

# EST-25134: Aprendizaje Estadístico

**Profesor:** Alfredo Garbuno Iñigo — Primavera, 202— Modelos aditivos.

**Objetivo:** En esta sección estudiaremos modelos que empiezan desviarse de los supuesto lineales. Retendremos el componente aditivo por cuestiones de interpretabilidad. Estudiaremos métodos lineales por regiones, funciones polinomiales y la familia de modelos basados en *splines* con fines predictivos y como mecanismos de pre-procesamiento.

**Lectura recomendada:** Capítulo 7 de [2]. Finalmente, el libro [3] presenta un tratamiento completo sobre modelos aditivos generalizados.

## 1. INTRODUCCIÓN

Hasta ahora hemos utilizado modelos predictivos basados en **supuestos lineales**. Aunque el modelo lineal es un modelo sencillo y bastante útil en la práctica tiene sus **limitaciones**. Hemos explorado la idea de hacer más complejo el modelo lineal —utilizando más covariables— utilizando el concepto de **regularización** para mantener cierta garantía en la generalización del modelo.

Un modelo lineal es una aproximación que muchas veces usamos por conveniencia. El modelo es conveniente pues representa una aproximación interpretable a primer orden. Es decir, corresponde a una aproximación lineal de primer orden de Taylor.

En esta sección del curso estudiaremos modelos que rompen el supuesto de linealidad por medio de **regresión polinomial**, **funciones simples**, **splines**, **regresión local**, y **modelos aditivos**.

## 2. REGRESIÓN POLINOMIAL

Empezamos con un modelo de regresión

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad (1)$$

el cual *extendemos* a través de una expresión

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \dots + \beta_d x_i^d + \epsilon_i. \quad (2)$$

Creamos una colección nueva de **atributos**  $x_i \mapsto x_i, x_i^2, x_i^3, \dots, x_i^d$  y ajustamos los coeficientes utilizando herramientas de regresión lineal múltiple.

### 2.1. Ejemplo:

Datos demográficos y de ingreso para individuos que viven en región central del Atlántico en Estados Unidos.

	year	age	wage	education	hi.income
1	2003	53	82	4. College Grad	0
2	2008	50	83	4. College Grad	0
3	2006	35	155	4. College Grad	0
4	2004	37	148	4. College Grad	0
5	2007	29	88	4. College Grad	0

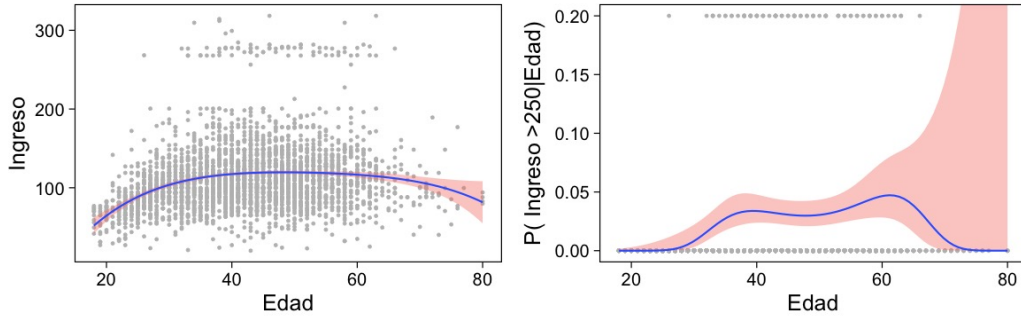


FIGURA 1. Predicción con polinomio de grado 4 con bandas predictivas.

## 2.2. Incertidumbre en predicciones (regresión)

Nuestro interés es en la capacidad predictiva del modelo (no en los coeficientes). Nos interesa evaluar

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

donde notamos la respuesta es una combinación lineal de los coeficientes. Esto nos ayuda a calcular

$$\mathbb{V}(\hat{f}(x_0)). \quad (4)$$

En el panel de regresión (Fig. 1) calculamos intervalos de la forma

$$\hat{f}(x_0) \pm 2 \cdot \text{SE}(\hat{f}(x_0)), \quad (5)$$

donde  $\text{SE}(\hat{\theta})$  denota el error estándar del estimador  $\hat{\theta}$ .

**2.2.1. Detalles** Para calcular la varianza de la predicción utilizamos la combinación lineal de los predictores

$$\mathbb{V}(\hat{f}(x_0)) = \mathbb{V}(x_0^\top \hat{\beta}) = x_0^\top \mathbb{V}(\hat{\beta}) x_0. \quad (6)$$

## 2.3. Incertidumbre en predicciones (clasificación)

Para el modelo logístico (Fig. 1) lo que tenemos es

$$\mathbb{P}(\text{Ingreso} > 250 | x_i) = \frac{\exp(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \cdots + \beta_d x_i^d)}{1 + \exp(\beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \cdots + \beta_d x_i^d)}. \quad (7)$$

Los intervalos de confianza se calculan en dos pasos: 1) calculando las cotas en **escala logit** y 2) convirtiendo las cotas en **escala probabilística**.

## 2.4. Observaciones

- Regresión polinomial requiere una elección para  $d$ , la cual se puede establecer por medio de validación cruzada.
- Los polinomios, aunque son mas flexibles y ajustan relaciones no lineales, son muy malos para extrapolar.
- Los modelos se ajustan con expresiones en el componente de **fórmula**

```
1 y ~ poly(x, degree = 4)
```

### 3. FUNCIONES SIMPLES

Otra forma de crear transformaciones de variables es considerar distintas regiones y ajustar las medias. Por ejemplo, considerar **grupos de edad**:  $[18, 35)$ ,  $[35, 50)$ ,  $[50, 65)$ ,  $[65, 100)$ .

Para lograr esto, generamos una nueva colección de predictores por medio de variables categóricas o *dummy*

$$C_1(X) = I(X < 35), \quad C_2(X) = I(35 \leq X < 50), \quad \dots, \quad C_4(X \geq 65). \quad (8)$$

con los que ajustaremos el modelo de regresión (clasificación) lineal

$$f(x) = \beta_0 + \beta_1 C_1(x) + \dots + \beta_4 C_4(x). \quad (9)$$

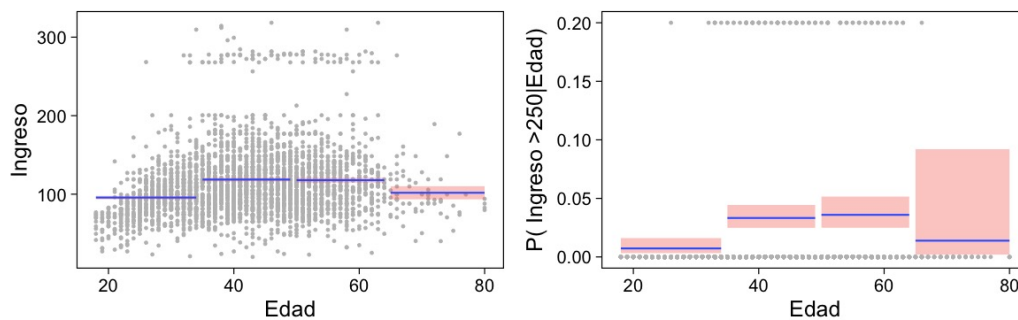


FIGURA 2. Predicción con regresión local.

#### 3.1. Observaciones

- Los parámetros se ajustan de manera *local*. Contrario con los polinomios que ajustan parámetros para todo el rango de los datos.
- Los modelos se ajustan en cada grupo de edad, donde `age.group` es una variable categórica que tiene las indicadoras de los grupos.

```
1 y ~ age.group
```

**Nota:** hay que tener cuidado al querer interpretar, pues en automático se crea un grupo *base* pues no queremos tener problemas de multicolinealidad.

Para graficar (`ggplot2`) basta con pedir la predicción constante con los gráficos agrupados por grupo de edad. Esto se logra con

```
1 ggplot(data, aes(age, wage, group = age.group)) +
2 geom_smoth(formula = y ~ 1)
```

#### 3.2. Extensiones

Una noción natural de incrementar la complejidad del modelo y al mismo tiempo mejorar la capacidad predictiva de éste sería ajustar una recta en cada región, ver Fig. 3.

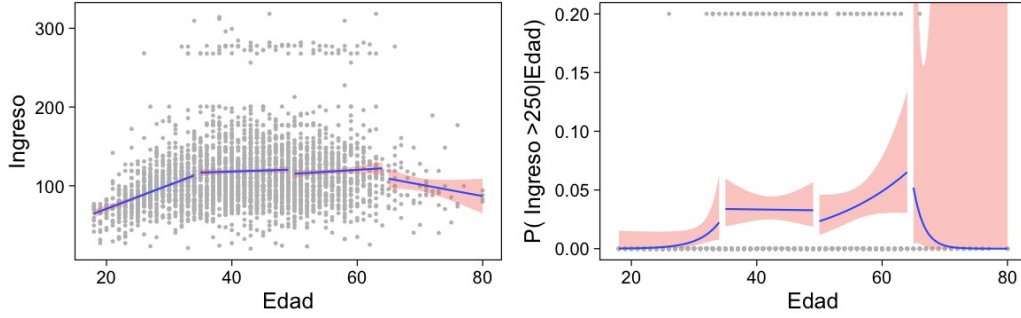


FIGURA 3. Predicción con regresión lineal local.

#### 4. MODELOS POR SEGMENTOS

Uno de los problemas del modelo anterior es que definimos la regresión con **modelos discontinuos**. Por ejemplo, si ajustamos un modelo donde las regiones utilicen distintos polinomios. No tendremos garantía de que el modelo completo se vea conectado. Además, si lo hacemos sin cuidado entonces tendremos modelos volátiles en las cotas de las regiones de ajuste, ver Fig. 4.

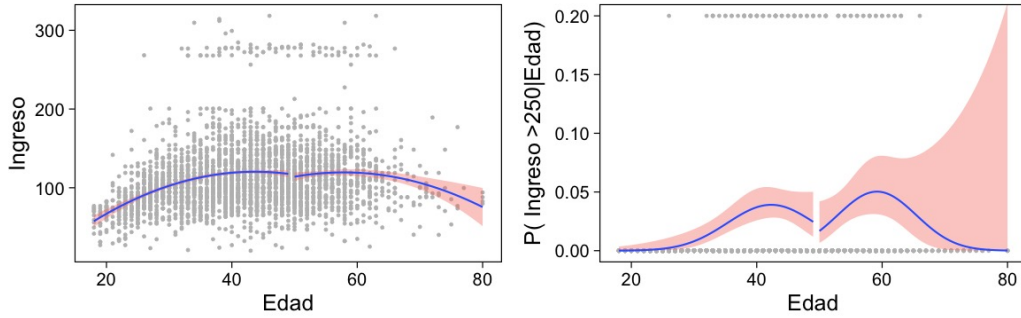


FIGURA 4. Predicción con regresión polinomial de grado 2 en localidades.

##### 4.1. Splines

Un modelo basado en *splines* es un modelo basado en polinomios donde se les añade la propiedad de **continuidad** (en las primeras dos derivadas), ver Fig. 5.

##### 4.2. Splines lineales

Los *splines* de grado 1 son funciones lineales continuas por segmentos. Se construyen a través de **funciones base**

$$b_1(x) = x \quad (10)$$

$$b_{k+1}(x) = (x - \xi_k)_+, \quad k = 1, \dots, K, \quad (11)$$

y una colección de **nodos**  $\xi_k$ , donde  $(\cdot)_+$  denota la **parte positiva** de la función.

De tal manera que el modelo predictivo queda en términos de

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \dots + \beta_{K+1} b_{K+1}(x_i) + \epsilon_i. \quad (12)$$

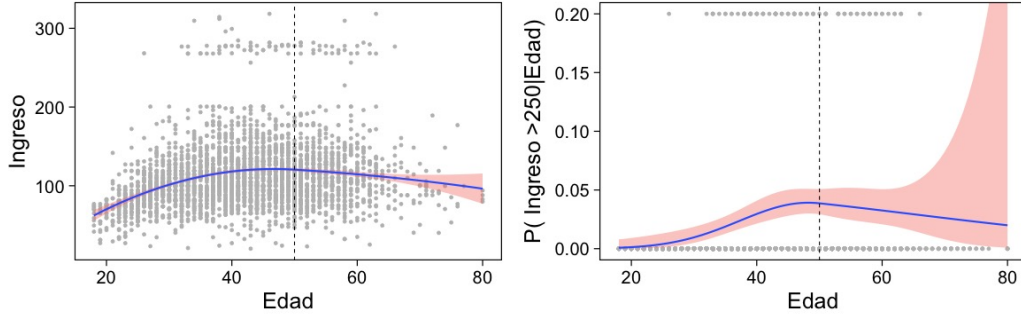


FIGURA 5. Predicción con regresión splines de grado 2. La línea punteada marca el punto donde se conectan los dos polinomios.

### 4.3. Splines cúbicos

Los *splines* de grado 3 son funciones continuas por segmentos. Se construyen a través de funciones base

$$b_1(x) = x, \quad (13a)$$

$$b_2(x) = x^2, \quad (13b)$$

$$b_3(x) = x^3, \quad (13c)$$

$$b_{k+3}(x) = (x - \xi_k)_+^3, \quad k = 1, \dots, K, \quad (13d)$$

y una colección de **nodos**  $\xi_k$ , donde  $(\cdot)_+^3$  denota la **parte positiva** de la función.

**Nota** que en cada nodo la función construida tiene a lo más 2 derivadas continuas.

De tal manera que el modelo predictivo queda en términos de

$$y_i = \beta_0 + \beta_1 b_1(x_i) + \dots + \beta_{K+3} b_{K+3}(x_i) + \epsilon_i. \quad (14)$$

### 4.4. Splines naturales

Un *spline natural* es un *spline* con la restricción adicional de considerar una **extrapolación lineal** fuera de los **nodos frontera**. Ver Fig. 6.

4.4.1. *Para pensar* Para el caso de regresión  $f : \mathbb{R} \rightarrow \mathbb{R}$ , incorporar un *spline natural* agrega  $4 = 2 \times 2$  restricciones adicionales, ¿por qué?

### 4.5. Selección de nodos

- Una estrategia es elegir el número de nodos  $K$  y después utilizar los percentiles correspondientes de  $X$ .
- Un *spline* cúbico con  $K$  nodos tiene  $K + 4$  parámetros (o grados de libertad).
- Un *spline natural* con  $K$  nodos tiene  $K$  parámetros (o grados de libertad).

Utilizando la noción de grados de libertad podemos comparar un polinomio con un grado predeterminado y un spline natural con un número de nodos fijo.

## 5. SUAVIZAMIENTO POR SPLINES

Consideremos el problema de **ajustar un función continua y diferenciable**  $g(\cdot)$  a un **conjunto de datos**. Lo cual logramos por medio de

$$\min_{g \in \mathcal{S}} \sum_{i=1}^n (y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt. \quad (15)$$

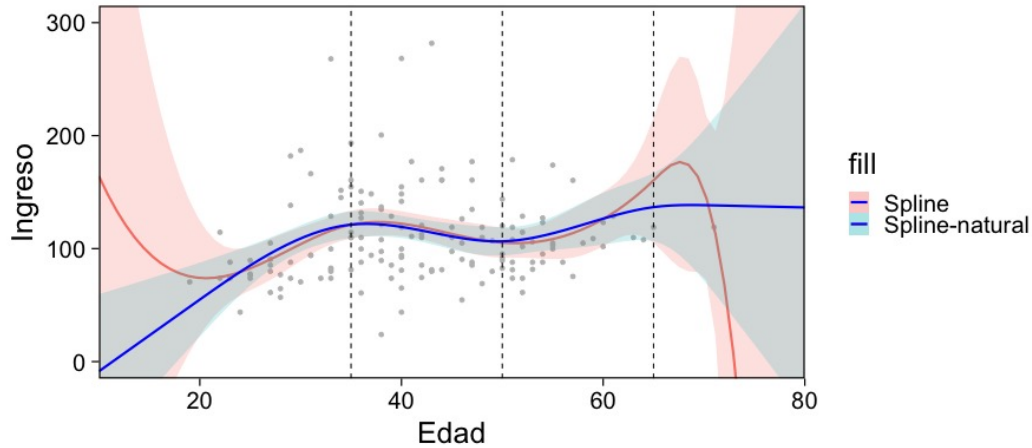


FIGURA 6. Predicción con regresión utilizando splines de grado 3. Las líneas punteadas representan los nodos ( $\xi_k$ ) del modelo.

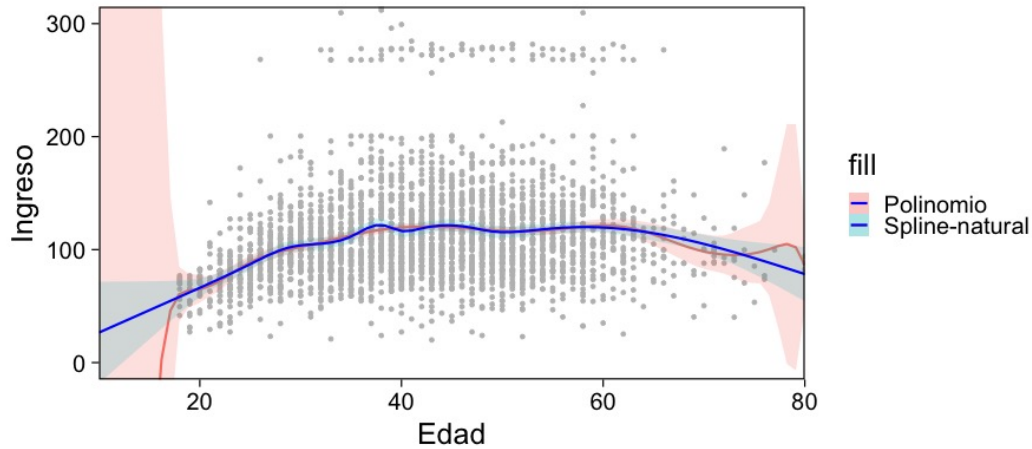


FIGURA 7. Ajuste con modelos con 15 grados de libertad. Polinomio de potencia 14, y spline natural (cúbico).

- ¿Qué rol juega  $\lambda$ ?

### 5.1. Solución

- La **solución** es un *spline* natural con polinomios cúbicos. Los nodos se localizan en cada uno de los datos de entrenamiento  $x_i$ . La suavidad del estimador es controlada por medio de  $\lambda$ .

El término de regularización afecta directamente en los coeficientes. Logrando así, eliminar la aparente complejidad de considerar tantos nodos como observaciones tengamos.

- El vector de  $n$  predicciones se puede escribir como

$$\hat{g}_\lambda = S_\lambda y. \quad (16)$$

La solución al problema de optimización se logra por medio de  $f(x) = \sum_{j=1}^N \beta_j N_j(x)$  donde  $N_j(\cdot)$  denota la base de funciones para el espacio de *splines* naturales. De esta manera la función objetivo se puede reescribir como

$$\mathcal{J}(\beta) = (y - N\beta)^\top (y - N\beta) + \lambda \beta^\top \Omega_N \beta, \quad (17)$$

donde  $\{\Omega_N\}_{jk} = \int N_j''(t) N_k''(t) dt$ , cuya solución se puede escribir de manera analítica.

- El número efectivo de grados de libertad se puede calcular a través de

$$\text{df}_\lambda = \sum_{i=1}^n \{S_\lambda\}_{ii}. \quad (18)$$

Los grados de libertad efectivos en el contexto de *splines* fueron definidos ([1]) en analogía con que  $M = \text{tr}(H) = \text{tr}(X(X^\top X)^{-1}X^\top)$  nos da la dimensión del espacio en donde se proyectan las predicciones de mínimos cuadrados para el modelo lineal.

5.1.1. Bonus: El error de validación cruzada se puede calcular por medio de

$$\text{RSS}_{\text{CV}}(\lambda) = \sum_{i=1}^n (y_i - \hat{g}_\lambda^{(-i)}(x_i))^2 = \sum_{i=1}^n \left[ \frac{y_i - \hat{g}_\lambda(x_i)}{1 - \{S_\lambda\}_{ii}} \right]^2. \quad (19)$$

## 5.2. Ajuste de suavizador

Para ajustar el suavizador podemos controlar por los grados de libertad ( $\text{df}_\lambda$ ) en lugar de utilizar el coeficiente de penalización de curvatura. Esto es por que existe una **relación inversa** entre  $\lambda$  y  $\text{df}_\lambda$ .

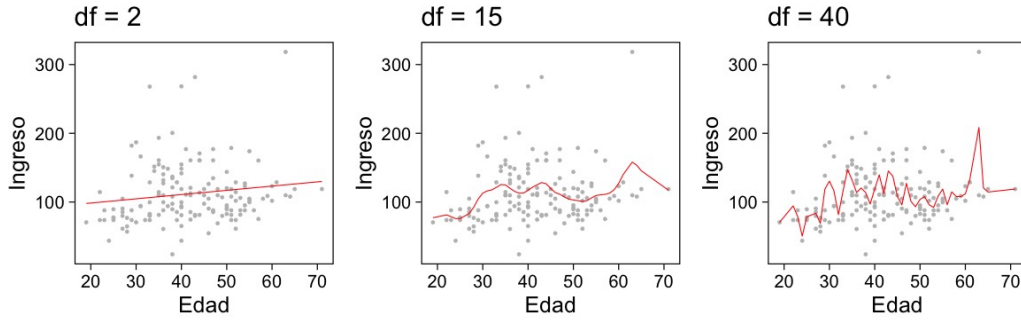
Establecer la relación entre los grados de libertad y el coeficiente de penalización escapa los intereses del curso. Se puede encontrar en el Capítulo 5 de [1]. La idea general la esbozamos a continuación. Utilizamos  $\text{df}_\lambda = \text{tr}(S_\lambda)$  donde además sabemos que la traza de una matriz es la suma de sus eigenvalores. Utilizaremos analogía con regresión y regularización  $L_2$ . La solución de mínimos cuadrados bajo regresión Ridge es  $\hat{\beta} = (X^\top X + \lambda I)^{-1} X^\top y$ . Las predicciones las realizamos por medio de  $X\hat{\beta}$ , de donde podemos escribir el vector de predicciones  $\tilde{y} = H_\lambda y$ . Utilizando la descomposición espectral de  $X \in \mathbb{R}^{n \times p}$  podemos encontrar los eigenvalores de  $H_\lambda$  y en consecuencia veremos una relación entre los grados de libertad ( $\text{df}_\lambda$ ) y el coeficiente de penalización ( $\lambda$ ).

## 6. REGRESIÓN LOCAL

El objetivo es: ajustar un **modelo por regiones** donde tengamos una función de peso que sólo considere una vecindad. El ajuste se realiza por medio de mínimos cuadrados **ponderados**. Los pesos alrededor de un punto base  $x_0$  usualmente están definidos por **funciones kernel**  $K_\lambda(x, x_0)$  donde  $\lambda$  determina el tamaño de la vecindad.

El problema de regresión que se resuelve es un problema puntual en cada  $x_0$  que queramos evaluar

$$\min_{\alpha(x_0), \beta(x_0)} \sum_{i=1}^N K_\lambda(x_0, x_i) [y_i - \alpha(x_0) - \beta(x_0)x_i]^2. \quad (20)$$

FIGURA 8. Suavizamiento por splines. Controlamos por grados de libertad ( $df_\lambda$ ).

La solución a este problema está dada por mínimos cuadrados **ponderados** donde en específico tenemos la expresión

$$\hat{f}(x_0) = b(x_0)^\top (B^\top W(x_0)B)^{-1} B^\top W(x_0)y, \quad (21a)$$

$$= \sum_{i=1}^N c_i(x_0)y_i. \quad (21b)$$

El planteamiento y la solución de este problema de regresión se conoce como el suavizador de Watson-Nadaraya suavizado [1]. Aunque parezca un problema con una vasta historia aún sigue siendo motivo de estudio. De hecho su aplicación se encuentra en el centro de los modelos del estado del arte en Procesamiento de Lenguaje Natural (NLP, por sus siglas en inglés) como BERT, GPT-3, etc. [4].

De modo que se ajusta un modelo lineal de manera local (Fig. 9). La vecindad está controlada por un parámetro **span** que dicta el porcentaje de puntos de entrenamiento alrededor de donde queremos hacer predicciones.

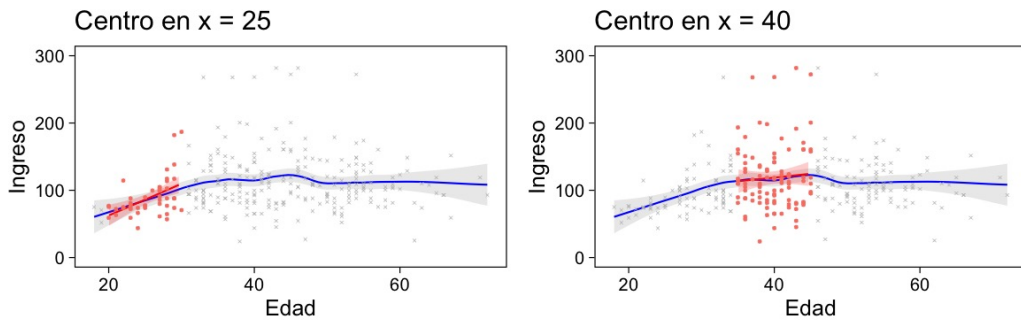


FIGURA 9. Regresión local con ventana móvil.

Cambiar el radio de las vecindades afecta la complejidad del modelo que usaremos para predecir (Fig. 10).

### 6.1. Observaciones

- En la práctica un suavizador por splines (`smooth.spline`) o un modelo de regresión local (`loess`) tienen un comportamiento similar.



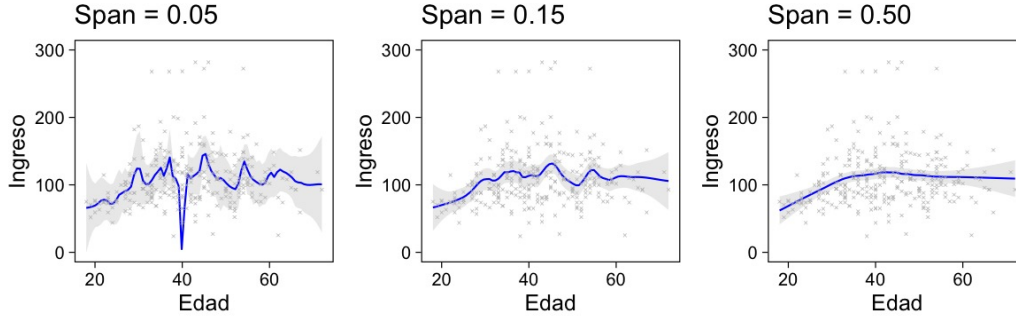


FIGURA 10. Regresión local con amplitud variable.

## 7. MODELO ADITIVOS GENERALIZADOS

La estructura aditiva se mantiene y nos permite incorporar una estructura predictiva en cada componente

$$y_i = \beta_0 + \beta_1 f_1(x_{i1}) + \cdots + \beta_p f_p(x_{ip}) + \epsilon_i. \quad (22)$$

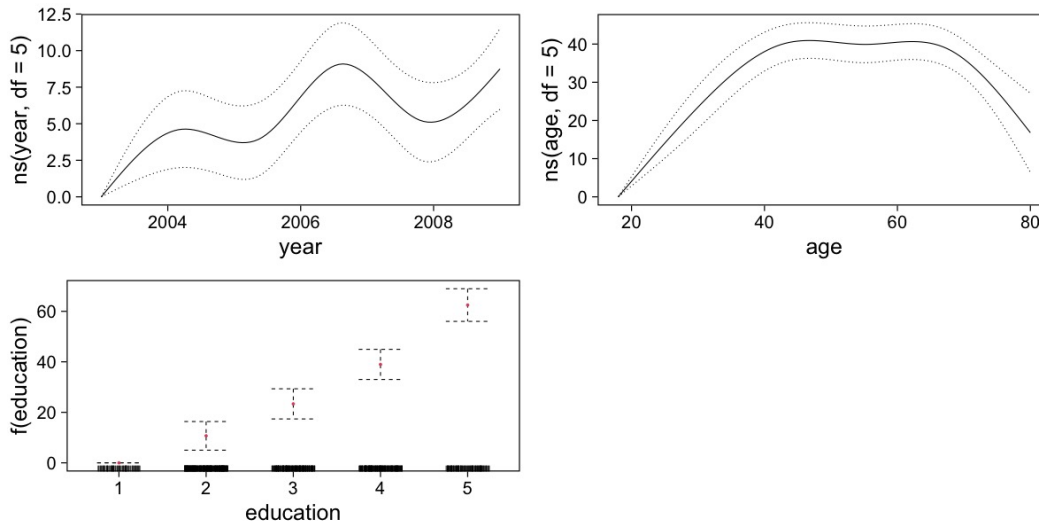


FIGURA 11. Regresión con modelo aditivo con tres componentes.

### 7.1. Clasificación

La linealidad se mantiene y se pueden explorar las contribuciones de cada término en escala logit:

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 f_1(x_{i1}) + \cdots + \beta_p f_p(x_{ip}). \quad (23)$$

## 8. CONCLUSIONES

Los modelos estudiados en esta sección son modelos que pertenecen a la familia de **modelos aditivos**. Son relativamente sencillos de explorar y criticar (al visualizar o estudiar cada uno de los términos). La sección siguiente del curso estudiaremos modelos que utilizan una estructura muy particular para los términos  $f_i(x)$ .

## REFERENCIAS

- [1] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York, New York, NY, 2009. ISBN 978-0-387-84857-0 978-0-387-84858-7. . [7](#), [8](#)
- [2] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: With Applications in R*. Springer Texts in Statistics. Springer US, New York, NY, 2021. ISBN 978-1-07-161417-4 978-1-07-161418-1. . [1](#)
- [3] S. N. Wood. *Generalized Additive Models*. CRC Press, 2017. [1](#)
- [4] A. Zhang, Z. C. Lipton, M. Li, and A. J. Smola. *Dive into Deep Learning*. In print, 2021. [8](#)