Selección de Modelos y Regularización

Machine Learning

Ignacio Sarmiento-Barbieri

Universidad de La Plata

Agenda

1. Recap: Predicción y Overfit
2. Selección de Modelos
3. Regularización
   - Recap: OLS Mechanics
   - Ridge
     - Detalles de Implementación
     - Ridge as Data Augmentation
   - Lasso
   - Ridge and Lasso: Pros and Cons
   - Familia de regresiones penalizadas
   - $k > n$
   - Elastic Net

Agenda

Recap: Predicción y Overfit

$y = f(x) + u$

$y = x\beta + u$

- Last Week:
  - Machine Learning is all about prediction
  - ML targets something different than causal inference, they can complement each other
  - Bias Variance trade-off: tolerating some bias is possible to reduce $V(\hat{f}(X))$ and lower MSE (ML best kept secret)
  - Overfit and Model Selecction
    - AIC y BIC
    - Validation Approach
    - LOOCV
    - K-fold Cross-Validation

Recap: Train and Test Sets. In-Sample and Out-of-Sample Prediction.

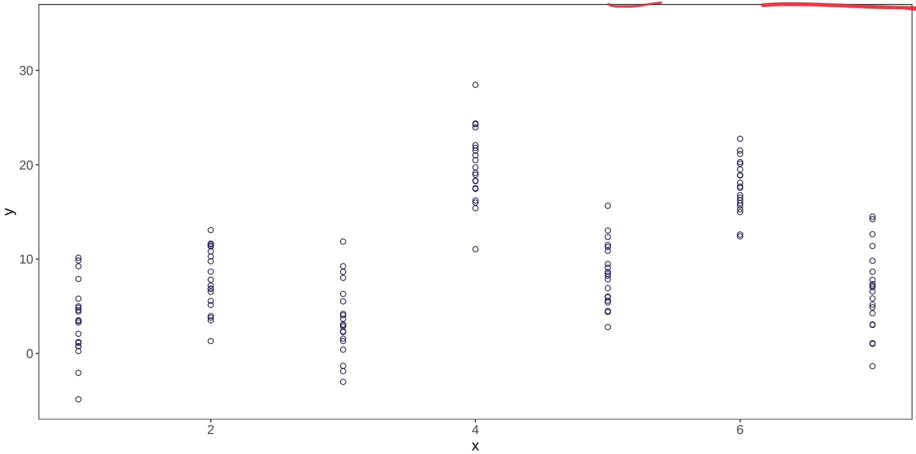
- El objetivo es predecir y dadas otras variables X. Ej: salario dadas las características del individuo
- Asumimos que el link entre y and X esta dado por el modelo:

$$y = f(X) + u \tag{1}$$

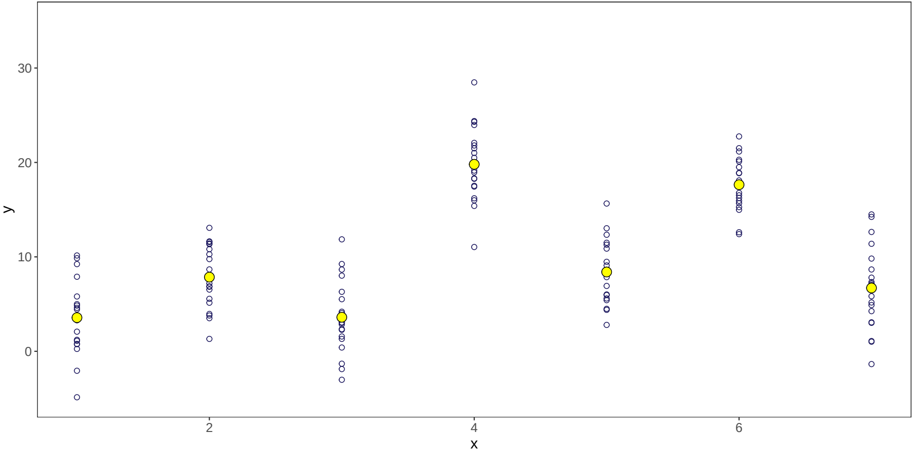
- donde $f(X)$ por ejemplo es $\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$
- u una variable aleatoria no observable $E(u) = 0$ and $V(u) = \sigma^2$

Recap: In-Sample Prediction and Overfit

Mejor predictor y' surge de min $\boxed{MSE} \rightarrow \boxed{E(y|x)}$

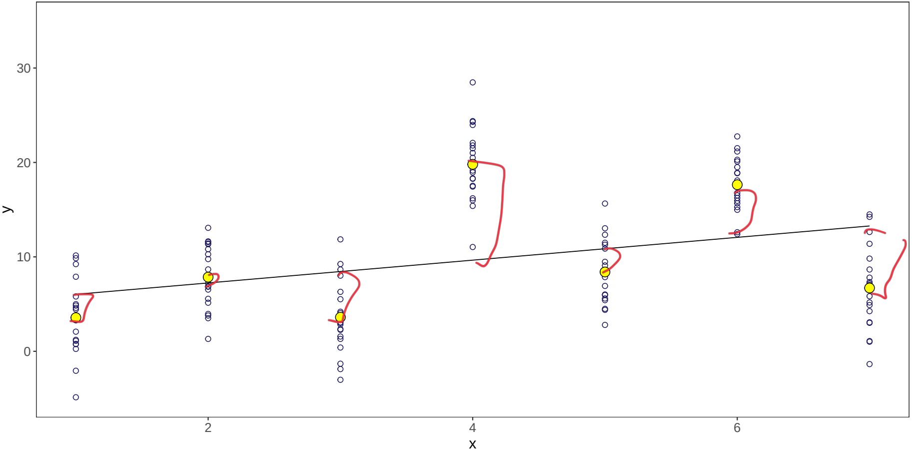


Recap: In-Sample Prediction and Overfit



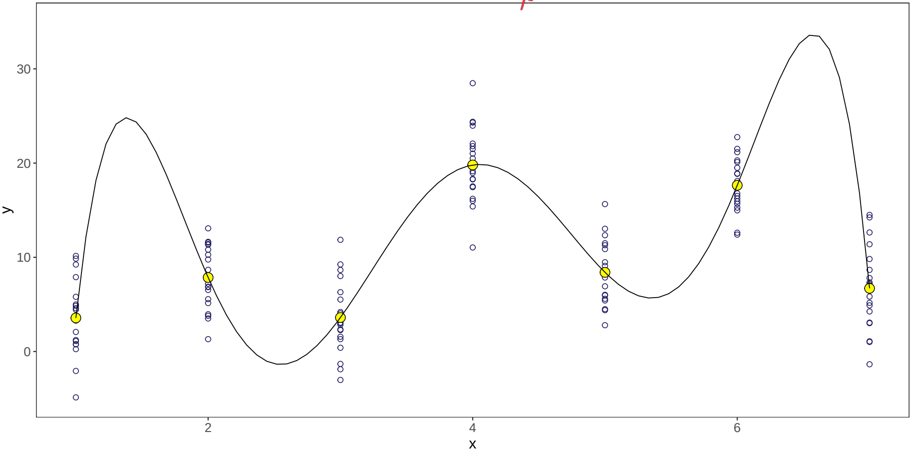
Recap: In-Sample Prediction and Overfit

$\hat{y} = \hat{\alpha} + \hat{\beta} x$

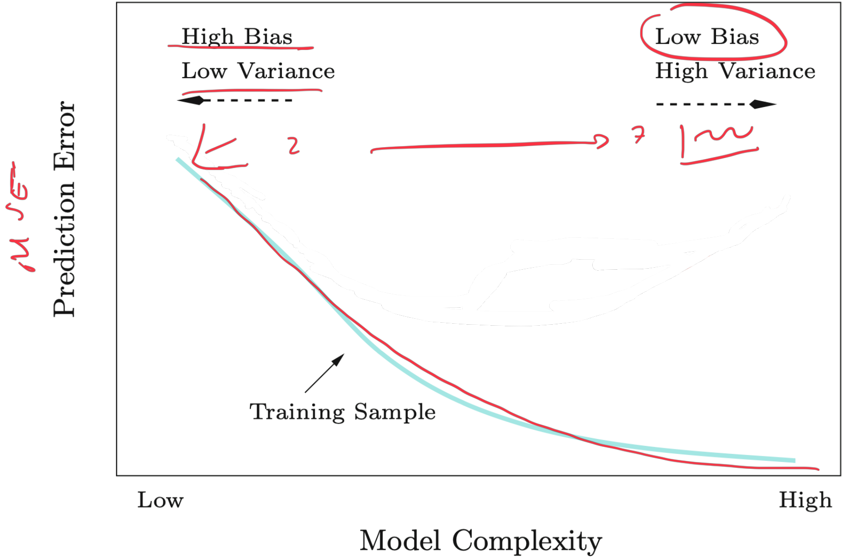


Recap: In-Sample Prediction and Overfit

$$y = \alpha + \sum_{p=1}^{6} \beta_p x^p$$



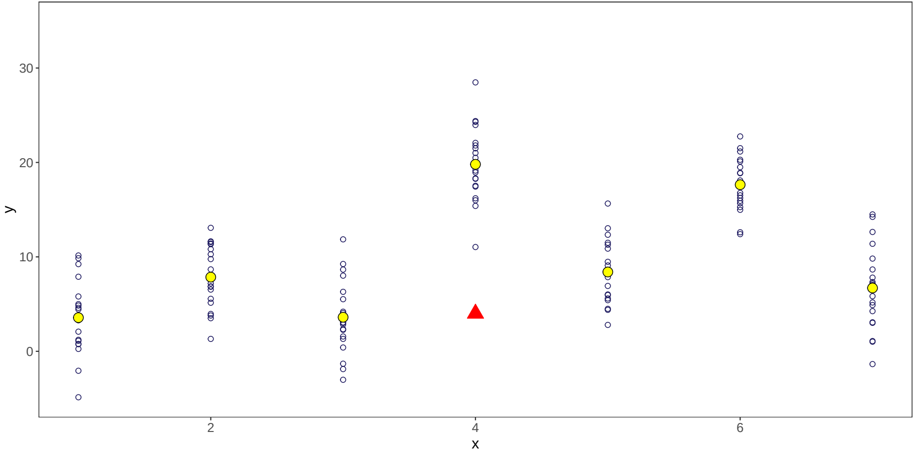
Recap: In-Sample Prediction and Overfit



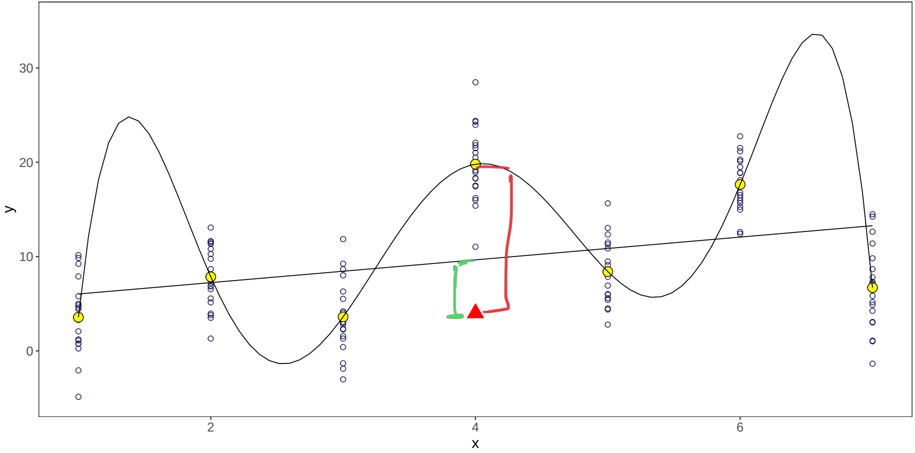
Recap: Out-of-Sample Prediction and Overfit

- ML nos interesa la predicción fuera de muestra

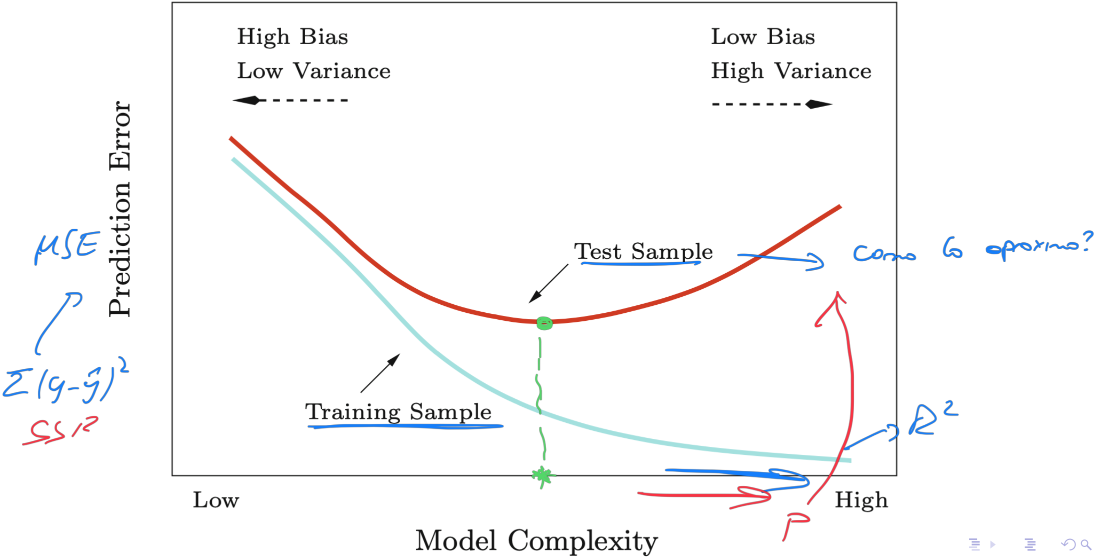
Recap: Out-of-Sample Prediction and Overfit



Recap: Out-of-Sample Prediction and Overfit



Recap: Overfit y Predicción fuera de Muestra



Recap: 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)

$\log(SSR) - \frac{p}{n}$

$- \log(n)\, p \frac{1}{2}$

Test set → Fuera de muestra

Training

Agenda

Model Subset Selection

k

$M_0 \quad M_1 \quad M_2 \quad \cdots \quad M_k$

- 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

$k = 20$ mas de 1 millon de modelos

$M_1^* \to$ Edad $\quad M_2^* =$ Edad + Edu.

$M_3 =$ Edad + Edu + ?

$M_0 \quad M_1^* \quad M_2^* \quad M_3^* \quad \cdots \quad M_k^*$

Agenda

1. Recap: Predicción y Overfit
2. Selección de Modelos
3. Regularización
   - Recap: OLS Mechanics
   - Ridge
     - Detalles de Implementación
     - Ridge as Data Augmentation
   - Lasso
   - Ridge and Lasso: Pros and Cons
   - Familia de regresiones penalizadas
   - $k > n$
   - Elastic Net

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

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

OLS 1 Dimension

$\bar{X} = 0 \quad V(x) = 1 = \sum x_i^2$

$\sum x_i^2 - \bar{x}^2$

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

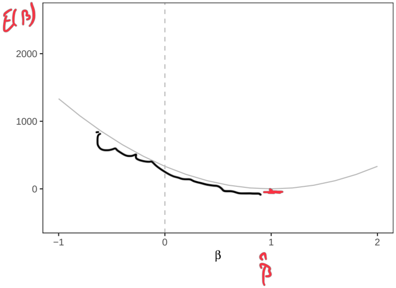
$\frac{\partial E}{\partial \beta} = 0$

$\sum 2(y_i - x_i\beta)(-)x_i = 0$

$\sum y_i x_i - \underbrace{\sum x_i^2}_{1} \beta = 0$

$\hat{\beta} = \sum y_i x_i$

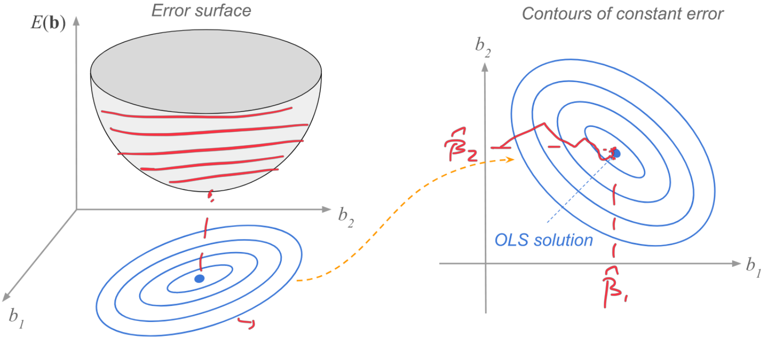
$\hat{\beta}^{OLS} = \sum y_i x_i$



App

OLS 2 Dimensiones

$$min_{\beta}E(\beta) = \sum_{i=1}^{n}(y_i - x_{i1}\beta_1 - x_{i2}\beta_2)^2 \quad (4)$$



Fuente: https://allmodelsarewrong.github.io

App

Regularización: Motivación

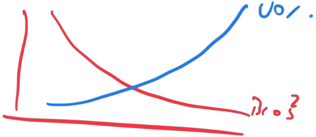
- Las técnicas econometricas estándar no están optimizadas para la predicción porque se enfocan en la insesgadez.
- OLS por ejemplo es el mejor estimador lineal *insesgado*
- OLS minimiza el error *"dentro de muestra"*, eligiendo β de forma tal que

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

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

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

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

- 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

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

$$(\beta_j - 0)^2$$

Ridge: Intuición en 1 Dimension

- 1 predictor estandarizado $\bar{x} = 0 \quad \sum x_i^2 = 1$
- El problema:

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

- La solución?

$$2\sum (y_i - x_i\beta)(-x_i) + 2\lambda\beta = 0$$

$$-\sum y_i x_i + \beta + \lambda\beta = 0$$

$$-\sum y_i x_i + (1+\lambda)\beta = 0$$

$$\hat{\beta} = \frac{\sum y_i x_i}{(1+\lambda)}.$$

$$\hat{\beta}^{Ridge} = \frac{\hat{\beta}_{OLS}}{(1+\lambda)}$$

$$\lambda = 0 \to \hat{\beta}_{OLS}$$

$$\lambda \uparrow \to \hat{\beta}^{ridge} \downarrow$$

App

Ridge: Intuición en 1 Dimension

Problema como optimización restringida

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

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

sujeto a

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

Ridge: Intuición en 2 Dimensiones

- Al problema en 2 dimensiones podemos escribirlo como

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_{i1}\beta_1 - x_{i2}\beta_2 + \lambda \left(\beta_1^2 + \beta_2^2\right) \tag{9}$$

- podemos escribirlo como un problema de optimización restringido

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

sujeto a

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

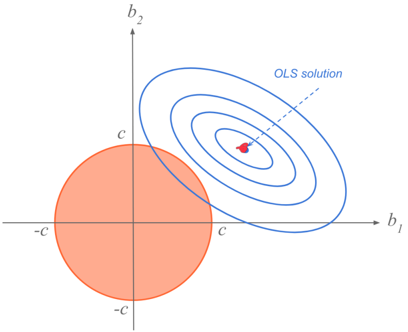
$$(\beta_1 - 0)^2 + (\beta_2 - 0)^2 \leq c$$

Ridge: Intuición en 2 Dimensiones

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

$\lambda \uparrow \rightarrow c \downarrow$

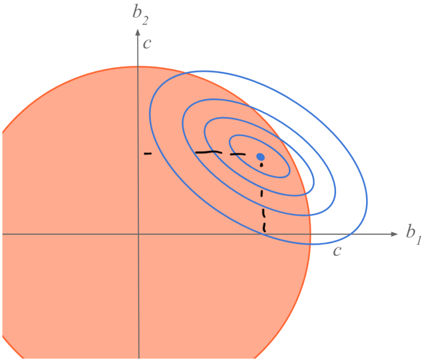
$\lambda \downarrow \rightarrow c \uparrow$



Fuente: https://allmodelsarewrong.github.io

Ridge: Intuición en 2 Dimensiones

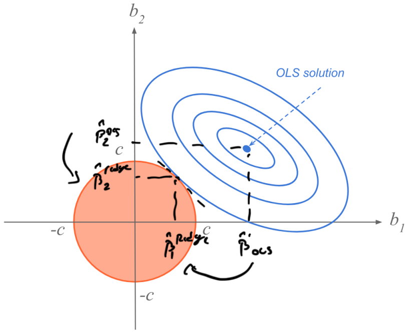
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \text{ s.a } \left((\beta_1)^2 + (\beta_2)^2\right) \leq c \tag{12}$$



Fuente: https://allmodelsarewrong.github.io

Ridge: Intuición en 2 Dimensiones

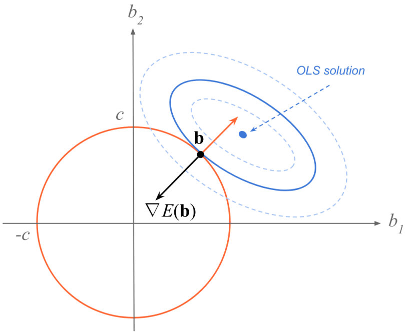
$$min_{\beta}E(\beta) = \sum_{i=1}^{n}(y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \text{ s.a } \left((\beta_1)^2 + (\beta_2)^2\right) \leq c \tag{13}$$



Fuente: https://allmodelsarewrong.github.io

Ridge: Intuición en 2 Dimensiones

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



Fuente: https://allmodelsarewrong.github.io

Términos generales

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

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

$$\hat{\beta}^{ridge} = \frac{\sum y_i x_i}{(1+\lambda)}$$

- Ridge

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

Ridge vs OLS

$E(\hat{\beta}_{OLS}) = \beta$

$E(\hat{\beta}^{ridge}) = E\left(\frac{\hat{\beta}^{OLS}}{1+\lambda}\right) = \frac{1}{1+\lambda}\beta$

- Ridge es sesgado $E(\hat{\beta}_{ridge}) \neq \beta$
- Pero la varianza es menor que la de OLS

$V(\hat{\beta}^{OLS}) = \sigma^2$

- Para ciertos valores del parámetro $\boxed{\lambda \Rightarrow MSE_{OLS} > MSE_{ridge}}$

$V(\hat{\beta}^{ridge}) = \frac{V(\hat{\beta}^{OLS})}{(1+\lambda)^2} = \frac{\sigma^2}{(1+\lambda)^2}$

Escala de las variables

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

$$\sum \left(y_i - \beta_0 - \sum^{k} \beta_i x_i\right)^2 + \lambda \sum_{j=1}^{k} \beta_j^2$$

- La escala de las variables importa en Ridge, mientras que en OLS no.
- Es importante estandarizar las variables (la mayoría de los softwares lo hace automáticamente)

Pesos \$ $(\underline{1.000.000} - 0)^2$ ✓

Años $(\underline{32} - 0)^2$ ✓

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.

$$min_{\beta} E(\beta) = \sum_{i=1}^{n}(y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 + \lambda \sum_{j=1}^{p} R(\beta_j) \tag{15}$$

- 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 $\lambda_i = \lambda^*$

λ	MSE
0,1	MSE_1
0,1	MSE_2
0,1	MSE_3
0,1	MSE_4
0,1	MSE_5

0,1	$\frac{1}{k}\Sigma MSE_k$
0,2	MSE CV
$\vdots$	
λ_{max}	MSE CV

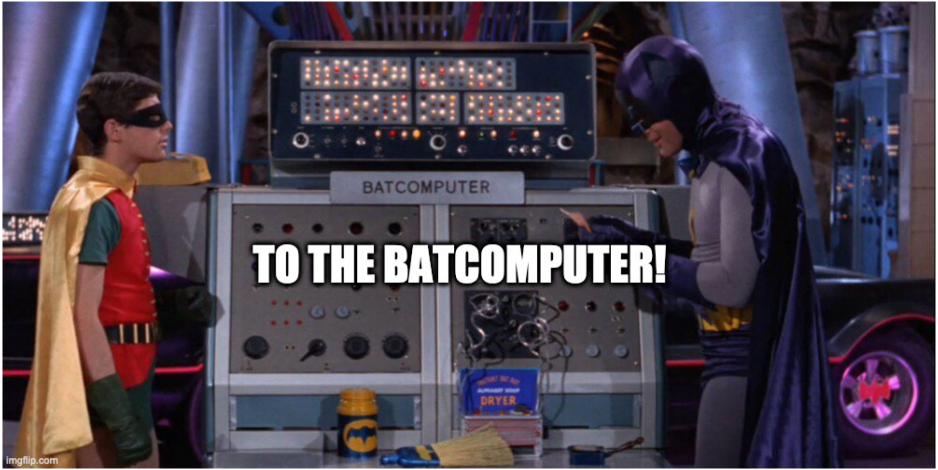


photo from https://www.dailydot.com/parsec/batman-1966-labels-tumblr-twitter-vine/

Ridge as Data Augmentation (1)

- Add λ additional points

$$\sum_{j=1}^{\lambda} \underbrace{(\beta - 0)^2} = \lambda(\beta - 0)^2$$

$$\lambda \beta^2$$

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

$$= \sum_{i=1}^{n}(y_i - x_i\beta)^2 + \underbrace{\sum_{j=1}^{\lambda}(\overset{y}{0} - \beta)^2} \tag{17}$$

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

$y_j = 0$

$x_j = 1$

λ puntos $\overset{x,y}{(1,0)}$

RidgeDataAug

Ridge as Data Augmentation (2)

$\lambda\beta^2 = (\sqrt{\lambda}\beta)^2$

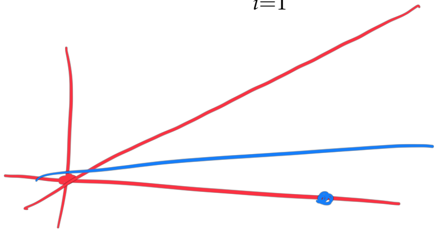
$(0 - \sqrt{\lambda}\beta)^2$

- Add a single point

$$\sum_{i=1}^{n}(y_i - x_i\beta)^2 + \lambda\beta^2 = \tag{19}$$

$$= \sum_{i=1}^{n}(y_i - x_i\beta)^2 + (0 - \sqrt{\lambda}\beta)^2 \tag{20}$$

$$= \sum_{i=1}^{n+1}(y_i - x_i\beta)^2 \tag{21}$$



$(0, \sqrt{\lambda})$

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 $rank(X_{n\times k}) = k \implies k \leq n$

$(X'X)^{-1}X y$

- 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 $k > n$

σ: autovalores

autovect, autovalores

$(X'X) = P'\Delta P$

$\det(X'X) = \prod_{i=1}^{k} \sigma$

$(X'X + \lambda I) = P'\Delta P + P'\lambda P$

$= P'(\Delta + \lambda) P$

$\sigma_j + \lambda$

$\lambda \geq 0$

$\rightarrow$ rango completo.

$\sqrt{\lambda}$ en la diag principal

$$\begin{pmatrix} X \\ \begin{matrix} \sqrt{\lambda} & 0 & 0 & \cdots & 0 \\ 0 & \ddots & & & \\ 0 & & & & \sqrt{\lambda} \end{matrix} \end{pmatrix}$$

$n \times k$, $n << k$, $(n+k) \times k$

$$min_{\beta}E(\beta) = \sum_{i=1}^{n}(y_i - \sum_{j=1}^{k} x'_{ij}\beta_j)^2 + \lambda(\sum_{j=1}^{k}\beta_j)^2 \tag{22}$$

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

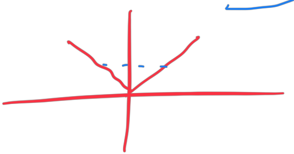
k puntos de coordenadas x, y: $(\sqrt{\lambda}, 0)$

Lasso

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

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - x_{i1}\overset{0}{\beta_1} - \cdots - x_{ip}\beta_p)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{23}$$

$$|\beta_j - 0|$$



Lasso

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

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - x_{i1}\beta_1 - \cdots - x_{ip}\beta_p)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{23}$$

  - "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

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda|\beta| \quad (24)$$

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

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 \quad (25)$$

- y la solución es

$$\hat{\beta}_{OLS} \quad (26)$$

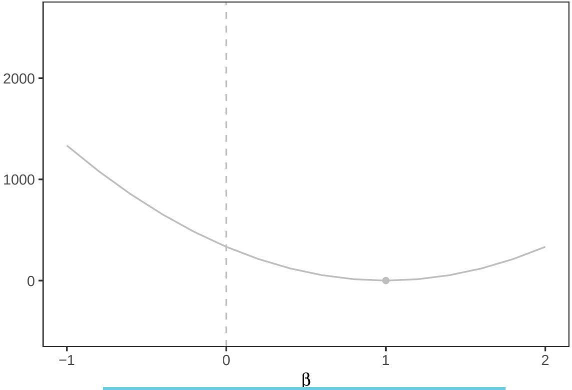
Intuición en 1 Dimension

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

Intuición en 1 Dimensión

$\hat{\beta} > 0$

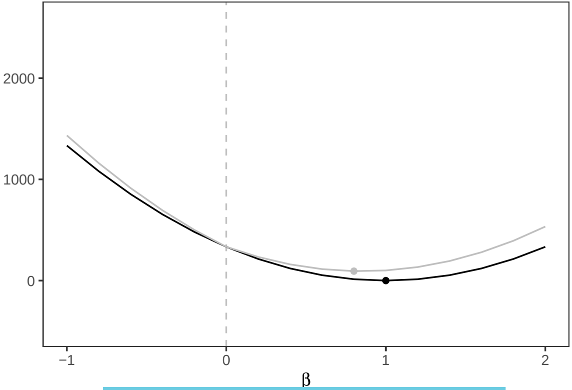
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda\beta \quad (28)$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

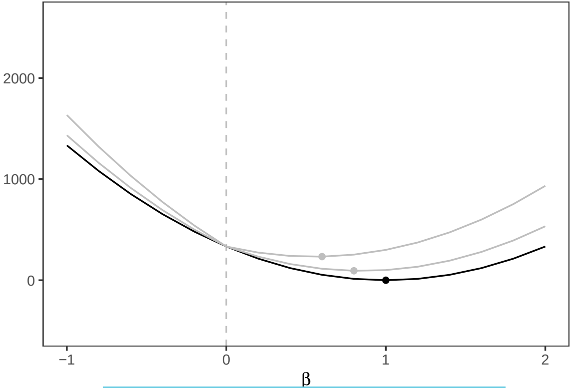
$$min_{\beta}E(\beta) = \sum_{i=1}^{n}(y_i - x_i\beta)^2 + \lambda\beta \tag{29}$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

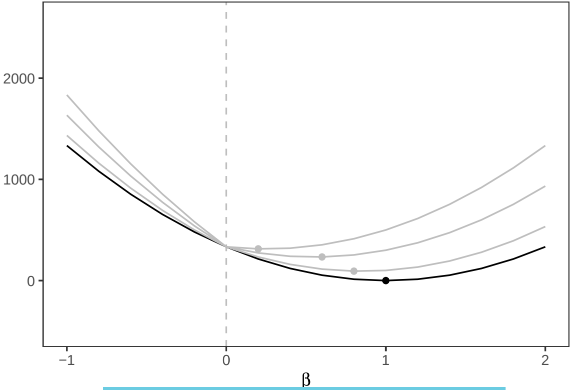
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda\beta \tag{30}$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

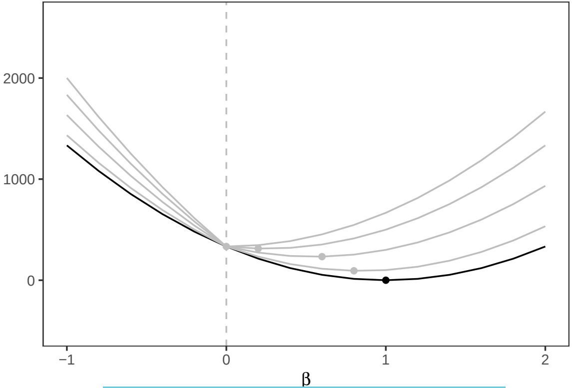
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda\beta \tag{31}$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

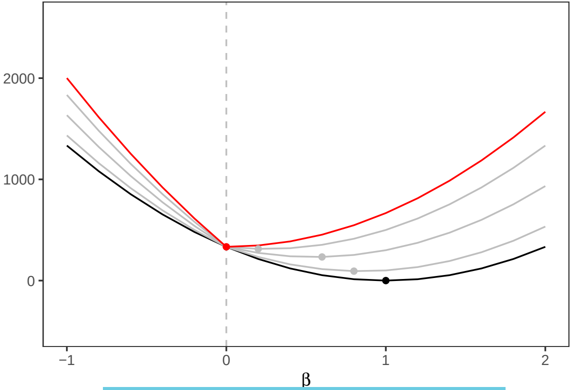
$$min_\beta E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda\beta \tag{32}$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

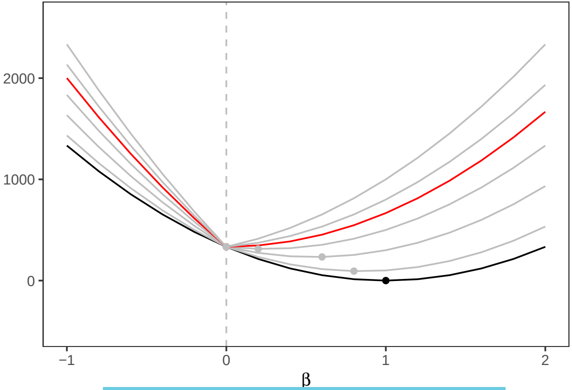
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda\beta \tag{33}$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

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



Intuición en 1 Dimension

Solución analitica

$$min_\beta E(\beta) = \sum_{i=1}^{n}(y_i - x_i\beta)^2 + \lambda|\beta| \tag{35}$$

Intuición en 1 Dimension

Solución analitica

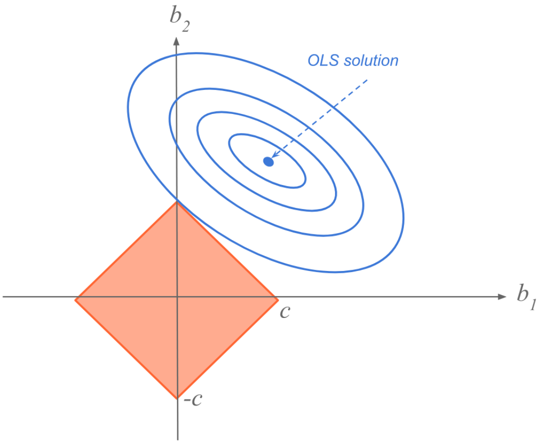
$$min_{\beta}E(\beta) = \sum_{i=1}^{n}(y_i - x_i\beta)^2 + \lambda|\beta| \tag{35}$$

- la solución analítica es

$$\hat{\beta}_{lasso} = \begin{cases} 0 & \textbf{si } \lambda \geq \lambda^* \\ \hat{\beta}_{OLS} - \frac{\lambda}{2} & \textbf{si } \lambda < \lambda^* \end{cases} \tag{36}$$

Intuición en 2 Dimensiones (Lasso)

$$min_{\beta}E(\beta) = \sum_{i=1}^{n}(y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \text{ s.a } (|\beta_1| + |\beta_2|) \leq c \quad (37)$$



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
  - Como elegimos λ?

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
  - Como elegimos λ? $\rightarrow$ Validación cruzada

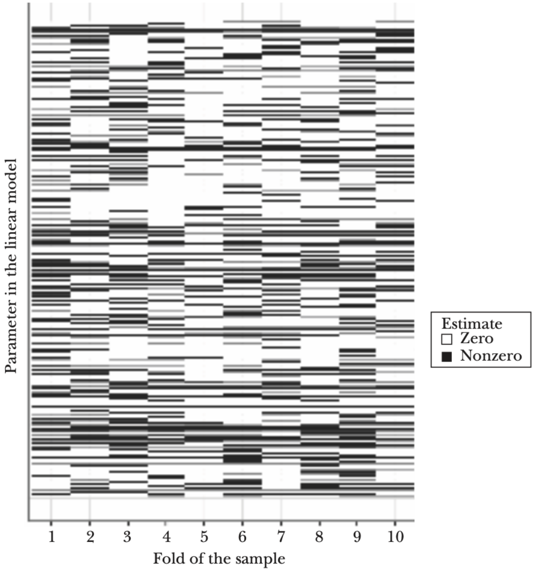
Ridge and Lasso: The good and the bad

- 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

- 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

Family of penalized regressions

$$min_\beta R(\beta) = \sum_{i=1}^{n}(y_i - x_i'\beta)^2 + \lambda \sum_{s=2}^{p} |\beta_s|^p \quad (38)$$

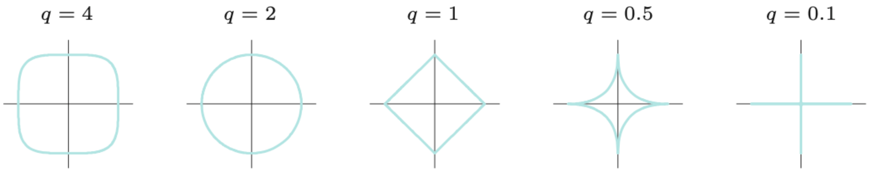


FIGURE 3.12. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for given values of* q.

More predictors than observations ($k > n$)

- Objective 1: Accuracy
  - Minimize prediction error (in one step) $\rightarrow$ Ridge, Lasso
- Objective 2: Dimensionality
  - Reduce the predictor space $\rightarrow$ Lasso's free lunch
- What happens when we have more predictors than observations ($k > n$)?
  - OLS fails
  - Ridge augments data
  - and Lasso?

Lasso when $k > n$

- Lasso works fine in this case
- However, there are some issues to keep in mind
  - When $k > n$ chooses at most n variables
  - When we have a group of highly correlated variables,
    - Lasso chooses only one. Makes it unstable for prediction. (Doesn't happen to Ridge)
    - Ridge shrinks the coefficients of correlated variables toward each other. This makes Ridge "work" better than Lasso. "Work" in terms of prediction error

Elastic net

$$min_\beta EN(\beta) = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + \frac{(1-\alpha)}{2} \sum_{j=1}^{p} (\beta_j)^2 \right) \quad (39)$$

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

Elastic Net

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

$$min_{\beta}EN(\beta) = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\left(\alpha\sum_{j=1}^{p}|\beta_j| + \frac{(1-\alpha)}{2}\sum_{j=1}^{p}(\beta_j)^2\right) \tag{40}$$

- 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