

Selección de Modelos y Regularización

Big Data y Machine Learning para Economía Aplicada

Ignacio Sarmiento-Barbieri

Universidad de los Andes

Agenda

1 Recap: Predicción y Overfit

2 Regularización

- Recap: OLS Mechanics
- Ridge

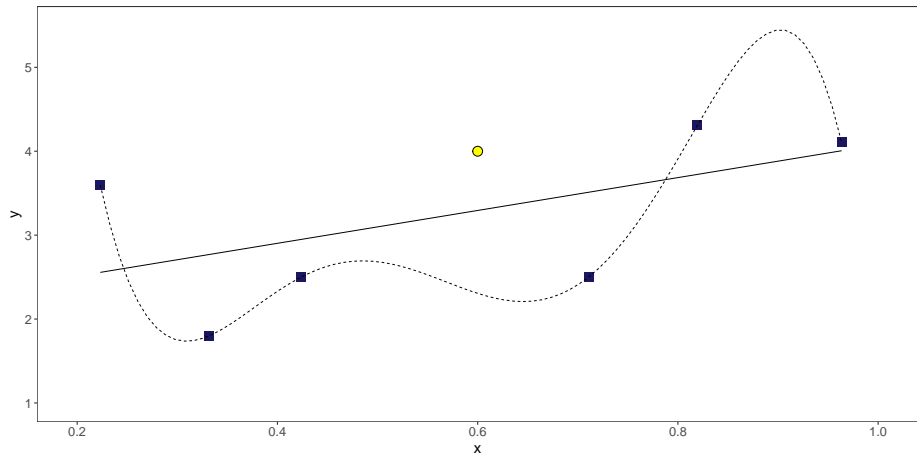
Agenda

1 Recap: Predicción y Overfit

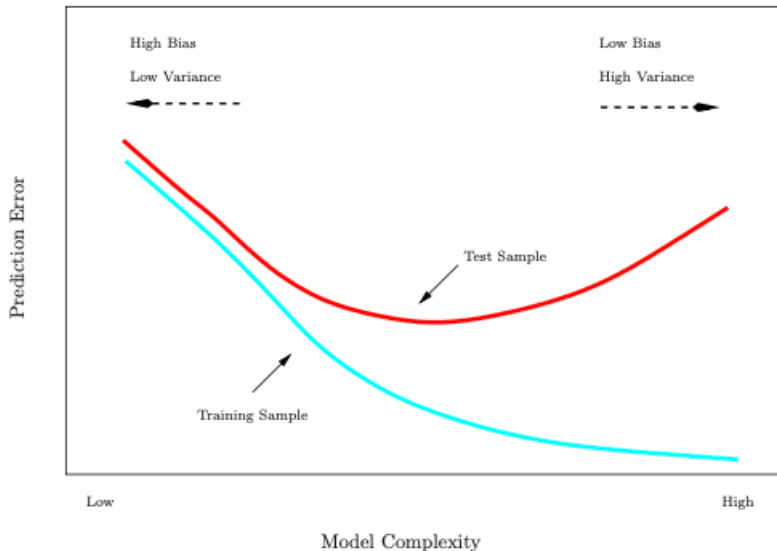
2 Regularización

- Recap: OLS Mechanics
- Ridge

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 fuera de muestra (out-of-sample)
 - ▶ Métodos de Remuestreo
 - ▶ Enfoque del conjunto de validación
 - ▶ LOOCV
 - ▶ Validación cruzada en K-partes (5 o 10)



photo from <https://www.dailydot.com/parsec/batman-1966-labels-tumblr-twitter-vine/>

Agenda

1 Recap: Predicción y Overfit

2 Regularización

- Recap: OLS Mechanics
- Ridge

Agenda

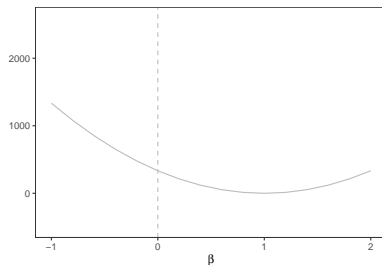
1 Recap: Predicción y Overfit

2 Regularización

- Recap: OLS Mechanics
- Ridge

OLS 1 Dimension

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



Solution

$$\frac{\partial E(\beta)}{\partial \beta} = \sum_{i=1}^n 2(y_i - x_i \beta)(-x_i) \quad (2)$$

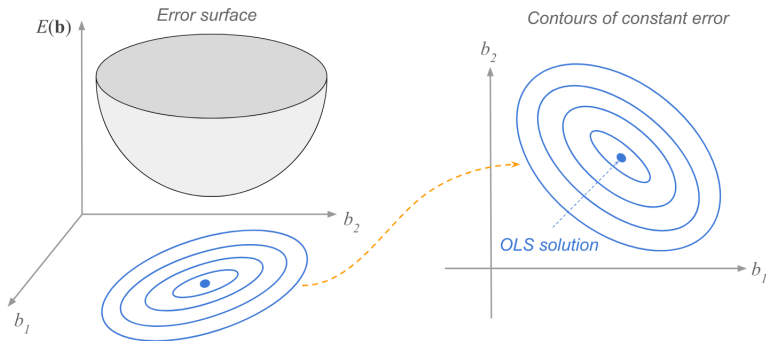
OLS 1 Dimension

$$\frac{\partial E(\beta)}{\partial \beta} = \sum_{i=1}^n 2(y_i - x_i\beta)(-x_i) \quad (3)$$

$$\hat{\beta} = \frac{\sum y_i x_i}{\sum x_i^2} \quad (4)$$

OLS 2 Dimensiones

$$\min_{\beta} E(\beta) = \sum_{i=1}^n (y_i - x_{i1}\beta_1 - x_{i2}\beta_2)^2 \quad (5)$$



Fuente: <https://allmodelsarewrong.github.io>

Agenda

1 Recap: Predicción y Overfit

2 Regularización

- Recap: OLS Mechanics
- Ridge

Ridge

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

$$\min_{\beta} E(\beta) = \sum_{i=1}^n (y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 + \lambda \sum_{j=1}^p (\beta_j)^2 \quad (6)$$

Ridge: Intuición en 1 Dimension

- ▶ 1 predictor estandarizado
- ▶ El problema entonces es

$$\min_{\beta} E(\beta) = \sum_{i=1}^n (y_i - x_i \beta)^2 + \lambda \beta^2 \quad (7)$$

- ▶ La solución?

Ridge: Intuición en 2 Dimensiones

- Al problema en 2 dimensiones podemos escribirlo como

$$\min_{\beta} E(\beta) = \sum_{i=1}^n (y_i - x_{i1}\beta_1 - x_{i2}\beta_2 + \lambda (\beta_1^2 + \beta_2^2)) \quad (8)$$

- podemos escribirlo como un problema de optimización restringido

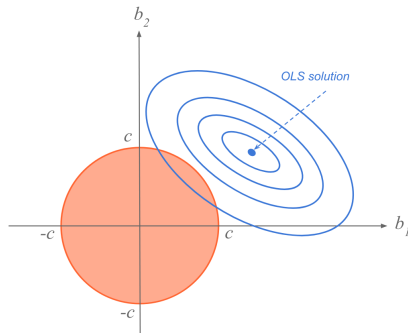
$$\min_{\beta} E(\beta) = \sum_{i=1}^n (y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \quad (9)$$

sujeto a

$$((\beta_1)^2 + (\beta_2)^2) \leq c$$

Ridge: Intuición en 2 Dimensiones

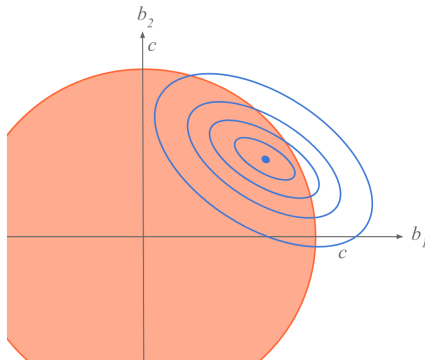
$$\min_{\beta} E(\beta) = \sum_{i=1}^n (y_i - x_{i1}\beta_1 - x_{i2}\beta_2)^2 \text{ s.a. } ((\beta_1)^2 + (\beta_2)^2) \leq c \quad (10)$$



Fuente: <https://allmodelsarewrong.github.io>

Ridge: Intuición en 2 Dimensiones

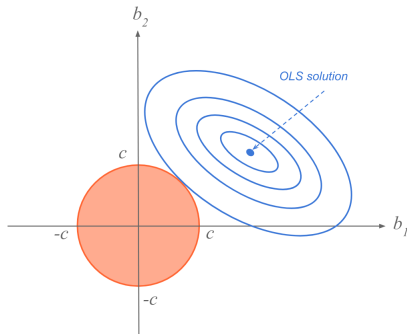
$$\min_{\beta} E(\beta) = \sum_{i=1}^n (y_i - x_{i1}\beta_1 - x_{i2}\beta_2)^2 \text{ s.a. } ((\beta_1)^2 + (\beta_2)^2) \leq c \quad (11)$$



Fuente: <https://allmodelsarewrong.github.io>

Ridge: Intuición en 2 Dimensiones

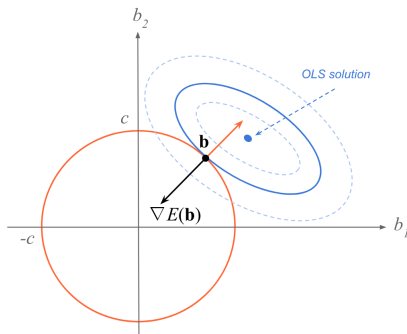
$$\min_{\beta} E(\beta) = \sum_{i=1}^n (y_i - x_{i1}\beta_1 - x_{i2}\beta_2)^2 \text{ s.a. } ((\beta_1)^2 + (\beta_2)^2) \leq c \quad (12)$$



Fuente: <https://allmodelsarewrong.github.io>

Ridge: Intuición en 2 Dimensiones

$$\min_{\beta} E(\beta) = \sum_{i=1}^n (y_i - x_{i1}\beta_1 - x_{i2}\beta_2)^2 \text{ s.a. } ((\beta_1)^2 + (\beta_2)^2) \leq c \quad (13)$$



Fuente: <https://allmodelsarewrong.github.io>

Formalmente

- ▶ Las X s están centradas, media cero.

- ▶ OLS

- ▶ Regresión Simple $y = \beta x + u$

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2}$$

- ▶ Regresión múltiple y álgebra matricial

$$\hat{\beta}_{OLS} = (X'X)^{-1}X'y$$

- ▶ Ridge

$$\hat{\beta}_{ridge} = (X'X + \lambda I_p)^{-1}X'y$$

Formalmente

$$\hat{\beta}_{ridge} = (X'X + \lambda I_p)^{-1} X'y$$

- ▶ Ridge es sesgado $E(\hat{\beta}_{ridge}) \neq \beta$
- ▶ Pero la varianza es menor que la de OLS
- ▶ Para ciertos valores del parámetro λ $MSE_{OLS} > MSE_{ridge}$

Consecuencias: Ridge no es invariante a las escala

- ▶ La escala de las variables importa en Ridge, mientras que en OLS no.
- ▶ Supongamos $z = c * x$
- ▶ Vamos a mostrar que $\hat{y}_i^z = \hat{y}_i^x$
- ▶ Partamos de hacer la regression

$$y_i = \beta_0^z + \beta_1^z z_i + u \quad (14)$$

$$\hat{\beta}_1^z = \frac{\sum (z_i - \bar{z})(y_i - \bar{y})}{\sum (z_i - \bar{z})^2} \quad (15)$$

- ▶ Pero $z = c * x$

$$\hat{\beta}_1^z = \frac{\sum (cx_i - c\bar{x})(y_i - \bar{y})}{\sum (cx_i - c\bar{x})^2} \quad (16)$$

$$= \frac{1}{c} \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \quad (17)$$

$$= \frac{1}{c} \hat{\beta}_1^x \quad (18)$$

Consecuencias: Ridge no es invariante a las escala

► Entonces

$$\hat{\beta}_1^z z_i = \hat{\beta}_1^z c x_i \quad (19)$$

$$= \frac{1}{c} \hat{\beta}_1^x c x_i \quad (20)$$

$$= \hat{\beta}_1^x x_i \quad (21)$$

$$(22)$$

Consecuencias: Ridge no es invariante a las escala

- En términos generales, si $Z = cX$

$$\begin{aligned}\hat{\beta}_{OLS}^Z &= (Z'Z)^{-1}Z'y \\ &= ((cX)'(cX))^{-1}(cX)'y \\ &= \frac{c}{c^2}(X'X)^{-1}X'y \\ &= \frac{1}{c}(X'X)^{-1}X'y \\ &= \frac{1}{c}\hat{\beta}_{OLS}^X\end{aligned}$$

Consecuencias: Ridge no es invariante a las escala

- ▶ Entonces

$$\begin{aligned}\hat{\beta}_{OLS}^Z Z &= \frac{1}{c} \hat{\beta}_{OLS}^X cX \\ &= \hat{\beta}_{OLS}^X X\end{aligned}$$

- ▶ Con Ridge esto no funciona

$$\hat{\beta}_{Ridge}^Z Z \neq \hat{\beta}_{Ridge}^X X$$

- ▶ Es importante estandarizar las variables (la mayoría de los softwares lo hace automáticamente)

Example: Ridge



photo from <https://www.dailydot.com/parsec/batman-1966-labels-tumblr-twitter-vine/>