Selección de Modelos y Regularización

Ciencia de Datos y Econometría Aplicada

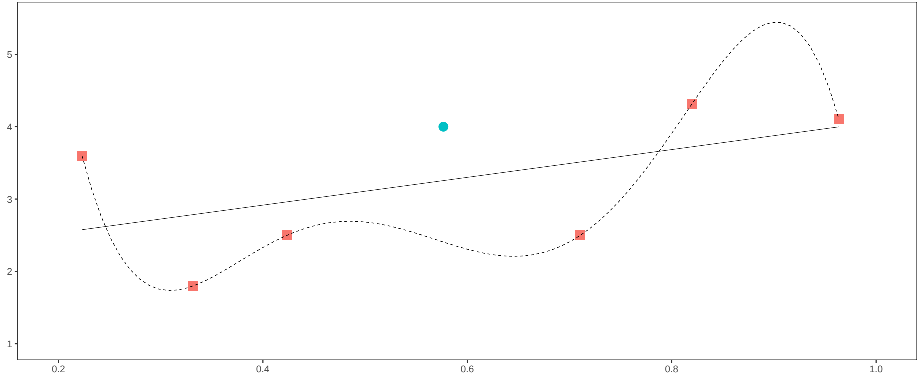
Ignacio Sarmiento-Barbieri

Universidad de los Andes

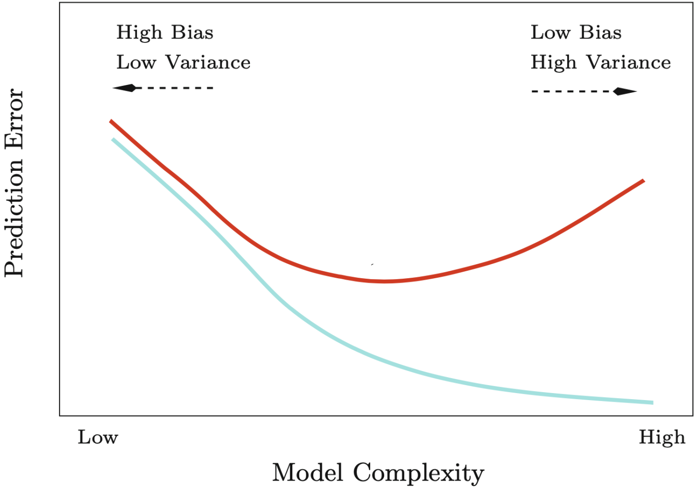
August 27, 2025

Agenda

Overfit y Predicción fuera de Muestra



Overfit y Predicción fuera de Muestra



Overfit y Predicción fuera de Muestra

- ML nos interesa la predicción fuera de muestra

Overfit y Predicción fuera de Muestra

- ML nos interesa la predicción fuera de muestra
- Overfit: modelos complejos predicen muy bien dentro de muestra, pero tienden a hacer un mal trabajo fuera de muestra
- Hay que elegir el modelo que "mejor" prediga
  - Métodos de Remuestreo
    - Enfoque del conjunto de validación
    - Loocv
    - Validación cruzada en K-partes (5 o 10)

Selección de Modelos: Motivación

- Tenemos M_k modelos
- Queremos encontrar el que mejor predice fuera de muestra
- Hay distintas formas de enfrentarlo
- Las clásicas
  - Elección del mejor conjunto
  - Elección por pasos
    - Hacia adelante (Forward selection)
    - Hacia atras (Backward selection)

Regularización

Lasso

- Para un $\lambda \geq 0$ dado, consideremos el siguiente problema de optimización

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{1}$$

Lasso

- Para un $\lambda \geq 0$ dado, consideremos el siguiente problema de optimización

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{1}$$

  - "LASSO's free lunch": selecciona automáticamente los predictores que van en el modelo ($\beta_j \neq 0$) y los que no ($\beta_j = 0$)
  - Por qué? Los coeficientes que no van son soluciones de esquina
  - $L(\beta)$ es no differentiable

Lasso Intuición en 1 Dimension

- Lasso Intuición

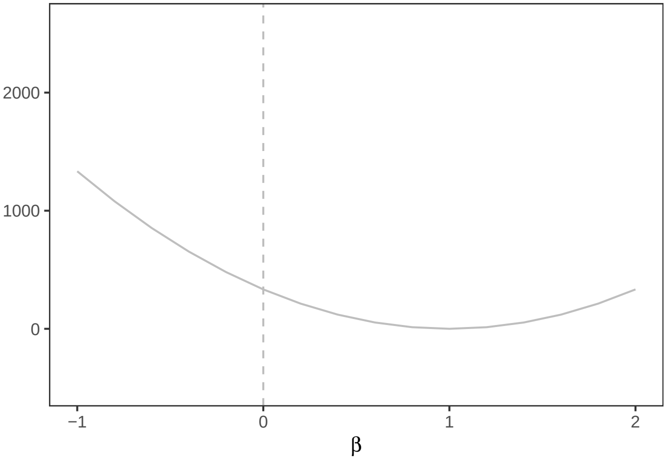
$$min_\beta E(\beta) = \sum_{i=1}^{n}(y_i - x_i\beta)^2 + \lambda|\beta| \tag{2}$$

- Un solo predictor, un solo coeficiente
- Si $\lambda = 0$

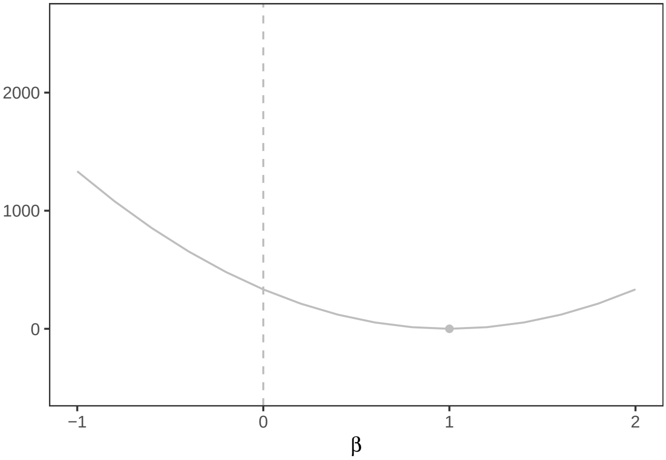
$$min_\beta E(\beta) = \sum_{i=1}^{n}(y_i - x_i\beta)^2 \tag{3}$$

- la solución es?

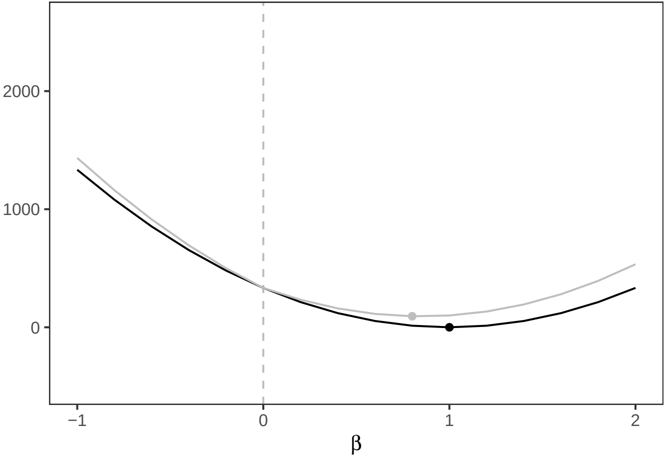
Intuición en 1 Dimension



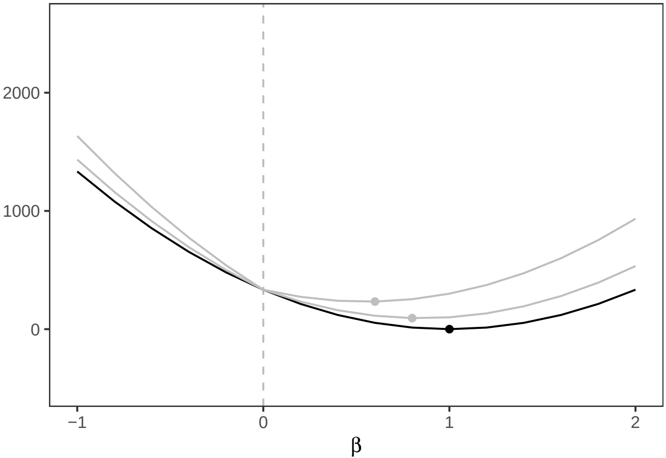
Intuición en 1 Dimension



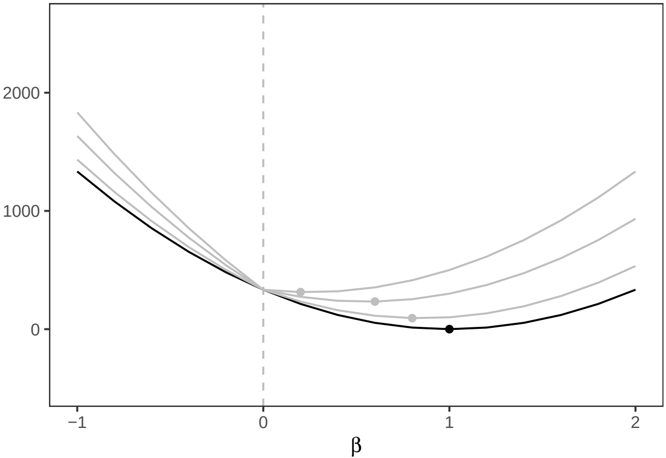
Intuición en 1 Dimension



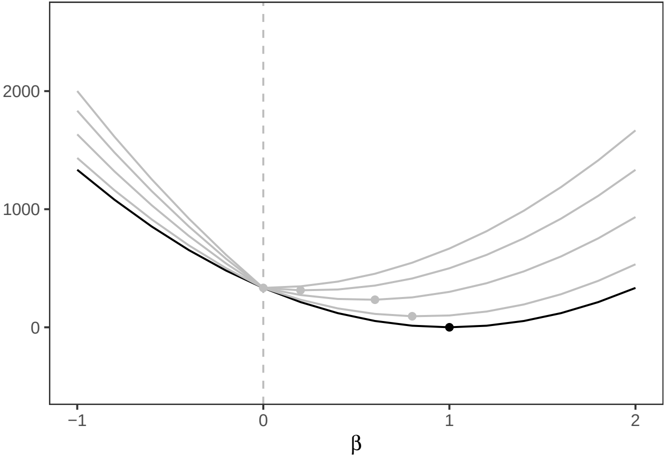
Intuición en 1 Dimension



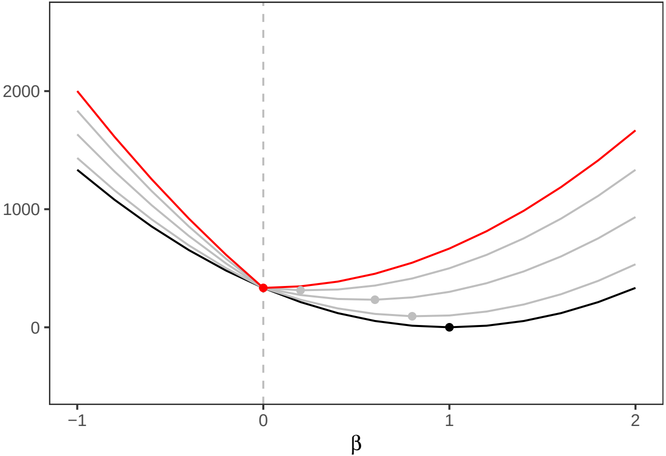
Intuición en 1 Dimension



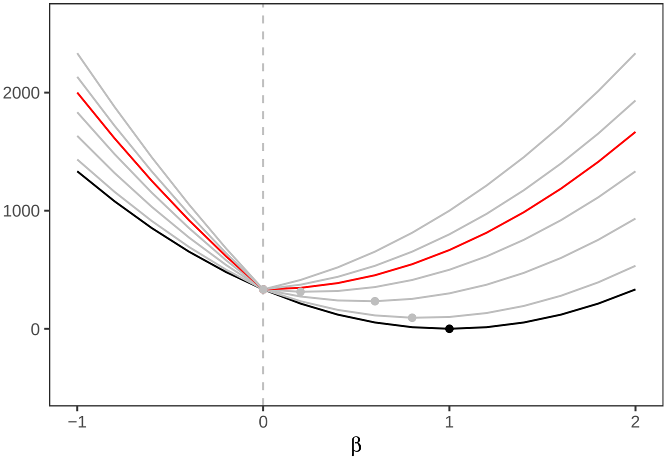
Intuición en 1 Dimension



Intuición en 1 Dimension

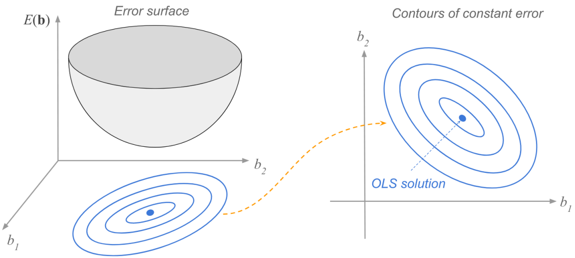


Intuición en 1 Dimension



Intuición en 2 Dimensiones (OLS)

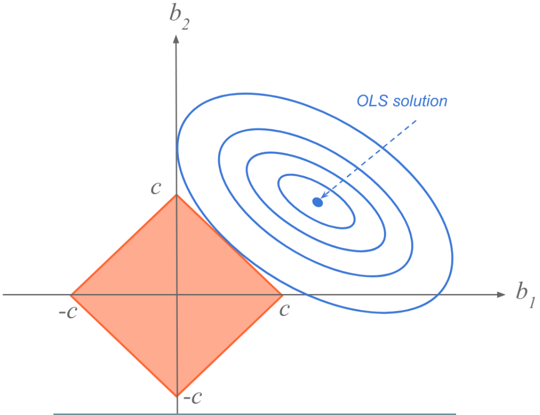
$$min_{\beta}E(\beta) = \sum_{i=1}^{n}(y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \quad (4)$$



Fuente: https://allmodelsarewrong.github.io

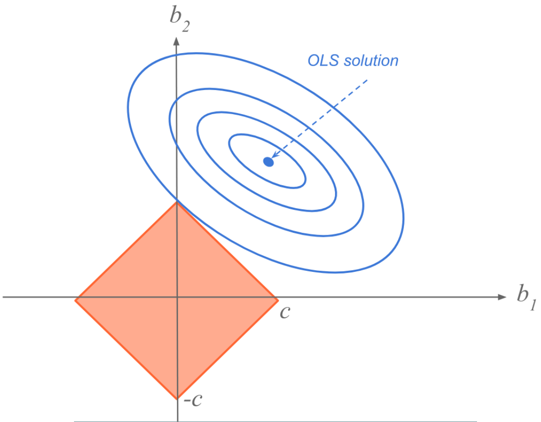
Intuición en 2 Dimensiones (Lasso)

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \text{ s.a } (|\beta_1| + |\beta_2|) \leq c \quad (5)$$



Intuición en 2 Dimensiones (Lasso)

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \text{ s.a } (|\beta_1| + |\beta_2|) \leq c \tag{6}$$



Comentarios técnicos

- Importante para aplicación:
  - Estandarizar los datos (media 0, y varianza 1)
  - Como elegimos λ?

Comentarios técnicos: selección de λ

- Como elegimos λ?
- λ es un parámetro y lo elegimos usando validación cruzada
  1. Partimos la muestra de entrenamiento en K Partes: $M_{train} = M_{fold\,1} \cup M_{fold\,2} \cdots \cup M_{fold\,K}$
  2. Cada conjunto $M_{fold\,K}$ va a jugar el rol de una muestra de evaluación $M_{eval\,k}$. Entonces para cada muestra
     - $M_{train-1} = M_{train} - M_{fold\,1}$
     - $\vdots$
     - $M_{train-k} = M_{train} - M_{fold\,k}$
  3. Luego hacemos el siguiente loop
     1. Para $\lambda_i = 0, 0.001, 0.002, \ldots, \lambda_{max}$
        - Para $k = 1, \ldots, K$
          - Ajustar el modelo $m_{i,k}$ con λ_i en $M_{train-k}$
          - Calcular y guardar el $MSE(m_{i,k})$ usando M_{eval-k}
        - fin para k
        - Calcular y guardar $MSE_i = \frac{1}{K} MSE(m_{i,k})$
     2. fin para λ
  4. Encontrar el menor MSE_i y usar ese $\lambda_i = \lambda^*$