Regularización: Lasso y Elastic Net

Big Data y Machine Learning para Economía Aplicada

Ignacio Sarmiento-Barbieri

Universidad de los Andes

Agenda

Agenda

Agenda

Regularización: Motivación

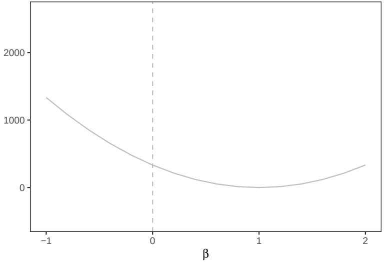
- Las técnicas econometricas estándar no están optimizadas para la predicción porque se enfocan en la insesgadez.
- OLS por ejemplo es el mejor estimador lineal *insesgado*
- OLS minimiza el error *"dentro de muestra"*, eligiendo β de forma tal que

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

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

OLS 1 Dimension

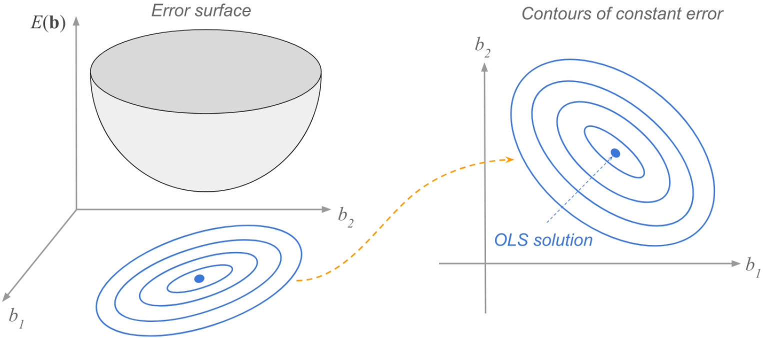
$$min_{\beta}E(\beta) = \sum_{i=1}^{n}(y_i - x_i\beta)^2 \tag{2}$$



App

OLS 2 Dimensiones

$$min_{\beta}E(\beta) = \sum_{i=1}^{n}(y_i - x_{i1}\beta_1 - x_{i2}\beta_2)^2 \tag{3}$$



Fuente: https://allmodelsarewrong.github.io

App

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

$$min_{\beta}E(\beta) = \sum_{i=1}^{n}(y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 + \lambda\sum_{j=1}^{p}R(\beta_j) \tag{4}$$

- donde R es un regularizador que penaliza funciones que crean varianza

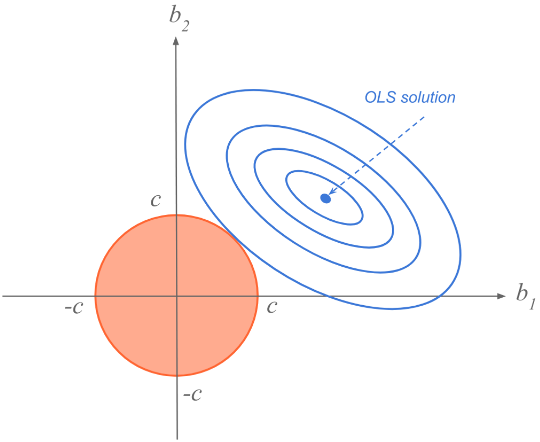
Ridge

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

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 + \lambda \sum_{j=1}^{p} (\beta_j)^2 \tag{5}$$

Intuición en 2 Dimensiones (Ridge)

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



Fuente: https://allmodelsarewrong.github.io

Ridge as Data Augmentation (1)

RidgeDataAug

- Add λ additional points

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

Ridge as Data Augmentation (2)

RidgeDataAug

- Add a single point

$$\sum_{i=1}^{n}(y_i - x_i\beta)^2 + \lambda\beta^2 = \tag{8}$$

Agenda

Lasso

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

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{9}$$

Lasso

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

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{9}$$

  - "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

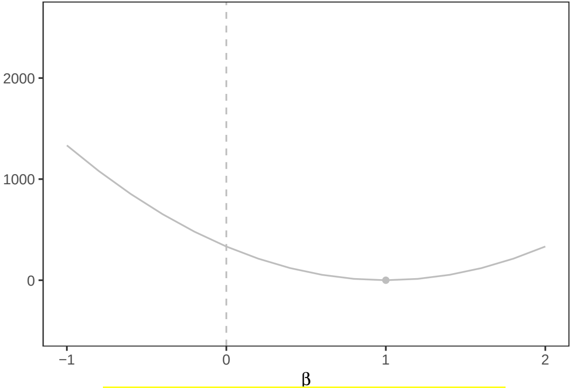
Intuición en 1 Dimensión

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda|\beta| \tag{10}$$

Intuición en 1 Dimensión

$\hat{\beta} > 0$

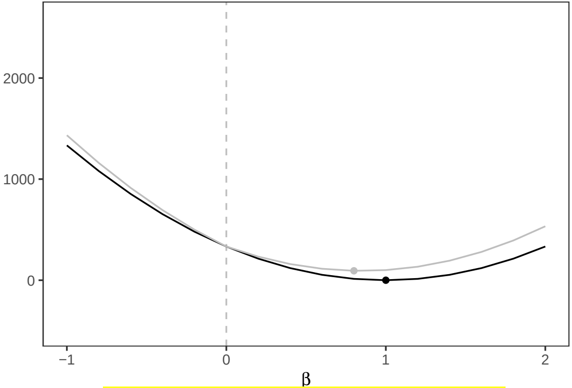
$$min_\beta E(\beta) = \sum_{i=1}^{n}(y_i - x_i\beta)^2 + \lambda\beta \tag{11}$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

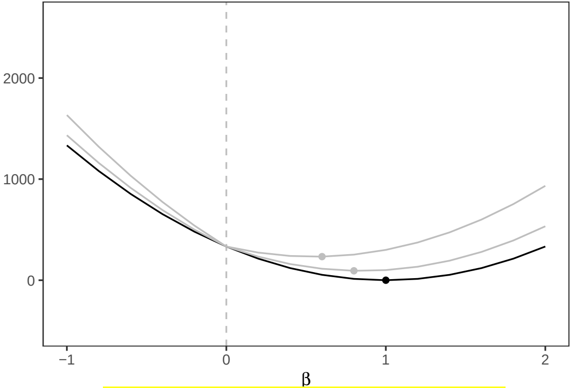
$$min_{\beta}E(\beta) = \sum_{i=1}^{n}(y_i - x_i\beta)^2 + \lambda\beta \tag{12}$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

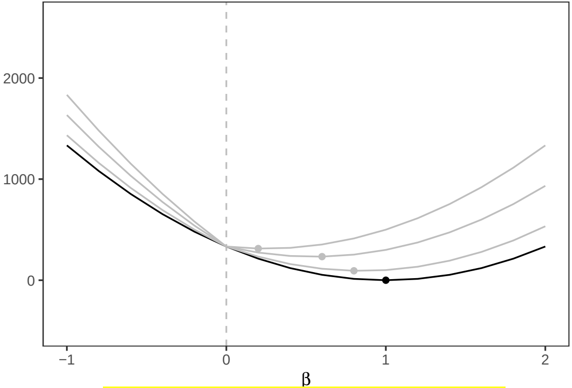
$$min_{\beta}E(\beta) = \sum_{i=1}^{n}(y_i - x_i\beta)^2 + \lambda\beta \quad (13)$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

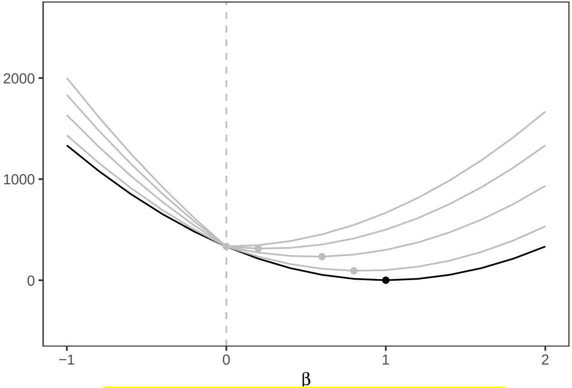
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i \beta)^2 + \lambda \beta \tag{14}$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

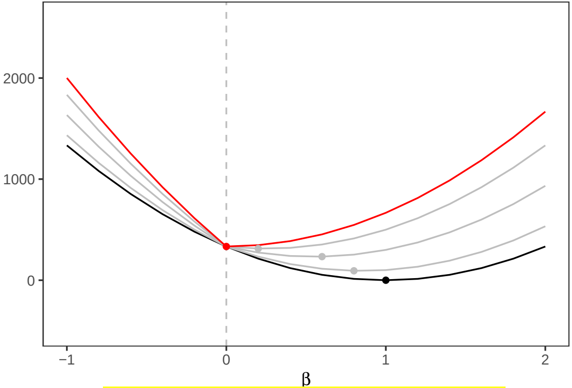
$$min_\beta E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda\beta \tag{15}$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

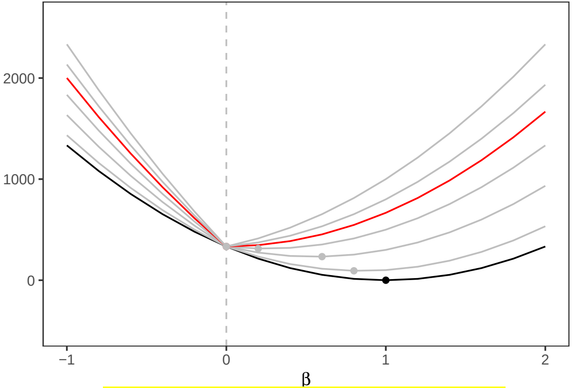
$$min_\beta E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda\beta \tag{16}$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

$$min_\beta E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda\beta \tag{17}$$



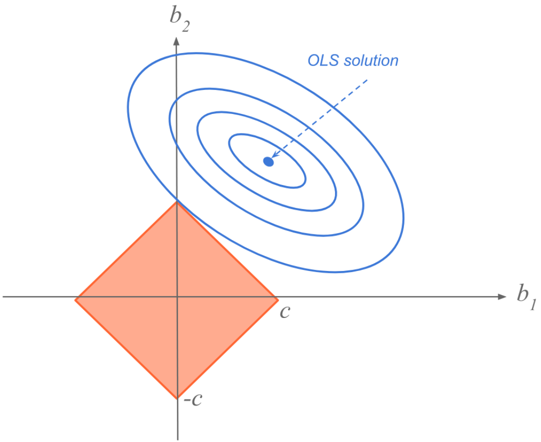
Intuición en 1 Dimension

Solución analitica

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda|\beta| \tag{18}$$

Intuición en 2 Dimensiones (Lasso)

$$min_{\beta}E(\beta) = \sum_{i=1}^{n}(y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \text{ s.a } (|\beta_1| + |\beta_2|) \leq c \quad (19)$$



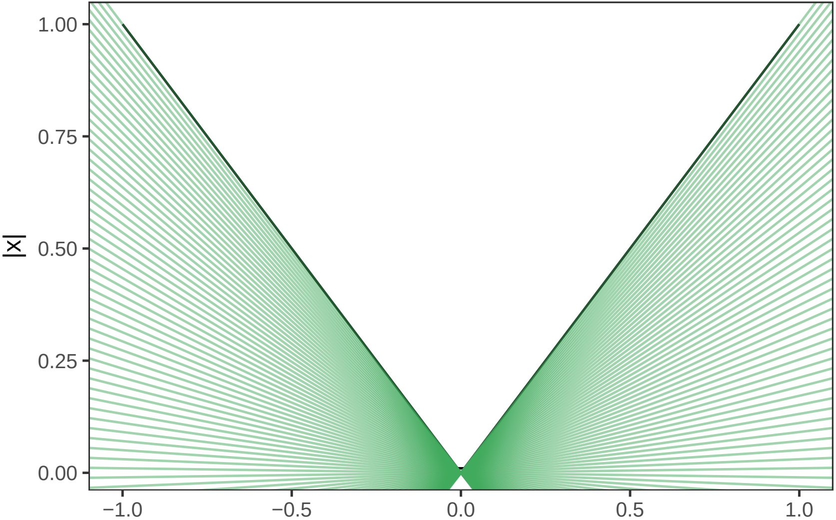
Example

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

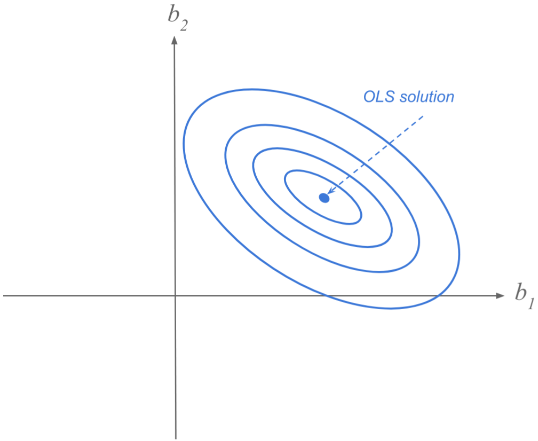
Resumen

- Ridge y Lasso son sesgados, pero las disminuciones en varianza pueden compensar estoy y llevar a un MSE menor
- Lasso encoje a cero, Ridge no tanto
- Importante para aplicación:
  - Estandarizar los datos
  - Elegimos $\lambda \rightarrow$ Validación cruzada

Subgradientes



Coordinate Descent



Agenda

Ridge and Lasso: The good and the bad

- Objective 1: Accuracy
  - Minimize prediction error (in one step) $\rightarrow$ Ridge, Lasso
- Objective 2: Dimensionality
  - Reduce the predictor space $\rightarrow$ Lasso's free lunch
- More predictors than observations ($k > n$)
  - OLS fails
  - Ridge augments data
  - Lasso chooses at most n variables

Ridge and Lasso: The good and the bad

OLS when $k > n$

- Rank? Max number of rows or columns that are linearly independent
  - Implies $rank(X_{n\times k}) \leq min(k,n)$
- MCO we need $rank(X_{n\times k}) = k \implies k \leq n$
- If $rank(X_{n\times k}) = k$ then $rank(X'X) = k$
- If $k > n$, then $rank(X'X) \leq n < k$ then $(X'X)$ cannot be inverted
- Ridge works when $k \geq n$

Ridge and Lasso: The good and the bad

Ridge when $k > n$

$$min_{\beta}E(\beta) = \sum_{i=1}^{n}(y_i - \sum_{j=1}^{k} x'_{ij}\beta_j)^2 + \lambda(\sum_{j=1}^{k}\beta_j)^2 \tag{20}$$

- Solution $\rightarrow$ data augmentation
- Intuition: Ridge "adds" k additional points.
- Allows us to "deal" with $k \geq n$

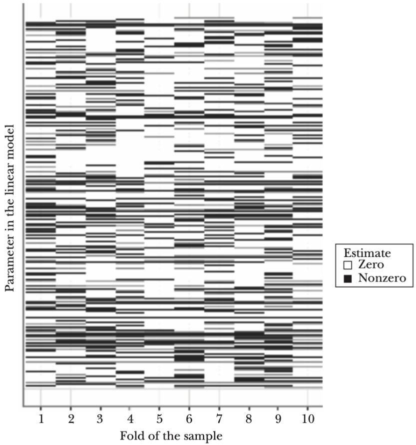
Ridge and Lasso: The good and the bad

Ridge when $k > n$

Ridge and Lasso: The good and the bad

- When we have a group of highly correlated variables,
  - Lasso chooses only one.

Ridge and Lasso: The good and the bad



Ridge and Lasso: The good and the bad

- When we have a group of highly correlated variables,
  - Lasso chooses only one. Makes it unstable for prediction.
  - Ridge shrinks the coefficients of correlated variables toward each other. This makes Ridge "work" better than Lasso. "Work" in terms of prediction error

Agenda

Family of penalized regressions

$$min_{\beta}R(\beta) = \sum_{i=1}^{n}(y_i - x_i'\beta)^2 + \lambda \sum_{s=2}^{p} |\beta_s|^q \quad (21)$$

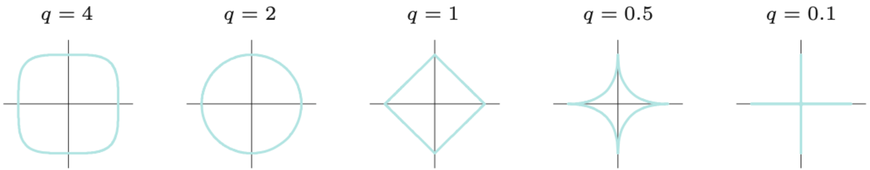


FIGURE 3.12. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for given values of* q.

Agenda

Elastic net

$$min_{\beta} EN(\beta) = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + \frac{(1-\alpha)}{2} \sum_{j=1}^{p} (\beta_j)^2 \right) \tag{22}$$

- Si $\alpha = 1$ Lasso
- Si $\alpha = 0$ Ridge

Elastic Net

- Elastic net: happy medium.
  - Good job at prediction and selecting variables

$$min_{\beta}EN(\beta) = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + \frac{(1-\alpha)}{2} \sum_{j=1}^{p} (\beta_j)^2 \right) \tag{23}$$

- Mixes Ridge and Lasso
- Lasso selects predictors
- Strict convexity part of the penalty (ridge) solves the grouping instability problem
- How to choose (λ, α)? $\rightarrow$ Bidimensional Crossvalidation
- Recomended lecture: Zou, H. & Hastie, T. (2005)

Example

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