Selección de Modelos y Regularización

Big Data y Machine Learning para Economía Aplicada

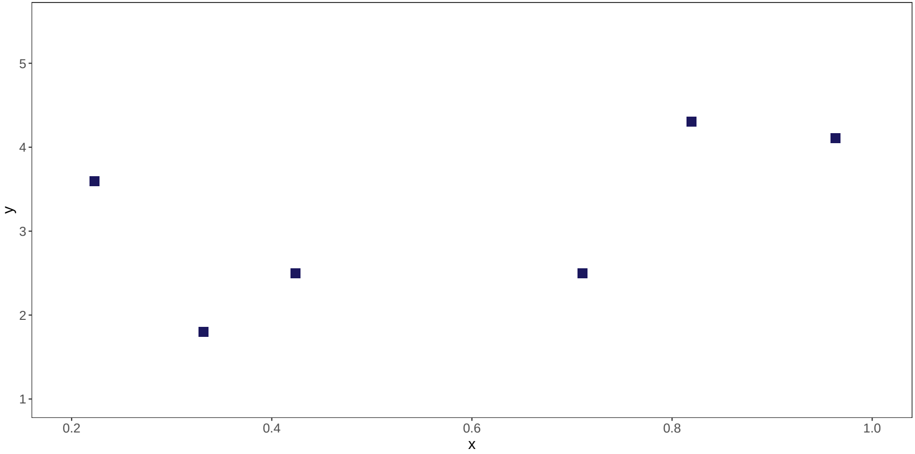
Ignacio Sarmiento-Barbieri

Universidad de los Andes

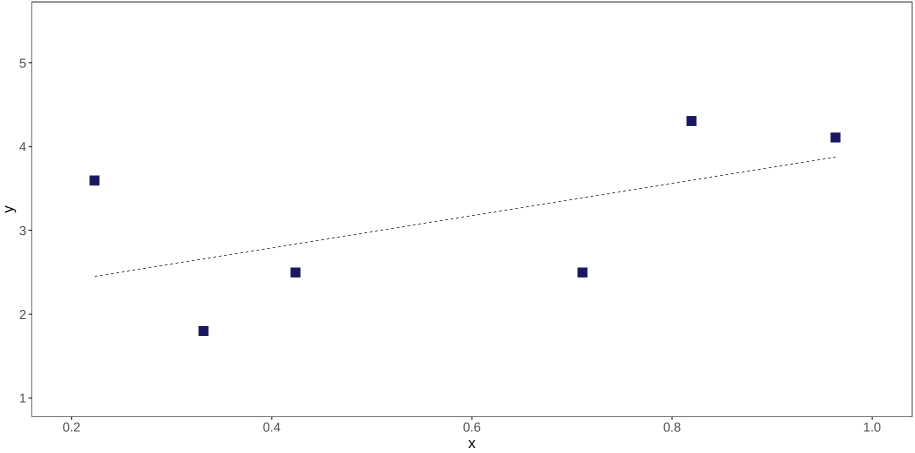
Agenda

Agenda

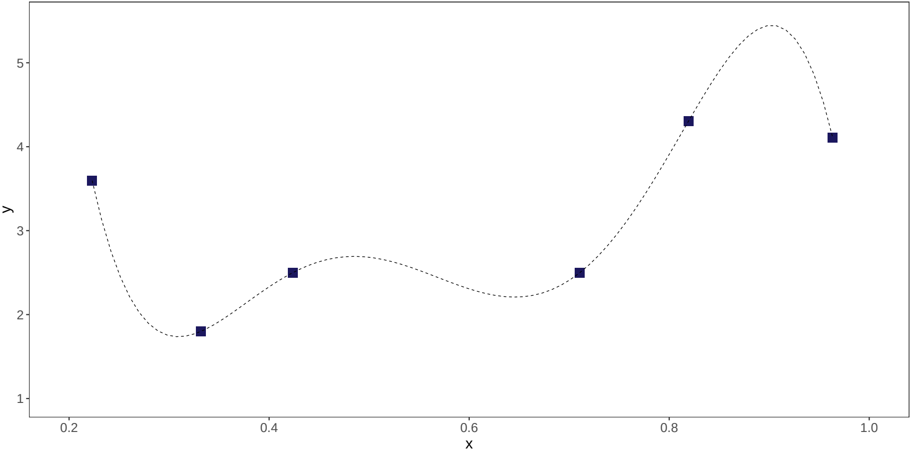
Overfit y Predicción fuera de Muestra



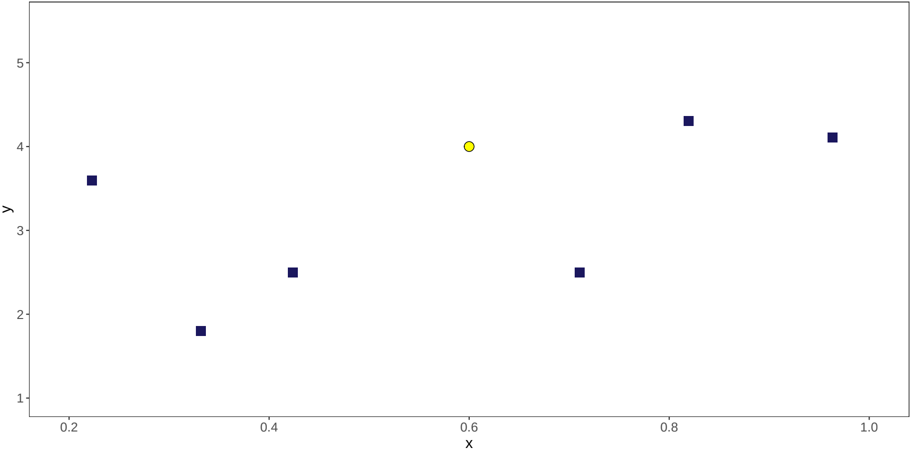
Overfit y Predicción fuera de Muestra



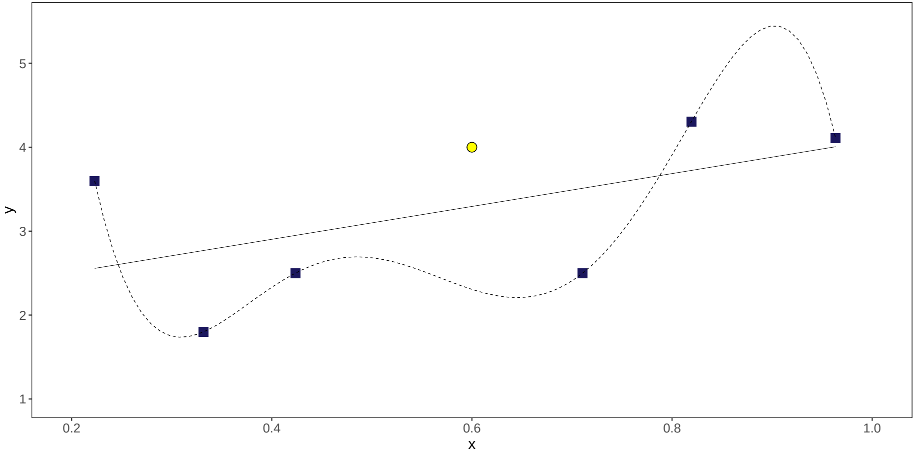
Overfit y Predicción fuera de Muestra



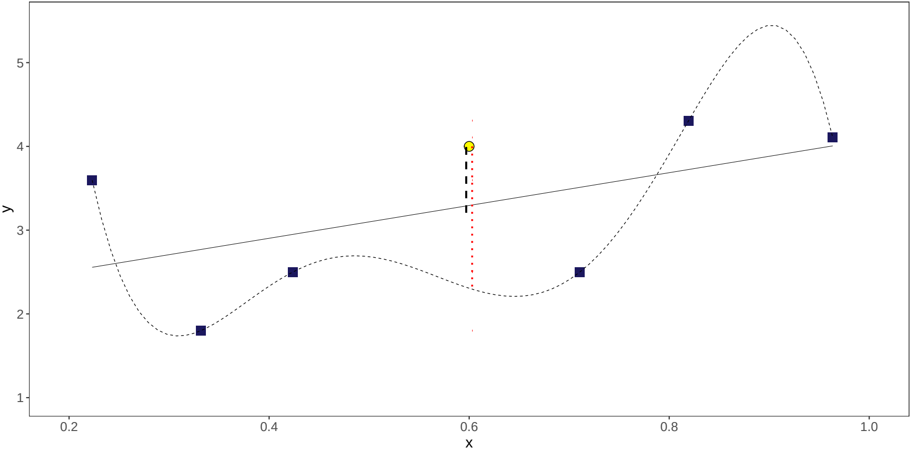
Overfit y Predicción fuera de Muestra



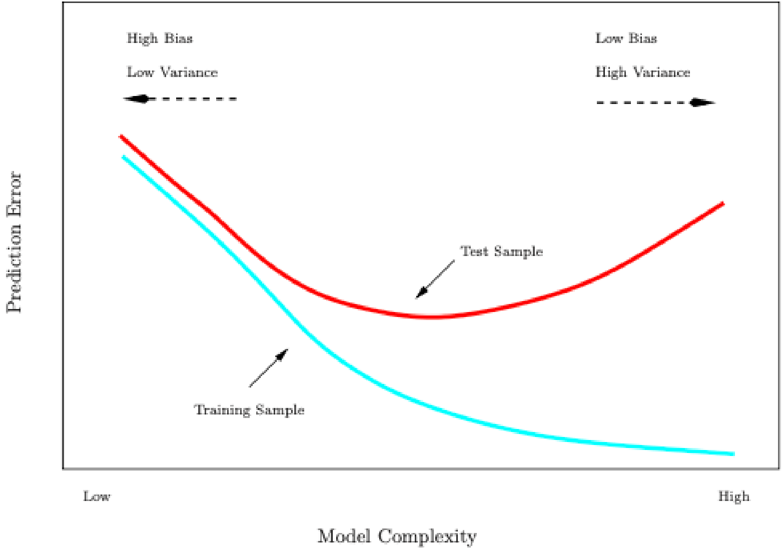
Overfit y Predicción fuera de Muestra



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: 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)

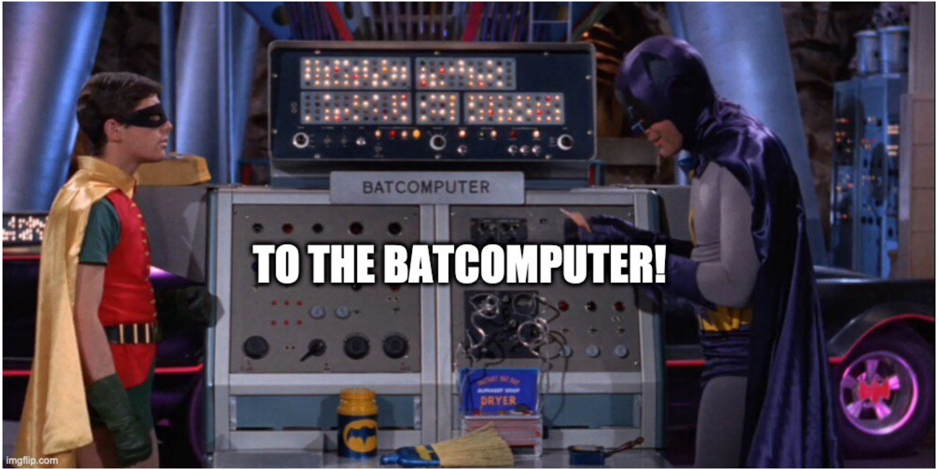


photo from https://www.dailydot.com/parsec/batman-1966-labels-tumblr-twitter-vine/

Agenda

1. Recap: Predicción y Overfit

2. Regularización
   - Recap: OLS Mechanics
   - Ridge

Regularización: Motivación

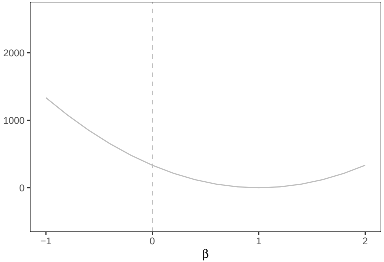
- Las técnicas econometricas estándar no están optimizadas para la predicción porque se enfocan en la insesgadez.
- OLS por ejemplo es el mejor estimador lineal *insesgado*
- OLS minimiza el error *"dentro de muestra"*, eligiendo β de forma tal que

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

Agenda

OLS 1 Dimension

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



Solución

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

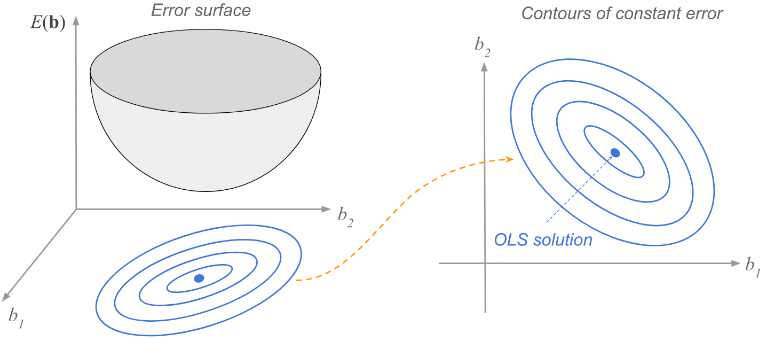
OLS 1 Dimension

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

$$\hat{\beta} = \frac{\sum y_i x_i}{\sum x_i^2} \tag{5}$$

OLS 2 Dimensiones

$$min_\beta E(\beta) = \sum_{i=1}^{n} (y_i - x_{i1}\beta_1 - x_{i2}\beta_2)^2 \tag{6}$$



Fuente: https://allmodelsarewrong.github.io

App

Regularización: Motivación

- Las técnicas econometricas estándar no están optimizadas para la predicción porque se enfocan en la insesgadez.
- OLS por ejemplo es el mejor estimador lineal *insesgado*
- OLS minimiza el error *"dentro de muestra"*, eligiendo β de forma tal que

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

- pero para predicción, no estamos interesados en hacer un buen trabajo dentro de muestra
- Queremos hacer un buen trabajo, **fuera de muestra**

Agenda

Ridge

- Asegurar cero sesgo dentro de muestra crea problemas fuera de muestra: trade-off Sesgo-Varianza
- Las técnicas de machine learning fueron desarrolladas para hacer este trade-off de forma empírica.
- Vamos a proponer modelos del estilo

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 + \lambda \sum_{j=1}^{p} R(\beta_j) \tag{7}$$

- donde R es un regularizador que penaliza funciones que crean varianza
- Explícitamente en la minimización incluimos un termino de sesgo y un termino de varianza.

Ridge

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

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 + \lambda \sum_{j=1}^{p} (\beta_j)^2 \tag{8}$$

Ridge: Intuición en 1 Dimension

- 1 predictor estandarizado
- El problema entonces es

$$min_{\beta}E(\beta) = \sum_{i=1}^{n}(y_i - x_i\beta)^2 + \lambda\beta^2 \tag{9}$$

- La solución?

Ridge: Intuición en 2 Dimensiones

- Al problema en 2 dimensiones podemos escribirlo como

$$min_\beta E(\beta) = \sum_{i=1}^{n} (y_i - x_{i1}\beta_1 - x_{i2}\beta_2 + \lambda \left(\beta_1^2 + \beta_2^2\right) \tag{10}$$

- podemos escribirlo como un problema de optimización restringido

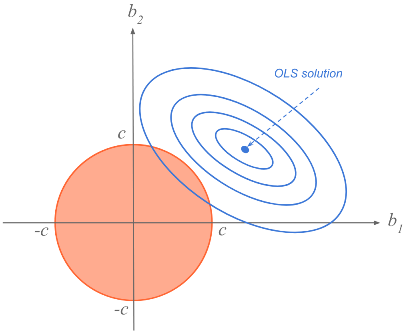
$$min_\beta E(\beta) = \sum_{i=1}^{n} (y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \tag{11}$$

sujeto a

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

Ridge: Intuición en 2 Dimensiones

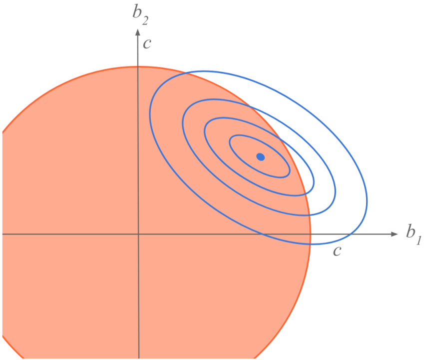
$$min_{\beta}E(\beta) = \sum_{i=1}^{n}(y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \text{ s.a } \left((\beta_1)^2 + (\beta_2)^2\right) \leq c \tag{12}$$



Fuente: https://allmodelsarewrong.github.io

Ridge: Intuición en 2 Dimensiones

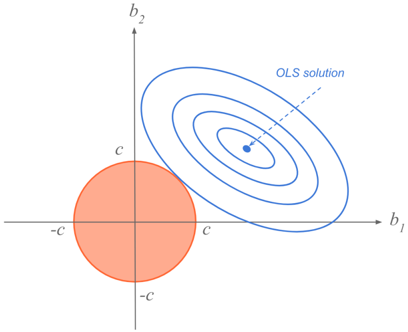
$$min_{\beta}E(\beta) = \sum_{i=1}^{n}(y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \text{ s.a } \left((\beta_1)^2 + (\beta_2)^2\right) \leq c \tag{13}$$



Fuente: https://allmodelsarewrong.github.io

Ridge: Intuición en 2 Dimensiones

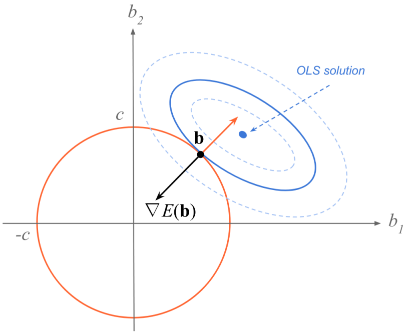
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \text{ s.a } \left((\beta_1)^2 + (\beta_2)^2\right) \leq c \quad (14)$$



Fuente: https://allmodelsarewrong.github.io

Ridge: Intuición en 2 Dimensiones

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \text{ s.a } \left((\beta_1)^2 + (\beta_2)^2\right) \leq c \tag{15}$$



Fuente: https://allmodelsarewrong.github.io

Recap

- Asegurar cero sesgo dentro de muestra crea problemas fuera de muestra: trade-off Sesgo-Varianza
- Las técnicas de machine learning fueron desarrolladas para hacer este trade-off de forma empírica.

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 + \lambda \sum_{j=1}^{p} R(\beta_j) \tag{16}$$

- λ es el precio al que hacemos este trade off
- Próxima clase: como elegimos este λ