Selección de Modelos y Regularización

Ciencia de Datos y Econometría Aplicada

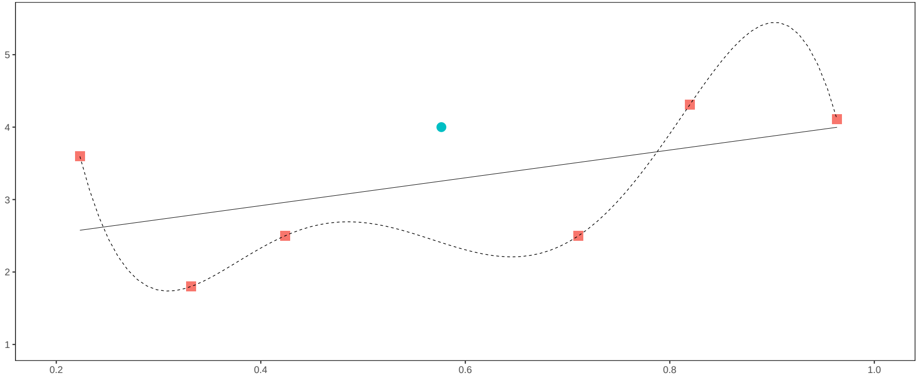
Ignacio Sarmiento-Barbieri

Universidad de los Andes

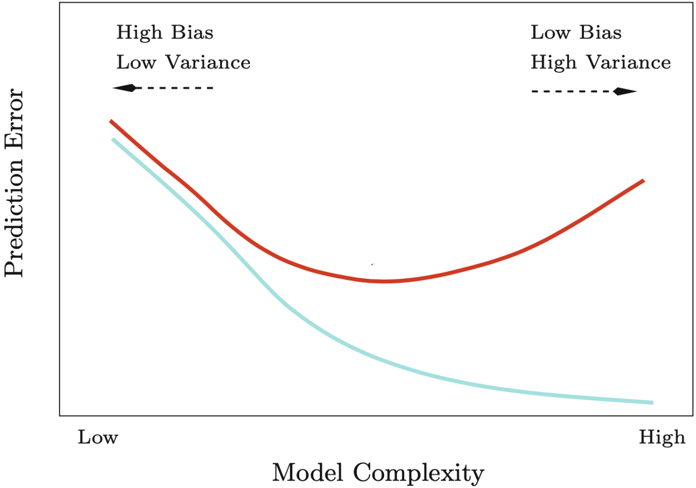
August 26, 2025

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)

Model Subset Selection

- We have M_k models
- We want to find the model that best predicts out of sample
- We have a number of ways to go about it
  - Best Subset Selection
  - Stepwise Selection
    - Forward selection
    - Backward selection

Best Subset Selection

$$y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + u \tag{1}$$

1. Estimate **all** possible models with $k = 0, 1, ..., p$ predictors.
2. Compute the prediction error using cross validation
3. Pick the one with the smallest prediction error

Best Subset Selection

1. Let M_0 denote the null model, which contains no predictors. This model simply predicts the sample mean for each observation.

2. For $k = 1, 2, \ldots, p$:

   1. Fit all $\binom{p}{k}$ models that contain exactly k predictors

   2. Pick the best among these $\binom{p}{k}$ models, and call it M_k. Where *best* is the one with the smallest SSR

3. Select a single best model from among $M_0, \ldots, M_p$ using cross-validated prediction error.

Stepwise Selection

- For computational reasons, best subset selection cannot be applied with very large p.
- Best subset selection may also suffer from statistical problems when p is large
- An enormous search space can lead to overfitting and high variance of the coefficient estimates.
- For both of these reasons, stepwise methods, which explore a far more restricted set of models, are attractive alternatives to best subset selection.

Stepwise Selection

1. Forward Stepwise Selection
   - Start with no predictors
   - Test all models with 1 predictor. Choose the best model
   - Add 1 predictor at a time, without taking away.
   - Of the p+1 models, choose the one with smallest prediction error using cross validation
2. Backward Stepwise Selection
   - Same idea but start with a complete model and go backwards, taking one at a time.

Forward Stepwise Selection

- Computational advantage over best subset selection is clear.
- It is not guaranteed to find the best possible model out of all 2^p models containing subsets of the p predictors.
- Drawback: once a predictor enters, it cannot leave.

Backward Stepwise Selection

- Like forward stepwise selection, the backward selection approach searches through only $1 + p(p+1)/2$ models
- However, unlike forward stepwise selection, it begins with the model containing all p predictors, and then iteratively removes the least useful predictor, one-at-a-time.
- Like forward stepwise selection, backward stepwise selection is not guaranteed to yield the best model containing a subset of the p predictors.
- Backward selection requires that the number of observations (samples) n is larger than the number of variables p (so that the full model can be fit).
- In contrast, forward stepwise can be used even when $n < p$, and so is the only viable subset method when p is very large.

Regularización

Lasso

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

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

Lasso

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

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

  - "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

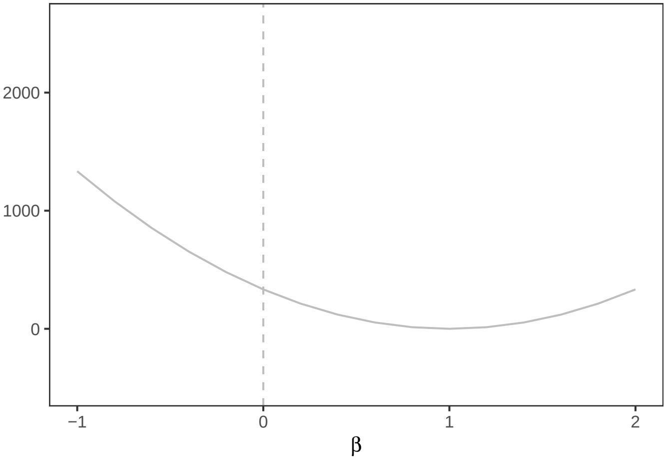
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i \beta)^2 + \lambda |\beta| \tag{3}$$

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

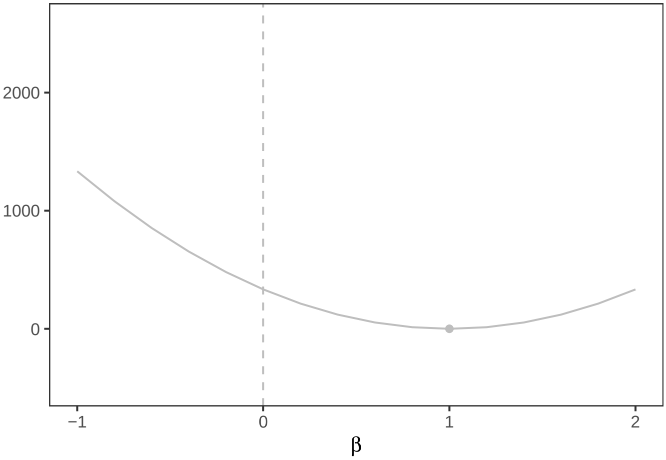
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i \beta)^2 \tag{4}$$

- la solución es?

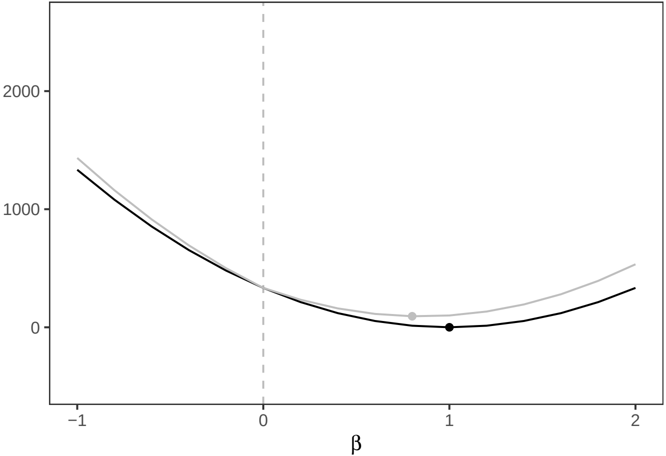
Intuición en 1 Dimension



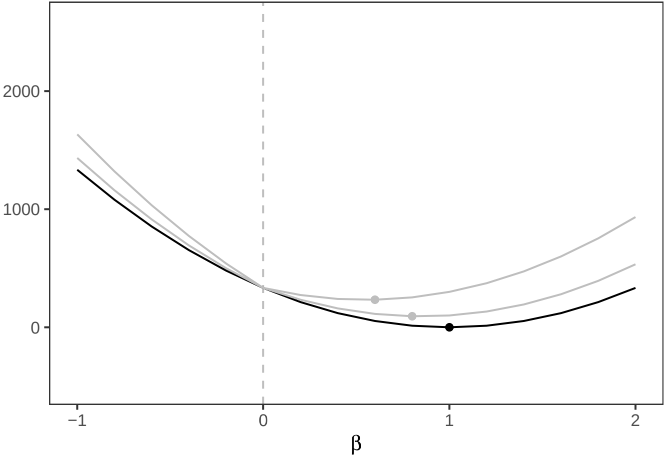
Intuición en 1 Dimension



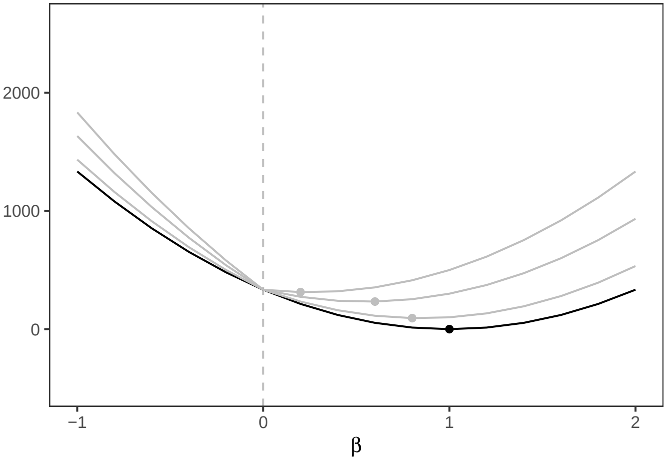
Intuición en 1 Dimension



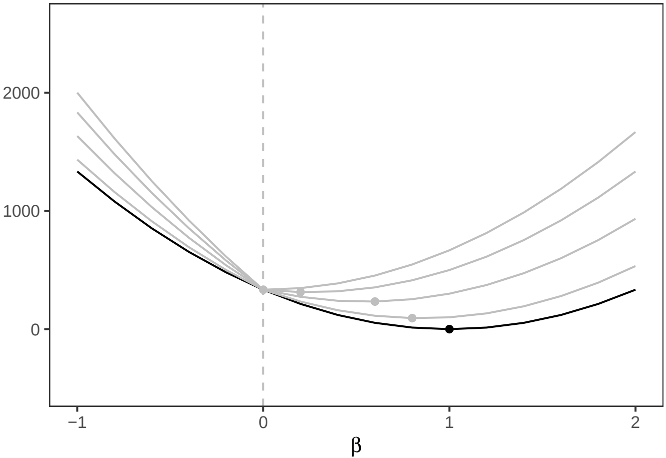
Intuición en 1 Dimension



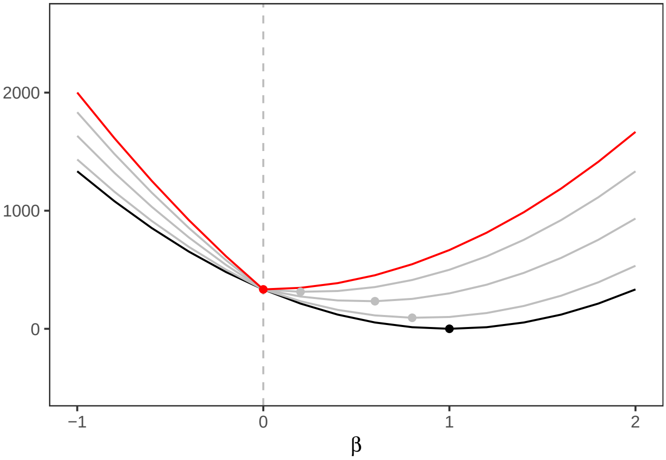
Intuición en 1 Dimension



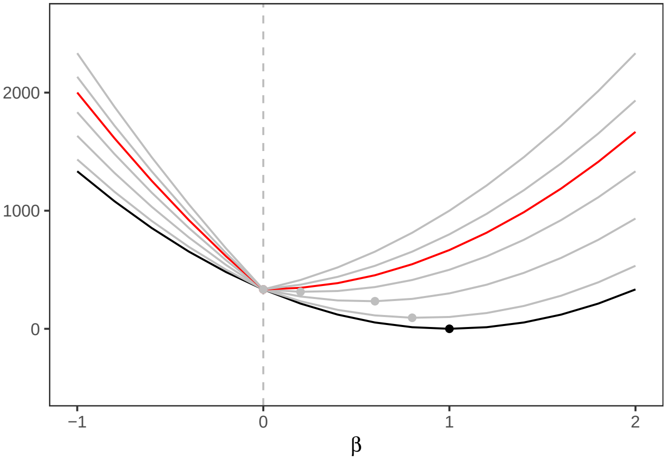
Intuición en 1 Dimension



Intuición en 1 Dimension



Intuición en 1 Dimension



Intuición en 1 Dimension

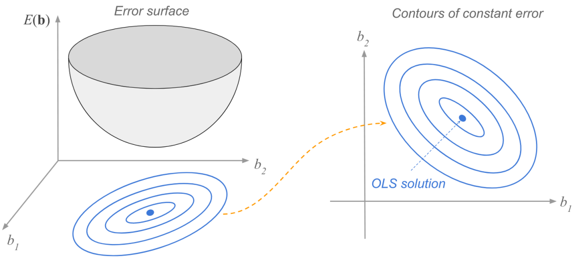
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda|\beta| \tag{5}$$

la solución analítica es

$$\hat{\beta}_{lasso} = \begin{cases} 0 & \text{si } \lambda \geq \lambda^* \\ \hat{\beta}_{OLS} - \frac{\lambda}{2} & \text{si } \lambda < \lambda^* \end{cases} \tag{6}$$

Intuición en 2 Dimensiones (OLS)

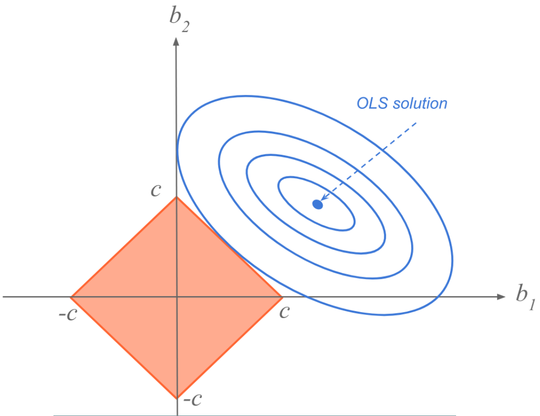
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \tag{7}$$



Fuente: https://allmodelsarewrong.github.io

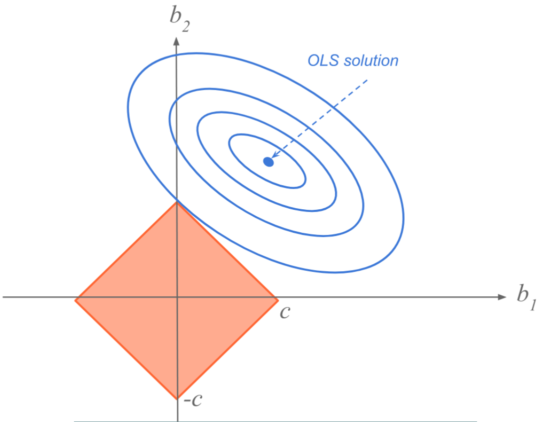
Intuición en 2 Dimensiones (Lasso)

$$min_{\beta}E(\beta) = \sum_{i=1}^{n}(y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \text{ s.a } (|\beta_1| + |\beta_2|) \leq c \quad (8)$$



Intuición en 2 Dimensiones (Lasso)

$$min_\beta E(\beta) = \sum_{i=1}^{n} (y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \text{ s.a } (|\beta_1| + |\beta_2|) \leq c \quad (9)$$



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 $\lambda_i = \lambda^*$