Selección de Modelos y Regularización

Big Data y Machine Learning para Economía Aplicada

Ignacio Sarmiento-Barbieri

Universidad de los Andes

Agenda

Agenda

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

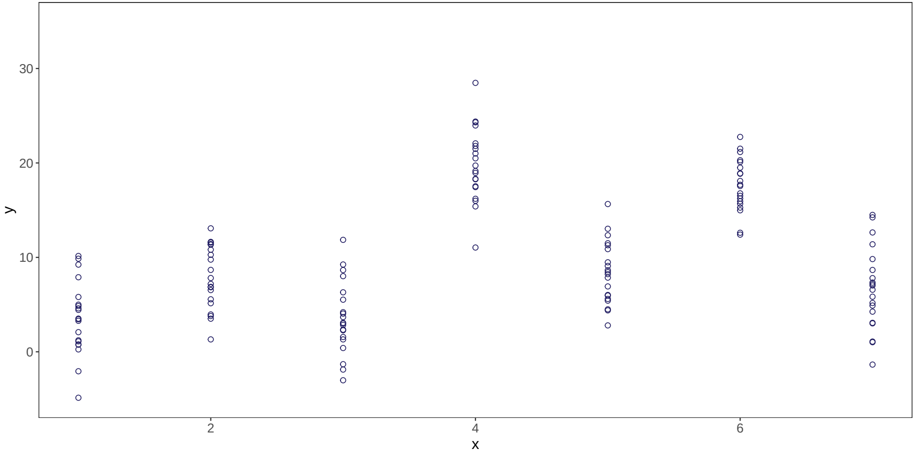
Recap: Train and Test Sets. In-Sample and Out-of-Sample Prediction.

- 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:

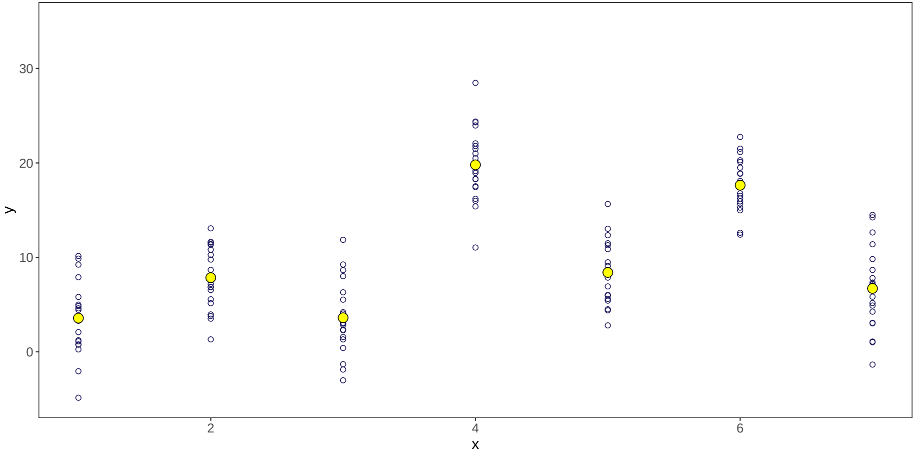
$$y = f(X) + u \tag{1}$$

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

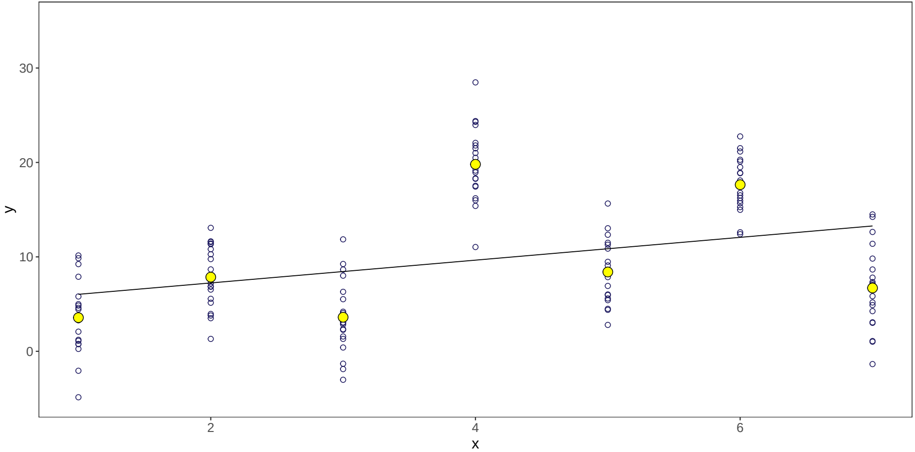
Recap: In-Sample Prediction and Overfit



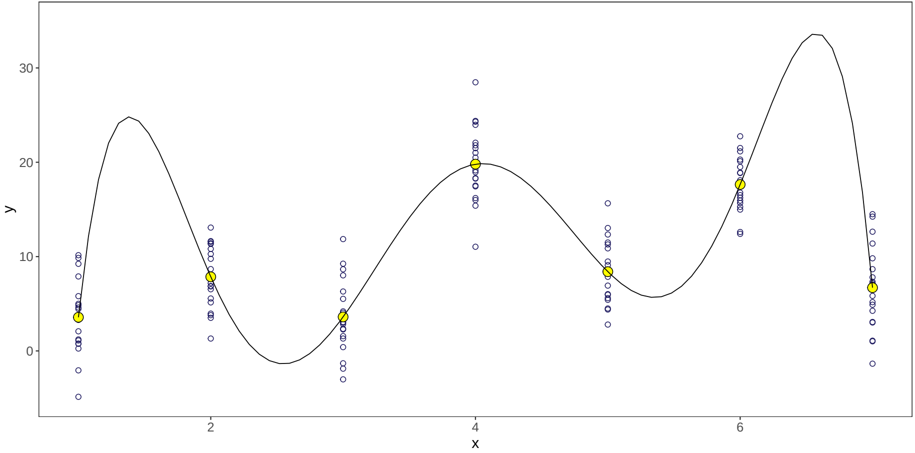
Recap: In-Sample Prediction and Overfit



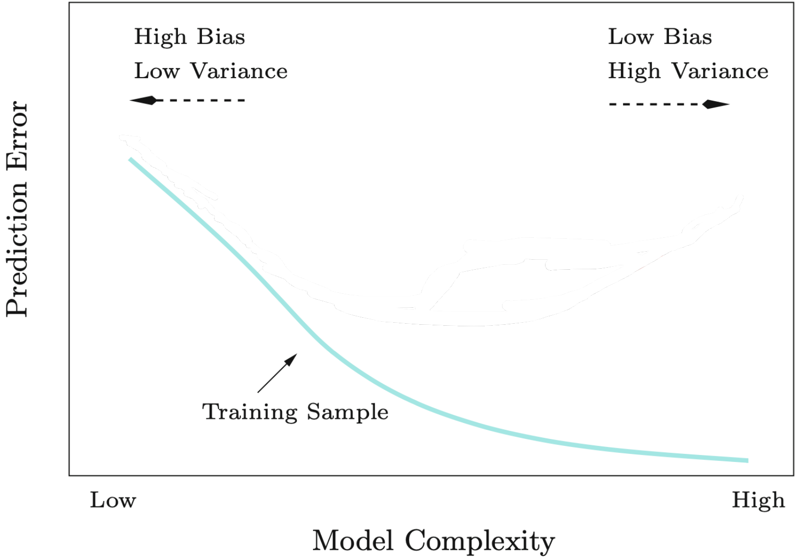
Recap: In-Sample Prediction and Overfit



Recap: In-Sample Prediction and Overfit



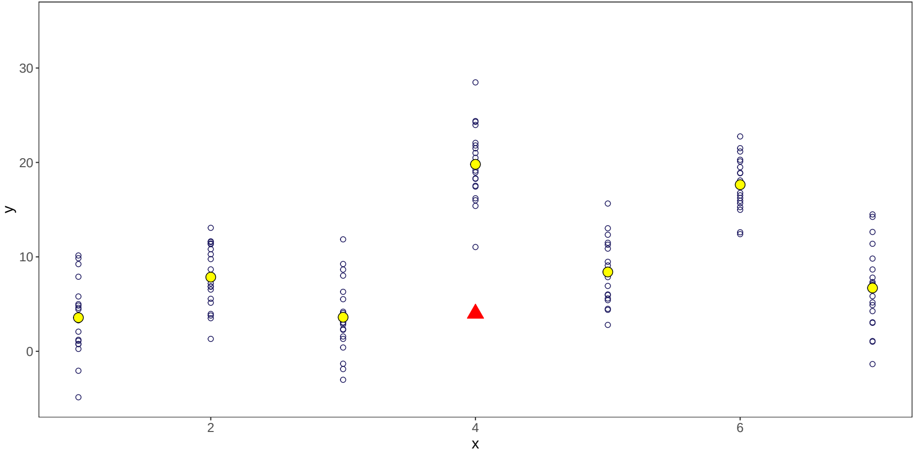
Recap: In-Sample Prediction and Overfit



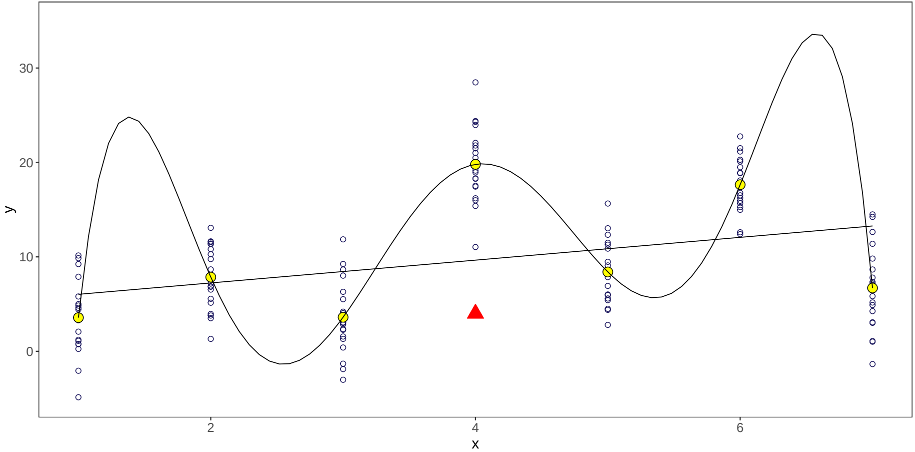
Recap: Out-of-Sample Prediction and Overfit

- ML nos interesa la predicción fuera de muestra

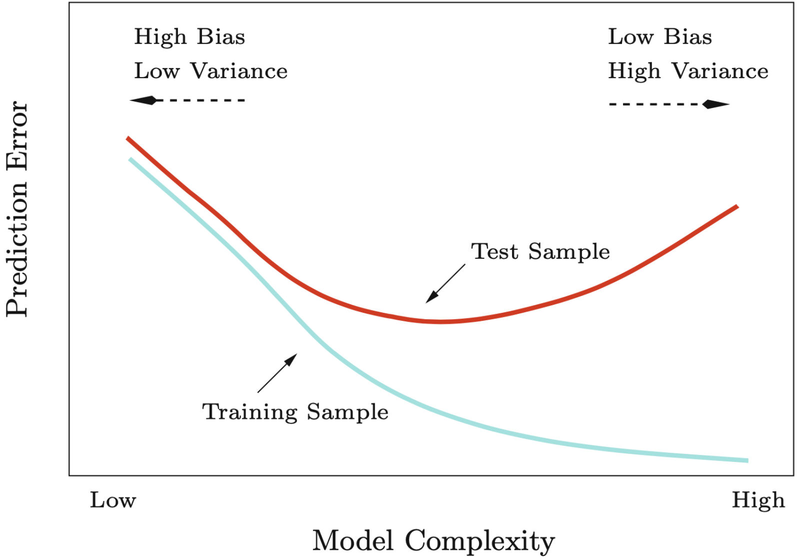
Recap: Out-of-Sample Prediction and Overfit



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

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

Demo

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

Agenda

Agenda

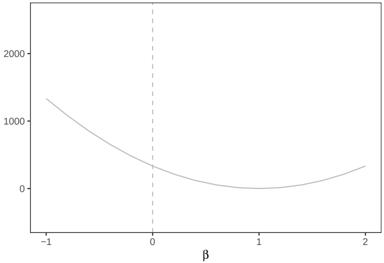
Regularización: Motivación

- 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

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

OLS 1 Dimension

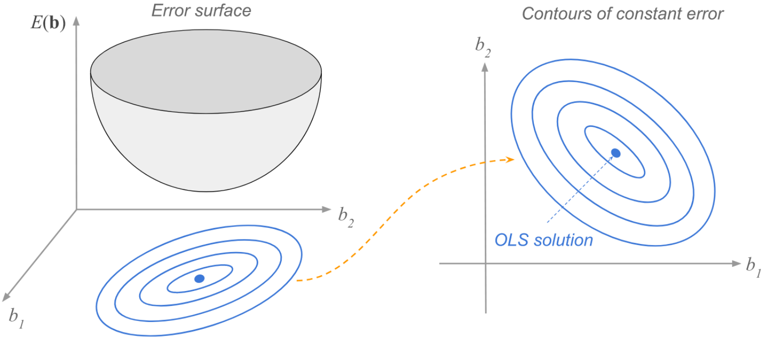
$$min_\beta E(\beta) = \sum_{i=1}^{n}(y_i - x_i\beta)^2 \tag{3}$$



App

OLS 2 Dimensiones

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



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**

Agenda

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 $c > 0$ tal que $\hat{\beta}(\lambda)$ es la solución a

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

$$\text{sujeto a}$$

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

Ridge: Intuición en 1 Dimension

Problema como optimización restringida

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

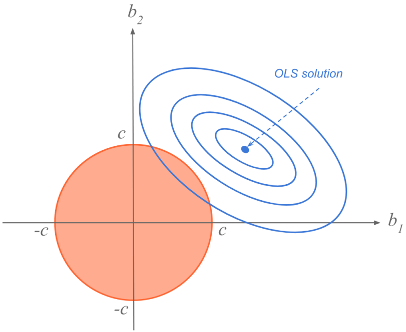
$$min_{\beta}E(\beta) = \sum_{i=1}^{n}(y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \tag{10}$$

sujeto a

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

Ridge: Intuición en 2 Dimensiones

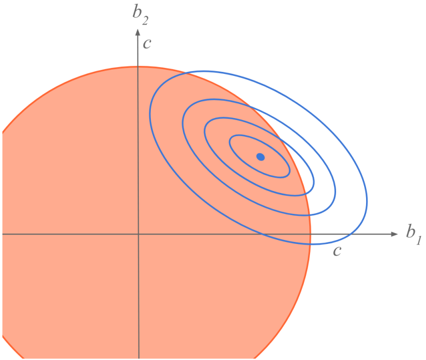
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \text{ s.a } \left((\beta_1)^2 + (\beta_2)^2\right) \leq c \quad (11)$$



Fuente: https://allmodelsarewrong.github.io

Ridge: Intuición en 2 Dimensiones

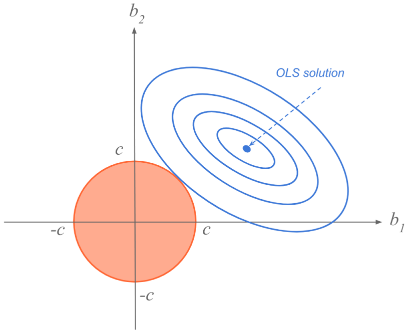
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \text{ s.a } \left((\beta_1)^2 + (\beta_2)^2\right) \leq c \tag{12}$$



Fuente: https://allmodelsarewrong.github.io

Ridge: Intuición en 2 Dimensiones

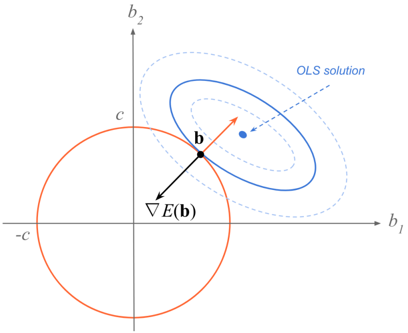
$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - x_{i1}\beta_1 - x_{i1}\beta_2)^2 \text{ s.a } \left((\beta_1)^2 + (\beta_2)^2\right) \leq c \tag{13}$$



Fuente: https://allmodelsarewrong.github.io

Ridge: Intuición en 2 Dimensiones

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



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

Términos generales

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

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

- Ridge

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

Ridge vs OLS

- Ridge es sesgado $E(\hat{\beta}_{ridge}) \neq \beta$

Ridge vs OLS

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

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

Escala de las variables

- La escala de las variables importa en Ridge, mientras que en OLS no.
- Por qué?

Escala de las variables

Escala de las variables

Ridge no es invariante a las escala

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

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - \beta_0^z - \beta_1^z z_i)^2 + \lambda (\beta_1^z)^2 \tag{15}$$

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

- Demo: baticomputer, math: HW

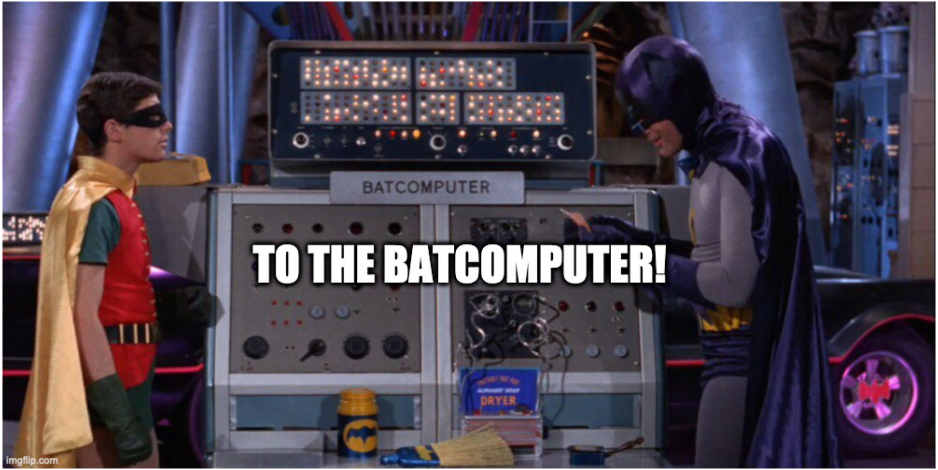


photo from https://www.dailydot.com/parsec/batman-1966-labels-tumblr-twitter-vine/

Agenda

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.

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

- 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 $i = \lambda_{min}, \ldots, \lambda_{max}$ {

    - 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 $\lambda_i = \lambda^*$

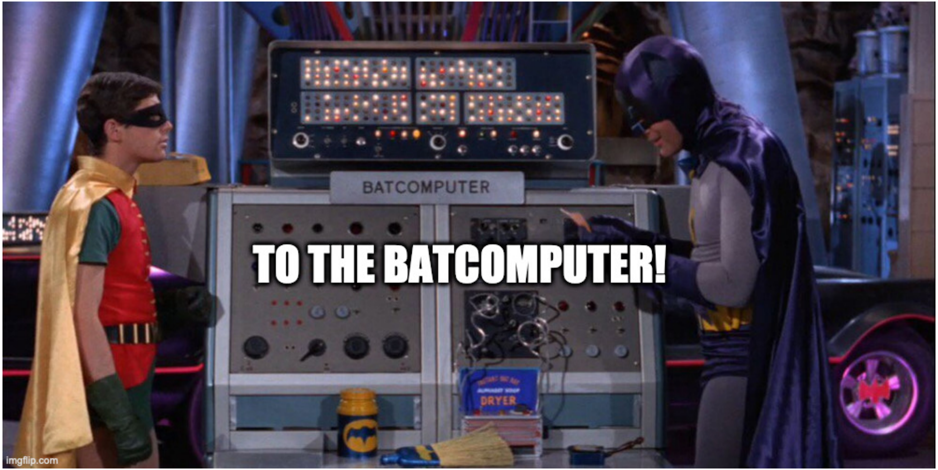


photo from https://www.dailydot.com/parsec/batman-1966-labels-tumblr-twitter-vine/

Agenda

Ridge as Data Augmentation (1)

RidgeDataAug

- Add λ additional points

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

Ridge as Data Augmentation (2)

RidgeDataAug

- Add a single point

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

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 $k > n$

$$min_{\beta} E(\beta) = \sum_{i=1}^{n} (y_i - \sum_{j=1}^{k} x'_{ij}\beta_j)^2 + \lambda(\sum_{j=1}^{k} \beta_j)^2 \tag{19}$$

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

Ridge when $k > n$