EST-25134: Aprendizaje Estadístico

Profesor: Alfredo Garbuno Iñigo — Primavera, 2022.
Objetivo: Splines. Troncos. Modelos aditivos.
Lectura recomendada: Capítulo 7 de [1]. [2].

1. INTRODUCCIÓN

Hasta ahora hemos utilizado modelos predictivos basados en `supuestos lineales`. Aunque el modelo lineal es un modelo sencillo. Hasta ahora hemos visto como `reducir la complejidad` del modelo utilizando el concepto de regularización, pues ayuda a reducirla controlando la `varianza` de los estimadores.

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

2. REGRESIÓN POLINOMIAL

Empezamos con un modelo de regresión

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \,, \tag{1}$$

el cual *extendemos* a través de una expresión

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \cdots + \beta_d x_i^d + \epsilon_i \,. \tag{2}$$

Creamos una colección nueva de `atributos` $x_i \mapsto x_i, x_i^2, x_i^3, \ldots x_i^d$ 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
```

2.2. Incertidumbre en predicciones (regresión)

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

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

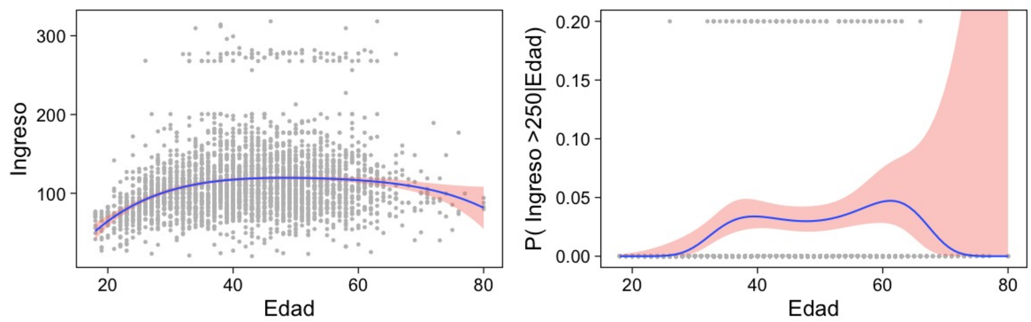

Figura 1. *Predicción con polinomio de grado 4 con bandas predictivas.*

Y dado que la respuesta es una combinación lineal de los coeficientes podemos evaluar

$$\mathbb{V}(\hat{f}(x_o))\,. \tag{4}$$

En la figura anterior (regresión) calculamos intervalos de la forma

$$\hat{f}(x_0) \pm 2 \cdot \mathsf{SE}(\hat{f}(x_0))\,, \tag{5}$$

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

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

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

2.3. Incertidumbre en predicciones (clasificación)

Para el modelo logístico lo que tenemos es

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

Los intervalos de confianza se calculan: 1) calculando las cotas en `escala logit` y 2) convertir 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`

```
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*

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

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

$$f(x) = \beta_0 + \beta_1 C_1(x) + \cdots + \beta_4 C_4(x)\,. \tag{9}$$

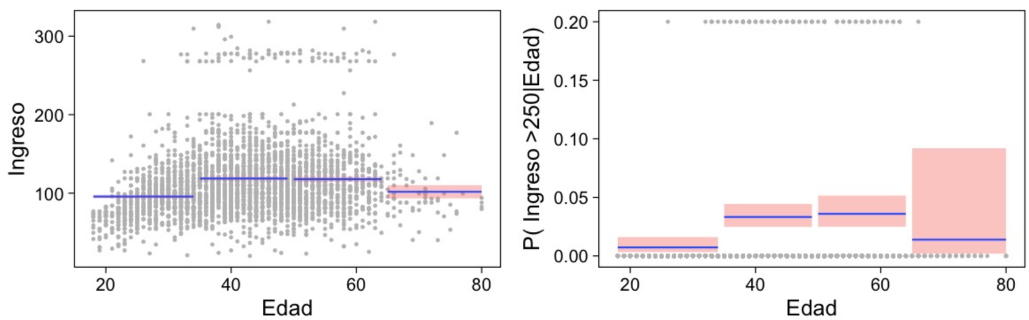

Figura 2. *Predicción con regresión local.*

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.

```
y ~ age.group
```

Nota: hay que tener cuidado, 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

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

4. MODELOS POR SEGMENTOS

Uno de los problemas del modelo anterior es que definimos la regresión con modelos discontinuos. Podemos ajustar un modelo donde las regiones utilicen distintos polinomios. Sólo que si lo hacemos sin cuidado entonces tendremos modelos volátiles en las cotas de las regiones de ajuste, ver Fig. 3.

4.1. Splines

Un modelo basado en *splines* es un modelo basado en polinomios donde se les añade la restricción de continuidad (en las primeras dos derivadas), ver Fig. 4.

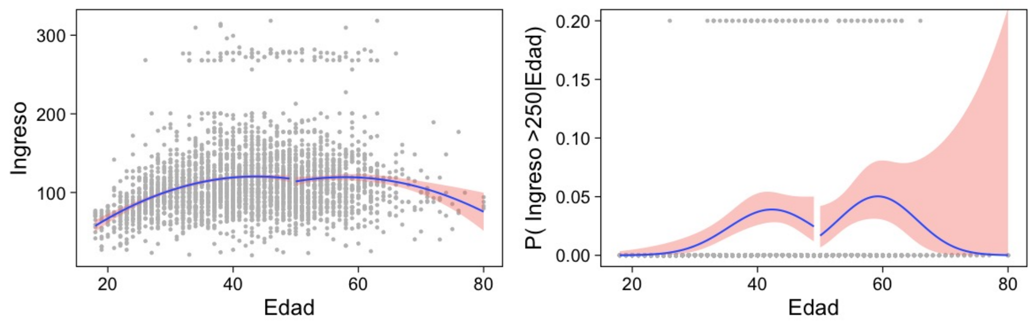

FIGURA 3. *Predicción con regresión polinomial de grado 2 en localidades.*

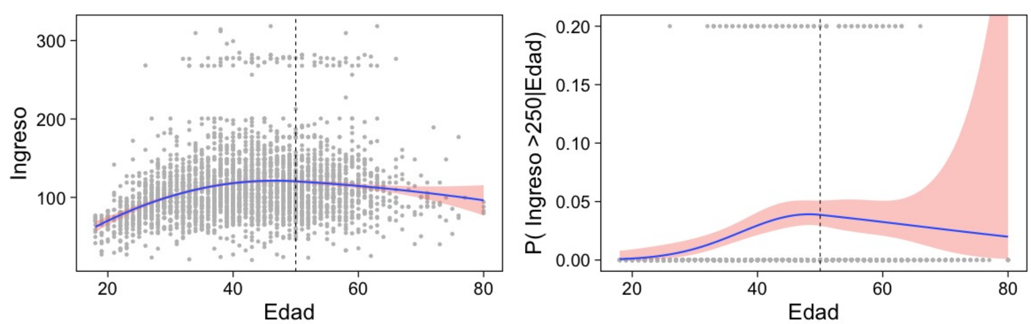

FIGURA 4. *Predicción con regresión* splines *de grado 2. La línea punteada marca el punto donde se* conectan *los dos polinomios.*

4.2. Splines lineales

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

$$b_1(x) = x \tag{10}$$
$$b_{k+1}(x) = (x - \xi_k)_+, \qquad k = 1, \ldots, K\,, \tag{11}$$

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

De tal manera que el modelo predictivo queda en términos de

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

4.3. Splines cúbicos

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

$$b_1(x) = x\,, \tag{13a}$$
$$b_2(x) = x^2\,, \tag{13b}$$
$$b_3(x) = x^3\,, \tag{13c}$$
$$b_{k+3}(x) = (x - \xi_k)^3_+, \qquad k = 1, \ldots, K\,, \tag{13d}$$

y una colección de `nodos` ξ_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

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

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. 5.

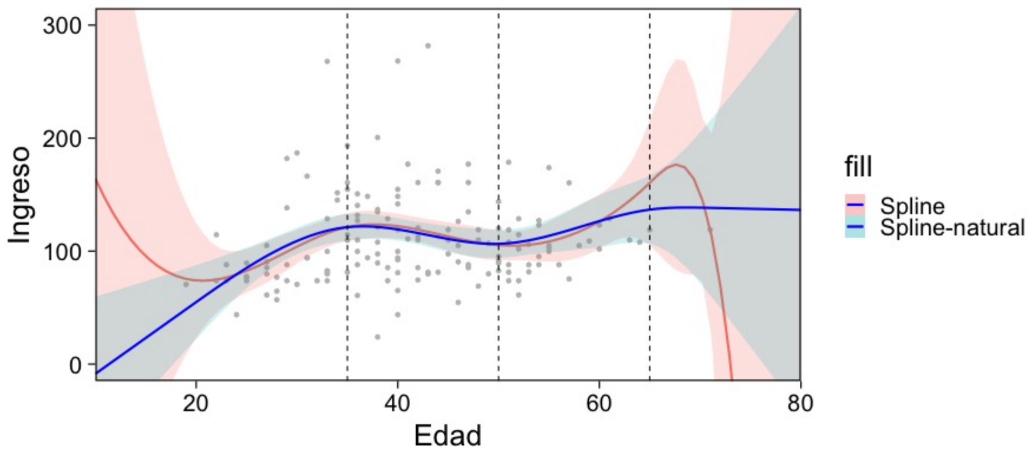

FIGURA 5. *Predicción con regresión utilizando* splines *de grado 3. Las líneas punteadas representan los nodos* (ξ_k) *del modelo.*

4.4.1. Para pensar Para el caso de regresión $f : \mathbb{R} \to \mathbb{R}$, 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 cuantiles de X.
- Un *spline* cúbico con K nodos tiene $K + 4$ parámetros.
- Un *spline* natural con K nodos tiene K parámetros.

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

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

- ¿Qué rol juega λ?

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 λ.
- El vector de n predicciones se puede escribir como

$$\hat{g}_\lambda = S_\lambda y\,. \tag{16}$$

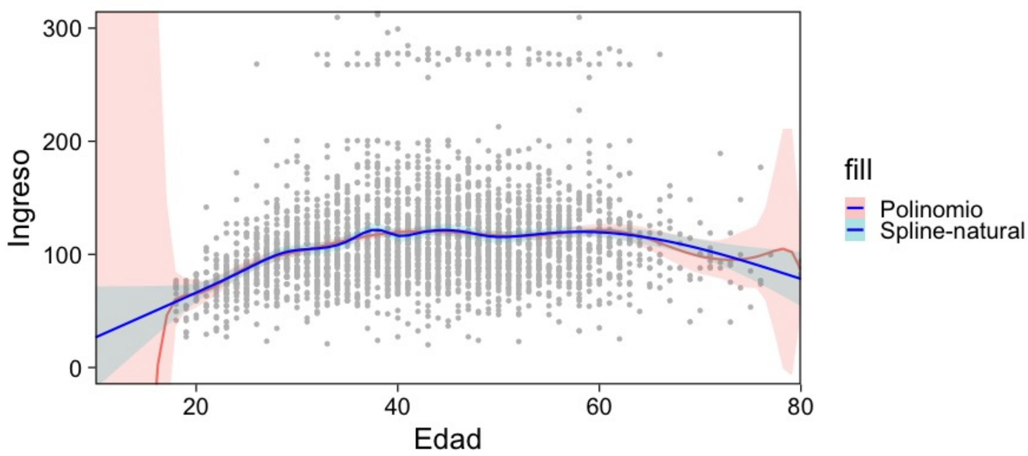

FIGURA 6. *Ajuste con modelos con 15 grados de libertad. Polinomio de potencia 14, y* spline *natural (cúbico).*

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

$$\mathsf{df}_\lambda = \sum_{i=1}^{n} \{S_\lambda\}_{ii}\,. \tag{17}$$

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

$$\mathsf{RSS}_{\mathsf{CV}}(\lambda) = \sum_{i=1}^{n} (y_i - \hat{g}_\lambda^{(-i)}(x_i))^2 = \sum_{i=1}^{n} \left[\frac{y_i - g_\lambda(x_i)}{1 - \{S_\lambda\}_{ii}}\right]^2 . \tag{18}$$

5.2. Ajuste de suavizador

Para ajustar el suavizador podemos controlar por los grados de libertad (df) en lugar de utilizar el coeficiente de penalización de curvatura.

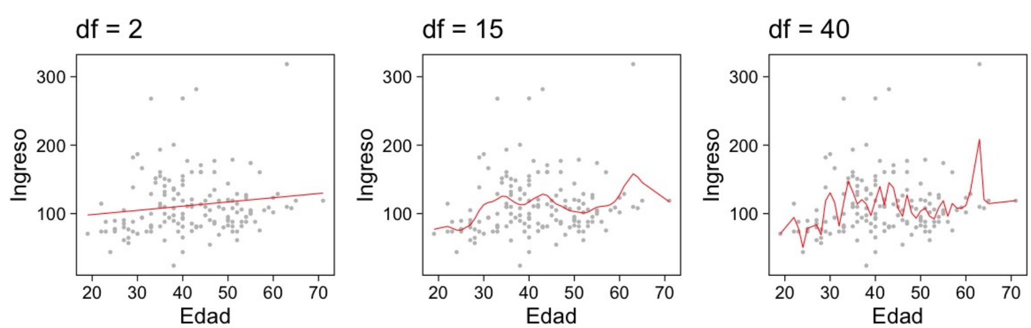

FIGURA 7. *Suavizamiento por* splines. *Controlamos por grados de libertad* (df).

6. REGRESIÓN LOCAL

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`.

6.1. Observaciones

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

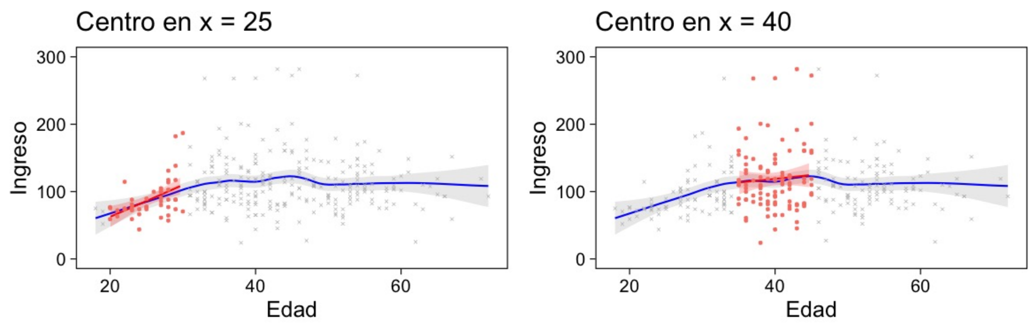

FIGURA 8. *Regresión local con ventana móvil.*

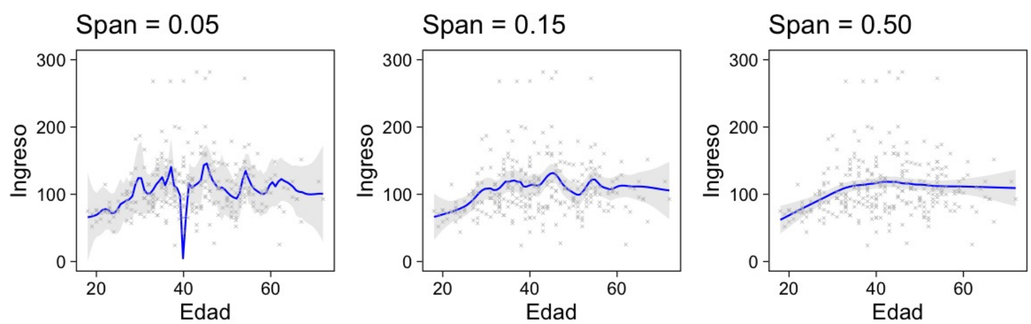

FIGURA 9. *Regresión local con amplitud variable.*

7. MODELO ADITIVOS GENERALIZADOS

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

$$y_i = \beta_0 + \beta_1 f_1(x_{i1}) + \cdots + \beta_p f_p(x_{ip}) + \epsilon_i \,. \tag{19}$$

7.1. Clasificación

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

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 f_1(x_{i1}) + \cdots + \beta_p f_p(x_{ip}) \,. \tag{20}$$

REFERENCIAS

[1] 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

[2] S. N. Wood. *Generalized Additive Models.* CRC Press, 2017. 1

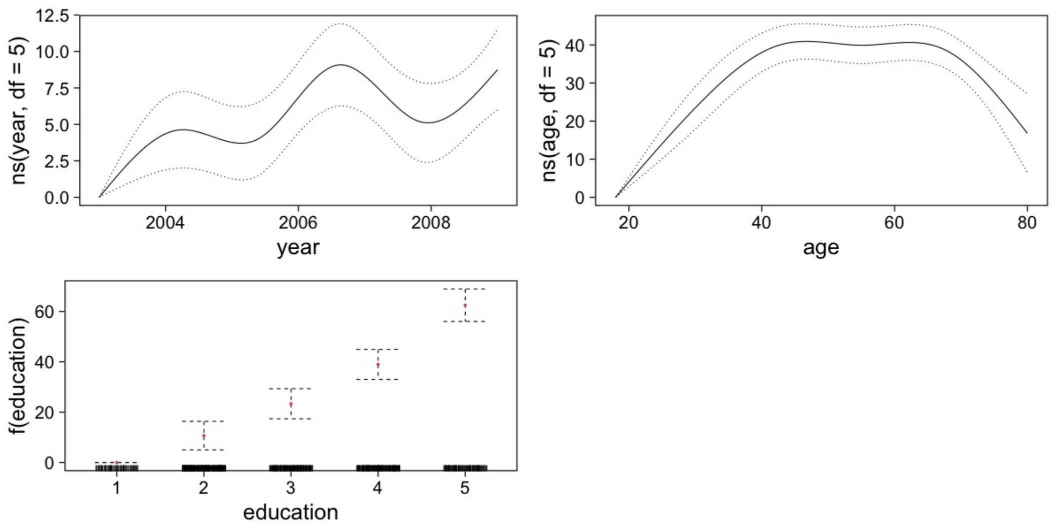

FIGURA 10. *Regresión con modelo aditivo con tres componentes.*