Selección de Modelos y Regularización

Big Data y Machine Learning para Economía Aplicada

Ignacio Sarmiento-Barbieri

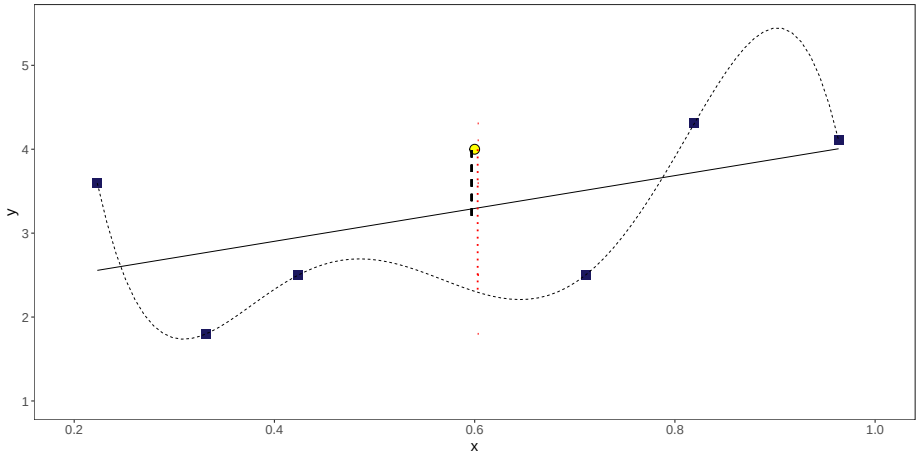
Universidad de los Andes

Agenda

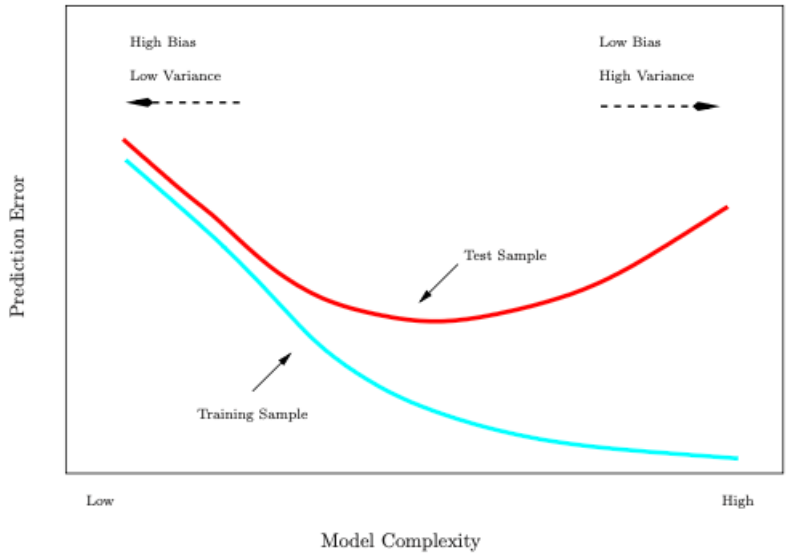
1. Recap: Predicción y Overfit
2. Regularización
   - Recap: OLS Mechanics
   - Ridge
   - Escala de las variables
   - Selección de λ
   - Ridge as Data Augmentation
3. Lasso

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: 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)
  - Penalización ex-post: AIC, BIC, R2 ajustado, etc
  - Métodos de Remuestreo
    - Enfoque del conjunto de validación
    - LOOCV
    - Validación cruzada en K-partes (5 o 10)

Agenda

1. Recap: Predicción y Overfit

2. Regularización
   - Recap: OLS Mechanics
   - Ridge
   - Escala de las variables
   - Selección de λ
   - Ridge as Data Augmentation

3. Lasso

Agenda

1. Recap: Predicción y Overfit
2. Regularización
   - Recap: OLS Mechanics
   - Ridge
   - Escala de las variables
   - Selección de λ
   - Ridge as Data Augmentation
3. Lasso

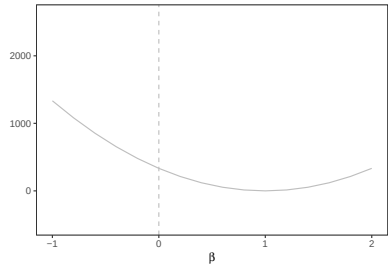
Regularización: Motivación

- 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

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

OLS 1 Dimension

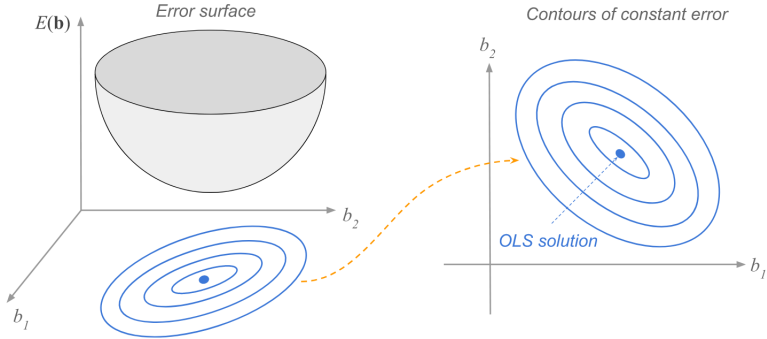
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 \tag{2}$$



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

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

1. Recap: Predicción y Overfit
2. Regularización
   - Recap: OLS Mechanics
   - Ridge
   - Escala de las variables
   - Selección de λ
   - Ridge as Data Augmentation
3. Lasso

Regularización

- 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

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

Ridge: Intuición en 1 Dimension

- 1 predictor estandarizado
- El problema:

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda\beta^2 \tag{6}$$

- La solución?

App

Ridge: Intuición en 1 Dimension

Problema como optimización restringida

- Existe un $c \geq 0$ tal que $\hat{\beta}(\lambda)$ es la solución a

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

$$\text{sujeto a}$$

$$(\beta)^2 \leq c$$

Ridge: Intuición en 2 Dimensiones

- Al problema en 2 dimensiones podemos escribirlo como

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

- podemos escribirlo como un problema de optimización restringido

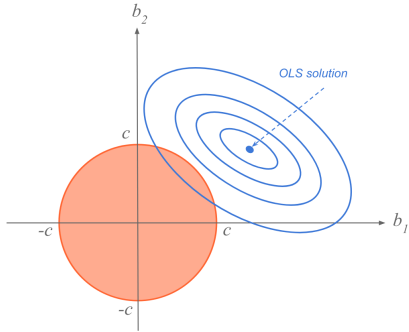
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \tag{9}$$

sujeto a

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

Ridge: Intuición en 2 Dimensiones

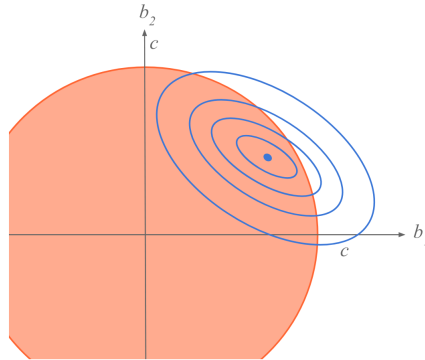
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \text{ s.a } \left((\beta_1)^2 + (\beta_2)^2\right) \leq c \tag{10}$$



Fuente: https://allmodelsarewrong.github.io

Ridge: Intuición en 2 Dimensiones

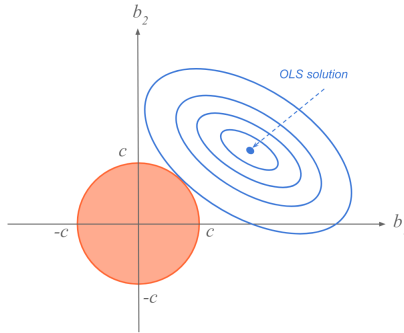
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \text{ s.a } \left((\beta_1)^2 + (\beta_2)^2\right) \leq c \tag{11}$$



Fuente: https://allmodelsarewrong.github.io

Ridge: Intuición en 2 Dimensiones

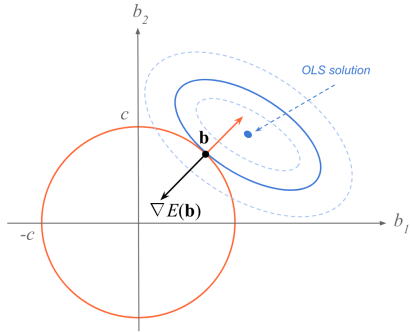
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \text{ s.a } \left((\beta_1)^2 + (\beta_2)^2\right) \leq c \tag{12}$$



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

Términos generales

- En regresión multiple (X es una matriz $n \times k$)
- Regresión: $y = X\beta + u$
- OLS

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

- Ridge

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

Ridge vs OLS

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

Agenda

1. Recap: Predicción y Overfit
2. Regularización
   - Recap: OLS Mechanics
   - Ridge
   - Escala de las variables
   - Selección de λ
   - Ridge as Data Augmentation
3. Lasso

Escala de las variables

- La escala de las variables importa en Ridge, mientras que en OLS no.
- Tiene consecuencias
  - En la solución ($\hat{\beta}$)
  - En la predicción ($\hat{y}$)

Escala de las variables

Ridge no es invariante a las escala

- Supongamos $z = c * x$
- Vamos a mostrar que $\hat{y}_i^z = \hat{y}_i^x$
- Partamos del modelo

$$y_i = \beta_0^z + \beta_1^z z_i + u \tag{14}$$

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

Escala de las variables

Ridge no es invariante a las escala

- Continuando

$$\hat{\beta}_1^z = \frac{\sum(z_i - \bar{z})(y_i - \bar{y})}{\sum(z_i - \bar{z})^2} \tag{16}$$

- Pero $z = c * x$

$$\hat{\beta}_1^z = \frac{\sum(cx_i - c\bar{x})(y_i - \bar{y})}{\sum(cx_i - c\bar{x})^2} \tag{17}$$

$$= \frac{1}{c}\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \tag{18}$$

$$= \frac{1}{c}\hat{\beta}_1^x \tag{19}$$

- En Ridge?

Escala de las variables

Ridge no es invariante a las escala

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

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - \beta_0^z - \beta_1^z z_i)^2 + \lambda (\beta_1^z)^2 \tag{20}$$

- Demo: baticomputer, math: Homework

Escala de las variables

Ridge no es invariante a las escala

- En la predicción

$$\hat{\beta}_1^z z_i = \hat{\beta}_1^z c x_i \tag{21}$$

$$= \frac{1}{c}\hat{\beta}_1^x c x_i \tag{22}$$

$$= \hat{\beta}_1^x x_i \tag{23}$$

Escala de las variables

Ridge no es invariante a las escala

- En términos generales, si $Z = cX$

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

Escala de las variables

Ridge no es invariante a las escala

- Entonces

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

- Con Ridge esto no funciona

$$\hat{\beta}^Z_{Ridge} Z \neq \hat{\beta}^X_{Ridge} X$$

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

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
   - Escala de las variables
   - Selección de λ
   - Ridge as Data Augmentation
3. Lasso

Selección de λ

- Asegurar cero sesgo dentro de muestra crea problemas fuera de muestra: trade-off Sesgo-Varianza
- Ridge hace este trade-off de forma empírica.

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 + \lambda \sum_{j=1}^{p} R(\beta_j) \tag{24}$$

- λ es el precio al que hacemos este trade off
- Como elegimos λ?

Selección de λ

- λ es un hiper-parámetro y lo elegimos usando validación cruzada
  - Partimos la muestra de entrenamiento en K Partes:
    $MUESTRA = M_{fold\,1} \cup M_{fold\,2} \cdots \cup M_{fold\,K}$
  - Cada conjunto $M_{fold\,K}$ va a jugar el rol de una muestra de evaluación $M_{eval\,k}$.
  - Entonces para cada muestra
    - $M_{train-1} = M_{train} - M_{fold\,1}$
    - $\vdots$
    - $M_{train-k} = M_{train} - M_{fold\,k}$

Selección de λ

- Luego hacemos el siguiente loop
  - Para $i = 0, 0.001, 0.002, \ldots, \lambda_{max}$ $\{$

    - 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 $MSE_i = \frac{1}{K} MSE(m_{i,k})$

    $\}$ # *fin para* λ

- Encontramos el menor MSE_i y usar ese $\lambda_i = \lambda^*$

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
   - Escala de las variables
   - Selección de λ
   - Ridge as Data Augmentation
3. Lasso

Ridge as Data Augmentation (1)

- Add λ additional points

$$\sum_{i=1}^{n}(y_i - x_i\beta)^2 + \lambda\beta^2 = \tag{25}$$

$$= \sum_{i=1}^{n}(y_i - x_i\beta)^2 + \sum_{j=1}^{\lambda}(0 - \beta)^2 \tag{26}$$

$$= \sum_{i=1}^{n+\lambda}(y_i - x_i\beta)^2 \tag{27}$$

RidgeDataAug

Ridge as Data Augmentation (2)

- Add a single point

$$\sum_{i=1}^{n}(y_i - x_i\beta)^2 + \lambda\beta^2 = \tag{28}$$

$$= \sum_{i=1}^{n}(y_i - x_i\beta)^2 + (0 - \sqrt{\lambda}\beta)^2 \tag{29}$$

$$= \sum_{i=1}^{n+1}(y_i - x_i\beta)^2 \tag{30}$$

RidgeDataAug

More predictors than observations ($k > n$)

- What happens when we have more predictors than observations ($k > n$)?
  - OLS fails
  - Ridge ?

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 when $k > n$

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - \sum_{j=1}^{k} x'_{ij}\beta_j)^2 + \lambda (\sum_{j=1}^{k} \beta_j)^2 \tag{31}$$

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

Agenda

1. Recap: Predicción y Overfit
2. Regularización
   - Recap: OLS Mechanics
   - Ridge
   - Escala de las variables
   - Selección de λ
   - Ridge as Data Augmentation
3. Lasso

Lasso

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

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{32}$$

Lasso

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

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{32}$$

  - "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

$$min_\beta E(\beta) = \sum_{i=1}^{n}(y_i - x_i\beta)^2 + \lambda|\beta| \tag{33}$$

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

$$min_\beta E(\beta) = \sum_{i=1}^{n}(y_i - x_i\beta)^2 \tag{34}$$

- y la solución es

$$\hat{\beta}_{OLS} \tag{35}$$

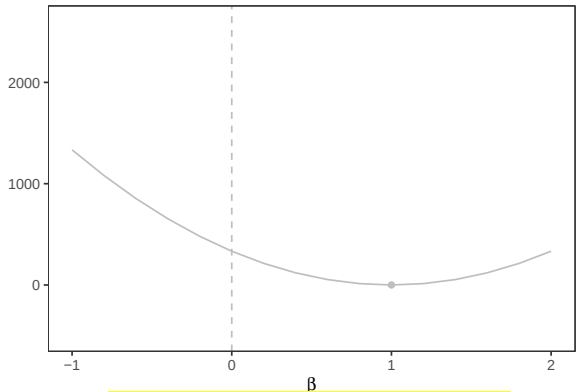
Intuición en 1 Dimension

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda|\beta| \tag{36}$$

Intuición en 1 Dimensión

$\hat{\beta} > 0$

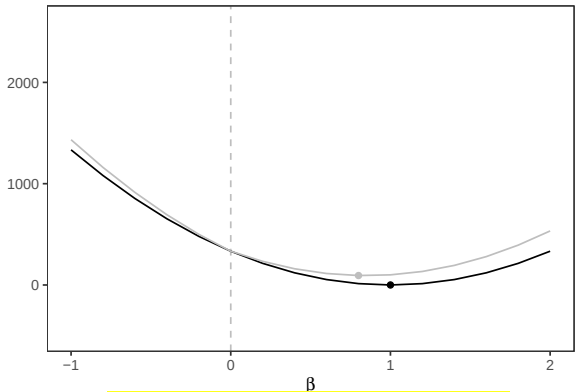
$$min_\beta E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda\beta \tag{37}$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

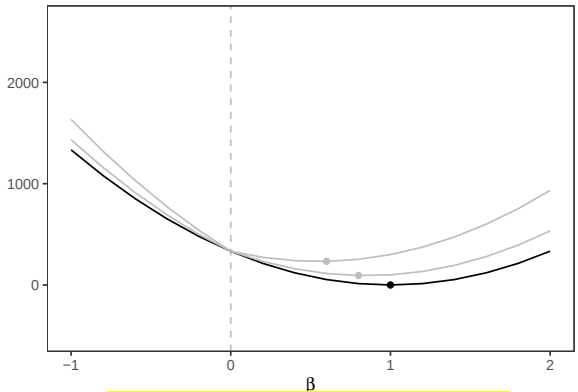
$$min_\beta E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda\beta \tag{38}$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

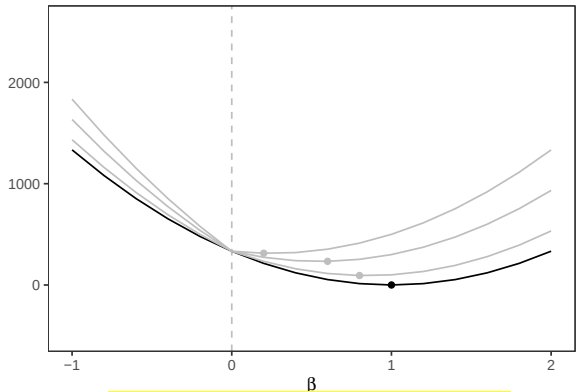
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda\beta \tag{39}$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

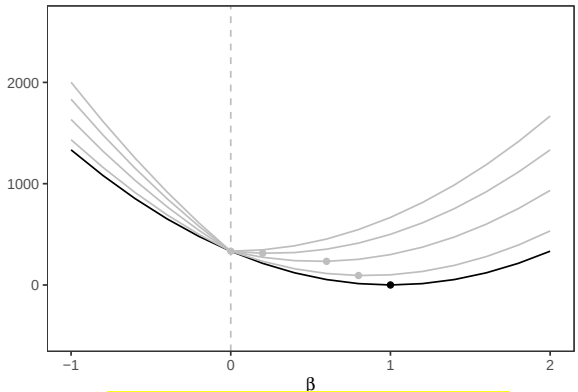
$$min_\beta E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda\beta \tag{40}$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

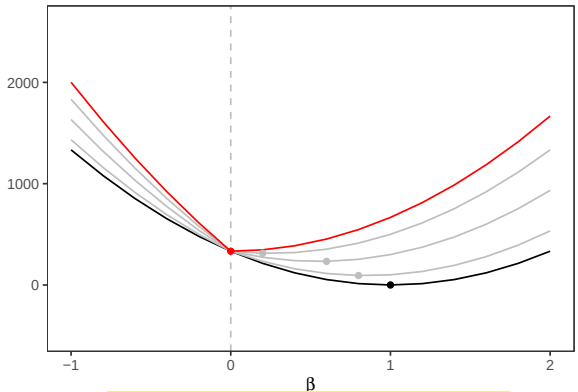
$$min_\beta E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda\beta \tag{41}$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

$$min_\beta E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda\beta \tag{42}$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

$$min_\beta E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda\beta \tag{43}$$

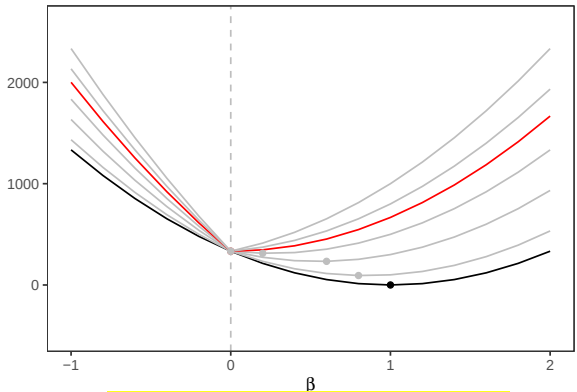


Ilustración en R

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