De los modelos lineales a los aditivos

Selección automática de modelos lineales y aditivos

Gabriel Martos Venturini
gmartos@utdt.edu

Universidad Torcuato Di Tella

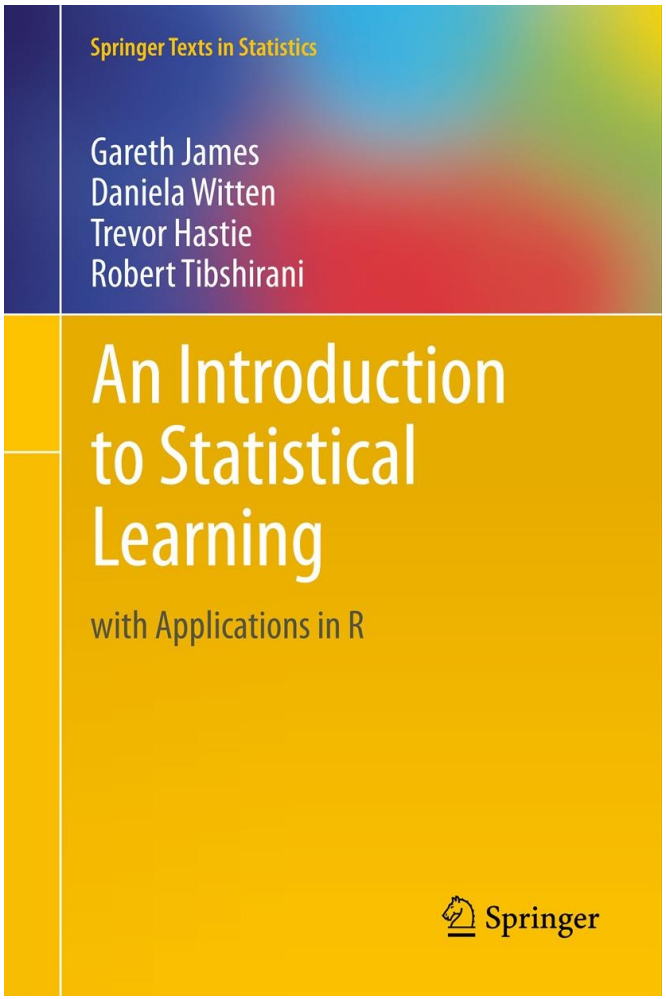
Agenda

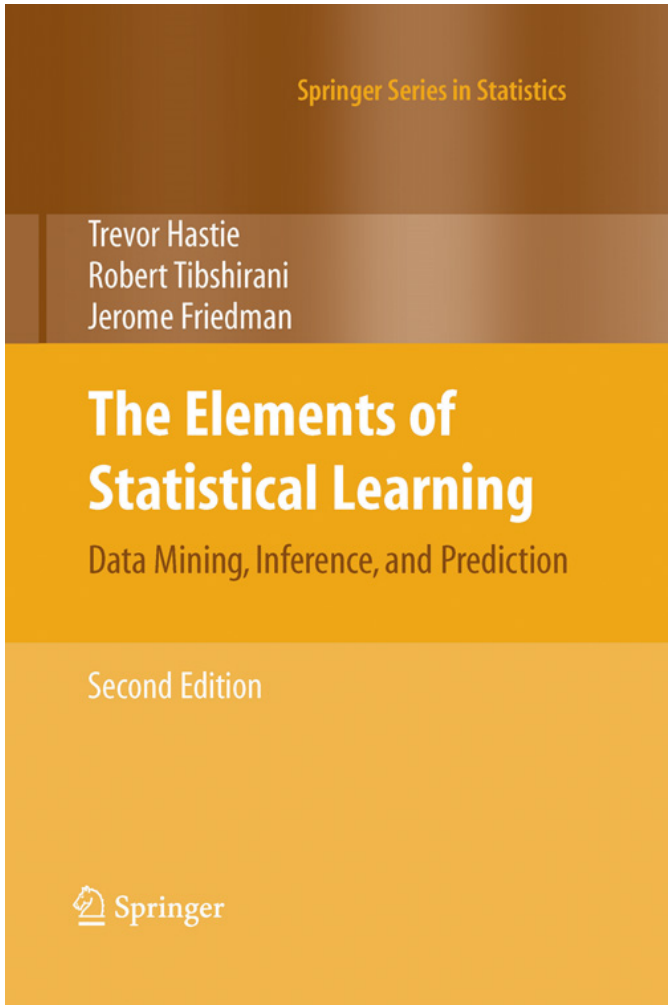
Modelo de regresión lineal

De los modelo lineales a los modelos aditivos

Técnicas de Selección Automática de Modelos Lineales (y Aditivos)

Bibliografía recomendada





ISL: 3.1 a 3.4, 6.1 y 7.1 a 7.5.
ESL: 3.1–3.3.

Agenda

Modelo de regresión lineal

De los modelo lineales a los modelos aditivos

Técnicas de Selección Automática de Modelos Lineales (y Aditivos)

Motivación: Ventas y gasto en publicidad por canal

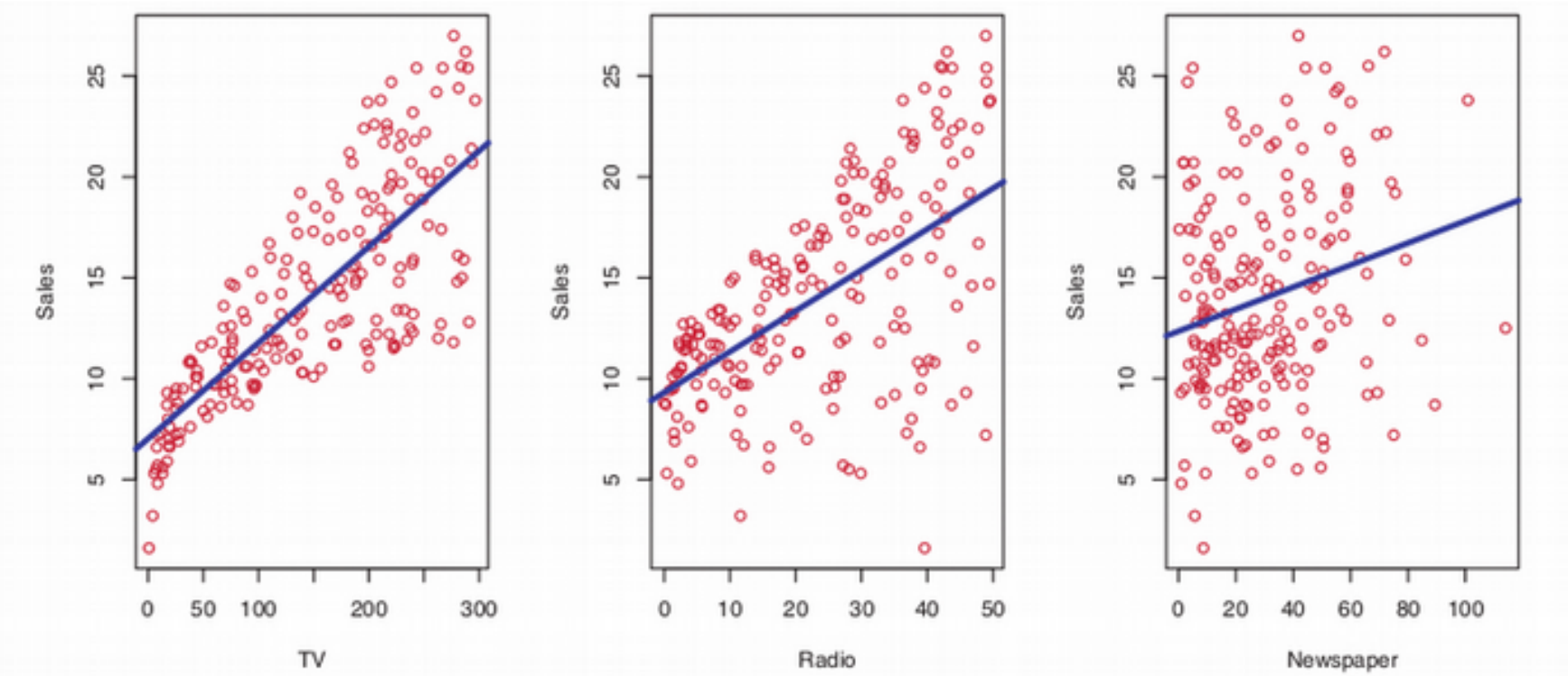


FIGURE 2.1. *The* `Advertising` *data set. The plot displays* `sales`*, in thousands of units, as a function of* `TV`*,* `radio`*, and* `newspaper` *budgets, in thousands of dollars, for* 200 *different markets. In each plot we show the simple least squares fit of* `sales` *to that variable, as described in Chapter 3. In other words, each blue line represents a simple model that can be used to predict* `sales` *using* `TV`*,* `radio`*, and* `newspaper`*, respectively.*

- El modelo de regresión con p covariables:

$$Y = \underbrace{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p}_{f(\mathbf{x};\beta_0,\ldots,\beta_p)} + \varepsilon,$$

$\mathbf{x} = (X_1, \ldots, X_p)$ y $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ vector parámetros.

- Es un modelo (lineal) para $E(Y|X_1, \ldots, X_p)$.

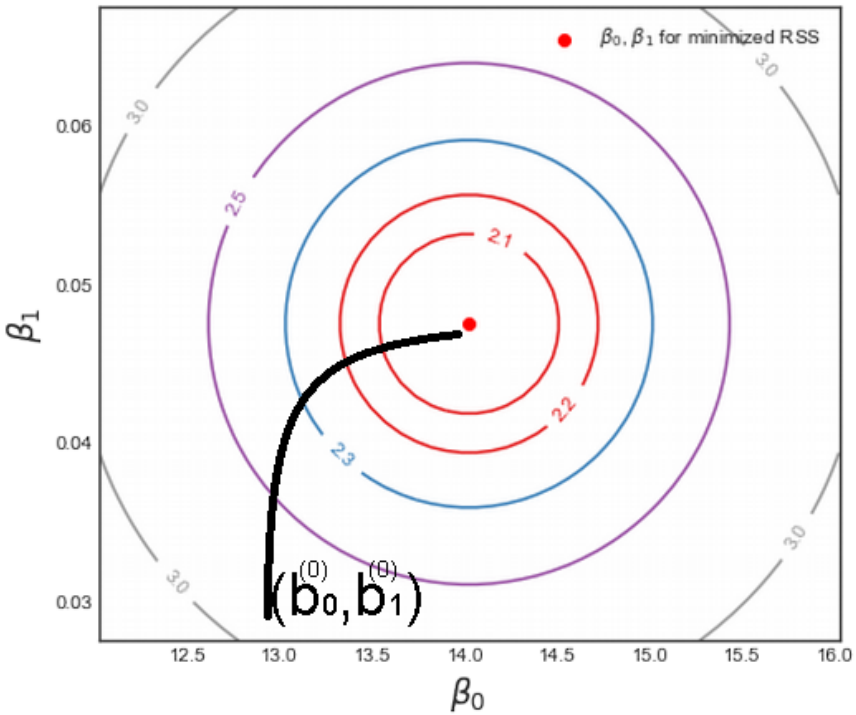
- Dada $S_n = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\}$, aprendemos $\boldsymbol{\beta}$ minimizando:

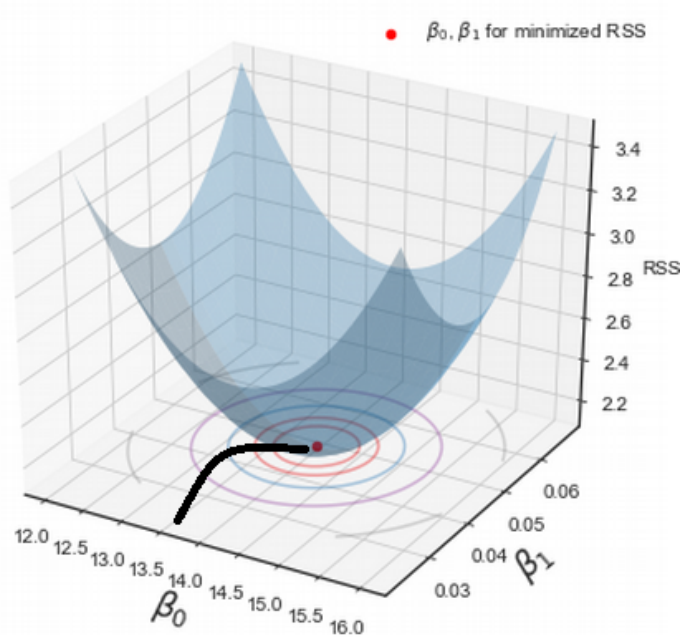
$$\text{RSS}(\mathbf{b}, S_n) = \underbrace{(\mathbf{y} - \mathbf{X}\mathbf{b})^T(\mathbf{y} - \mathbf{X}\mathbf{b})}_{\sum_{i=1}^n (y_i - b_0 - b_1 x_{1i} - \cdots - b_p x_{pi})^2}, \text{donde}$$

$$\mathbf{b} = \begin{bmatrix} b_0 \\ \vdots \\ b_p \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \text{y} \quad \mathbf{X} = \underbrace{\begin{bmatrix} 1 & x_{1,1} & \ldots & x_{p,1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1,n} & \ldots & x_{p,n} \end{bmatrix}}_{\text{Matriz de Diseño de } n \times (p+1)}$$

- Gradiente descendiente.

Gradiente descendiente cuando $n \gg 0$ y/o $p \gg 0$





- Solución exacta: $\widehat{\boldsymbol{\beta}}^{\text{ols}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.
- Si $n \gg 0$, es caro computar $\mathbf{X}^T\mathbf{X}$.
- Si $p \gg 0$, es caro calcular $(\mathbf{X}^T\mathbf{X})^{-1}$.
- Aproximamos solución exacta con GD / SGD.

- Predicciones: Para $\mathbf{x}_{\text{new}}$, estimamos $E(Y|\mathbf{x}_{\text{new}})$ con

$$\widehat{y}_{\text{new}} = \widehat{f}(\mathbf{x}_{\text{new}}) \equiv \mathbf{x}_{\text{new}}^T \widehat{\beta} = \mathbf{x}_{\text{new}}^T (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

- Bondad de ajuste:
  - In sample: $R_a^2 = 1 - (1 - R^2)(n-1)/(n-p-1)$.
  - Out of sample: Sobre una muestra de **validación** de tamaño m

$$\text{RSS} = \sum_{j=1}^{m}(y_j - \widehat{y}_j)^2, \text{ o ASR} = \sum_{j=1}^{m}|y_j - \widehat{y}_j|.$$

- Selección de Modelos: Con estas cantidades *comparas* entre diferentes modelos para seleccionar el "más adecuado".
  - Técnicas *automáticas* de selección (en breve).
  - Regularización (Ridge, Lasso, ENets).

Aproximaciones Estadísticas al problema de selección

- ¿El modelo en su conjunto es relevante para explicar Y?

$$H_0 : \beta_1 = \cdots = \beta_p = 0 \text{ vs } H_1 : \text{al menos una pendiente } \neq 0.$$

  - $F = (n - p - 1)(\text{SCT - SCR})/(p\text{SCR}) \sim F_{p,n-p-1}$.

- Test individuales de significación:

$$H_0 : \beta_i = 0 \text{ vs } H_1 : \beta_i \neq 0, \quad i = 1, \ldots, p.$$

  - $t = \widehat{\beta}_1/\text{se}(\widehat{\beta}_1) \sim t_{n-p-1}$.

- En ML la selección de modelos se hace atendiendo a la capacidad predictiva (discutimos estrategias en breve).

Cuantificando incertidumbres sobre las predicciones

- Característica relevante de los modelos lineales (extensible a los modelos aditivos) bajo supuestos estadísticos débiles.

- Para $\mathbf{x}_{\text{new}}$ el int. predictivo de confianza $1-\alpha$ tiene la forma:

  $$\widehat{y}_{\text{new}} \in [\widehat{\boldsymbol{\beta}}^T \mathbf{x}_{\text{new}} - \widehat{\sigma} C(\alpha, \mathbf{x}_{\text{new}}, S_n),\ \widehat{\boldsymbol{\beta}}^T \mathbf{x}_{\text{new}} + \widehat{\sigma} C(\alpha, \mathbf{x}_{\text{new}}, S_n)],$$

  donde $\alpha \in (0,1)$ determina el nivel de confianza y $C(\alpha, \mathbf{x}_{\text{new}}, \mathbf{X})$ es una cantidad conocida que depende de los datos, de la confianza y del punto donde quieres hacer tu predicción.

  - $\widehat{\sigma}^2$ es una estimación de la varianza de Y.

  - Con el comando `predic` en `R` computas estos intervalos.

- Los intervalos sobre los parámetros son poco relevante en ML.

Caso estudio: Regresión lineal en R

Figure: El objetivo de este ejercicio en clase es modelizar la calidad del vino como una función lineal de sus atributos químicos.

Recapitulación:

- El modelo lineal es interpretable (parámetros = pendientes).
- Escalable a contextos de Bigdata.
- Modelo probabilístico (incertidumbre cuantificable).
- Poco robustos: Tratar los datos atípicos antes de modelar o robustecer el mecanismo de aprendizaje de parámetros.
- En ISL § 3.3: Dummy variables, transformaciones log–lineales, autocorrelación y heterocedasticidad, colinealidad, etc.
- Reflexión: ¿Sólo podemos modelar relaciones lineales?

Agenda

Modelo de regresión lineal

De los modelo lineales a los modelos aditivos

Técnicas de Selección Automática de Modelos Lineales (y Aditivos)

- Para el problema de regresión $Y = f(X_1, \ldots, X_p) + \varepsilon$.
  - Asumamos momentaneamente que $X_i \in \mathbb{R}$ para $i = 1, \ldots, p$.
- Los modelos *aditivos* proponen que:

$$f(X_1, \ldots, X_p) = \beta_0 + g_1(X_1) + \cdots + g_p(X_p)$$

- donde $\{g_1, \ldots, g_p\}$ son funciones suaves que se pueden *representar* utilizando una **base de funciones** $\{\phi_1, \ldots, \phi_B\}$:

$$g_j(x) = \sum_{b=1}^{B} \beta_{jb}\phi_b(x), \quad j = 1, \ldots, p \text{ (con } B < \infty).$$

1. Bases de polinomios (regresión polinómica).
2. B–splines (Polinomios *locales*).
3. Fourier, Wavelets, RKHS, etc (fuera del curso).

- Parámetros del modelo: $(\beta_0, \beta_{11}, \beta_{12}, \ldots, \beta_{pB})$.
  - Notar que el modelo es lineal en los parámetros.

(BackUp)

- Las información cualitativa (que se corresponden con las dummy's: $D_1, \ldots, D_k$), solo puede ingresar de manera lineal en la expresión funcional de la media condicional del modelo:

$$\underbrace{f(D_1, \ldots, D_k, X_1, \ldots, X_p)}_{E(Y|D_1,\ldots,D_k,X_1,\ldots,X_p)} = \beta_0 + \sum_{i=1}^{k} \alpha_i D_i + \sum_{j=1}^{p} \underbrace{\sum_{b=1}^{B} \beta_{jb} \phi_b(x_j)}_{g_j(X_j)} .$$

- La cantidad de funciones en la base para cada coordenada $X_1, \ldots, X_p$ no tiene porque ser la misma (es decir que no tiene porqué ocurrir que $B_1 = B_2 = \cdots = B_p \equiv B$).

- En lo que sigue, y para simplificar la exposición, vamos a asumir que no hay variables cualitativas en el data set.

Regresión polinómica

- Stone–Weierstrass approximation theorem.

$$f(X_1, \ldots, X_p) = \beta_0 + \underbrace{\sum_{b=1}^{B_1} \beta_{1b} X_1^b}_{g_1(X_1)} + \underbrace{\sum_{b=1}^{B_2} \beta_{2b} X_2^b}_{g_2(X_2)} + \cdots + \underbrace{\sum_{b=1}^{B_p} \beta_{pb} X_p^b}_{g_p(X_p)}.$$

- Tendremos que aprender $1 + B_1 + \cdots + B_p$ parámetros.

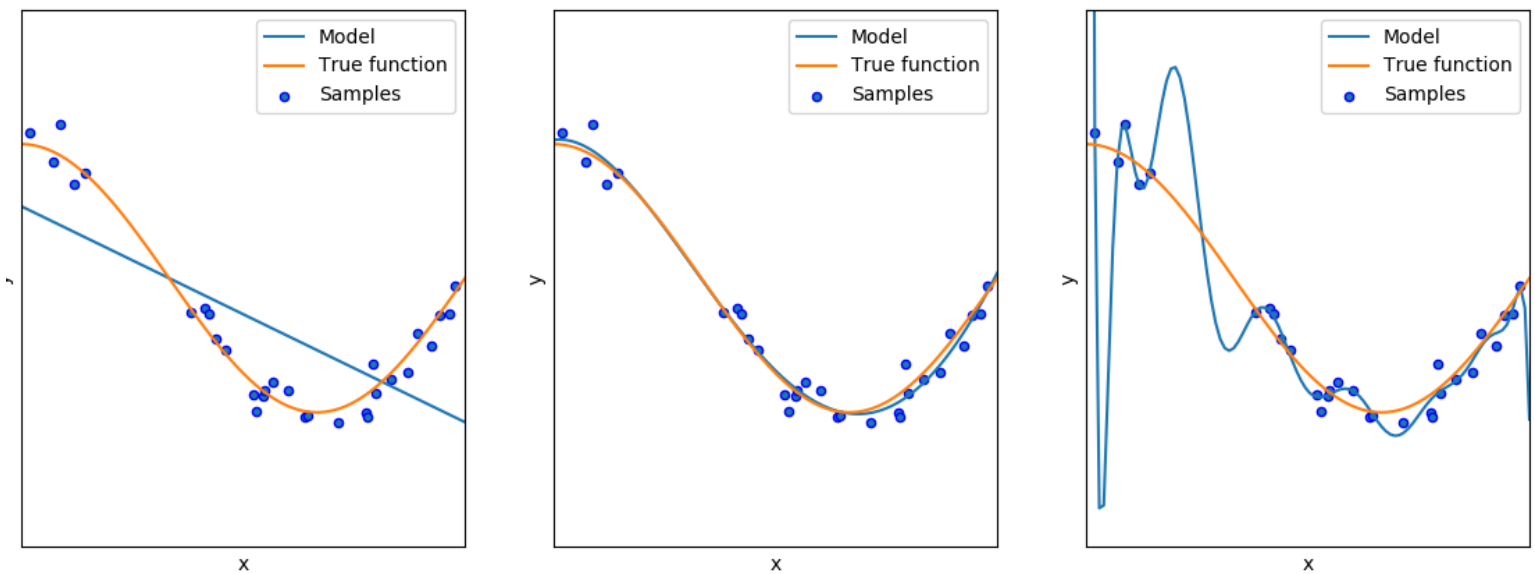


Figure: ↑ B ⇒ Mayores chances de hacer overfitting.

Aprendizaje de parámetros

- Es un modelo lineal en los parámetros donde asumiendo que todos los features son continuos y que $B_1 = \cdots = B_p = B$ luego:

$$\mathbf{b}_{(pB+1)\times 1} = \begin{bmatrix} b_0 \\ \vdots \\ b_{p,B} \end{bmatrix}, \ \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \ \mathbf{X}_{n\times(pB+1)} = \begin{bmatrix} 1 & x_{1,1} & x_{1,1}^2 & \cdots & x_{p,1}^B \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1,n} & x_{1,n}^2 & \cdots & x_{p,n}^B \end{bmatrix}.$$

- Fitting e inferencia: Idem a los modelos de regresión lineal.
- Automatizamos el *feature engineering*.
- Efectos cruzados e información cualitativa (D) en el modelo:

$$f(D, X_1, X_2) = \beta_0 + \alpha D + \beta_{11} X_1 + \beta_{12} X_1^2 + \beta_{21} X_2 + \beta_{22} X_2^2 + \gamma X_1 X_2.$$

- Drawback: No tenemos *control local*.

Piecewise linear regression

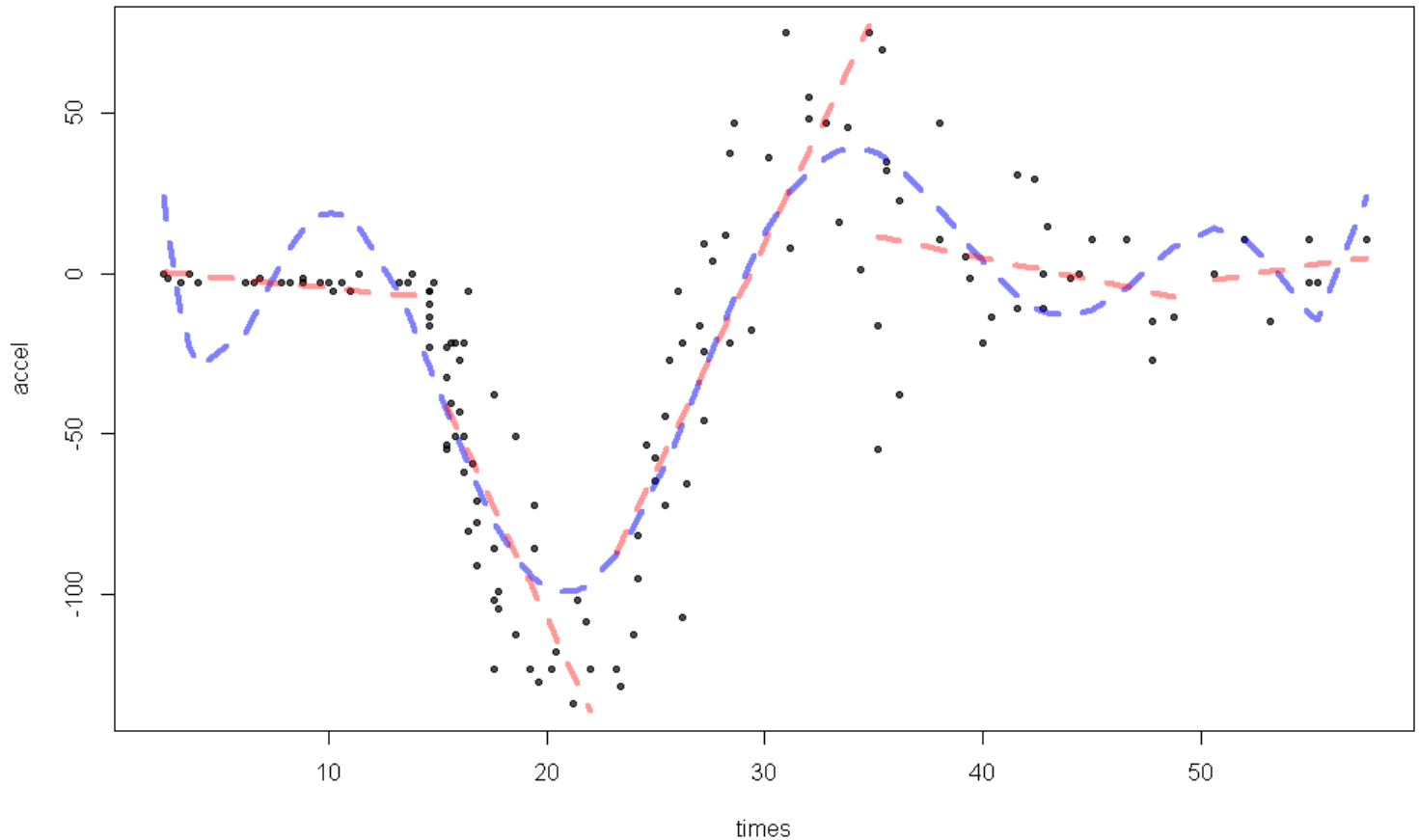


Figure: `Motorcycle crash data`: En azul el fitting de una regresión polinómica ($B = 9$) en rojo el fitting lineal por tramos (4 nodos).

- ¿Podemos introducir continuidad y escribir el modelo como aditivo?

Piecewise linear regression

- Llamemos nodos a los puntos en donde la función de regresión tiene discontinuidades: $\{\nu_1 = 15, \nu_2 = 22, \nu_3 = 35, \nu_4 = 50\}$.
- Definimos $(x - \nu)_+ \equiv \max\{x - \nu, 0\}$.
- Fiteamos el modelo de regresión (continuo en X):

$$Y = \underbrace{\beta_0 + \beta_1 X + \sum_{b=1}^{4} \beta_{b+1}(x - \nu_b)_+}_{g(X)} + \varepsilon.$$

- Esto equivale a un modelo aditivo en donde:

$$\{\phi_1(x) = x, \phi_2(x) = (x - \nu_1)_+, \ldots, \phi_5(x) = (x - \nu_4)_+\}$$

- Veamos como luce nuestra regresión.

Piecewise linear regression

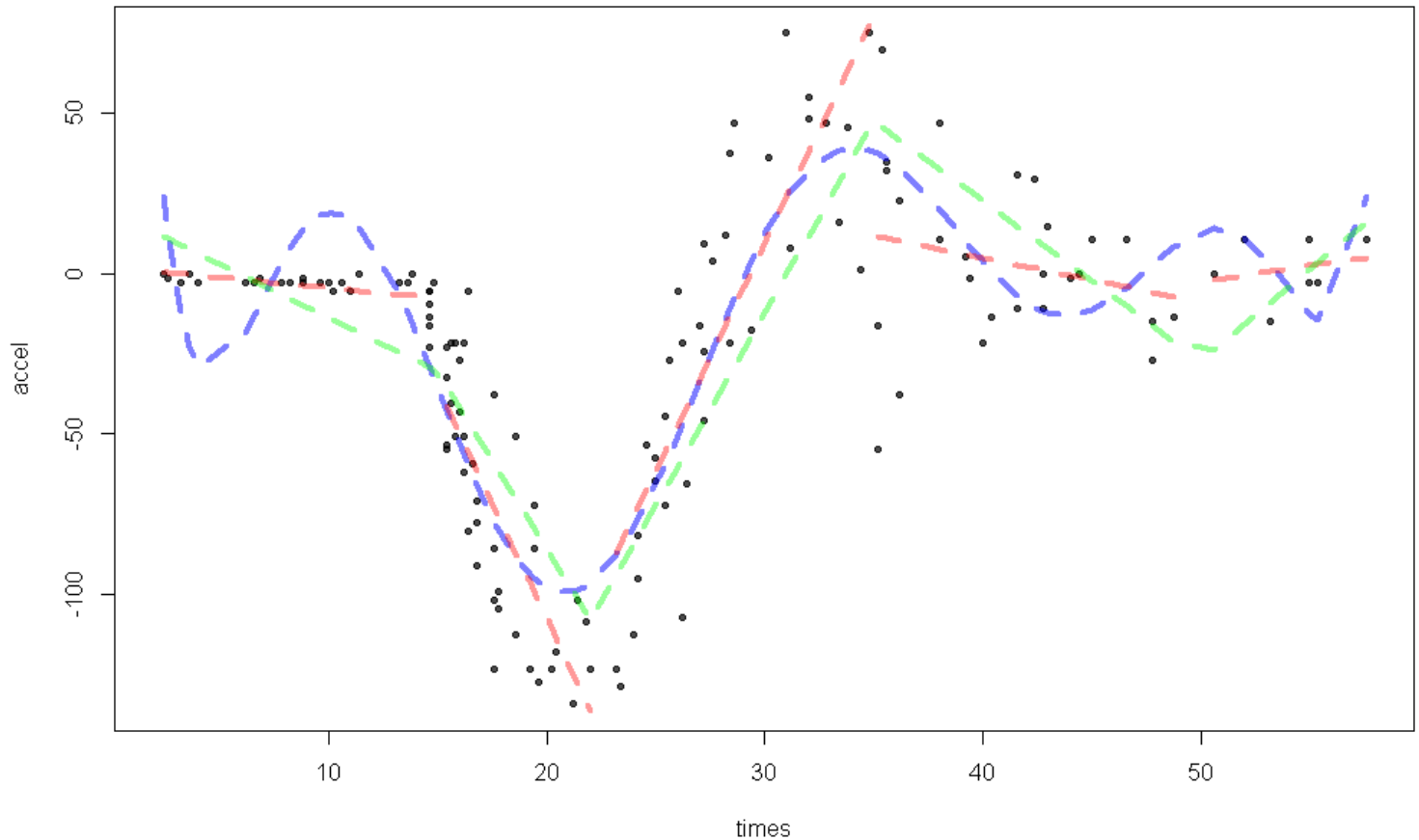


Figure: En verde el fitting de una regresión lineal continua a trozos con nodos en las coordenadas $\{\nu_1 = 15, \nu_2 = 22, \nu_3 = 35, \nu_4 = 50\}$.

- ¿Podemos utilizar polinómios continuos a trozos?

Splines (un solo feature)

- Estimamos $f(X)$ en $X \in [a, b]$ con polim. continuos a trozos.
  - Controlamos la cantidad de derivadas continuas del modelo.
- Definimos k nodos $\{a \leq \nu_1 \leq \nu_2 \leq \cdots \leq \nu_k \leq b\}$.
  - k es un "Hiperparámetro" del modelo (VC / Regularización).
- En cada intervalo $[\nu_i, \nu_{i+1}]$, determinado por dos nodos contiguos, aproximamos $f(X)$ con un *polinomio local*.
  - El grado de ese polinomio local (generalmente cúbico) determina el número de derivadas continuas que tendrá el modelo de regresión.
  - Veamos un ejemplo ilustrativo asumiendo (para simplificar la exposición) que sólo disponemos de 1 covariable.

- Consideremos la base[1]:

$$\{\phi_1(x) = x, \phi_2(x) = x^2, \phi_3(x) = x^3, \phi_4(x) = (x-\nu_1)^3_+, \ldots, \phi_{k+3}(x) = (x-\nu_k)^3_+\}$$

- El modelo de regresión de Splines cúbicos plantea que:

$$Y = \overbrace{\beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \underbrace{\sum_{i=4}^{k+3} \beta_i \phi_i(X)}_{\text{polinomios locales}}}^{g(X)} + \varepsilon.$$

- $k + 4$ β's se aprenden minimizando RSS.
  - $\uparrow k \Rightarrow$ Mayores chances de hacer overfitting.
  - Ahora $\mathbf{X}_{n\times(k+4)}$ (en columnas van las $\phi(x)$'s).
- Modelo lineal (fitting e inferencia).
- Regularización para controlar under/overfitting.

[1] Para evitar problemas numéricos generalmente ortonormalizamos la base (puedes ver más detalles en S. Wood, *Generalized Additive Models*, **pp–201**.)

Polinomios cúbicos locales (BackUp)

- Para cada $\nu \in \{\nu_1, \nu_2, \ldots, \nu_k\}$, consideremos

$$h(x, \nu) = \begin{cases} (x - \nu)^3 \text{ si } x > \nu, \\ 0 \text{ en otro caso.} \end{cases}$$

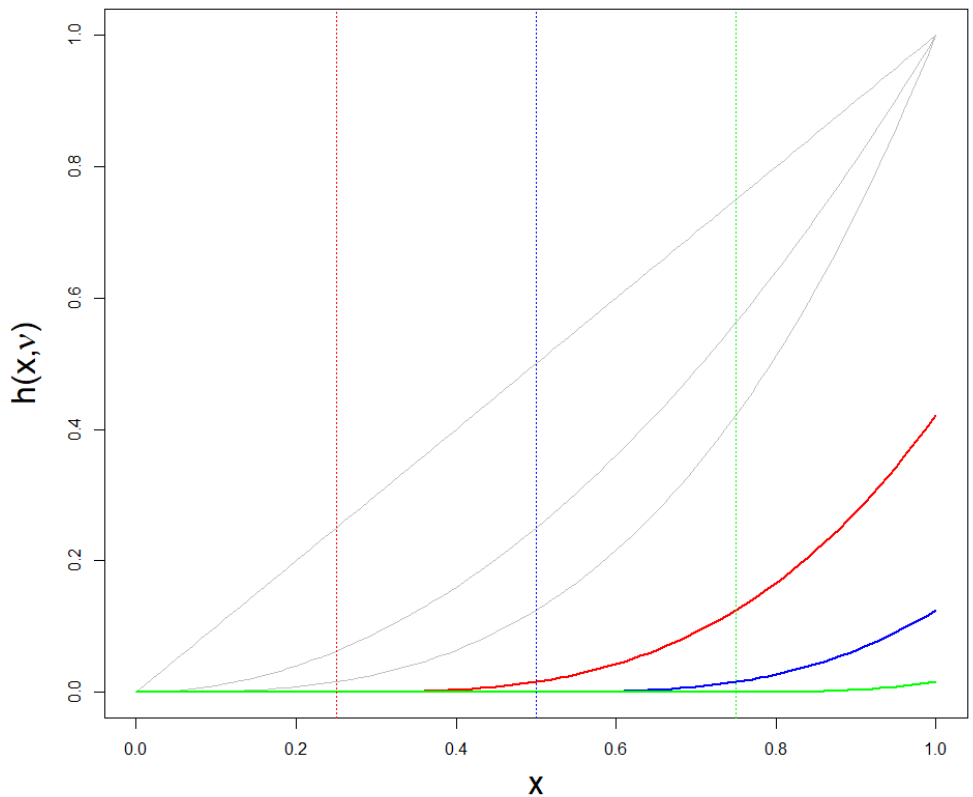


Figure: Ejemplo para $x \in [0, 1]$ y $\nu_1 = 0.25$, $\nu_2 = 0.50$, $\nu_3 = 0.75$.

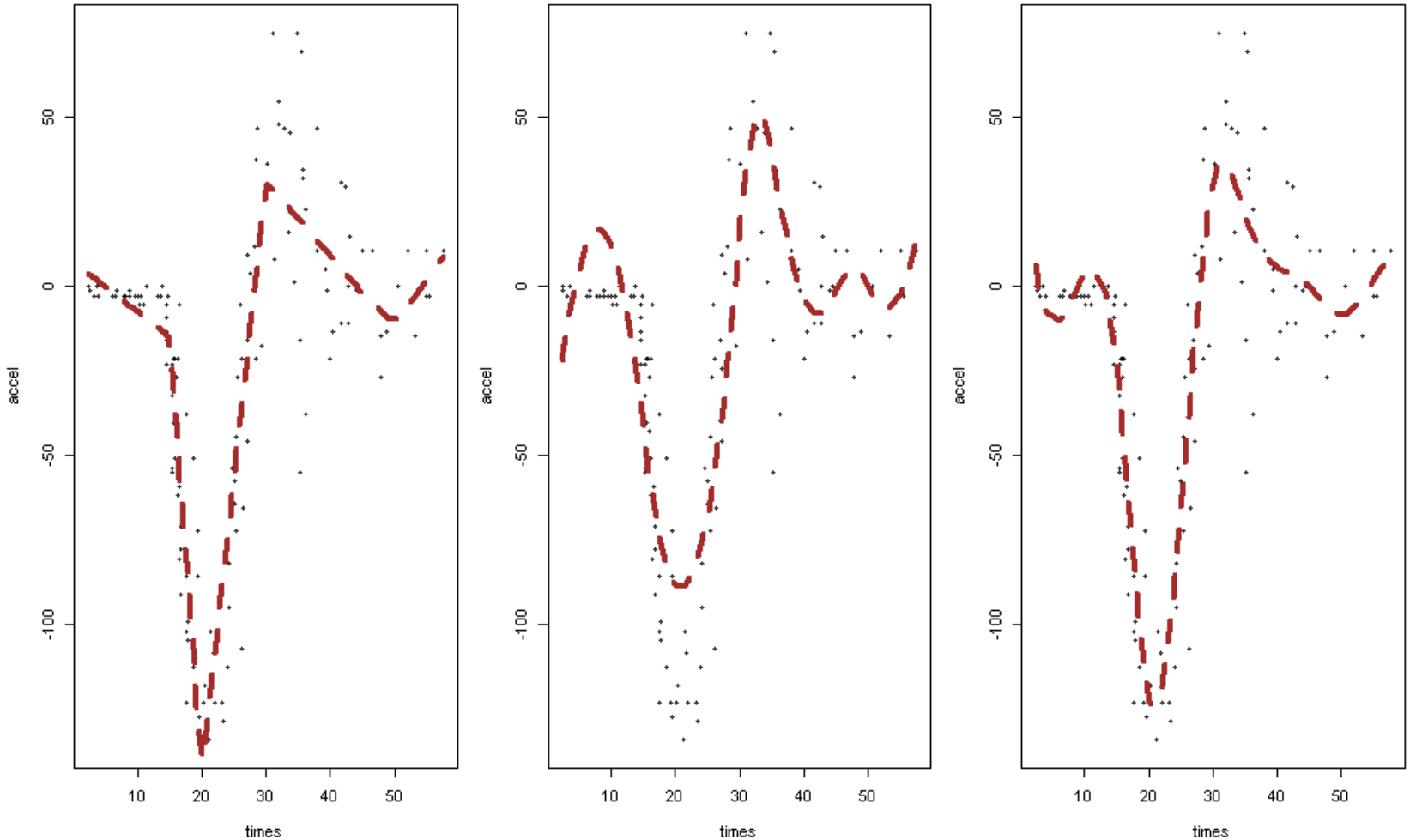


Figure: Spline lineal, cuadrático y cúbico (izquierda a derecha) con nodos en los puntos: $\{\nu_1 = 15, \nu_2 = 20, \nu_3 = 30, \nu_4 = 35, \nu_5 = 45, \nu_6 = 50\}$.

(BackUp slide)

- Es fácil ver que las dos primeras derivadas del modelo que asumimos para la media condicional de Y son continuas.

- Por ejemplo para $k = 1$ (un solo nodo), se tiene que:

$$g'(X) = \beta_1 + 2\beta_2 X + 3\beta_3 X^2 + 3\beta_4 (X - \nu)_+^2,$$

- Luego: $\lim_{x \to \nu^-} g'(x) = \beta_1 + 2\beta_2 \nu + 3\beta_3 \nu^2 = \lim_{x \to \nu^+} g'(x)$.

- Cuando $k > 1$ ocurre lo mismo cuando $x \to \nu_i$ para todo $i = 1, \ldots, k$. Mismo análisis para la segunda derivada.

- ¿Cómo procedemos si tenemos más de un feature numérico?

Caso multivariante

- Para cada coordenada de la función de regresión consideramos

$$\{a^{(j)} \leq \nu_1^{(j)} \leq \nu_2^{(j)} \leq \cdots \leq \nu_k^{(j)} \leq b^{(j)}\}, \text{ para } j = 1, \ldots, p.$$

  - La cantidad de nodos en cada coordenada puede ser diferente.

- Sea $\{\phi_{4,j}(x_j) = (x_j - \nu_1^{(j)})_+^3, \ldots, \phi_{k+3,j}(x_j) = (x_j, \nu_k^{(j)})_+^3\}$, luego:

$$g_j(X_j) = \beta_{1j}X_j + \beta_{2j}X_j^2 + \beta_{3j}X_j^3 + \underbrace{\sum_{i=4}^{k+3} \beta_{ij}\phi_{ij}(X_j)}_{\text{efectos locales}}, \text{ para } j = 1, \ldots, p.$$

- Y el modelo de regresión multivariante se escribe como:

$$Y = \beta_0 + g_1(X_1) + \cdots + g_p(X_p) + \varepsilon.$$

- Aprendemos $(k+3)p + 1$ parámetros minimizando la RSS.

Modelos aditivos en la práctica

- Los dos modelos aditivos discutidos se encuentran perfectamente integrados al comando `lm` de `R`.

- `poly(x,d)`: Creamos una matriz de diseño con monomios de hasta grado d para el vector/matriz de inputs x.

- `bs(x,knots,degree)`: (cargar la librería *splines*).
  - nknots / df: controla el número y localización de los nodos.
  - degree: controla el grado polinomial del spline (3 por defecto).

- Introducimos el uso de estas técnicas con `mcycle`. NO discutimos como elegir los grado del polinomio (ni la cantidad de nodos) que deben ser seleccionados por VC o utilizando técnicas automáticas de selección de modelos (en breves).

Regresión Polinómica vs Splines vs Natural Splines

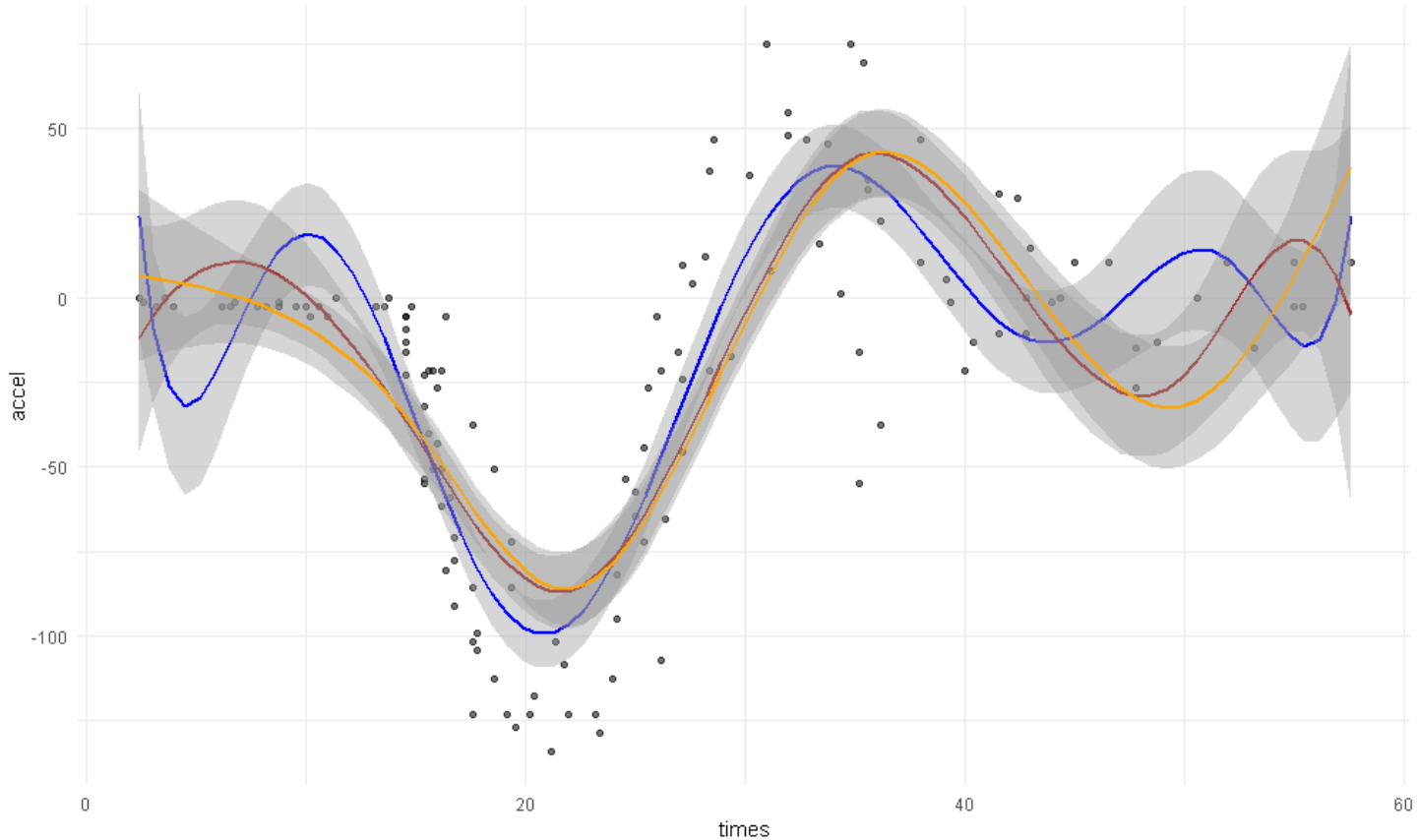


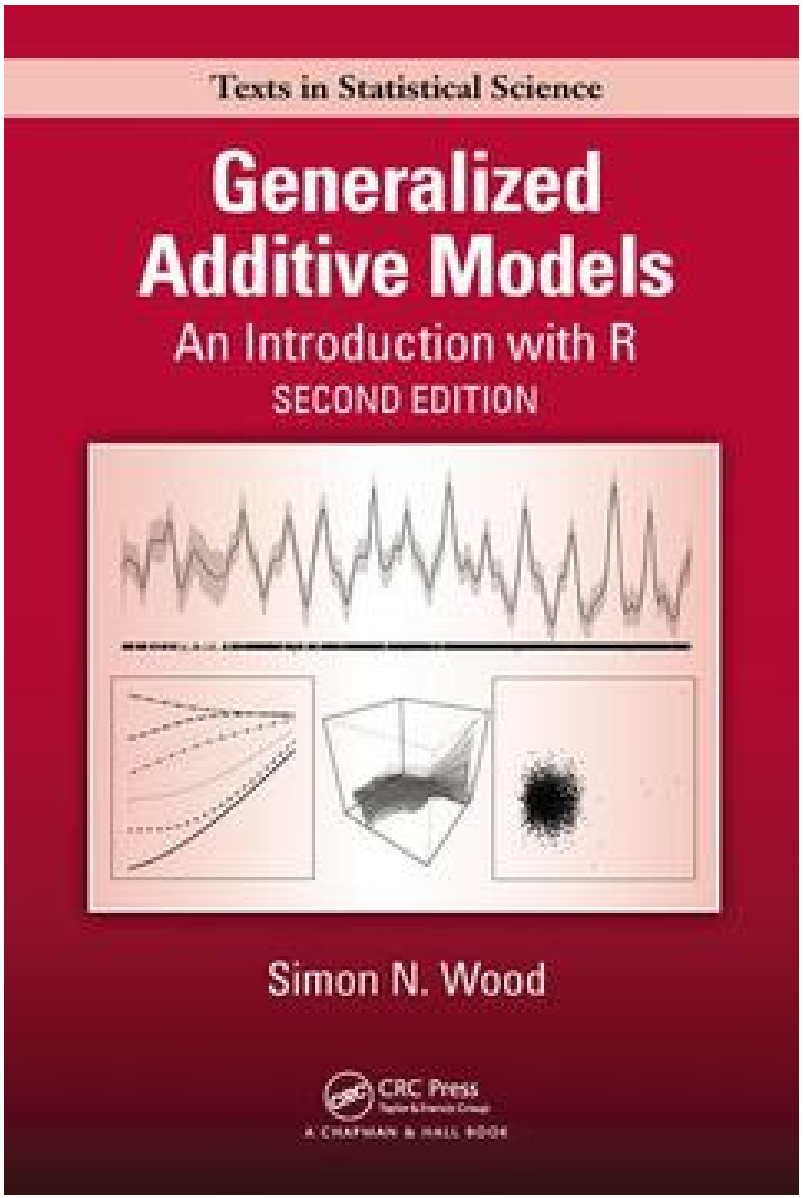
Figure: En color azul la regresión polinómica de grado 9, en marrón el spline cúbico (4 nodos) y en naranja el spline natural (mismos 4 nodos).

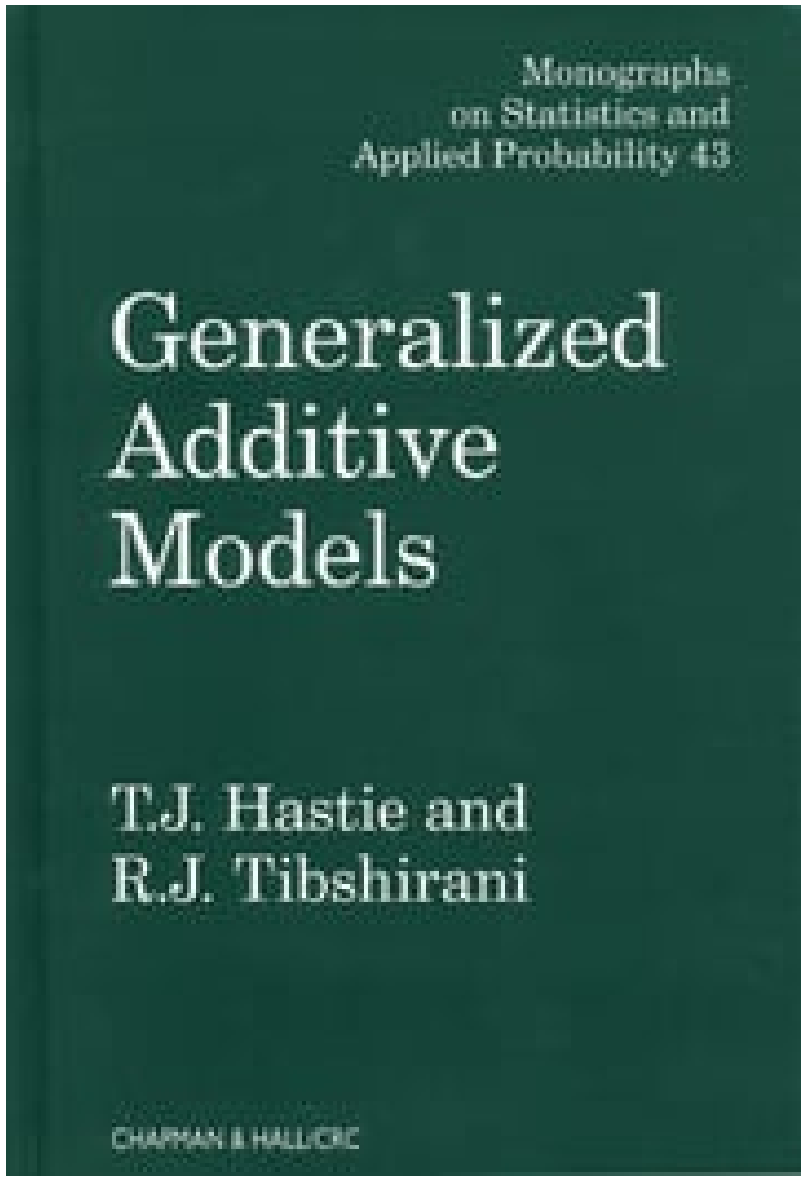
- Caso multivariante con el data set `wines`.

Consideraciones finales

- Como elegir la ubicación y cantidad de *knots*:
  - Ubicación: Equi-espaciados vs. Cuantiles.
  - Cantidad: VC / Selección Automática / Regularización.
- Ventajas de modelos aditivos:
  - Permiten cuantificar el efecto aislado de cada regresor.
  - Podemos cuantificar la incertidumbre de las predicciones.
  - Computacionalmente escalable.
  - Admiten información cualitativa (dummy's).
- Desventajas:
  - Riesgo de overfitting si no cross-validas / regularizás.
  - Sensibles a datos atípicos y perdidos.

Bibliografía recomendada

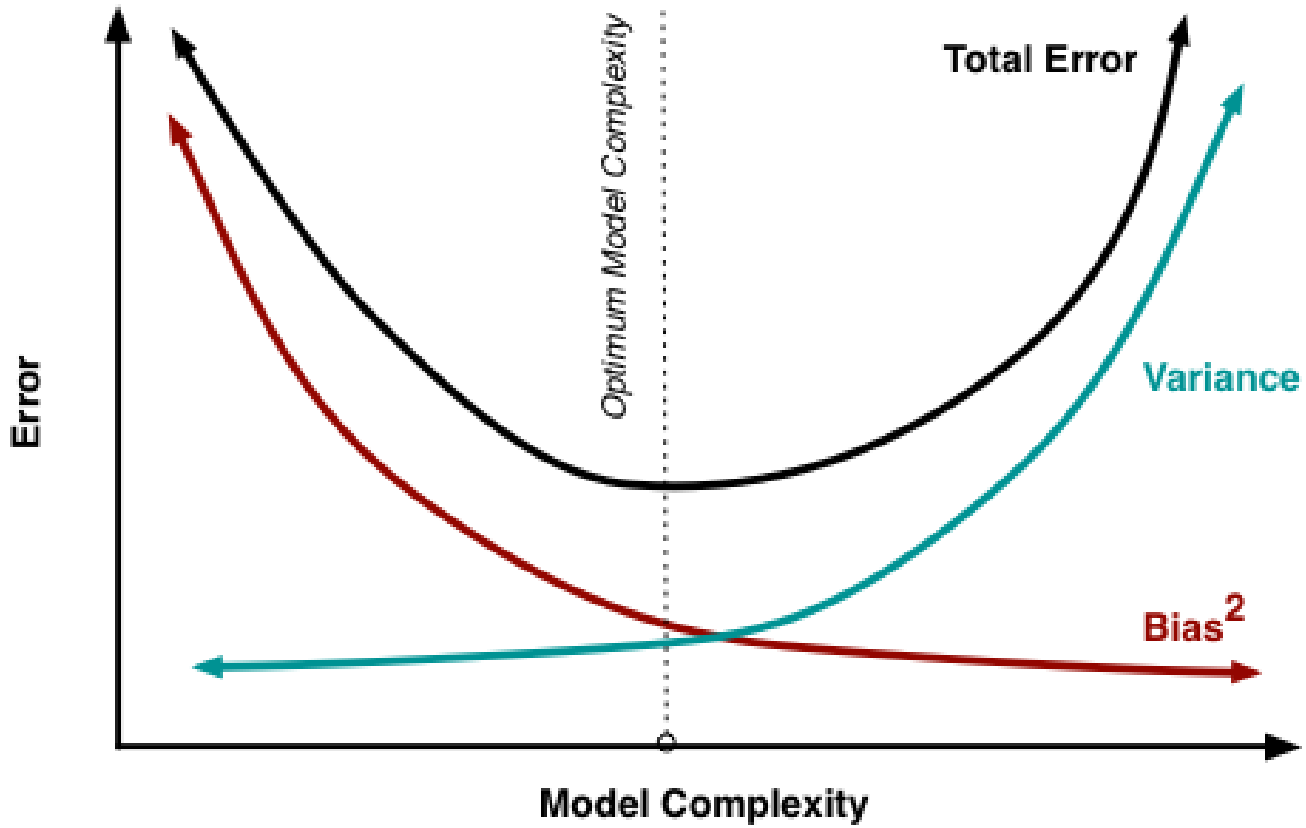




Agenda

Agenda

- ¿Por qué es importante aprender a seleccionar modelos/variables de manera automática?



- En contextos de alta dimensión ($p \gg 0$) es crucial!

- En tanto p/n sea pequeño y $f(x)$ linealmente aproximable; el modelo de regresión lineal tendrá poco sesgo y variabilidad.

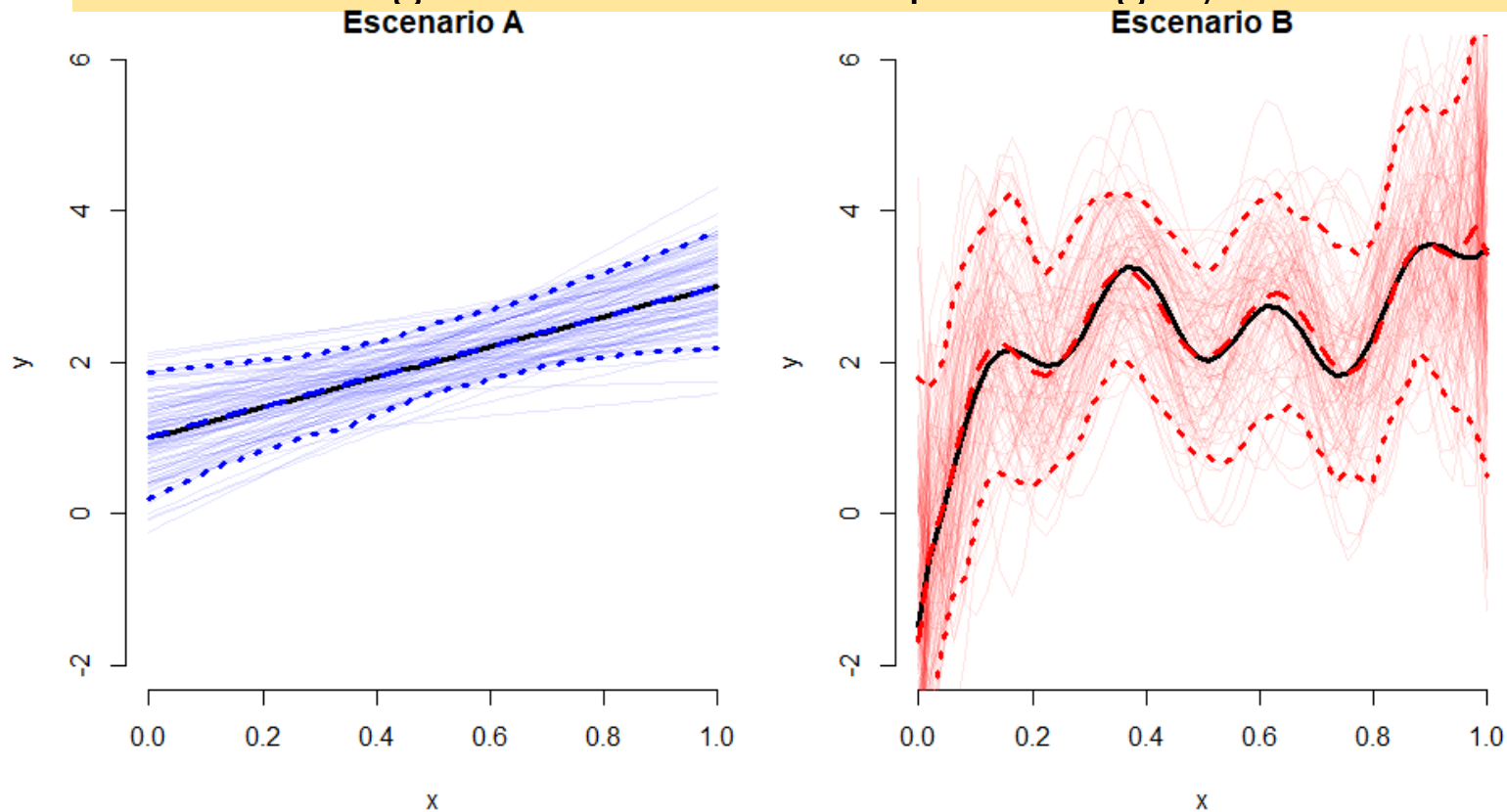


Figure: Escenario A: $f(X) = 1 + 2X$; Escenario B: $f(X) =$ Pol. de grado 10 en X. Generamos 100 muestras de entrenamiento ($n = 50$) y fiteamos.

- $\uparrow p/n \Rightarrow \uparrow \text{Var}(\widehat{f}) \Rightarrow \uparrow$ Error promedio del modelo.
- Quitar features para intercambiar (un poco de) sesgo por (una reducción considerable de) varianza (mejores predicciones).

¿Por qué *selección de modelos*?

- **Incrementar la calidad predictiva**.
- **Interpretabilidad**: Eliminar variables poco significativas.
- Cuando $p > n$ resulta imposible estimar el modelo.
- Técnicas habituales:
  - Selección manual: VIF (cuando p es relativamente pequeño).
  - Métodos automáticos de selección de modelos:
    1. Exhaustivo o Bruto (inviable cuando $p > 30$).
    2. Stepwise: Forward y Backward.
  - Shrinkage (regularización): RIDGE, LASSO y ENETS.

Notación

- La discusión vale para modelos aditivos en general:

$$Y = \beta_0 + \sum_{b=1}^{B_1} \beta_{1b}\phi_b(X_1) + \sum_{b=1}^{B_2} \beta_{2b}\phi_b(X_2) + \cdots + \sum_{b=1}^{B_p} \beta_{pb}\phi_b(X_p) + \varepsilon$$

- Pero para simplificar la notación, siempre se plantea que:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon.$$

- Notar que se trata simplemente de renombrar los features:
  - $X_1 \equiv \phi_1(X_1)$ y $\beta_1 \equiv \beta_{11}$.
  - $X_2 \equiv \phi_2(X_1)$ y $\beta_2 \equiv \beta_{12}$.
  - ...
  - $X_p \equiv \phi_{B_p}(X_p)$ y $\beta_p \equiv \beta_{pB_p}$.

Agenda

VIF

- Computar y analizar la matriz de $p \times p$ correlaciones entre pares de features no resulta práctico cuando $p \gg 2$.
  - Relaciones de colinealidad importantes entre grupos de 3 o más covariables, aún no siendo las correlaciones de a pares elevadas.

- VIF: *Variance Inflation Factor*

$$\text{VIF}(X_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}} \geq 1, \quad j = 1, \ldots, p.$$

- Cuando $R^2_{X_j|X_{-j}}$ es relativamente grande, la información de la covariable j-ésima es redundante ya que esta incluida en las demás covariables (en el límite, cuando $R^2_{X_j|X_{-j}} = 1$, la covariable j es una combinación lineal de las restantes $p - 1$).

- Rule of thumb: Descartar covariable con VIF > 5.

Agenda

Método Exhaustivo

- ¿Qué features entre $X_1, \ldots, X_p$ incluir en el modelo?
- Exploramos todo el *espacio de modelos*.

Algorithm 1 pseudocódigo para hacer selección exhaustiva

1: Partimos de $\mathcal{M}_0$ el modelos sin features.
2: **for** (i in $1 : p$) **do**
3: Ajustamos los $\binom{p}{i}$ modelos con i–features.
4: Elegimos de entre estos el que tenga el mayor R^2 o (equivalentemente) el menor RSS in–sample y lo llamamos $\mathcal{M}_i$.
5: **end for**
6: Construida $\{\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p\}$, utilizamos un conjunto de validación (VS) para elegir mejor modelo (menor RSS).

- Costo computacional: $\sum_{i=0}^{p} \frac{p!}{(p-i)!i!} = 2^p$ (ej: $2^{20} \approx 1M$).
- Existen implementaciones eficientes del método que permiten utilizarlo en situaciones donde tienes hasta $p \approx 35$ features.

Forward–Stepwise

Algorithm 2 pseudocódigo para hacer selección Forward–Stepwise

1: Partimos de $\mathcal{M}_0$ el modelos sin features.
2: **for** (i in $0 : p-1$) **do**
3: Consideramos los $p - i$ modelos que contienen un features adicional con respecto al modelo estimado en $\mathcal{M}_i$.
4: De estos, elegimos el que tiene el mayor R^2 o el menor RSS (in–sample) y lo llamamos $\mathcal{M}_{i+1}$.
5: **end for**
6: Una vez construida $\{\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p\}$, utilizamos VS para elegir mejor modelo (menor RSS o métrica equivalente).

- Costo computacional: $1 + p(p+1)/2$. Si por ejemplo $p = 20 \Rightarrow$ fiteamos 211 modelos (vs. 1M exhaustivo).

- Si $p > n$, podemos emplear este método para computar el mejor modelo entre $\{\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_{n-1}\}$.

Backward–Stepwise

Algorithm 3 pseudocódigo para hacer selección Backward–Stepwise

1: Partimos de $\mathcal{M}_p$ el modelo con todos los features.
2: **for** (i in $p : 1$) **do**
3: Consideramos los i modelos de $i-1$ features que podemos construir con las covariables incluidas en el modelo $\mathcal{M}_i$.
4: De estos, elegimos de entre estos el que tiene el mayor R^2 o el menor RSS (in–sample) y lo llamamos $\mathcal{M}_{i-1}$.
5: **end for**
6: Una vez construida $\{\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p\}$, utilizamos VS para elegir mejor modelo (menor RSS o métrica equivalente).

- Método exhaustivo cuando p es relativamente pequeño.
- Utilizamos Backward y/o Forward con p grande.
- Existen métodos híbridos que combinan Backward y Forward.

Alternativas a validation set approach

- Data Scarcity: Utilizar métricas in–sample (R_a^2, C_M, AIC o BIC) para seleccionar el modelo entre $\{\mathcal{M}_0, \mathcal{M}_1, \ldots, \mathcal{M}_p\}$

- C_M: Coeficiente de Mallows

$$C_M(\mathcal{M}_k) = \frac{1}{n}(\text{RSS}_k + 2k\widehat{\sigma}_\varepsilon^2).$$

  - $\hat{\sigma}_\varepsilon^2$ estimado con el modelo *saturado* (con p-features).

- AIC: Akaike Information Criterion

$$\text{AIC}(\mathcal{M}_k) = \frac{1}{n\hat{\sigma}_\varepsilon^2}(\text{RSS}_k + 2k\hat{\sigma}_\varepsilon^2).$$

- BIC: Bayesian Information Criterion

$$\text{BIC}(\mathcal{M}_k) = \frac{1}{n\hat{\sigma}_\varepsilon^2}(\text{RSS}_k + \log(n)k\hat{\sigma}_\varepsilon^2).$$

Implementación en R

```
library(leaps)

regsubsets(formula, data, nbest=1, nvmax=8,
       method=c("exhaustive", "backward", "forward")....)

nbest: number of subsets of each size to record (RSS).
nvmax: maximum size of subsets to examine.

Output:
rsq: The r-squared for each model.
rss: Residual sum of squares for each model.
adjr2: Adjusted r-squared.
cp: Mallow C (AIC).
bic: Bayes information criterion.
```

- Caso de estudio `wines`.

Recapitulación

- Estos métodos son recomendables cuando te interesa mantener la interpretabilidad de los parámetros / quieres explicar y cuantificar efectos con el modelo estimado.
  - Los métodos de shrinkage (siguiente) distorsionan los parámetros y no son recomendables si te interesa interpretar.
- Podes utilizar estos métodos para seleccionar covariables en contextos de modelos aditivos. Recordá que en este contexto (asumamos que tenemos un solo feature para simplificar):

$$\mathbf{X}_{n\times(B+1)} = \begin{bmatrix} 1 & \phi_1(x_1) & \phi_2(x_1) & \dots & \phi_B(x_1) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \phi_1(x_n) & \phi_2(x_n) & \dots & \phi_B(x_n) \end{bmatrix}.$$

  - Llamá $X_1 = \phi_1(X), X_2 = \phi_2(X), \dots, X_p = \phi_B(X)$, y aplicá el método tal y como lo presentamos en las filminas anteriores.

Tarea y ping–pong de preguntas

- En `R`: Divide los datos de `wines` en (train + validación) y test y contrasta los resultados a los que arribas seleccionando modelos con los 3 métodos de selección automáticos vistos en clase (en clase resolvimos solo para el primero).

- En genereal: ¿Qué método te parece que tiene las mayores posibilidades de elegir al modelo con mayor capacidad predictiva? ¿Puedes explicar porqué?

- Resuelve los ejercicios: Conceptuales 1 y Aplicados 8 planteados en ISLR en la sección 6.8.