Estimación No Paramétrica

- En muchos casos los supuestos de linealidad no son realistas.
- Hay propuestas no lineales, pero típicamente se pierde interpretabilidad.
- Muchas de las técnicas estudiadas dependen de la linealidad de los datos $\rightsquigarrow$ vamos a relajar este supuesto manteniendo la posibilidad de interpretar los modelos.

Estimación Polinomial

Comenzamos considerando el modelo lineal

$$f_X(x_i) = \beta_0 + \beta_1 x_i + \epsilon_i$$

Una primera aproximación para capturar la no linealidad es proponer un modelo

$$f_X(x_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \epsilon_i$$

la llamamos regresión polinomial.

- sigue siendo un modelo lineal.
- las interpretación de los coeficientes es más compleja en muchos casos.
- si el orden del polinomio es alto aparecen los *problemas de frontera*.

Estimación Polinomial

Consideramos el conjunto de datos `Wage` de la librería `ISLR`, que consta 3000 datos de hombres trabajadores la región del Atlántico medio de EEUU.

wage salario anual en miles de dólares.

year año en que la información fue tomada.

age edad del trabajador.

education variable categórica con 5 niveles diferentes niveles (<HS, HS, Some College, College Grad y Advanced Degree).

Estimación Polinomial

Comenzamos haciendo un ajuste polinomial de orden 4.
Representamos el `wage` en función de la `edad`

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4 + \epsilon_i$$

```
fit=lm(wage~poly(age ,4) ,data=Wage)
coef(summary(fit))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	111.70	0.73	153.28	0.00
poly(age, 4)1	447.07	39.91	11.20	0.00
poly(age, 4)2	-478.32	39.91	-11.98	0.00
poly(age, 4)3	125.52	39.91	3.14	0.00
poly(age, 4)4	-77.91	39.91	-1.95	0.05

Estimación No Paramétrica

Las varianzas en cada punto se calculan a partir de las varianzas de los $\widehat{\beta}$ y de las covarianzas entre ellos, para el caso lineal por ejemplo está dada por,

$$var\left(\widehat{f}(x_0)\right) = \widehat{\beta}_0 + \sum_{j=1}^{p} x_0'^j var(\widehat{\beta}_j) x_0^j + 2\sum_{j\neq j'} x_0'^j cov(\widehat{\beta}_j, \widehat{\beta_{j'}}) x_0^{j'}$$

Desventaja: Se impone una estructura global, que al ganar flexibilidad toma formas *extrañas* que no ajustan bien en algunos puntos.

Estimación Polinomial

Este ajuste es en una base de polinomios ortogonales de age, age^2, age^3 y age^4 se obtienen el mismo si consideramos que la base sea age, age^2, age^3 y age^4.

```
fit2=lm(wage~poly(age ,4,raw =T),data=Wage)
coef(summary(fit2))
```

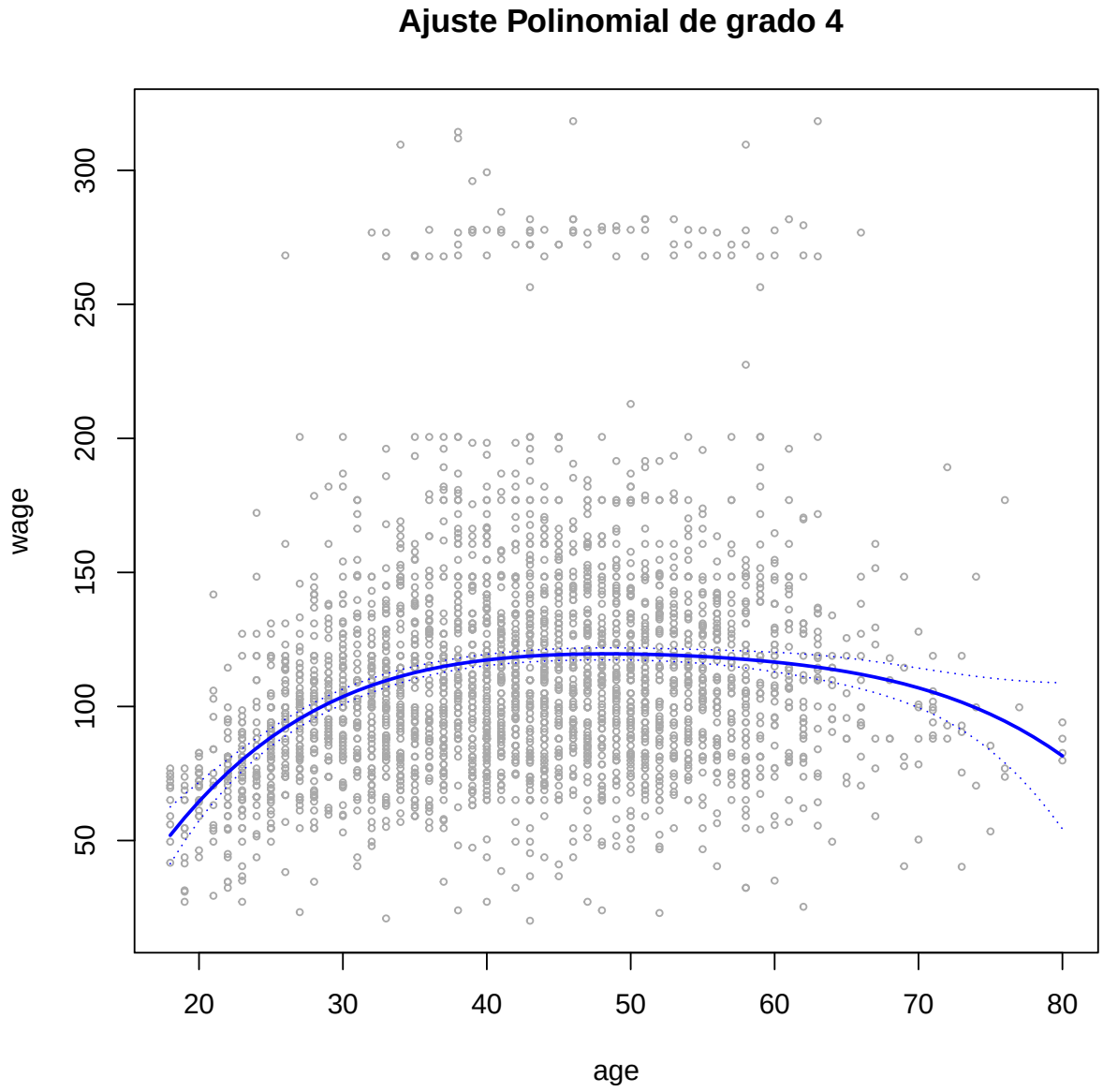
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-184.15	60.04	-3.07	0.00
poly(age, 4, raw = T)1	21.25	5.89	3.61	0.00
poly(age, 4, raw = T)2	-0.56	0.21	-2.74	0.01
poly(age, 4, raw = T)3	0.01	0.00	2.22	0.03
poly(age, 4, raw = T)4	-0.00	0.00	-1.95	0.05

Estimación Polinomial

Una vez realizado este ajuste podemos hacer las predicciones con sus desvíos estándares, que se pueden volcar al gráfico.

```
agelims =range(age)
age.grid=seq (from=agelims[1], to=agelims[2])
preds=predict(fit ,newdata
=list(age=age.grid),se=TRUE)
se.bands=cbind(preds$fit+2*preds$se.fit,
preds$fit-2*preds$se.fit)
```

Estimación Polinomial



Estimación Polinomial

Cómo decidir que grado tiene que tener el polinomio?
Podemos probar varios grados tenemos luego modelos anidados.
Realizamos un analisis de la varianza

H_0 : El modelo $\mathcal{M}_1$ es suficiente para explicar los datos.

vs

H_A : Es necesario un modelo de orden superior$\mathcal{M}_2$ para explicar los datos

```
fit.1= lm(wage~age,data=Wage)
fit.2= lm(wage~poly(age,2),data=Wage)
fit.3= lm(wage~poly(age,3),data=Wage)
fit.4= lm(wage~poly(age,4),data=Wage)
fit.5= lm(wage~poly(age,5),data=Wage)
```

Estimación Polinomial

```
anova(fit.1,fit.2,fit.3,fit.4,fit.5)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	2998	5022216.10				
2	2997	4793430.09	1	228786.01	143.59	0.0000
3	2996	4777674.40	1	15755.69	9.89	0.0017
4	2995	4771604.25	1	6070.15	3.81	0.0510
5	2994	4770321.69	1	1282.56	0.80	0.3697

Luego, hay que ajustar un polinomio de orden 3 ó 4.

Otra opción es elegir el grado del polinomio por cross validation

Estimación Polinomial

A continuación utilizaremos esta técnica en un modelo de regresión logística.
Queremos predicir la probabilidad de que el sueldo sea superior a 250000$, es decir,

$$Y = \mathcal{I}\left(wage > 250\right)$$

luego, proponemos el siguiente modelo

$$P(y_i > 250) = \frac{exp\left(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4\right)}{1 + exp\left(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 x_i^4\right)}.$$

```
fit=glm(I(wage >250)~ poly(age ,4) ,data=Wage
,family =binomial )
```

y realizamos las correspondientes predicciones

```
preds=predict (fit ,newdata =list(age=age.grid),se=T)
```

Estimación Polinomial

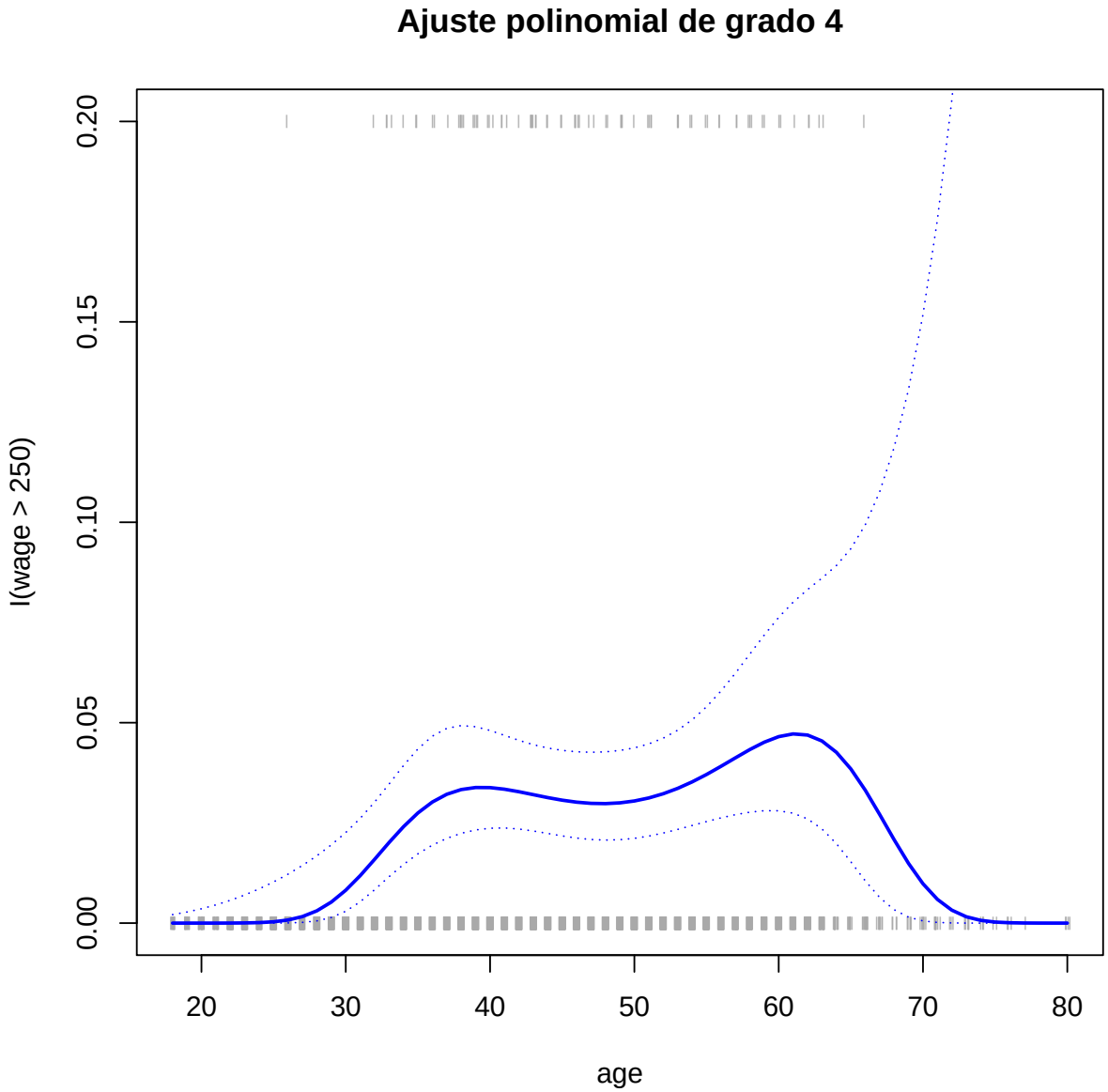
Encontrar las bandas de confianza en este caso es más complicado nos basamos en la idea de que la función de enlace es

$$\log\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) = X'\beta.$$

Luego, las predicciones son del tipo $X'\beta$, luego se pueden calcular de ese modo los errores estándares, luego las bandas de confianza se encuentran aplicando,

$$P(Y=1|X) = \frac{exp\,(X'\beta)}{1+exp\,(X'\beta)}$$

Estimación Polinomial



Funciones Escalera

Se divide al recorrido de la variable X en segmentos donde a cada uno de ellos se le asigna un valor constante. Es decir, determinar $c_1, \ldots, c_K$ puntos en el recorrido de X tales que

$$\begin{aligned} C_0(X) &= \mathcal{I}(X < c_1) \\ C_1(X) &= \mathcal{I}(c_1 \leq X < c_2) \\ C_2(X) &= \mathcal{I}(c_2 \leq X < c_3) \\ \vdots & \quad \vdots \\ C_{k-1}(X) &= \mathcal{I}(c_{k-1} \leq X < c_k) \\ C_k(X) &= \mathcal{I}(c_k \leq X) \end{aligned}$$

donde

$$C_0(X) + C_1(X) + \cdots + C_K(X) = 1$$

Funciones Escalera

Ajustando el modelo

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

para cada x_1 un solo $C_j(x_i) \neq 0$.
β_0 se puede interpretar como el valor medio de y si $X < c_1$ y $\beta_0 + \beta_j$ como el valor medio de y si $c_{j-1} \leq X < c_j$, es decir β_j representa el cambio medio de y para $c_{j-1} \leq X < c_j$ en relación a $X < c_1$.

Funciones Escalera

```
table(cut(age,4))
```

	age
(17.9,33.5]	750
(33.5,49]	1399
(49,64.5]	779
(64.5,80.1]	72

```
fit=lm(wage~cut(age,4),data=Wage)
coef(summary(fit))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	94.16	1.48	63.79	0.00
cut(age, 4)(33.5,49]	24.05	1.83	13.15	0.00
cut(age, 4)(49,64.5]	23.66	2.07	11.44	0.00
cut(age, 4)(64.5,80.1]	7.64	4.99	1.53	0.13

Funciones Escalera

De manera análoga se puede plantear la regresión logística,

$$P(y_i > 250) = \frac{exp\left(\beta_0 + \beta_1 C_1(x_i) + \cdots + \beta_K C_K(x_i)\right)}{1 + exp\left(\beta_0 + \beta_1 C_1(x_i) + \cdots + \beta_K C_K(x_i)\right)}$$

sin analizar los coeficientes.
Los estimadores de β se pueden estimar como se hace usualmente por OLS o máxima verosimilitud.

Funciones Escalera

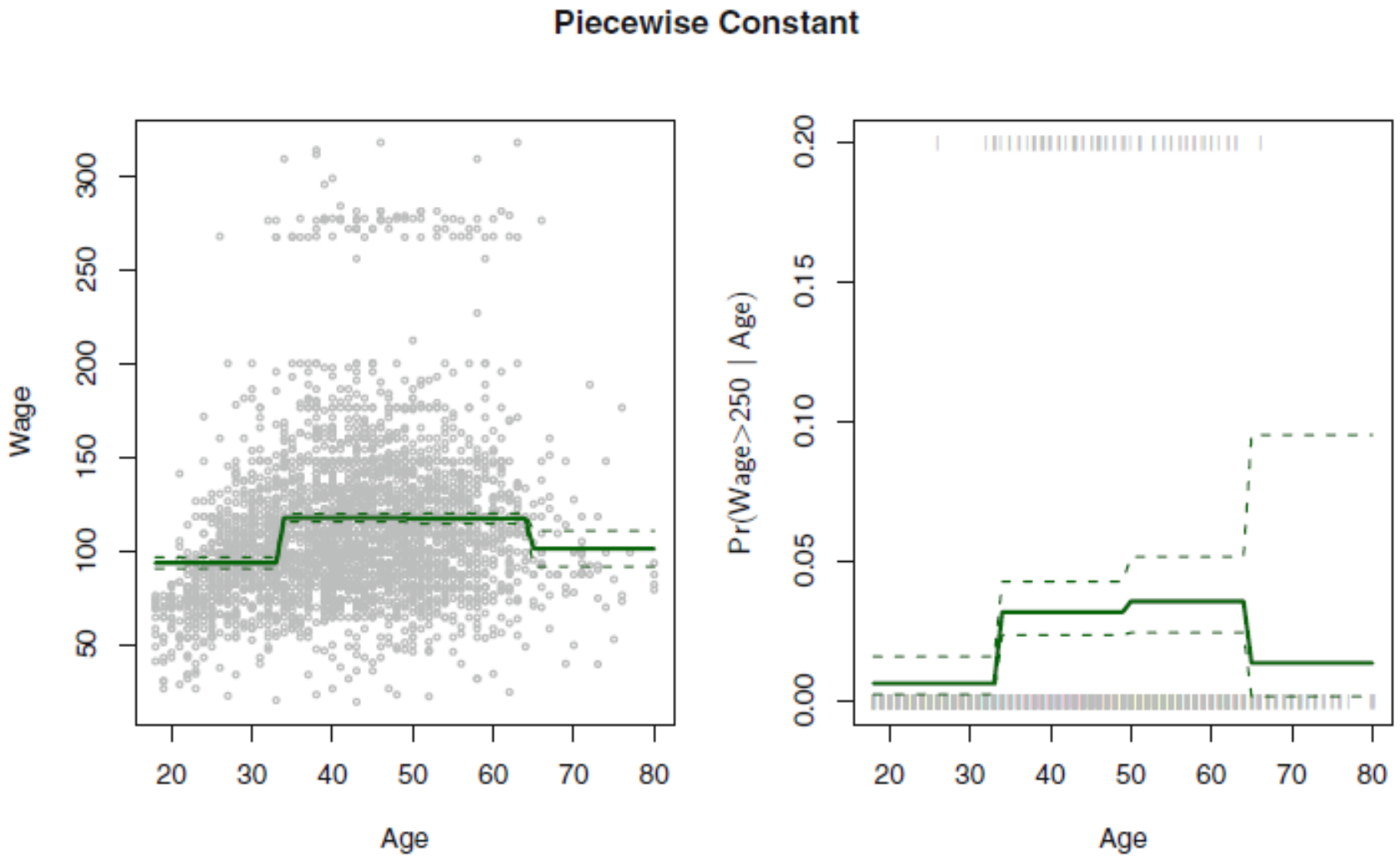


Figure: ISLR

Funciones Escalera

Ventajas

- Modelo muy sencillo de interpretar.

Desventajas

- Si no se conocen los cortes de los segmentos suele perder la señal principal. Ej: en el gráfico de la izquierda para edades bajas no captura el crecimiento.

Modelos definidos mediante una base de funciones

Supongamos que tenemos una base de funciones
$b_1(X), b_2(X), \ldots, b_k(X)$ una base de funciones.
Proponemos el modelo

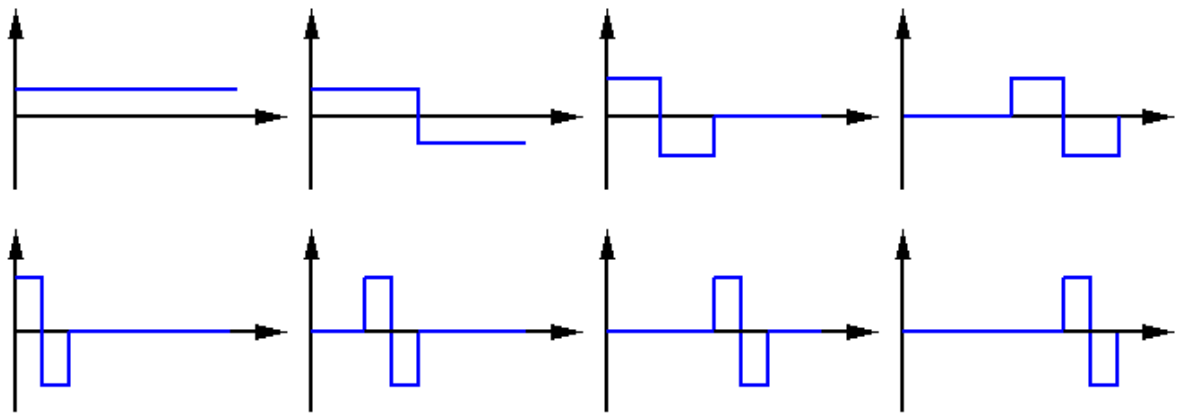
$$y_i = \beta_0 + \beta_1 b_1(x_i) + \cdots + \beta_K b_K(x_i) + \epsilon_i.$$

las funciones $b_1(\cdot), b_2(\cdot), \ldots, b_k(\cdot)$ son fijas y conocidas.
En los casos anteriores son $X, X^2, \ldots, X^K$ o $\mathcal{I}(c_{j-1} \leq X < c_j)$
Luego se estiman los coeficientes en forma usual y el análisis de los modelos se hereda en forma inmediata.

Modelos definidos mediante una base de funciones

Otras bases de funciones:

- Base de Fourier. $\{\cos(nx), \sin(nx)\}$ (`https://upload.wikimedia.org/wikipedia/commons/2/2b/Fourier_series_and_transform.gif`)
- Base de Wavelet, base de Haar.



- Splines

Splines

Combina de manera astuta las dos primeras propuestas

polinómico + constantes trozos =polinomios a trozos

Se segmenta X y en cada segmento se ajusta un polinomio de orden bajo, típicamente de orden 3.
Luego

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

Los β son diferentes en diferentes partes del ajuste. Los puntos donde cambian se llaman *knots*.

Splines

- A mayor cantidad de *knots* más flexibilidad tiene el polinomio.
- Si determinamos k knots tenemos que ajustar $k + 1$ polinomios de orden 3, es decir $(k + 1) \times 4$ parámetros.

- En principio queda discontinuo.

Splines

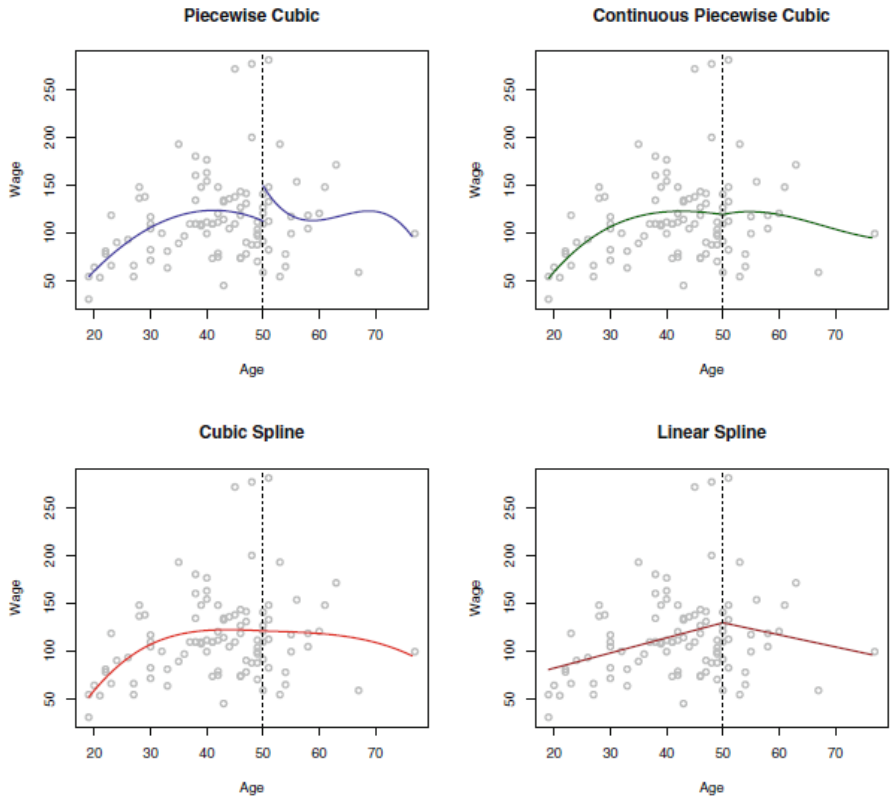


Figure: ISLR

Splines

Hay que imponer condiciones para que el ajuste quede bien

- Se puede pensar en imponer continuidad, ajuste superior derecho
- Más adecudado continuidad en las derivadas primera y segunda.
- Cada restricción que imponemos libera un grado de libertad (continuidad en f, f' y f'') quita 3 grados de libertad.

Splines

polinómico + constantes a trozos+suavidad =splines

La cantidad de parámetros a ajustar son $K + 4$, donde K es la cantidad de knots.
Un *spline cúbico* de orden K se puede luego modelar como

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

donde $b_1, \ldots, b_{K+3}$ son las bases del spline convenientemente elegidas, luego los estimadores de los coeficientes se pueden ajustar por mínimos cuadrados.
Típicamente la base que se considera son *polinomios truncados*

$$h(x, \psi) = (x - \xi)_+^3 = (x - \xi)^3 \mathcal{I}(x > \xi)$$

Splines

La base que se considera está dada por

$$X, X^2, X^3, h(x, \psi_1), \ldots, h(x, \psi_K)$$

donde

$$h(x, \psi) = (x - \xi)_+^3 = (x - \xi)^3 \mathcal{I}(x > \xi)$$

y $\xi_1, \ldots, \xi_K$ son los *knots.*

Splines

- Solamente utiliza $K + 4$ grados de libertad.
- Presenta problemas de frontera, donde aparece alta varianza.
- Adicionar restricciones de frontera $\rightsquigarrow$ imponer linealidad en los bordes, splines naturales.
- Se pierde la posibilidad de interpretar.

Splines

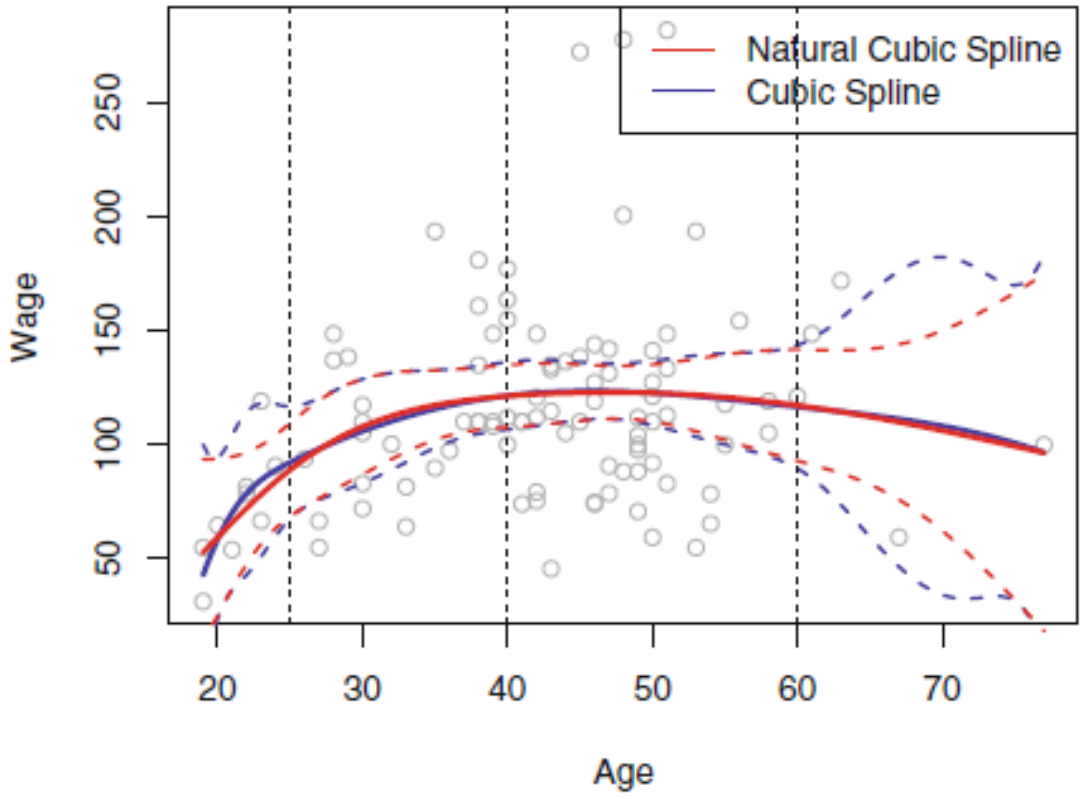


Figure: ISLR

Splines

Cuántos nudos considerar y donde situarlos?

- Teoría $\rightsquigarrow$ donde tenga crecimiento más pronunciado.
- Práctica $\rightsquigarrow$ en forma uniforme siguiendo los cuantiles de los datos.
- K se puede elegir por cross validation, como medida de desajuste se puede usar RSS o deviance, dependiendo del problema.
- En general las estimaciones por splines cúbicos son más estables que las estimaciones por polinomios y tienen menos problemas de fronteras.

Splines

El mismo problema se puede ver desde otra perspectiva. Buscamos la función g, tal que minimice

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

- Evitar el sobreajuste.
- Imponer regularidad.
- g'' da una medida de rugosidad de la función g
- $\sum_{i=1}^{n} (y_i - g(x_i))^2$ función de pérdida.
- $\lambda \int g''(t)^2 dt$ penalidad.

Splines

La solución de este problema es un **spline cúbico** con knots en $x_1, \ldots, x_n$ pero no es el mismo que se construye con la primer formulación que dimos.

- Muchos knots $\rightsquigarrow$ mucha rugosidad.
- g'' $\rightsquigarrow$ controla la varianza.
- λ $\rightsquigarrow$ cross validation.
- Hay una formulación explícita para LOOCV.

Regresión No Paramétrica

Consideramos el problema de regresión general en términos de la esperanza condicional. Sea $(X, Y) \in \mathbb{R}^p \times \mathbb{R}$, llamamos función de regresión a

$$f_0(X) = E(Y|X)$$

el objetivo es estimar f_0, mediante una función $\widehat{f}$, a partir de una muestra iid $(x_1, y_1), \ldots, (x_n, y_n)$ en $\mathbb{R}^p \times \mathbb{R}$, que tengan la misma distribución que (X, Y). No se asume ninguna forma específica para f_0.

Regresión No Paramétrica

Para una muestra $(x_1, y_1), \ldots, (x_n, y_n)$ siempre podemos escribir

$$y_i = f_0(x_i) + \epsilon_i \text{ para } i = 1, \ldots, n.$$

donde $\epsilon_1, \ldots, \epsilon_n$ son errores iid, con media cero. Asumimos también que los errores son independientes de las observaciones, x_i. Podemos considerar regresores fijos o aleatorios, esto no modifica los resultados teóricos a grandes rasgos.

k Vecinos más cercanos

Caso unidimensional, $d = 1$

Fijamos $k \in \mathbb{N}_{>1}$,

$$\widehat{f}(x) = \frac{1}{k} \sum_{i \in \mathcal{N}_k} y_i,$$

donde $\mathcal{N}_k$ es el conjunto de índices correspondientes a las k observaciones entre $x_1, \ldots, x_n$ más próximos a x.

- Estimador muy sencillo.
- Si k es pequeño la estimación suele ser muy rugosa, es muy flexible.
- Si k es pequeño la estimación suele sobresuavizar, es rígido.
- El valor de k se suele elegir por cross validation.

k Vecinos más cercanos

Una de las principales limitaciones es que la función estimada suele tener apariencia dentada.

$$\widehat{f}(x) = \frac{1}{k} \sum_{i \in \mathcal{N}_k} y_i = \sum_{i=1}^{n} w_i(x) y_i,$$

donde

$$w_i(x) = \begin{cases} 1/k & \text{si } x_i \text{ es uno de los } k \text{ vecinos más cercanos de } x. \\ 0 & \text{si no.} \end{cases}$$

La función $w_i(x)$ es discontinua y esta propiedad es heredada por la $\widehat{f}$.

k Vecinos más cercanos

- Bajo el modelo de regresores aleatorios, la estimación de vecinos más cercanos es consistente, siempre y cuando $k_n \to \infty$ y $k_n/n \to 0$. Por ej. $\sqrt{n}$ satisface esto.
- Asumiendo que la función f_0 es Lipschitz continua el estimador de k vecinos más cercanos con $k = \Theta(n^{2/(2+p)})$ satisface que

$$E\left(\left(\widehat{f}(X) - f_0(X)\right)^2\right) = \mathcal{O}\left(n^{2/(2+p)}\right)$$

k Vecinos más cercanos

- Fuerte dependencia de la dimensión del espacio. Para $\epsilon > 0$ pequeño, el tamaño muestral tiene que satisfacer que

$$\begin{aligned} (1/n)^{2/(2+p)} &\leq \epsilon \\ (1/\epsilon)^{(2+p)/2} &\leq n \end{aligned}$$

Luego el tamaño muestral aumenta exponencialmente con el aumento de la dimensión del espacio. Este fenómeno es característico de las regresiones no paramétricas y se conoce como maldición de la dimensionalidad (curse of dimensionality).

Estimación por núcleos

Caso unidimensional, $d = 1$

Consideramos una función de nucleo o *kernel*, $K : \mathbb{R} \to \mathbb{R}$ no negativa tal que

- $\int K(x)dx = 1.$
- $\int xK(x)dx = 0.$
- $0 < \int x^2 K(x)dx < \infty.$

Estimación por núcleos

Los núcleos más utilizados son el núcleo Gaussiano

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp -(x^2/2).$$

y el núcleo de Epanechnicov

$$K(x) = \begin{cases} 3/4(1-x^2) & \text{si } |x| \leq 1. \\ 0 & \text{si no.} \end{cases}$$

Estimación por núcleos

Luego dado un ancho de ventana h, el estimador de núcleos de Naradaya-Watson está dado por,

$$\widehat{f}(x) = \frac{\sum_{i=1}^{n} K(\frac{x-x_i}{h}) y_i}{\sum_{i=1}^{n} K(\frac{x-x_i}{h})} = \sum_{i=1}^{n} w_i(x) y_i,$$

donde

$$w_i(x) = \frac{K(\frac{x-x_i}{h})}{\sum_{j=1}^{n} K(\frac{x-x_j}{h})}$$

Estimación por núcleos

- Al igual que en el caso de vecinos más cercanos es un promedio ponderado de observaciones, la principal diferencia es que en este caso al ser continuo el núcleo también lo es la $\widehat{f}$.
- Presenta problemas de sesgo cerca de las fronteras en los cuales los datos están medidos, esto se debe a la simetría en los núcleos.

Estimación por núcleos

- Se extienden en forma inmediata a contextos multidimensionales considerando la distancia euclidea entre x_i y x. En estos casos los problemas de estimación de la frontera aumentan.
- Sufre también la maldición de la dimensionalidad.
- Se puede ver que el estimador es consistente si $h \to 0$ y $nh^d \to \infty$.
- Se puede encontrar por cv la ventana óptima.

Para profundizar un poco en estos temas
`http://www.stat.cmu.edu/~larry/=sml/nonpar.pdf`

Estimación por núcleos

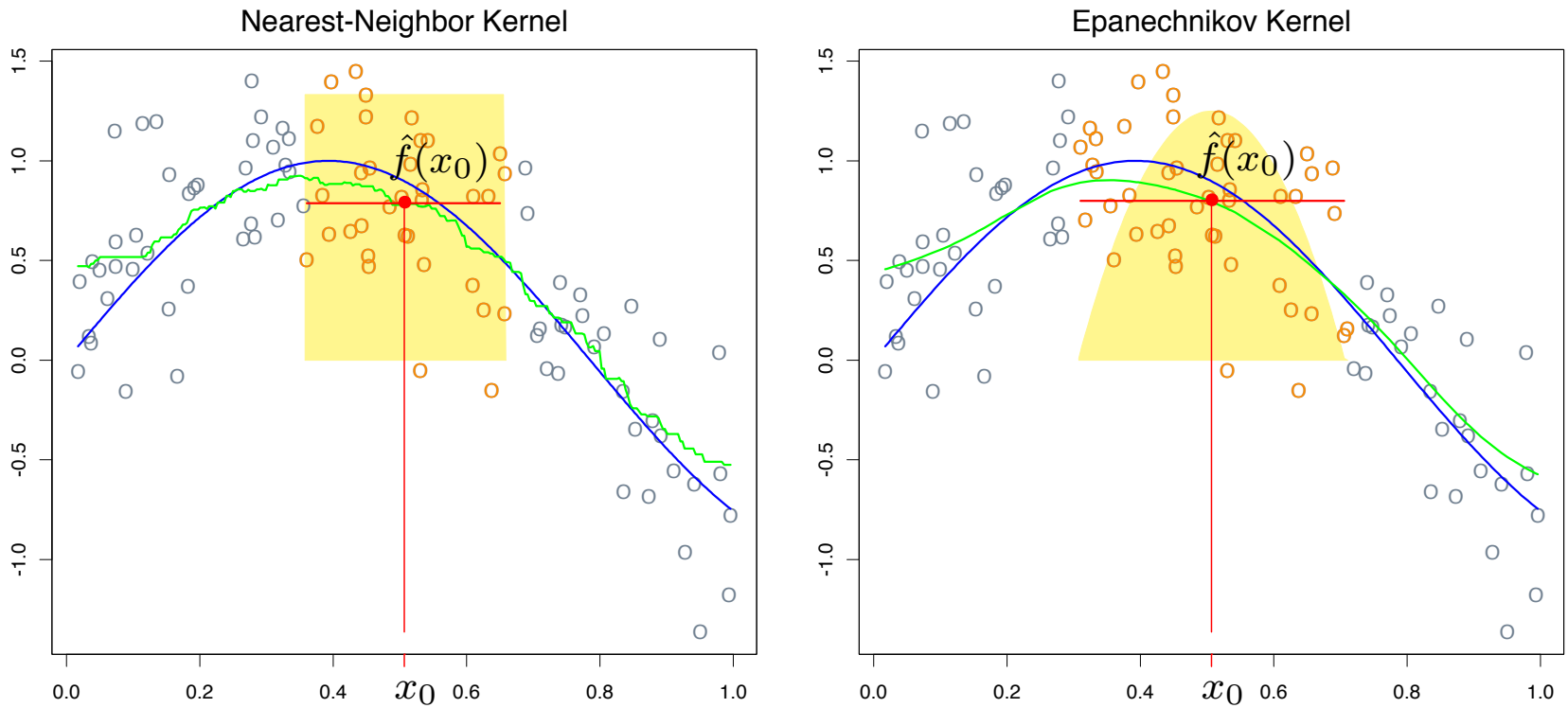
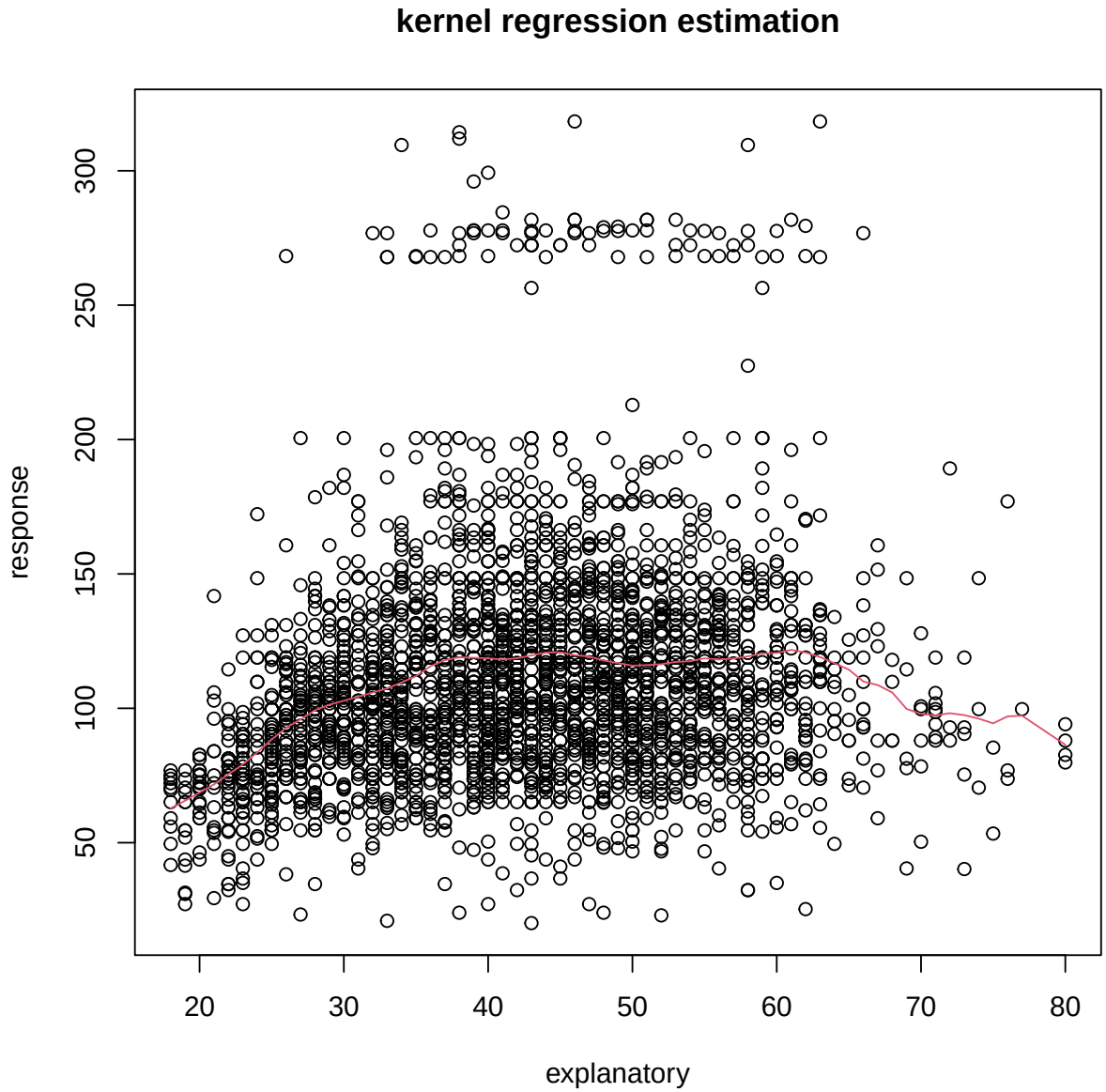


Figure 1: *Comparing k-nearest-neighbor and Epanechnikov kernels. From Chapter 6 of Hastie et al. (2009)*

Estimación por núcleos

```
library(snpar)
library(np)
bw <- npscoefbw(xdat=age, ydat=wage,
data=Wage,ckertype = "epanechnikov")
fit<- kre(age,wage, h=bw$bw, kernel = "epan", plot =
TRUE)
```

Estimación por núcleos



Polinomios locales

Para evitar los *problemas de frontera* que tienen los núcleos. Para eso en lugar de ajustar una constante en la regresión por núcleos se propone ajustar un polinomio. Asumimos por ejemplo, que $\widehat{f}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x$, que minimicen

$$y_i = \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)(y_i - \beta_0 - \beta_1 x_i)^2$$

esta es una regresión lineal local, se puede ver que se puede escribir como mínimos cuadrados ponderados.

Polinomios locales

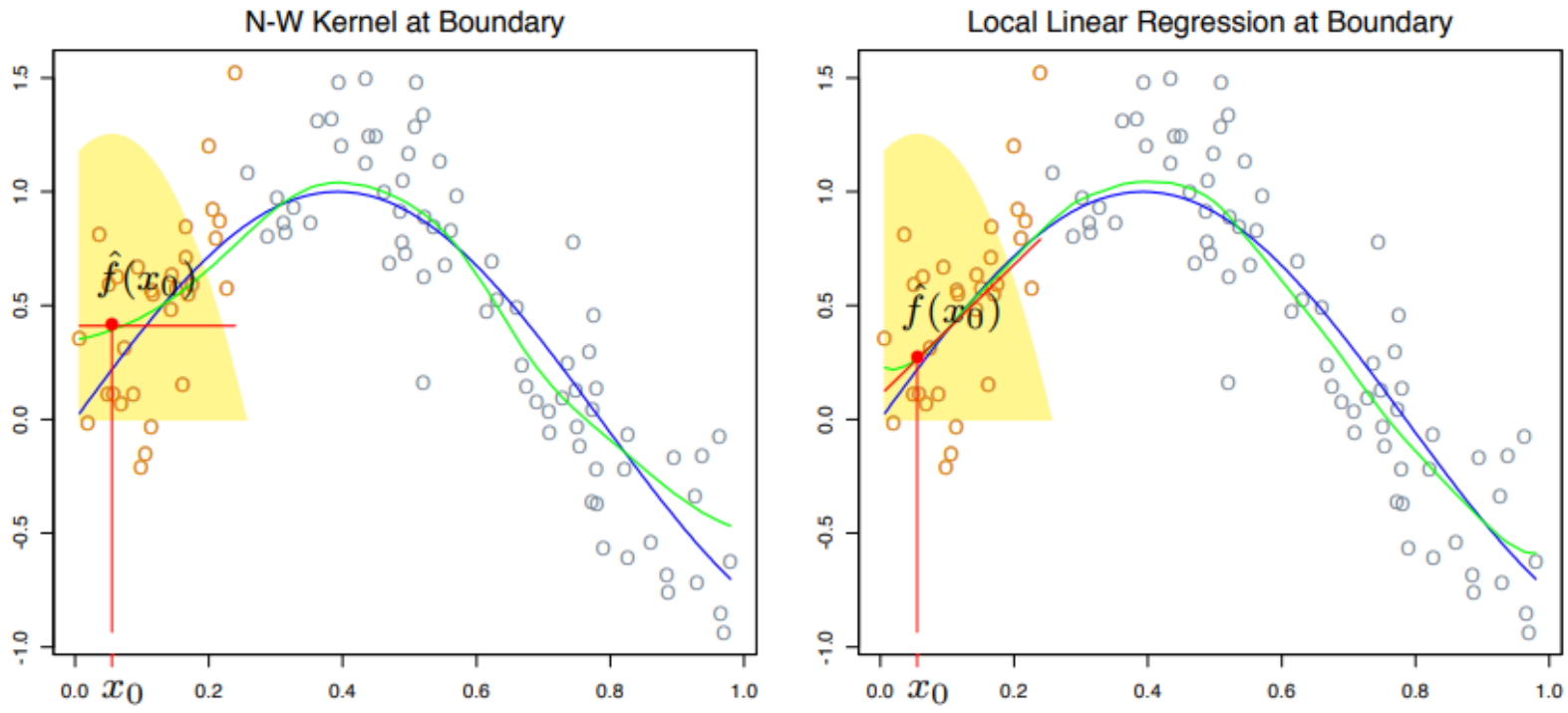


Figure 2: *Comparing (Nadaraya-Watson) kernel smoothing to local linear regression; the former is biased at the boundary, the latter is unbiased (to first order). From Chapter 6 of Hastie et al. (2009)*

Polinomios locales

Parámetros a elegir

- Kernel.
- h ancho de ventana juega el rol del parámetro de suavizado.
- grado del polinomio.

Polinomios locales

Se puede extender a dimensiones superiores pero nuevamente los procedimientos sufren la maldición de la dimensionalidad.
Utiliza mucha memoria.

Modelos aditivos generales

Ahora buscamos extender las ideas presentadas hasta acá a un contexto de regresión multivariada.
Es decir queremos predecir Y en función de $X_1, \ldots, X_p$.

- permitir la flexibilidad en todas las variables regresoras.
- Mantiene la aditividad.
- la variable de respuesta puede ser cuantitativa o cualitativa.

Modelos aditivos generales

Modelos de regresión

$$y_i = \beta_0 + \sum_{i=1}^{p} f_i(x_{ij}) + \epsilon_i.$$

Es un modelo aditivo porque estimamos por separado cada una de las funciones y luego las sumamos.
Las funciones f_i se pueden estimar utilizando cualquiera de las técnicas que describimos anteriormente.

Modelos aditivos generales

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

```
gam.m3=gam(wage~s(year,4)+s(age,5)+education,data=Wage)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
s(year, 4)	1	27162.12	27162.12	21.98	0.0000
s(age, 5)	1	195338.46	195338.46	158.08	0.0000
education	4	1069726.09	267431.52	216.42	0.0000
Residuals	2986	3689770.38	1235.69		

Modelos aditivos generales

```
plot(gam.m3, se=TRUE ,col ="blue")
```

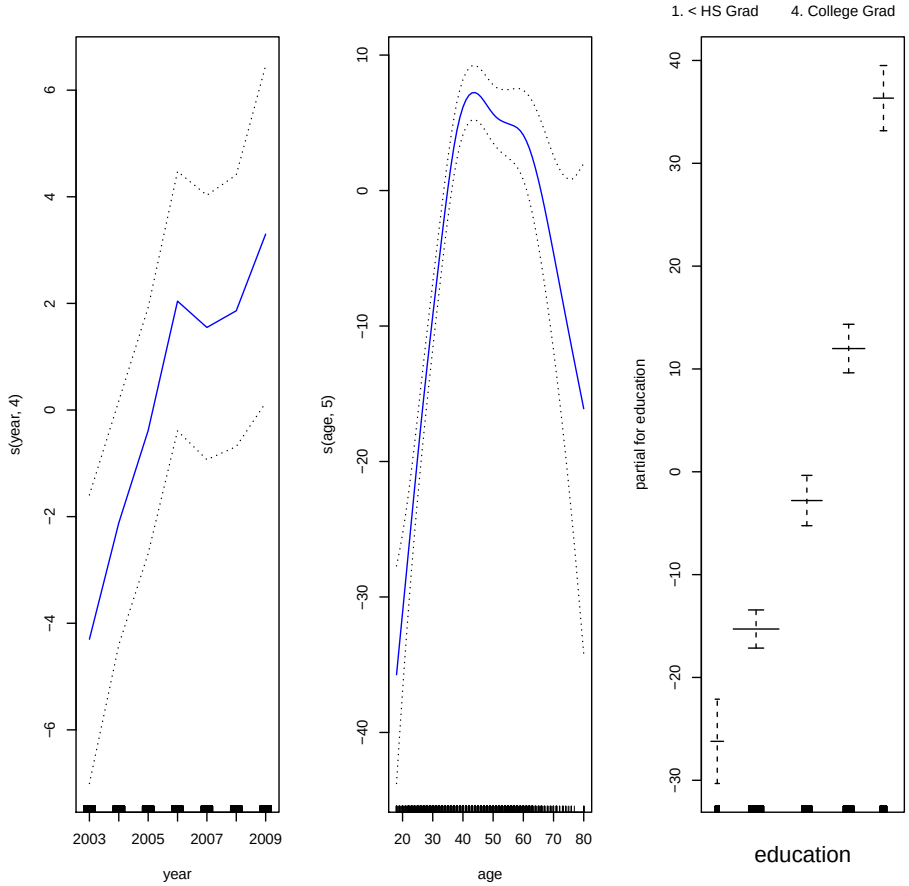


Figure: age y year ajustada con splines cúbicos y education con función escalera.

Modelos aditivos generales

Podemos ver que la variable `year` parece tener un ajuste bastante lineal, podemos probar los siguientes modelos:

- Sacar la variable year.
- Considerarla linealmente.
- Considerarla con un ajuste no paramétrico.

Modelos aditivos generales

```
gam.m1=gam(wage~s(age ,5)+education,data=Wage)
gam.m2=gam(wage~year+s(age,5)+education,data=Wage)
anova(gam.m1,gam.m2,gam.m3,test="F")
```

	Resid. Df	Resid. Dev	Df	Deviance	F	Pr(>F)
1	2990	3711730.76				
2	2989	3693841.51	1	17889.24	14.48	0.0001
3	2986	3689770.38	3	4071.13	1.10	0.3486

Es conveniente incorporar la variable `year`, sin embargo no hay beneficio en modelarla en forma noparámetrica.

Modelos aditivos generales

- Se puede modificar la forma de hacer el ajuste no paramétrico.
- Se pueden incorporar interacciones.

Modelos aditivos generales

Ventajas

- Se pueden modelar relaciones no lineales (marginales) en forma sencilla sin tener que intentar muchos modelos.
- Predicciones más precisas.
- El modelo aditivo permite estudiar efectos individuales dejando fijas las otras variables.
- La suavidad de las funciones f_j se resume mediante los grados de libertad.

Desventajas

- Se pierden las interacciones, aunque se podría modelar alguna $f(x_i x_j)$ que sea considerada importante.

Modelos aditivos generales

GAM para clasificación

Podemos aplicarlo en forma *plugin* en el modelo logístico,

$$\log\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) = \beta_0 + \sum_{i=1}^{p} f_i(x_{ij})$$

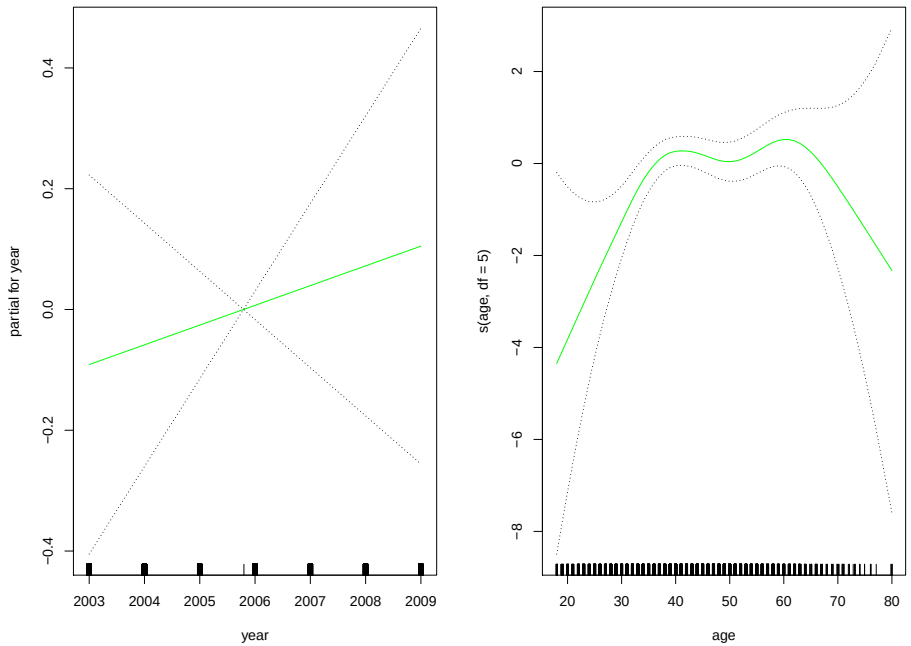
Modelos aditivos generales

```
gam.lr=gam(I(wage >250)~year+s(age,df=5),family
=binomial ,data=Wage)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
year	1	0.49	0.49	0.53	0.4682
s(age, df = 5)	1	4.02	4.02	4.28	0.0387
Residuals	2993	2813.65	0.94		

Pareciera no ser necesaria la variable `year`.

Modelos aditivos generales



Modelos aditivos generales

Ahora podemos pensar en incorporar la variable `aducation`

	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

Quitamos la primer categoria porque no hay registros de personas que ganen más de $250.000 en esa categoría.

Modelos aditivos generales

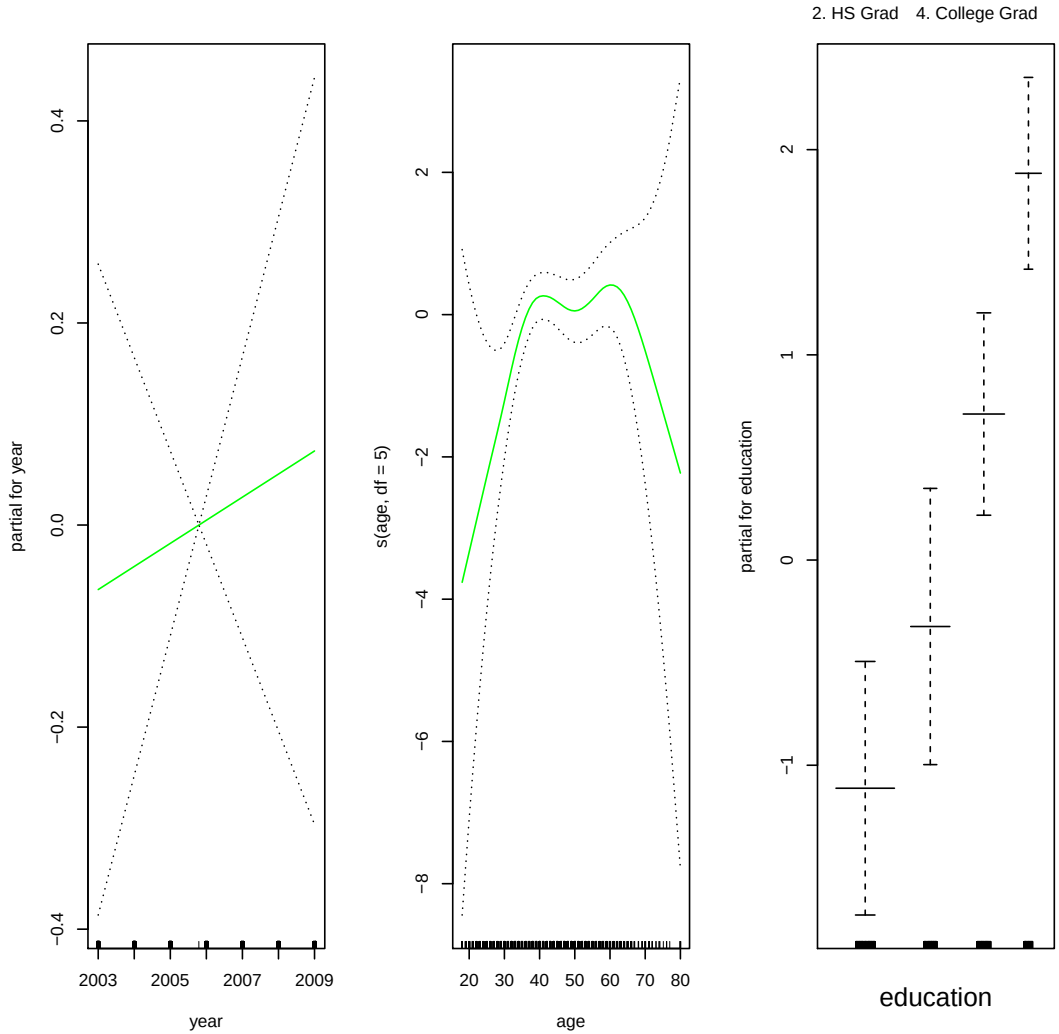
```
gam.lr.s=gam (I(wage >250) year+s(age
,df=5)+education ,family = binomial ,data=Wage
,subset =(education !="1.  < HS Grad"))
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
year	1	0.48	0.48	0.55	0.4600
s(age, df = 5)	1	3.83	3.83	4.31	0.0379
education	3	65.80	21.93	24.72	0.0000
Residuals	2722	2415.55	0.89		

Nuevamente pareciera no se necesaria la variable `year`.

Modelos aditivos generales

```
plot(gam.lr.s,se=T,col =" green ")
```




T. Hastie, R. Tibshirani, J. Friedman.
Elements of Statistical Learning, 2nd Ed.
Springer-Verlag, 2009.
Capítulo 6 .


J. Garret, D. Witten, T. Hastie, R. Tibshirani.
An Introduction to Statistical Learning.
Springer-Verlag, 2013.
Capítulo 7.