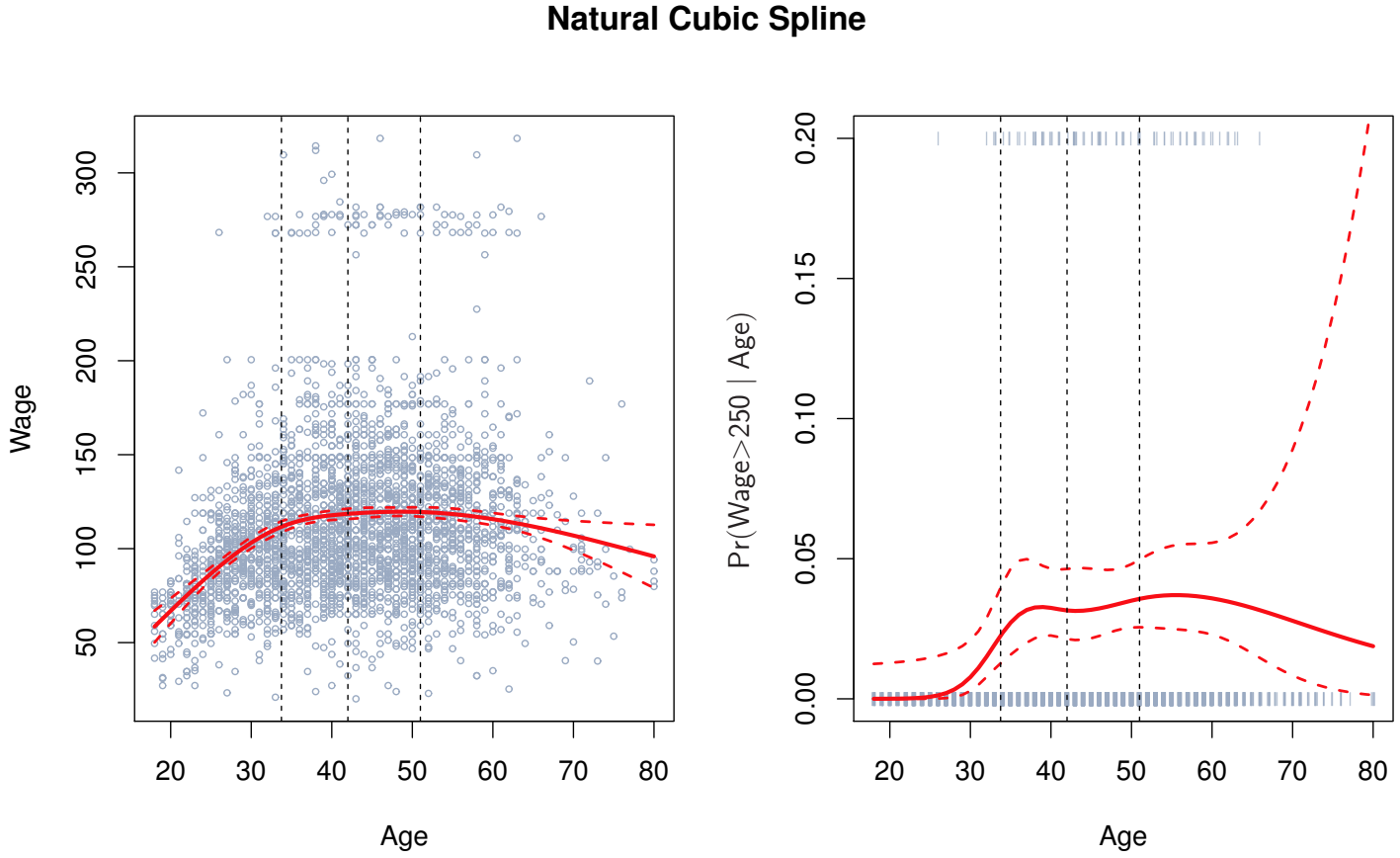
- La spline de regresión es más flexible en regiones que contienen muchos nudos, porque en esas regiones los coeficientes polinomiales pueden cambiar rápidamente.

- Por lo tanto, una opción es colocar más nudos en los lugares donde sentimos que la función podría variar más rápidamente y colocar menos nudos donde parece más estable.

- Si bien esta opción puede funcionar bien, en la práctica es común colocar los nudos de manera uniforme.

- Una forma de hacerlo es especificar los grados de libertad deseados y luego hacer que el software coloque automáticamente el número correspondiente de nudos en cuantiles uniformes de los datos.

Spline Cúbica Natural
Datos de Salario

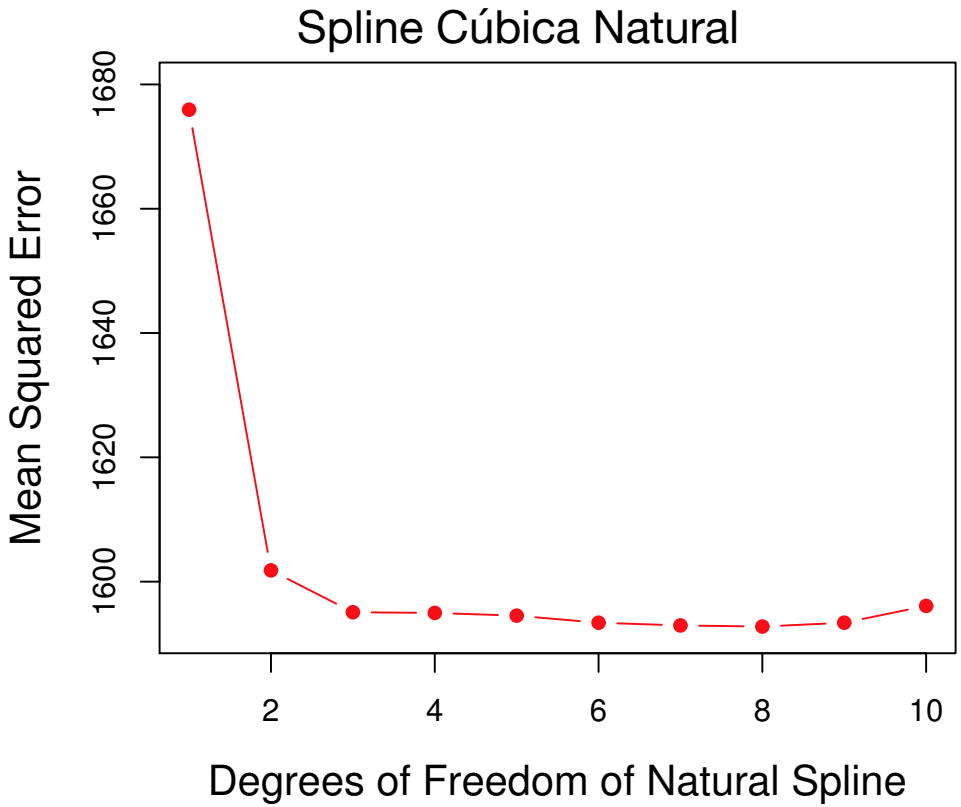


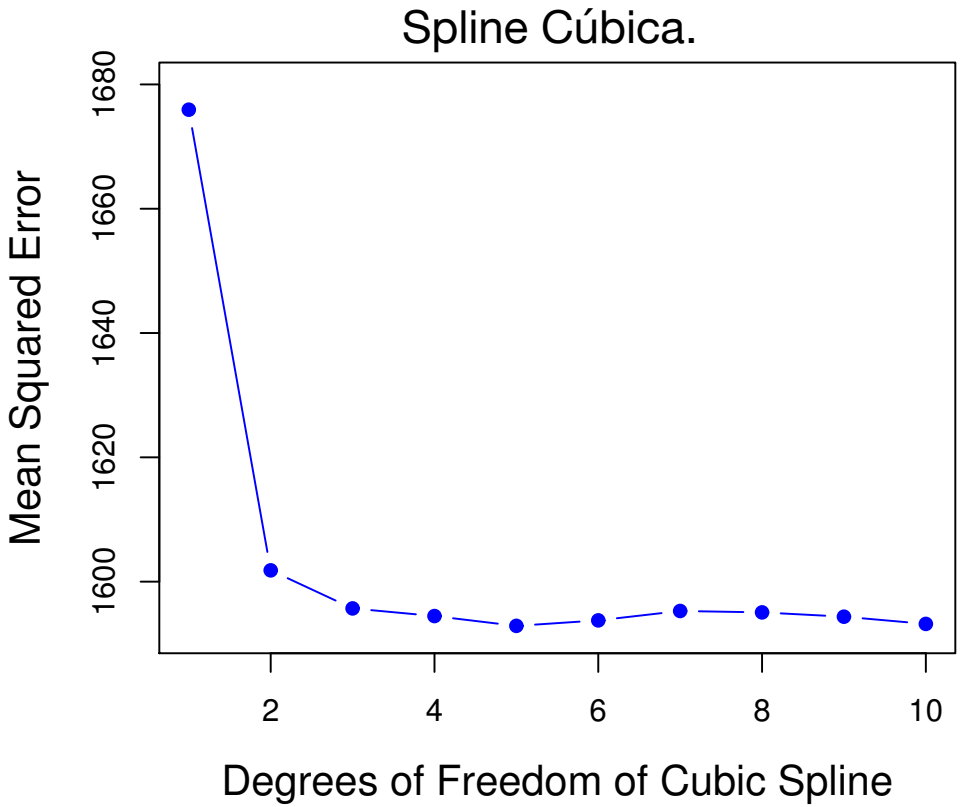
- Ajuste de una spline cúbica natural con tres nudos en los datos salariales
- Las ubicaciones de los nudos se eligieron automáticamente como los percentiles de edad 25, 50 y 75. Esto se especifica solicitando cuatro grados de libertad

¿Cuántos Nudos?

- ¿Cuántos nudos debemos usar o, equivalentemente, cuántos grados de libertad debe contener nuestra spline?
- Una opción es probar diferentes números de nudos y ver cuál produce la curva más atractiva :)
- Un enfoque algo más objetivo es utilizar *validación cruzada*

Validación Cruzada y Número de Nodos



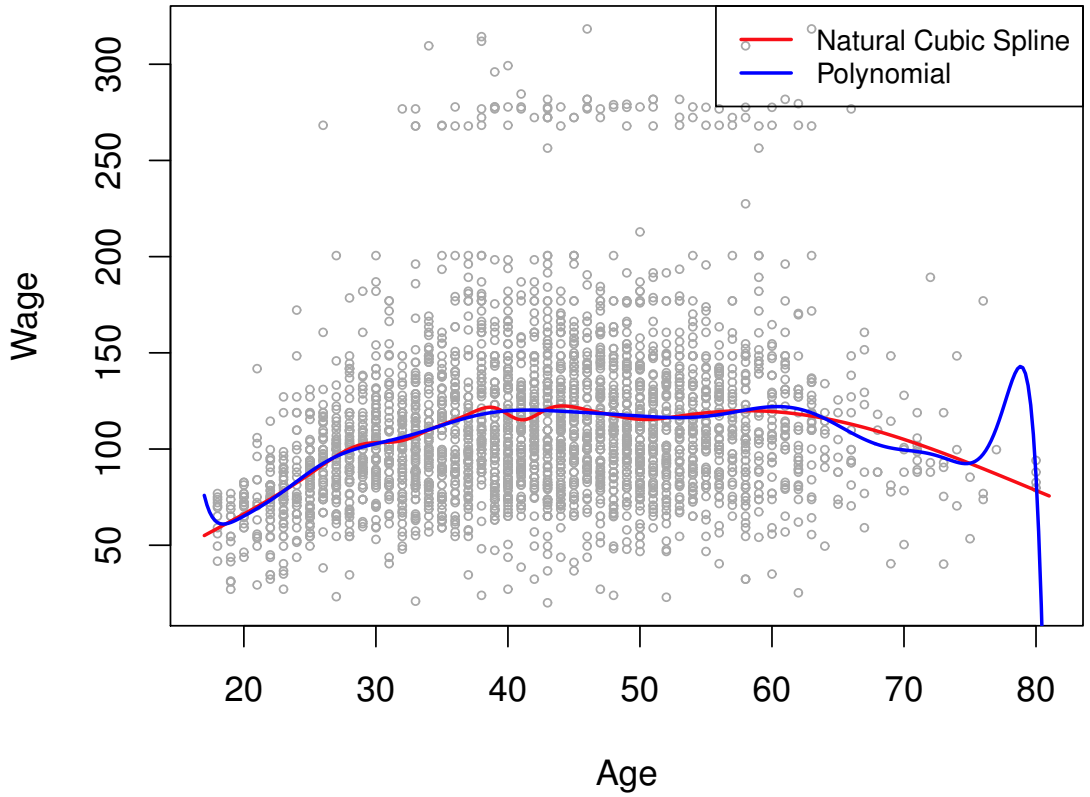


Errores cuadráticos medios con 10-veces Validación Cruzada para *seleccionar los grados de libertad* al ajustar splines a los datos salariales. La respuesta es el *salario* y el predictor de *edad*

7.4.5

Comparación con Regresión Polinomial

Datos Salariales



Los splines introducen flexibilidad al aumentar el número de nudos pero manteniendo el grado fijo del polinomio

- *Spline Cúbico Natural* con 15 grados de libertad y *Polinomio* de grado 15
- Flexibilidad adicional en el *Polinomio* → resultados no deseados en los límites
- *Spline Cúbico Natural* aún proporciona un ajuste razonable a los datos

Splines Suavizados

- Descripción general de las splines suavizados
- Elección del parámetro de suavizado λ

7.5.1

Descripción General de Splines Suavizados

- Ajustar una curva suave implica encontrar una función $g(x)$ que se ajuste bien a los datos observados

$$\text{minimice } RSS = \sum_{i=1}^{n} (y_i - g(x_i))^2$$

- Problema: Si no ponemos ninguna restricción a $g(x)$, $RSS = 0$ si eligimos *g que interpola todos los y_i*
  - Función demasiado flexible que se sobre ajustaría a los datos
- Queremos una *función g que minimice RSS y que además sea suave*

¿Cómo nos asegurarnos que g sea Suave?

- Muchas maneras de hacerlo
- Un enfoque natural es encontrar la *función g que*

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

- donde λ es un parámetro de sintonización no negativo
- La función g se conoce como *spline suavizada (smoothing spline)*

Concepto: Spline Suavizado

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

Pérdida + *Penalización*

- *Función de pérdida:* anima a g a ajustarse bien a los datos
- *Término de penalización:* anima a g a ser suave
  - penaliza la variabilidad en g
  - g'' denota la segunda derivada de la función g

Concepto: Spline Suavizado(2)

- La primera derivada $g'(t)$ mide la pendiente de una función en t
- La segunda derivada $g''(t)$ corresponde a la cantidad en la que cambia la pendiente
- La segunda derivada de una función es una *medida de su rugosidad*:
  - es grande en valor abs si $g(t)$ se mueve mucho cerca de t
  - en caso contrario es cercana a cero
  - La segunda derivada de una línea recta es cero → una línea es perfectamente suave

Parámetro λ

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

- $\lambda \int g''(t)dt$ anima a g a ser suave
  - ➡ Cuanto mayor sea el valor de λ, más suave será g
- Cuando $\lambda = 0$, el término de penalización no tiene efecto
  - la función g tendrá muchos saltos e interpolará exactamente las observaciones de entrenamiento
- Cuando $\lambda \to \infty$,
  - g será perfectamente suave: será una línea recta que pasa lo más cerca posible de los puntos de entrenamiento
- Para un valor intermedio de λ,
  - g se aproximará a las observaciones de entrenamiento pero será algo suave
- **λ controla el equilibrio entre sesgo y varianza del spline suavizado**

Observaciones

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

- Se puede demostrar que la función $g(x)$ que minimiza la expresión tiene algunas propiedades especiales:
  - es un *polinomio cúbico por partes* con nudos en los valores únicos de $x_1, \cdots, x_n$
  - primera y segunda derivadas continuas en cada nudo
  - además, es lineal en la región fuera de los nudos extremos

➡ La función $g(x)$ que minimiza la expresión es un spline cúbico natural con nudos en $x_1, \cdots, x_n$!

Observaciones

- Sin embargo, no es el mismo *spline cúbico natural* que se obtendría si se aplicara el *enfoque de función base* descrito anteriormente con nudos en $x_1, \cdots, x_n$
- Más bien, es una *versión reducida de un spline cúbico natural*, donde el valor del parámetro de ajuste λ *controla el nivel de contracción*

7.5.2

Elección del Parámetro de Suavizado λ

- Un spline suavizado es simplemente un *spline cúbico natural con nudos en cada valor único de* x_i
  - un nudo en cada punto de datos permite una gran flexibilidad
  - tendrá demasiados grados de libertad?
- El parámetro de ajuste λ *controla la rugosidad* del spline de suavizado y, por tanto, *los grados de libertad efectivos*
- A medida que λ aumenta de 0 a ∞, los grados de libertad efectivos df_λ disminuye de n a 2.

Grados de Libertad Efectivos

- En el contexto del suavizado de splines, ¿por qué hablamos de *grados de libertad efectivos* en lugar de *grados de libertad*?
- Los *grados de libertad* se refieren a la cantidad de parámetros libres, como la cantidad de coeficientes que caben en un polinomio o spline cúbico
  - Aunque un spline de suavizado tiene n parámetros y, por tanto, n *grados de libertad nominales*, estos están muy restringidos o reducidos
- Los *grados de libertad efectivos* df_λ es una medida de la flexibilidad de la spline de suavizado: cuanto más alta sea, más flexible (y con menor sesgo pero mayor variación) será la spline de suavizado

Definición Técnica

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

- $\hat{g}_\lambda$ es la solución pora una elección particular de λ

$$\hat{g}_\lambda = S_\lambda y$$

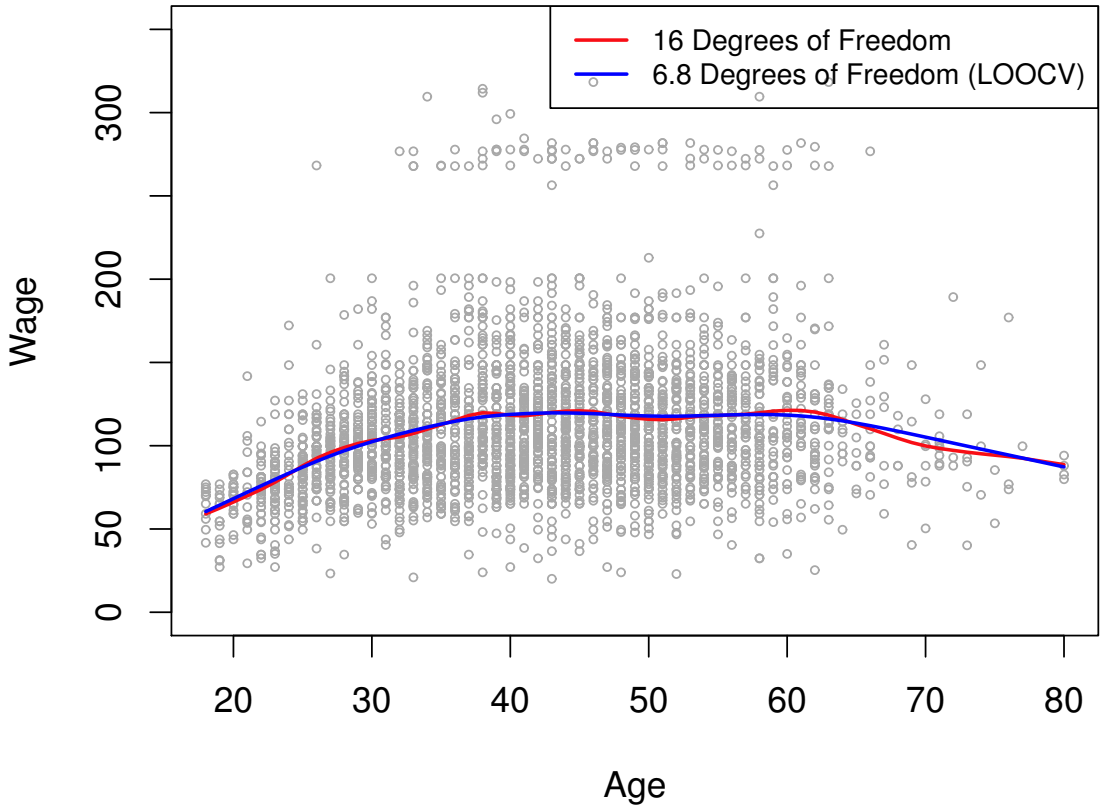
- vector (n) que contiene los valores ajustados del spline suavizado en los puntos de entrenamiento $x_1, \cdots, x_n$
  - se puede escribir como una matriz S_λ ($n \times n$), para la cual existe una fórmula, multiplicada por el vector de respuesta y
- Los *grados de libertad efectivos*: la suma de los elementos diagonales de S_λ

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

Validación Cruzada para Seleccionar λ

- Al ajustar un spline suavizado
  - no necesitamos seleccionar el número o la ubicación de los nudos → un nudo en cada observación de entrenamiento, $x_1, \cdots, x_n$
  - *necesitamos elegir el valor de λ*
- Una posible solución a este problema es usar *validación cruzada*
  - encontrar el λ que minimice el RSS con VC
- El error de Validación Cruzada Excluir Uno (LOOCV) se puede calcular de manera muy eficiente para splines suavizados, con esencialmente el mismo costo que calcular un ajuste único

Spline Suavizado
Datos de Salario



La curva roja resulta de especificar 16 grados de libertad efectivos. Para la curva azul, λ se encontró automáticamente mediante Validación Cruzada de dejar uno fuera, lo que resultó en 6.8 grados de libertad efectivos

Regresión Local

- Es un enfoque diferente para *ajustar funciones no lineales flexibles*, que implica
  - calcular el ajuste en un punto objetivo x_0 utilizando solo las observaciones de entrenamiento cercanas

Algoritmo

Algorithm 7.1 *Local Regression At $X = x_0$*

1. Gather the fraction $s = k/n$ of training points whose x_i are closest to x_0.

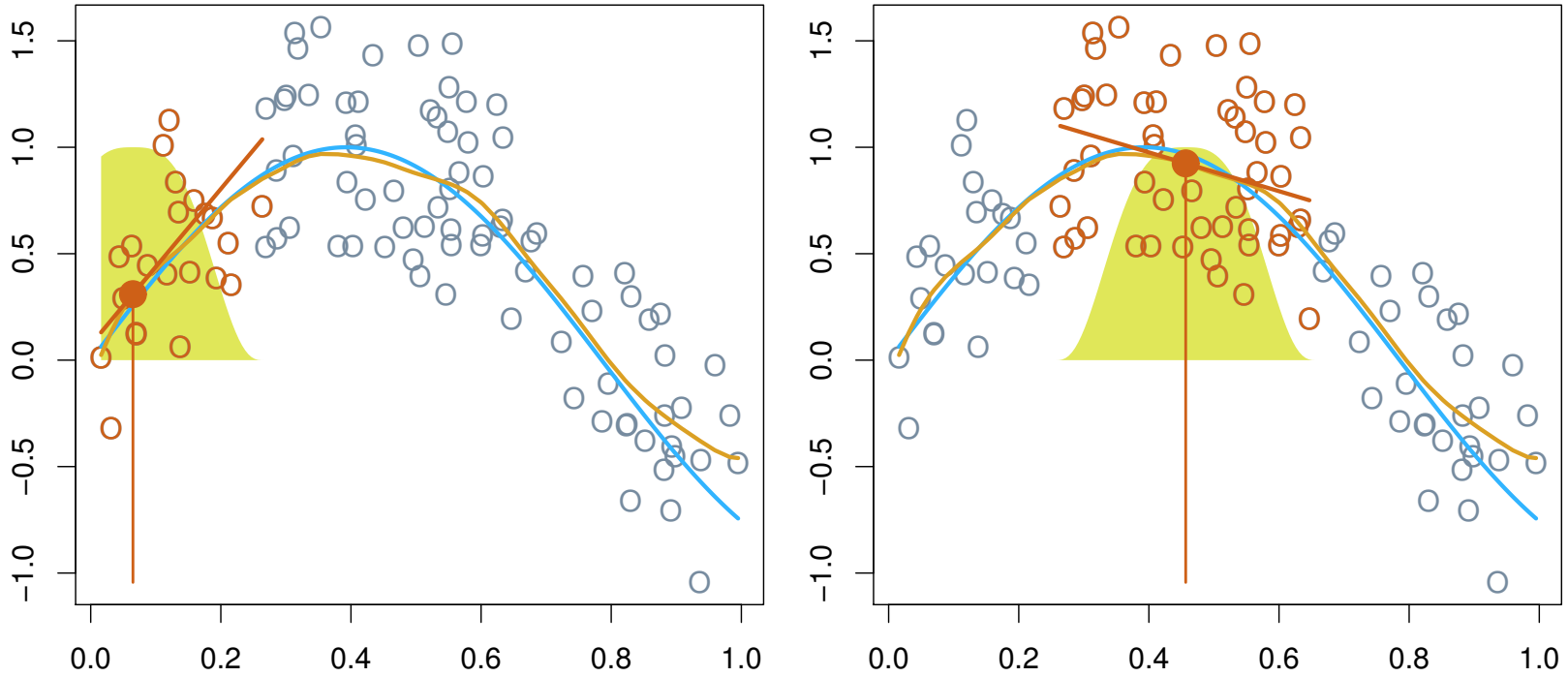
2. Assign a weight $K_{i0} = K(x_i, x_0)$ to each point in this neighborhood, so that the point furthest from x_0 has weight zero, and the closest has the highest weight. All but these k nearest neighbors get weight zero.

3. Fit a *weighted least squares regression* of the y_i on the x_i using the aforementioned weights, by finding $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize

$$\sum_{i=1}^{n} K_{i0}(y_i - \beta_0 - \beta_1 x_i)^2. \tag{7.14}$$

4. The fitted value at x_0 is given by $\hat{f}(x_0) = \hat{\beta}_0 + \hat{\beta}_1 x_0$.

Regresión Local Ilustrada



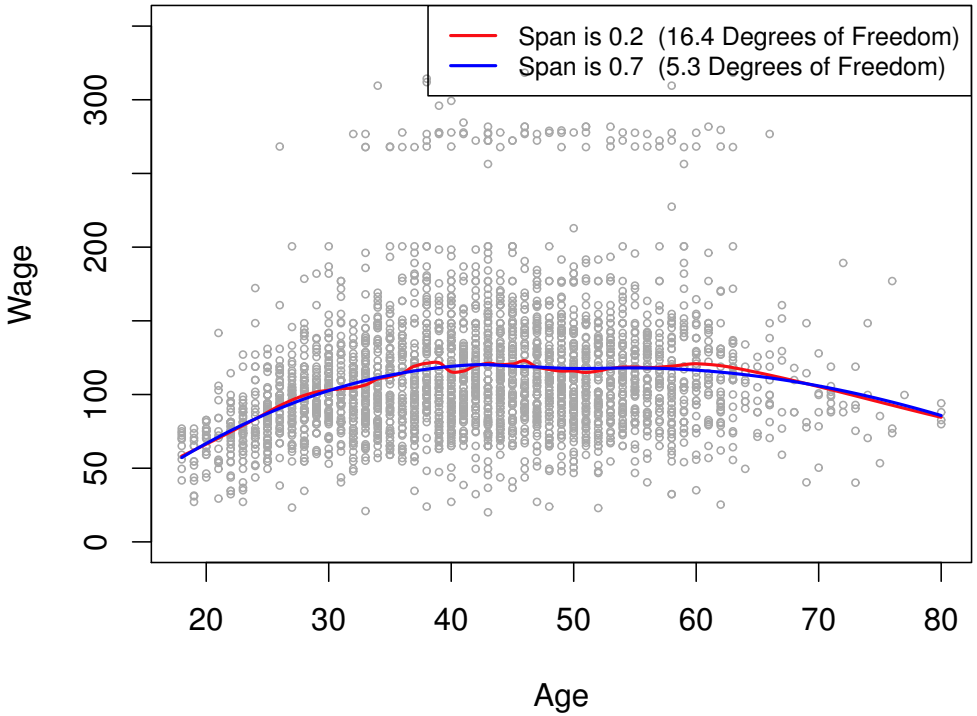
- Regresión Local ilustrada en datos simulados, con un punto objetivo cerca de 0.4 y otro cerca del límite en 0.05.

- La línea azul representa la función $f(x)$ a partir de la cual se generaron los datos, y la línea naranja corresponde a la estimación de regresión local $\hat{f}(x)$

Parámetros

- Para realizar una regresión local, se deben tomar varias decisiones
  - el *intervalo* s, proporción de puntos utilizados para la regresión local en x_0 (1)
  - definir la *función de ponderación* K (2)
  - Que regresión se ajusta: una *regresión lineal*, *constante* o *cuadrática*
- La opción más importante es el *intervalo* s,
  - Desempeña un papel similar al del parámetro de ajuste λ en los splines suavizados: controla la flexibilidad del ajuste no lineal
  - Cuanto menor sea el valor de s, más local y oscilante será el ajuste. Un valor muy grande de s conducirá a un ajuste global de los datos utilizando todas las observaciones de entrenamiento
- Se puede usar la validación cruzada para elegir s, o especificarlo directamente

Regresión Local
Datos de Salario



Ajustes de regresión lineal local en los datos de salarios, utilizando dos valores de s : 0.7 y 0.2

Como se esperaba, el ajuste obtenido usando $s = 0.7$ es más suave que el obtenido usando $s = 0.2$

Modelos Aditivos Generalizados (GAMs)

- GAMs para problemas de regresión
- GAMs para problemas de clasificación

Introducción

- Exploramos varios enfoques para predecir de manera flexible una respuesta Y sobre la base de un único predictor X
  - *Extensiones de la regresión lineal simple*
- Ahora exploramos el problema de predecir Y de manera flexible sobre la base de varios predictores, $X_1, \cdots, X_p$
  - *Extensión de* la *regresión lineal múltiple*
- Los *modelos aditivos generalizados (GAM)* proporcionan un marco general para ampliar un modelo lineal estándar al permitir *funciones no lineales de cada una de las variables*, manteniendo la aditividad.
- Al igual que los modelos lineales, los GAM se pueden aplicar con respuestas tanto *cuantitativas* como *cualitativas.*

7.7.1

GAMs
Problemas de Regresión

- Modelo de *regresión lineal múltiple*

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

- Modelo *aditivo generalizado* (GAM)

$$y_i = \beta_0 + \sum_{j=1}^{p} f_j(x_{ij}) + \epsilon_i$$
$$= \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \epsilon_i$$

β_j ?

- Se llama modelo aditivo porque calculamos un f_j separado para cada X_j y luego sumamos todas sus contribuciones.

GAMs
Componentes Básicos

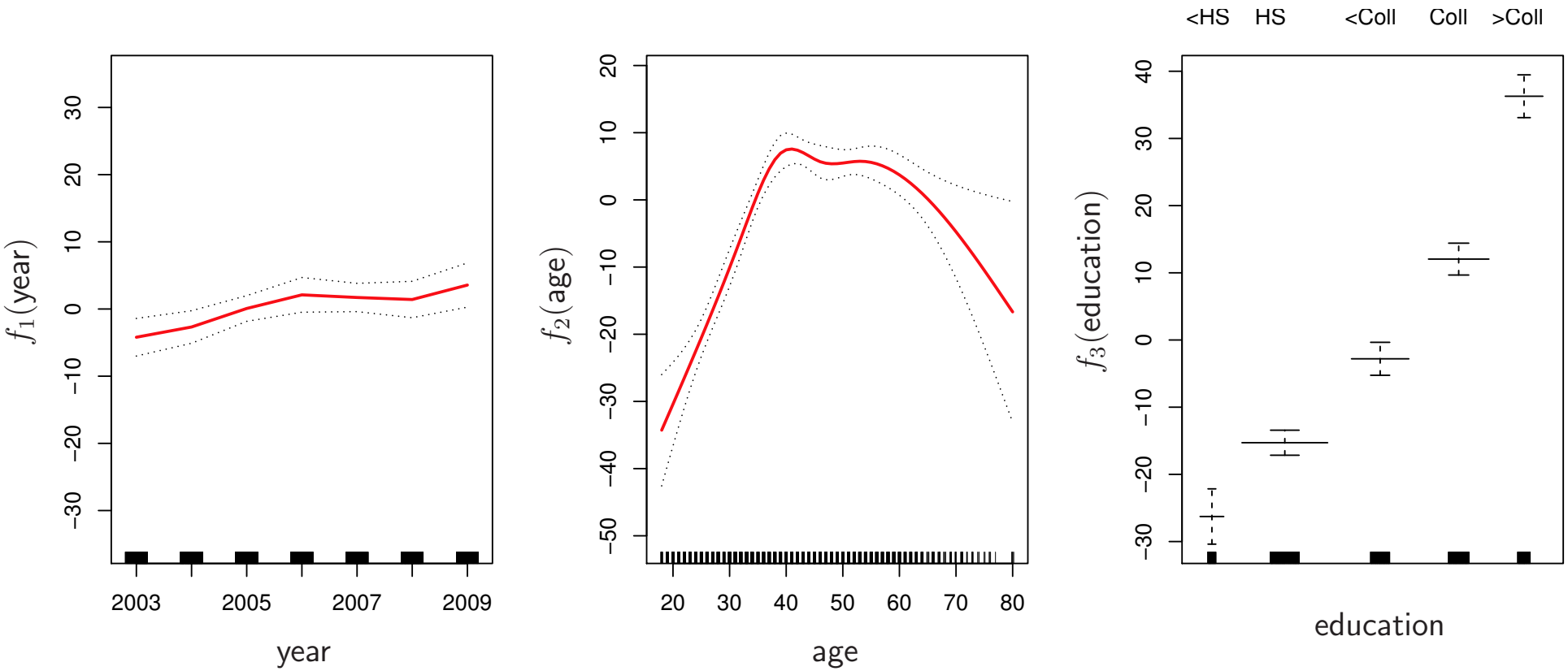
- Analizamos muchos métodos para ajustar funciones a una sola variable
- La belleza de los GAM es que podemos utilizar estos métodos como componentes básicos para ajustar un modelo aditivo
- Para la mayoría de los métodos que hemos visto hasta ahora en este capítulo, esto se puede hacer de manera bastante trivial

GAM : Datos de Salario

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

- *year* y *age* : variables cuantitativas
- *education* : variable qualitativa con 5 niveles
  - <HS, HS, <Coll, Coll, >Coll
- f_1 y f_2: splines naturales
- f_3 se ajusta usando una constante separada para cada nivel, a través del enfoque habitual de variable ficticia

GAM : Datos de Salario (2)

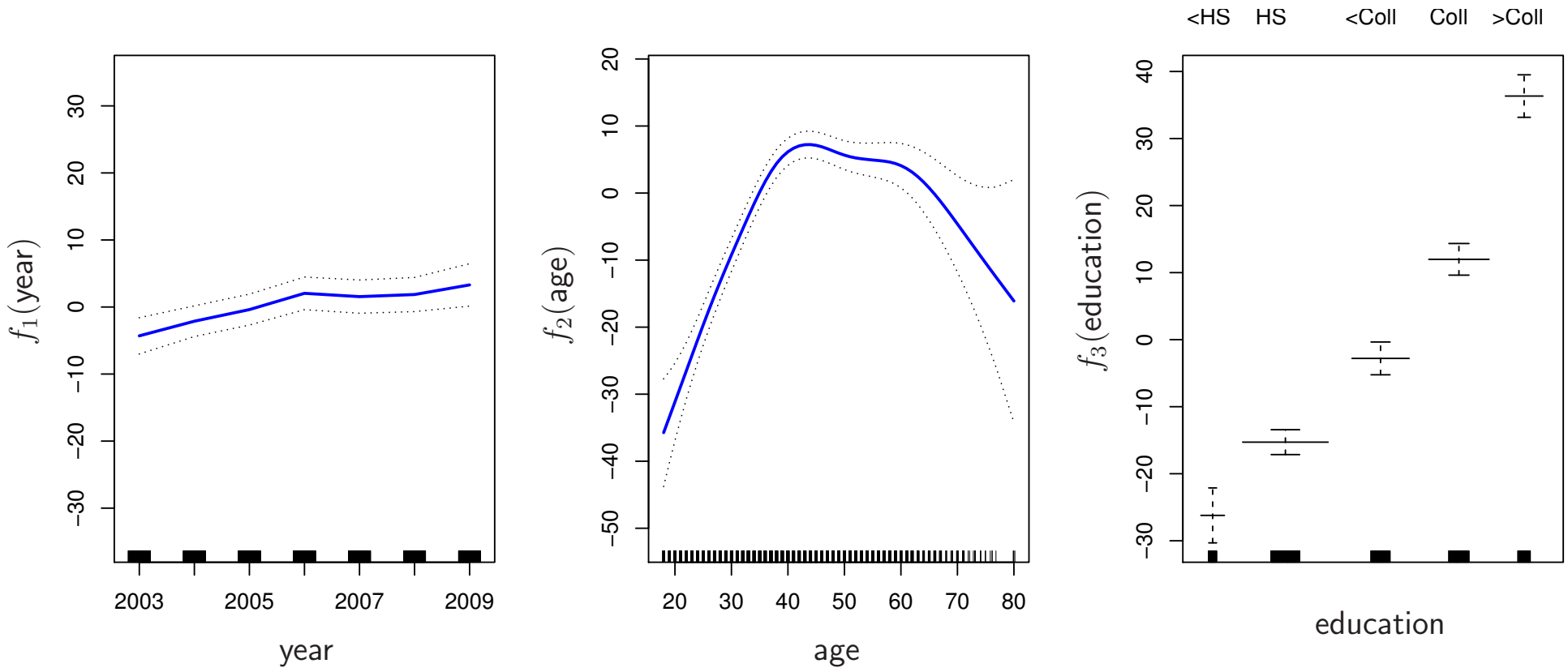


Relación entre cada característica y salario en el modelo ajustado. Cada gráfico muestra la función ajustada y los errores estándar. f_1 y f_2 son *splines naturales* en año y edad, con *4 y 5 grados de libertad*. $f_3$es una *función escalonada*, ajustada a la variable cualitativa educación.

Interpretación

- Izquierda: *manteniendo fijas la edad y la educación*, el *salario* tiende a aumentar ligeramente con el *año*; esto puede deberse a la inflación
- Centro: *manteniendo fijos la educación y el año*, el salario tiende a ser más alto para los valores intermedios de *edad* y más bajo para los muy jóvenes y los muy mayores
- Derecha: *manteniendo fijos el año y la edad*, el salario tiende a aumentar con la *educación*: cuanto más educada es una persona, mayor es su salario, en promedio.
- Todos estos hallazgos son intuitivos

GAM : Datos de Salario (3)



f_1 y f_2 son *splines suavizados* con 4 y 5 grados de libertad. Las funciones ajustadas son bastante similares a la figura anterior. En la mayoría de las situaciones, las diferencias en los GAM obtenidos usando *splines suavizados* versus *splines naturales* son pequeñas

Bloques de Construcción

- No tenemos que usar únicamente splines como bloques para construir los GAM
- Para crear un GAM también podemos usar
  - regresión local
  - regresión polinomial o
  - cualquier combinación de los enfoques vistos anteriormente en este capítulo

GAMs : Ventajas

- Los GAM nos permiten ajustar un f_j no lineal a cada X_j, de modo que podamos modelar automáticamente relaciones no lineales que la regresión lineal estándar pasará por alto
    - No necesitamos probar manualmente muchas transformaciones diferentes en cada variable individualmente
- Los ajustes no lineales pueden potencialmente hacer predicciones más precisas para la respuesta Y
- Como el modelo es aditivo, podemos examinar el efecto de cada X_j sobre Y individualmente manteniendo fijas todas las demás variables
- La suavidad de la función f_j para la variable X_j se puede resumir en grados de libertad

GAMs : Desventajas

- La principal limitación de los GAM es que el modelo está restringido a ser aditivo
- Con muchas variables, se pueden perder interacciones importantes
- Sin embargo, al igual que con la regresión lineal, podemos agregar manualmente términos de interacción al modelo GAM incluyendo predictores adicionales de la forma $X_j \times X_k$
- Además, podemos agregar funciones de interacción de baja dimensión de la forma $f_{jk}(X_j, X_k)$ al modelo
  - dichos términos se pueden ajustar utilizando suavizadores bidimensionales como la *regresión local* o *splines bidimensionales* (no cubiertos aquí)

Observaciones

- Para *modelos completamente generales*, tenemos que buscar enfoques aún más flexibles, como
  - random forest (bosques aleatorios) y
  - boosting (refuerzo ?)
- Los GAM proporcionan un compromiso útil entre *modelos lineales* y *modelos completamente no paramétricos*

7.7.2

GAMs
Problemas de Clasificación

- Regresión logística (Y toma valores 0 o 1)

$$log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

- Modelo aditivo generalizado (GAM)

β_j ?

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

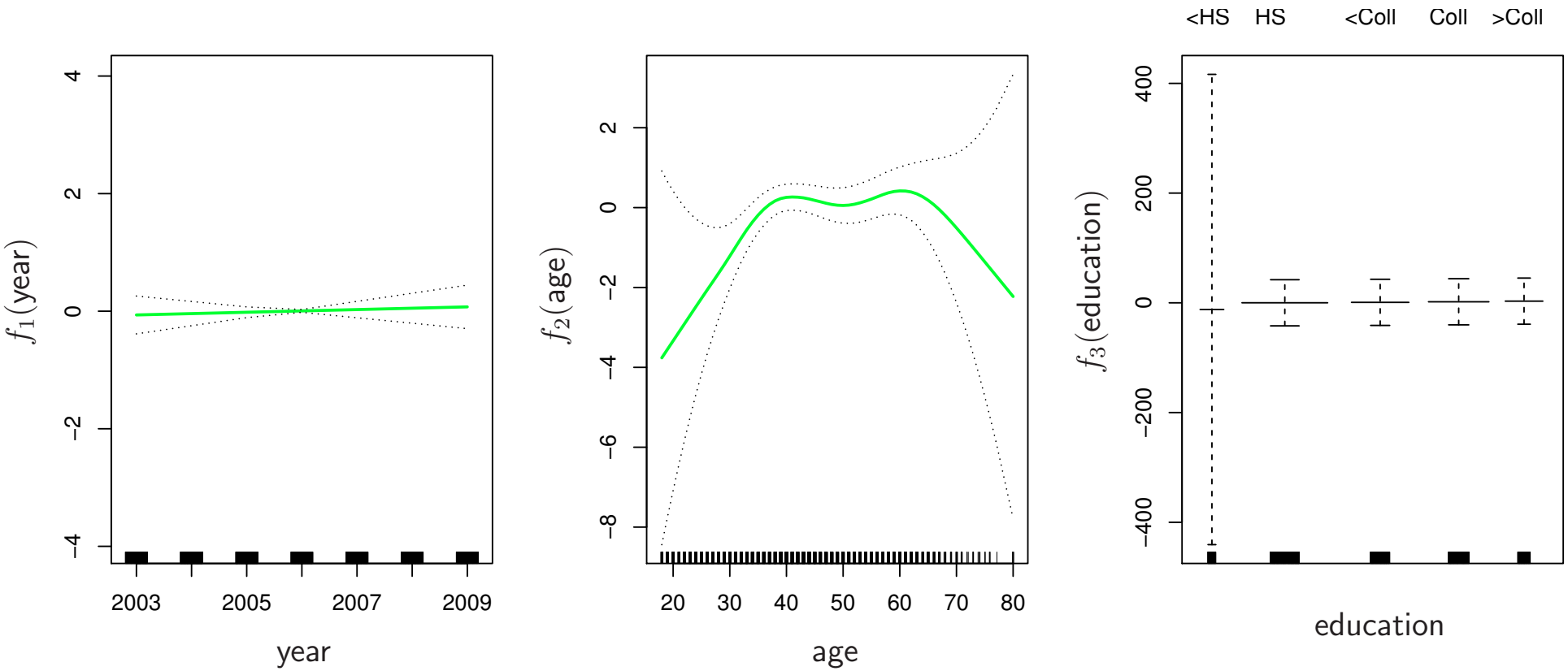
RL GAM: Datos de Salario

$$log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 \times year + f_2(age) + \cdots + f_3(education)$$

$$p(X) = Pr(wage > 250 \,|\, year, age, education)$$

- f_1 es lineal en el año
- f_2 se ajusta usando un spline suavizada con 5 grados de libertad
- f_3 se ajusta como una función escalonada, creando variables ficticias para cada nivel de educación

RL GAM: Datos de Salario (2)

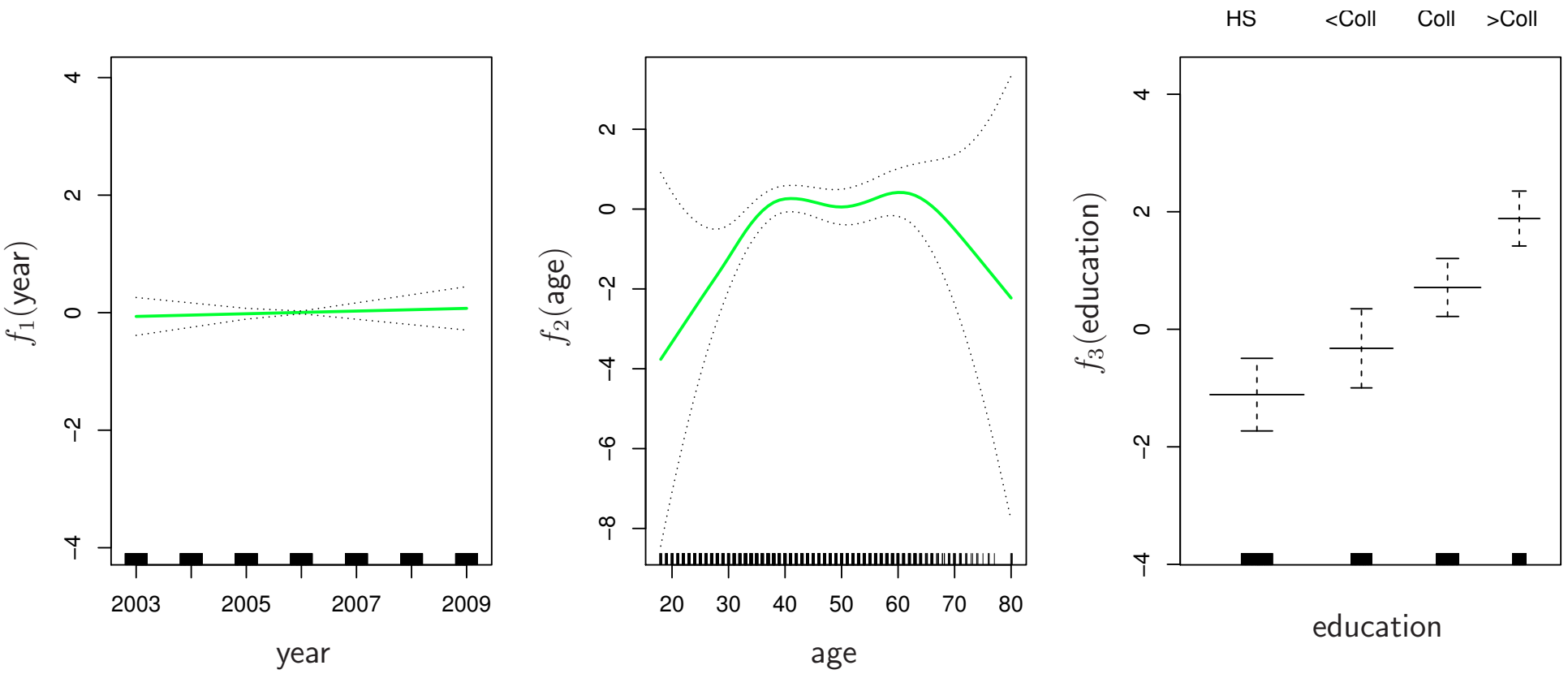


RL GAM se ajusta a la respuesta binaria I(*salario*>250). Cada gráfico muestra la función ajustada y los errores estándar. f_1 es *lineal* en *año*, $f_2$es un *spline suavizado* con 5 grados de libertad en *edad* y f_3 es una *función escalonada* para *educación*

Observaciones

- El último panel parece sospechoso, con intervalos de confianza muy amplios para el nivel <HS
- De hecho, ningún valor de respuesta equivale a uno para esa categoría: ninguna persona con *educación* inferior a la *secundaria* gana más de 250.000 dólares al año
- Por lo tanto, reajustamos el GAM, excluyendo a las personas con educación inferior a la secundaria.

RL GAM: Datos de Salario (3)



Se ajusta el GAM anterior excluyendo las observaciones para las cuales la *educación* es <HS

Se puede evaluar visualmente las contribuciones relativas de cada una de las variables

Ahora se ve que una mayor *educación* tiende a estar asociada con *salarios* más altos. La *edad* y la *educación* tienen un efecto mucho mayor que el *año* en la probabilidad de tener *ingresos altos*

7.8

Lab: Modelado no Lineal

- Regresión Polinómica y Funciones Escalonadas
- Splines (junquillo, empalme)
- Splines Suavizadas y GAMs
- Regresión Local

Ejercicios