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
     - Escala de las variables
     - Selección de λ
     - Ridge as Data Augmentation
   - Lasso
   - Ridge and Lasso: Pros and Cons
   - Familia de regresiones penalizadas
   - $k > n$
   - Elastic Net

Agenda

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

Recap: Predicción y Overfit

- 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

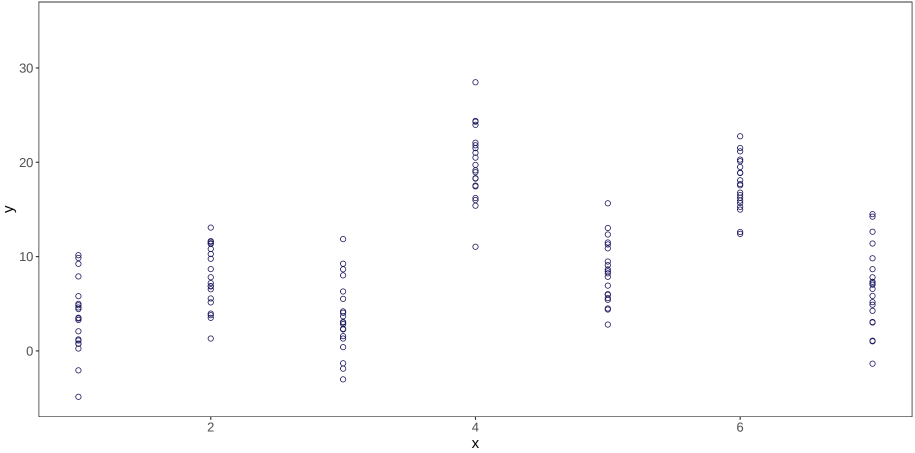
Recap: Train and Test Sets. In-Sample and Out-of-Sample Prediction.

- 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:

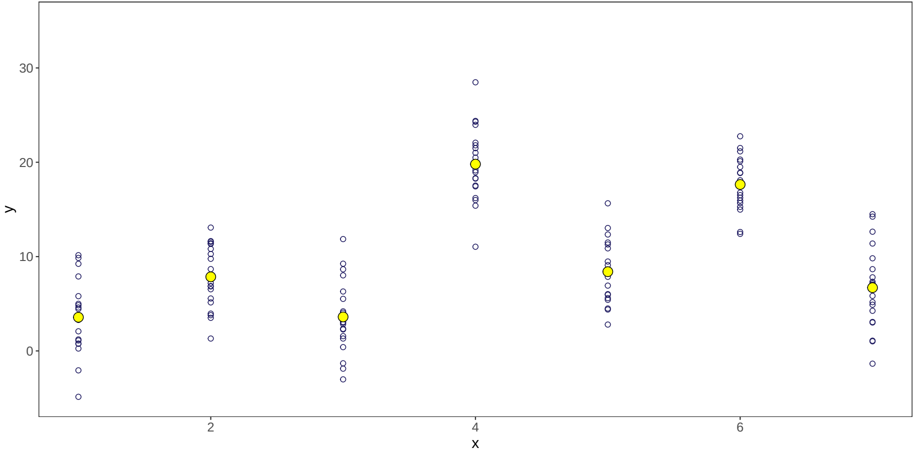
$$y = f(X) + u \tag{1}$$

- 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 $V(u) = \sigma^2$

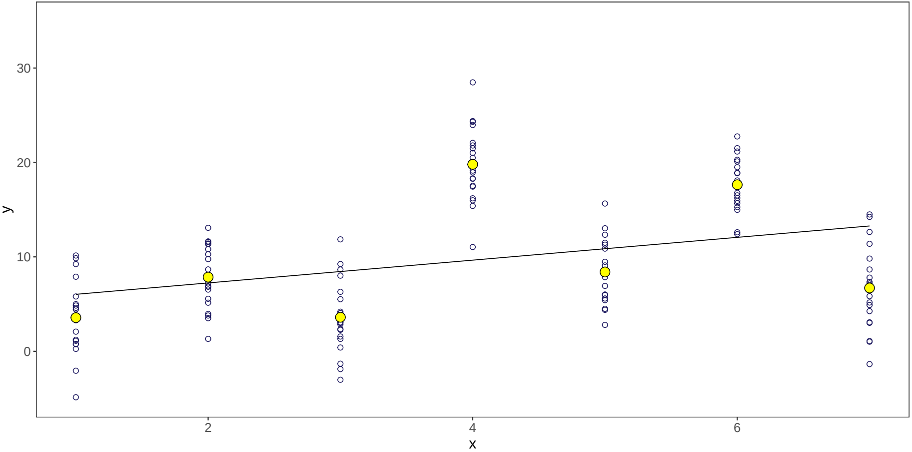
Recap: In-Sample Prediction and Overfit



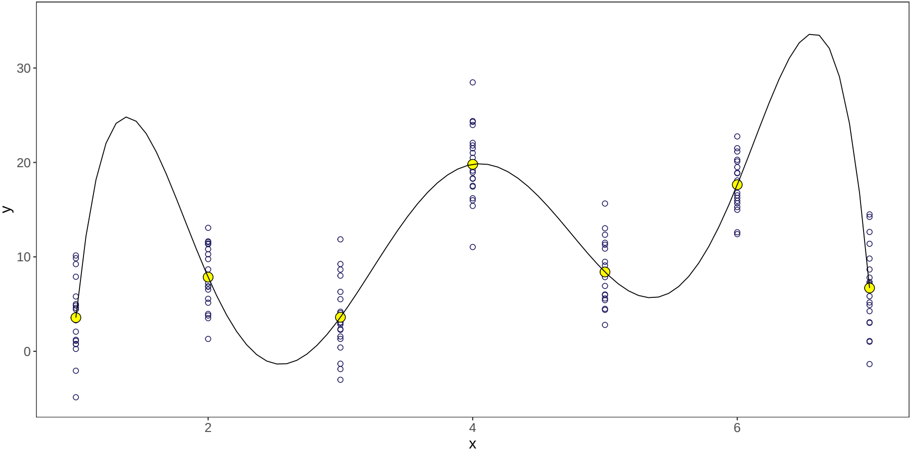
Recap: In-Sample Prediction and Overfit



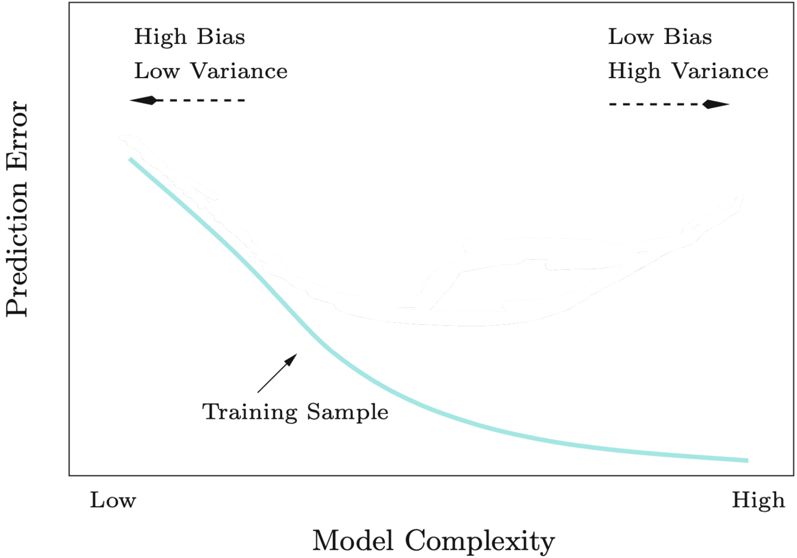
Recap: In-Sample Prediction and Overfit



Recap: In-Sample Prediction and Overfit



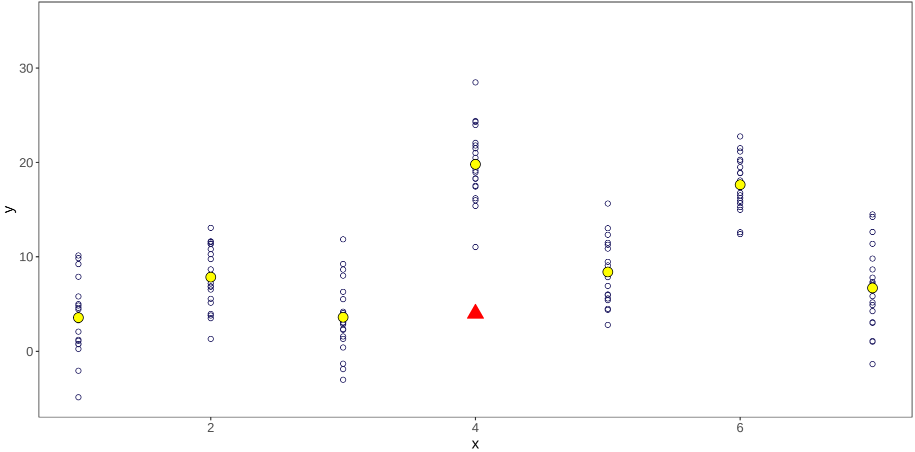
Recap: In-Sample Prediction and Overfit



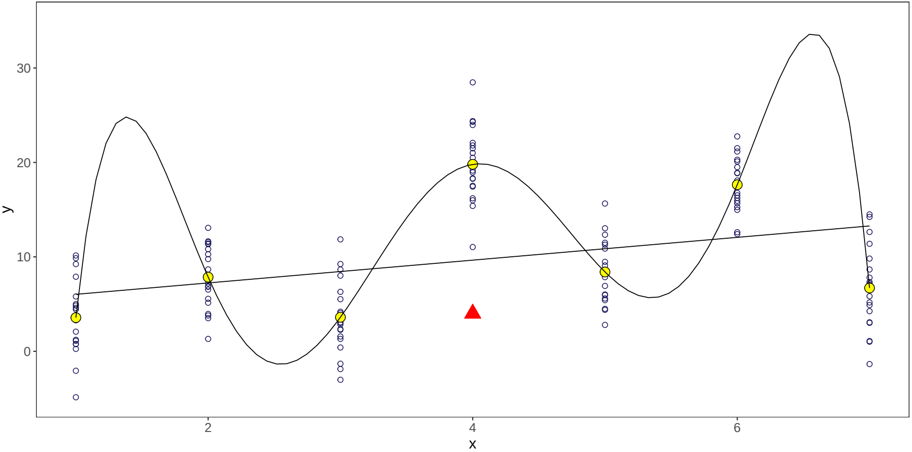
Recap: Out-of-Sample Prediction and Overfit

- ML nos interesa la predicción fuera de muestra

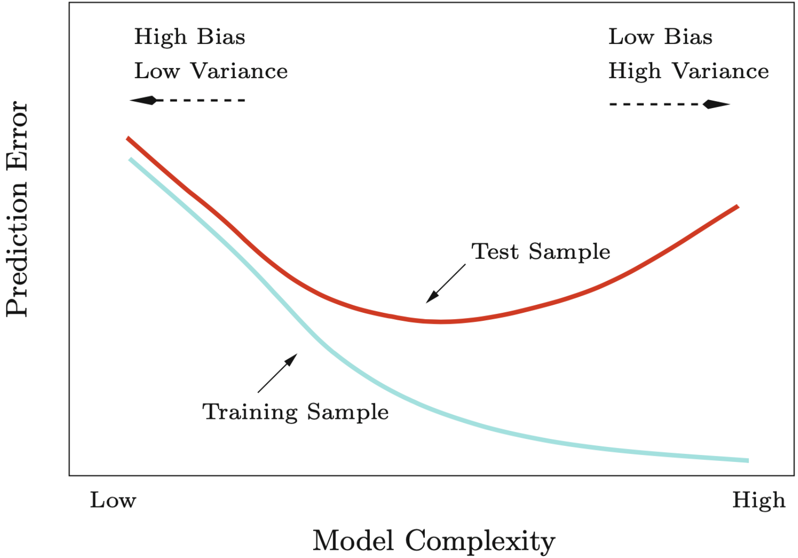
Recap: Out-of-Sample Prediction and Overfit



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)

Agenda

1 Recap: Predicción y Overfit

2 Selección de Modelos

3 Regularización
- Recap: OLS Mechanics
- Ridge
  - Escala de las variables
  - Selección de λ
  - Ridge as Data Augmentation
- Lasso
- Ridge and Lasso: Pros and Cons
- Familia de regresiones penalizadas
- $k > n$
- Elastic Net

Model Subset Selection

- 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

Agenda

1 Recap: Predicción y Overfit

2 Selección de Modelos

3 Regularización
- Recap: OLS Mechanics
- Ridge
  - Escala de las variables
  - Selección de λ
  - Ridge as Data Augmentation
- Lasso
- Ridge and Lasso: Pros and Cons
- Familia de regresiones penalizadas
- $k > n$
- Elastic Net

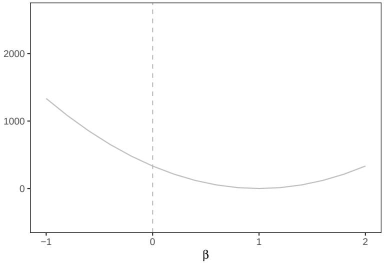
Regularización: Motivación

- 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

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

OLS 1 Dimension

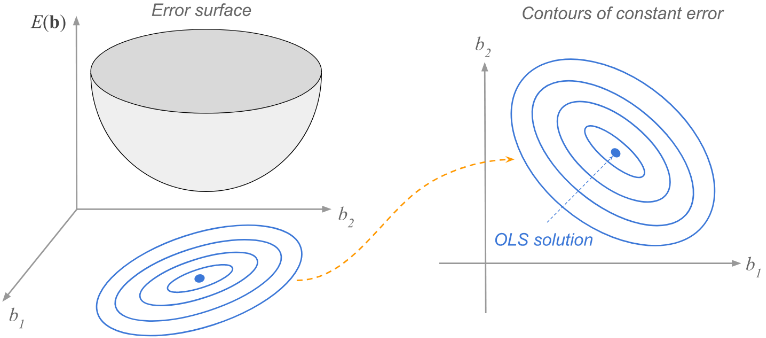
$$min_\beta E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 \tag{3}$$



App

OLS 2 Dimensiones

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



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

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

- 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

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

Ridge: Intuición en 1 Dimension

- 1 predictor estandarizado
- El problema:

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

- 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

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

$$\text{sujeto a}$$

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

Ridge: Intuición en 2 Dimensiones

- Al problema en 2 dimensiones podemos escribirlo como

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

- podemos escribirlo como un problema de optimización restringido

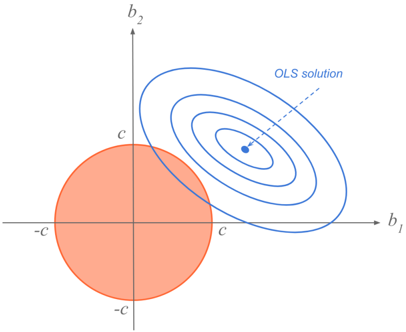
$$min_{\beta}E(\beta) = \sum_{i=1}^{n}(y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \tag{10}$$

sujeto a

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

Ridge: Intuición en 2 Dimensiones

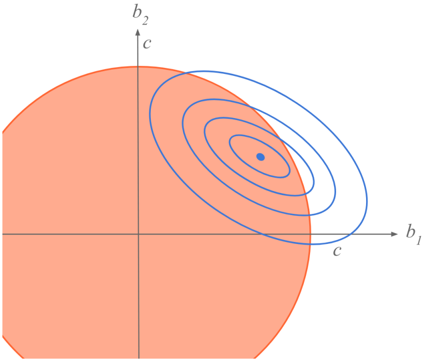
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \text{ s.a } \left((\beta_1)^2 + (\beta_2)^2\right) \leq c \tag{11}$$



Fuente: `https://allmodelsarewrong.github.io`

Ridge: Intuición en 2 Dimensiones

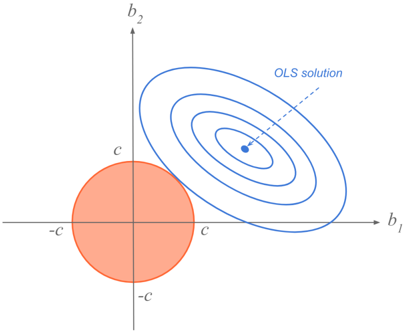
$$min_{\beta}E(\beta) = \sum_{i=1}^{n}(y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \text{ s.a } \left((\beta_1)^2 + (\beta_2)^2\right) \leq c \tag{12}$$



Fuente: https://allmodelsarewrong.github.io

Ridge: Intuición en 2 Dimensiones

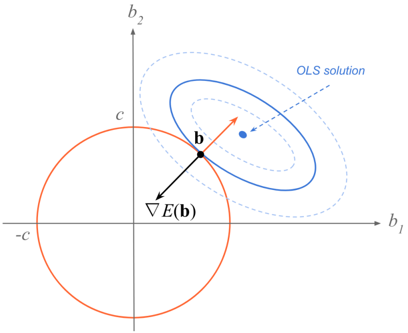
$$min_{\beta}E(\beta) = \sum_{i=1}^{n}(y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \text{ s.a } \left((\beta_1)^2 + (\beta_2)^2\right) \leq c \tag{13}$$



Fuente: `https://allmodelsarewrong.github.io`

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

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 $\lambda \Rightarrow MSE_{OLS} > MSE_{ridge}$

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

$$y_i = \beta_0^z + \beta_1^z z_i + u \tag{15}$$

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

Escala de las variables

Ridge no es invariante a las escala

- Continuando

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

- Pero $z = c * x$

$$\hat{\beta}_1^z = \frac{\sum(cx_i - c\bar{x})(y_i - \bar{y})}{\sum(cx_i - c\bar{x})^2} \tag{18}$$

$$= \frac{1}{c}\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \tag{19}$$

$$= \frac{1}{c}\hat{\beta}_1^x \tag{20}$$

- En Ridge?

Escala de las variables

Ridge no es invariante a las escala

- Continuando

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

- Pero $z = c * x$

$$\begin{aligned}
\hat{\beta}_1^z &= \frac{\sum(cx_i - c\bar{x})(y_i - \bar{y})}{\sum(cx_i - c\bar{x})^2} && (18) \\
&= \frac{1}{c}\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} && (19) \\
&= \frac{1}{c}\hat{\beta}_1^x && (20)
\end{aligned}$$

- En Ridge? $\rightarrow$ Ridge no es invariante a las escala

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

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

- 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)

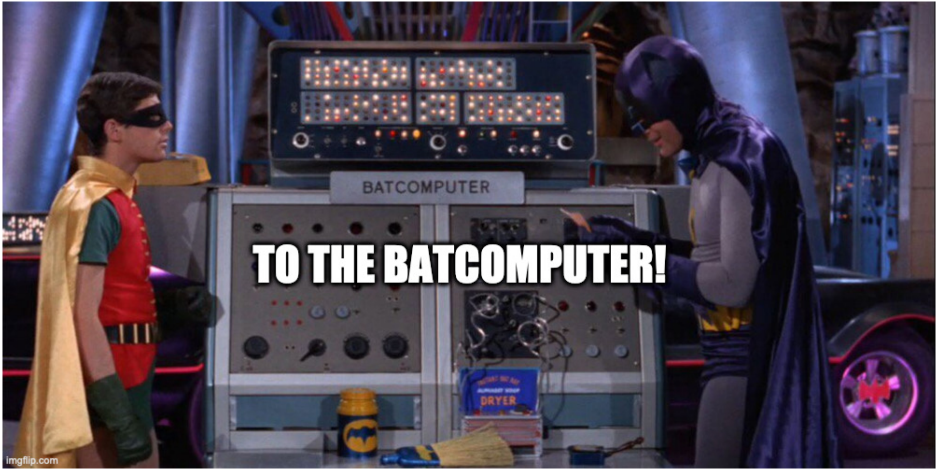


photo from https://www.dailydot.com/parsec/batman-1966-labels-tumblr-twitter-vine/

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 $\lambda_i = \lambda^*$

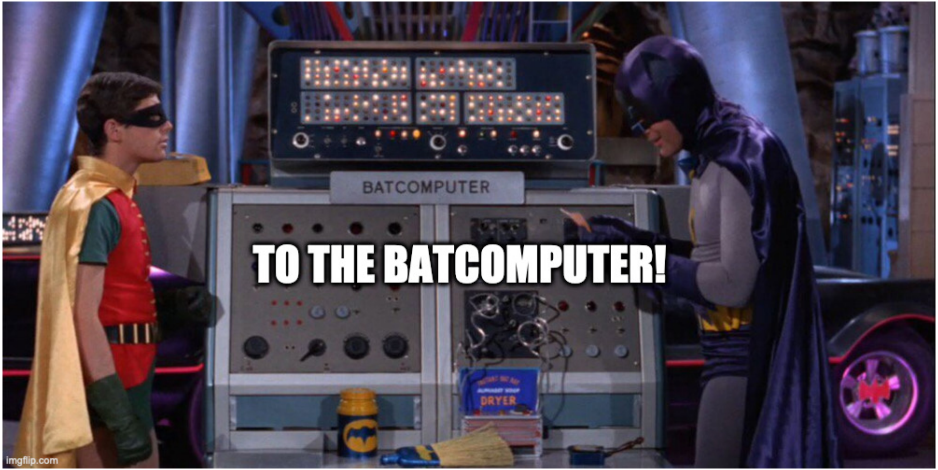


photo from https://www.dailydot.com/parsec/batman-1966-labels-tumblr-twitter-vine/

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 $k \geq n$

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

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

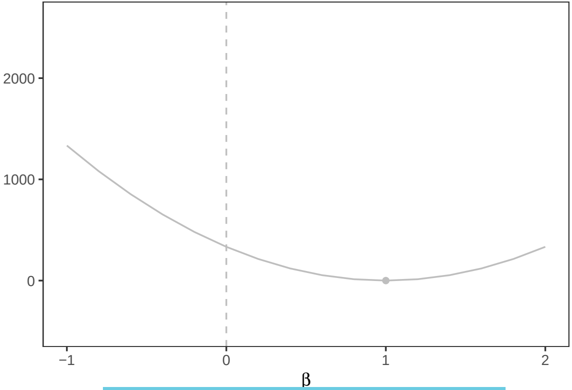
Intuición en 1 Dimension

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

Intuición en 1 Dimensión

$\hat{\beta} > 0$

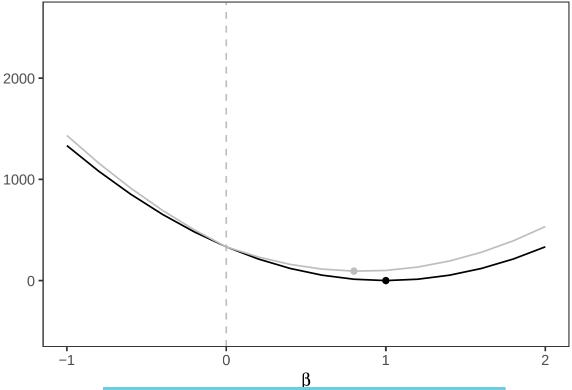
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda\beta \tag{37}$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

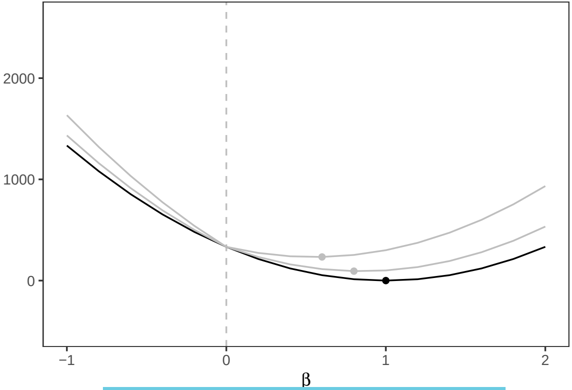
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda\beta \tag{38}$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

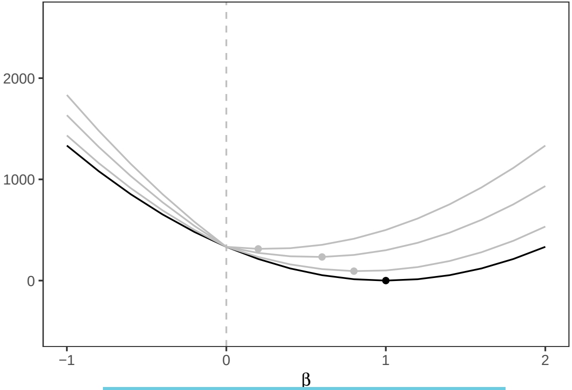
$$min_\beta E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda\beta \tag{39}$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

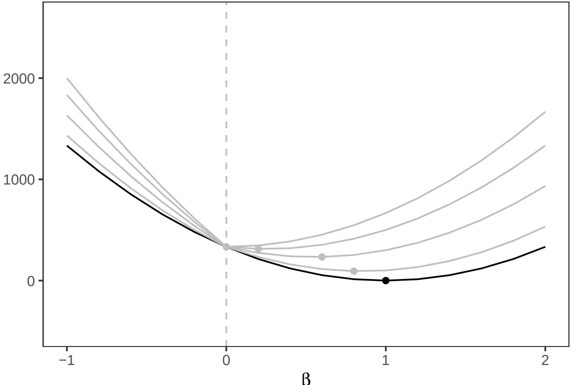
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda\beta \tag{40}$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

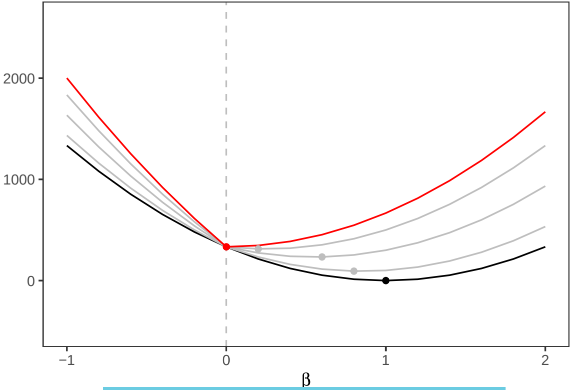
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda\beta \tag{41}$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

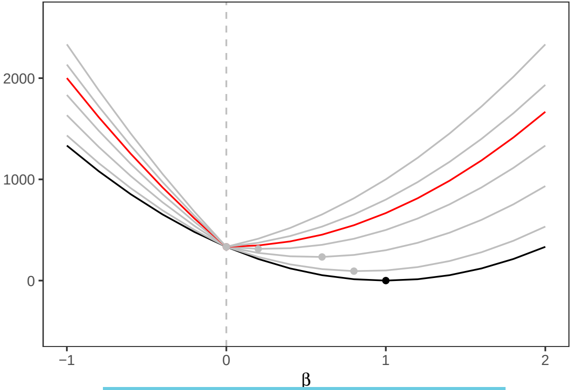
$$min_\beta E(\beta) = \sum_{i=1}^{n}(y_i - x_i\beta)^2 + \lambda\beta \tag{42}$$



Intuición en 1 Dimensión

$\hat{\beta} > 0$

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



Intuición en 1 Dimension

Solución analitica

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda|\beta| \tag{44}$$

Intuición en 1 Dimension

Solución analitica

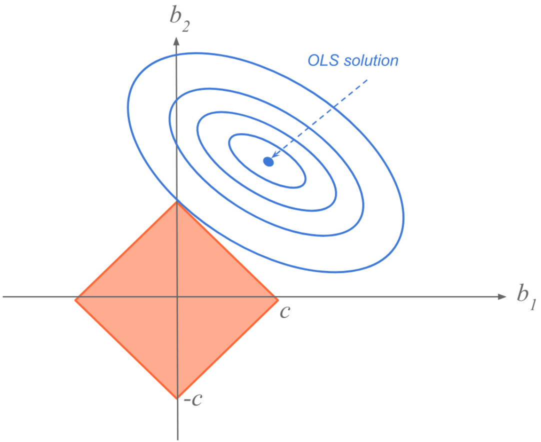
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_i\beta)^2 + \lambda|\beta| \tag{44}$$

- la solución analítica es

$$\hat{\beta}_{lasso} = \begin{cases} 0 & \text{si } \lambda \geq \lambda^* \\ \hat{\beta}_{OLS} - \frac{\lambda}{2} & \text{si } \lambda < \lambda^* \end{cases} \tag{45}$$

Intuición en 2 Dimensiones (Lasso)

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \text{ s.a } (|\beta_1| + |\beta_2|) \leq c \tag{46}$$



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

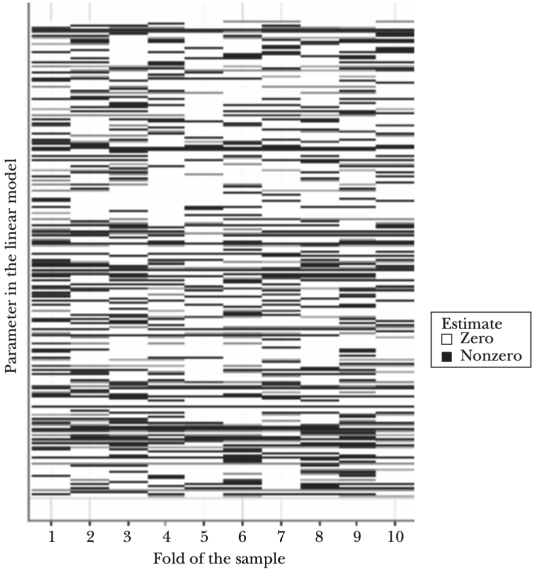
Ridge and Lasso: The good and the bad

- 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

$$min_{\beta} R(\beta) = \sum_{i=1}^{n} (y_i - x_i'\beta)^2 + \lambda \sum_{s=2}^{p} |\beta_s|^p \tag{47}$$

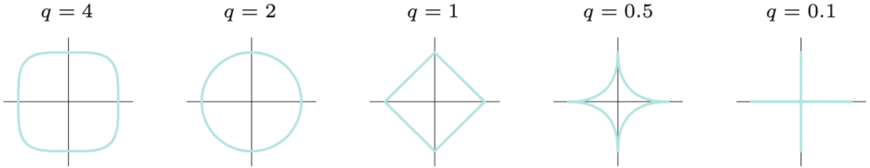


FIGURE 3.12. *Contours of constant value of* $\sum_j |\beta_j|^q$ *for given values of* q.

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

$$min_{\beta} EN(\beta) = \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + \frac{(1-\alpha)}{2} \sum_{j=1}^{p} (\beta_j)^2 \right) \tag{48}$$

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

Elastic Net

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

$$min_\beta EN(\beta) = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \left(\alpha \sum_{j=1}^{p} |\beta_j| + \frac{(1-\alpha)}{2} \sum_{j=1}^{p} (\beta_j)^2 \right) \tag{49}$$

- 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)
- H.W.: $\beta_{OLS} > 0$ one predictor standarized

$$\hat{\beta}_{EN} = \frac{\left(\hat{\beta}_{OLS} - \frac{\lambda_1}{2}\right)_+}{1+\lambda_2} \tag{50}$$