DMC ONLINE

#YoMeQuedoEnCasa
#YoMeSumo

MODELOS LINEALES REGULARIZADOS

DMC Perú

Soy

Victor Acevedo

Sub Gerente en Modelos de Riesgo Crediticio en Banco de Crédito BCP

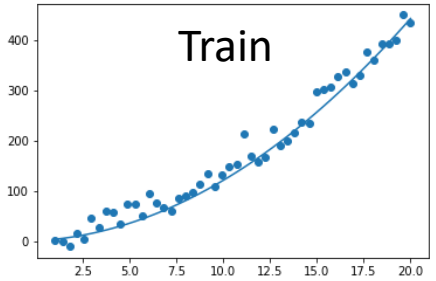
Agenda

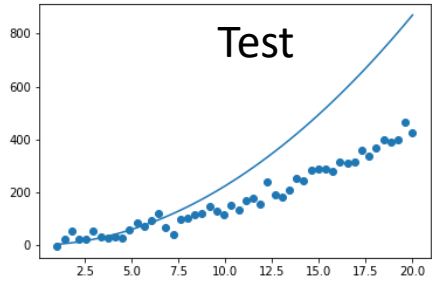
1. Trade off: sesgo – varianza
2. Regresión Ridge
3. Regresión Lasso

Trade off: sesgo - varianza

Name of Employee	Sales	Quarter	Country
Jon	1000	1	US
Mark	300	1	Japan
Tina	400	1	Brazil
Maria	500	1	UK
Bill	800	1	US
Jon	1000	2	Brazil
Mark	500	2	Japan
Tina	700	2	Brazil
Maria	50	2	US
Bill	60	2	US
Jon	1000	3	US
Mark	900	3	Japan
Tina	750	3	Brazil
Maria	200	3	UK
Bill	300	3	Brazil
Jon	1000	4	Japan
Mark	900	4	Japan
Tina	250	4	Brazil
Maria	750	4	UK
Bill	50	4	US

Modelamiento

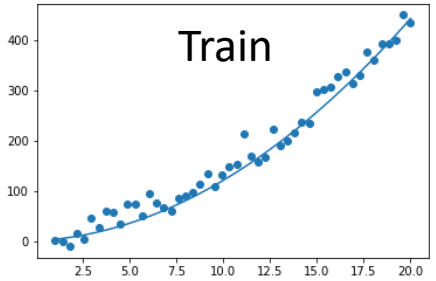


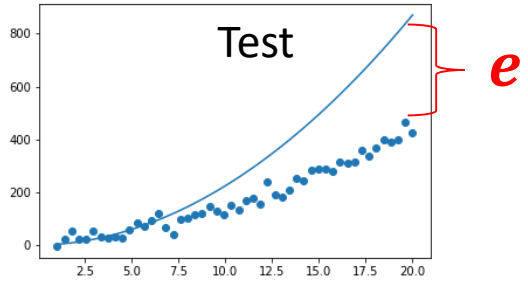


Trade off: sesgo - varianza

Name of Employee	Sales	Quarter	Country
Jon	1000	1	US
Mark	300	1	Japan
Tina	400	1	Brazil
Maria	500	1	UK
Bill	800	1	US
Jon	1000	2	Brazil
Mark	500	2	Japan
Tina	700	2	Brazil
Maria	50	2	US
Bill	60	2	US
Jon	1000	3	US
Mark	900	3	Japan
Tina	750	3	Brazil
Maria	200	3	UK
Bill	300	3	Brazil
Jon	1000	4	Japan
Mark	900	4	Japan
Tina	250	4	Brazil
Maria	750	4	UK
Bill	50	4	US

Modelamiento





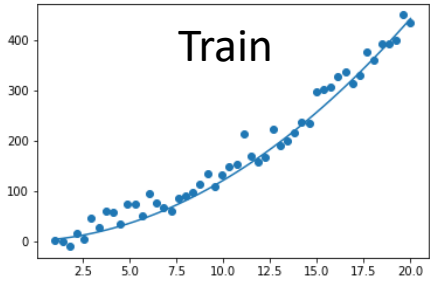
e

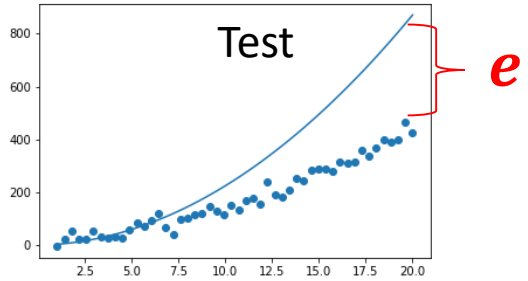
- Durante el entrenamiento, el modelo aparentaba explicar muy bien el PGD.
- En producción, se evidencia que el modelo sobreestima el efecto de la variable explicativa.

Trade off: sesgo - varianza

Name of Employee	Sales	Quarter	Country
Jon	1000	1	US
Mark	300	1	Japan
Tina	400	1	Brazil
Maria	500	1	UK
Bill	800	1	US
Jon	1000	2	Brazil
Mark	500	2	Japan
Tina	700	2	Brazil
Maria	50	2	US
Bill	60	2	US
Jon	1000	3	US
Mark	900	3	Japan
Tina	750	3	Brazil
Maria	200	3	UK
Bill	300	3	Brazil
Jon	1000	4	Japan
Mark	900	4	Japan
Tina	250	4	Brazil
Maria	750	4	UK
Bill	50	4	US

Modelamiento





e

- Durante el entrenamiento, el modelo aparentaba explicar muy bien el PGD.
- En producción, se evidencia que el modelo sobreestima el efecto de la variable explicativa.

$e =$

Bias + **Variance** + **Irreductible**

Originado por supuestos erróneos.

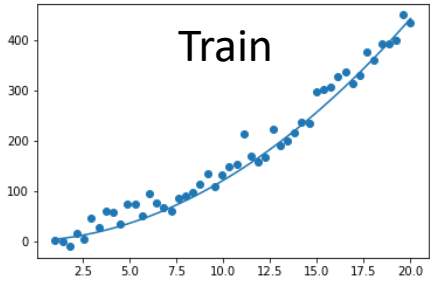
Sensibilidad excesiva a pequeñas variaciones de los datos.

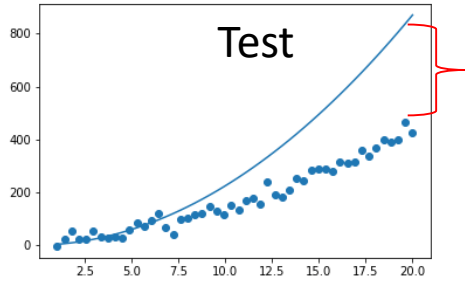
Ruidos intrínsecos de los datos de entrenamiento.

Trade off: sesgo - varianza

Name of Employee	Sales	Quarter	Country
Jon	1000	1	US
Mark	300	1	Japan
Tina	400	1	Brazil
Maria	500	1	UK
Bill	800	1	US
Jon	1000	2	Brazil
Mark	500	2	Japan
Tina	700	2	Brazil
Maria	50	2	US
Bill	60	2	US
Jon	1000	3	US
Mark	900	3	Japan
Tina	750	3	Brazil
Maria	200	3	UK
Bill	300	3	Brazil
Jon	1000	4	Japan
Mark	900	4	Japan
Tina	250	4	Brazil
Maria	750	4	UK
Bill	50	4	US

Modelamiento





e

- Durante el entrenamiento, el modelo aparentaba explicar muy bien el PGD.
- En producción, se evidencia que el modelo sobreestima el efecto de la variable explicativa.

$e =$

Bias + **Variance** + **Irreductible**

Originado por supuestos erróneos.

Sensibilidad excesiva a pequeñas variaciones de los datos.

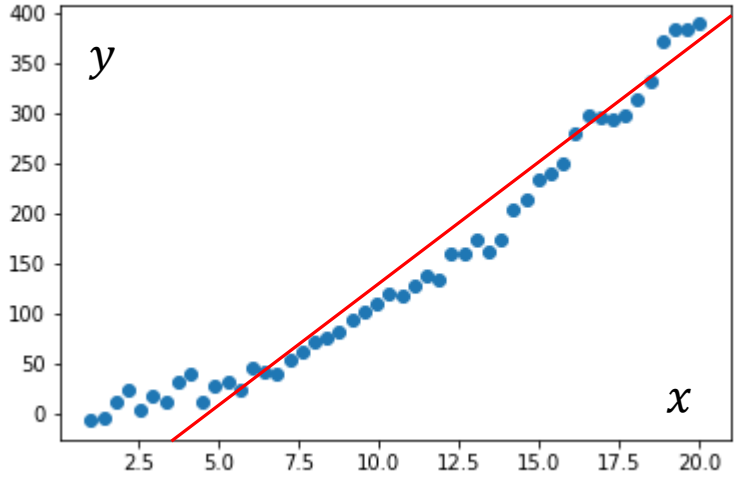
Ruidos intrínsecos de los datos de entrenamiento.

Técnicas para evitar el overfitting*:

- **Regresión Ridge**
- **Regresión Lasso**
- GNN
- Elastic Net
- Alasso
- ReLasso
- SCAD

* En modelos lineales

Regresión Ridge

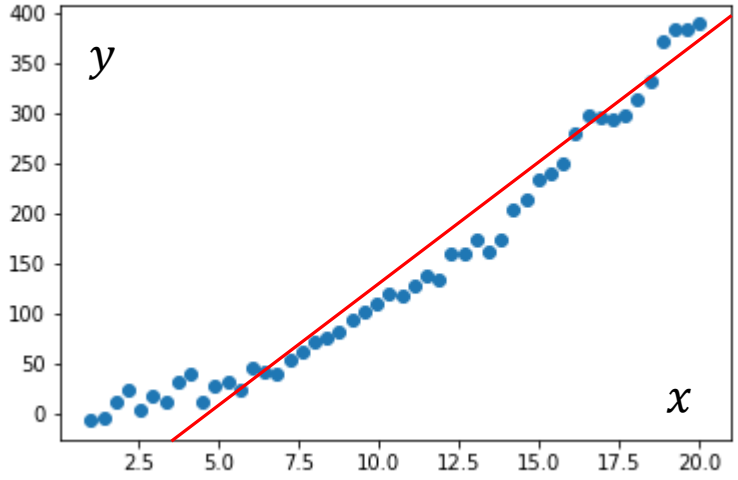


$$\hat{y}_{n1} = x_{nk}\hat{\beta}_{k1}$$

MCO:

$$Argmín\ L\colon\ (y_{n1} - \hat{y}_{n1})'(y_{n1} - \hat{y}_{n1})$$

Regresión Ridge

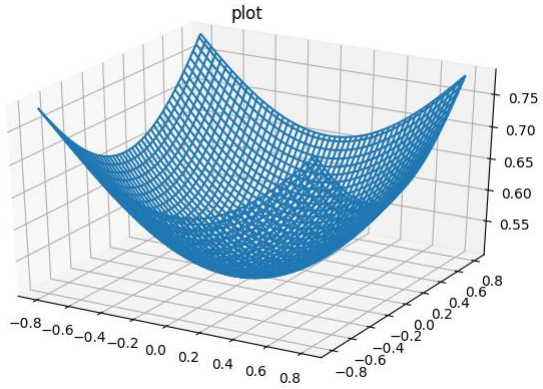


$$\hat{y}_{n1} = x_{nk}\hat{\beta}_{k1}$$

MCO:

$$Argmín\ L\colon\ (y_{n1} - \hat{y}_{n1})'(y_{n1} - \hat{y}_{n1})$$

$$\hat{\beta}_{mco} = ({x_{nk}}' x_{nk})^{-1} x'_{nk} y_{n1}$$



Mientras se cumplan los supuestos clásicos, este estimador es MELI:

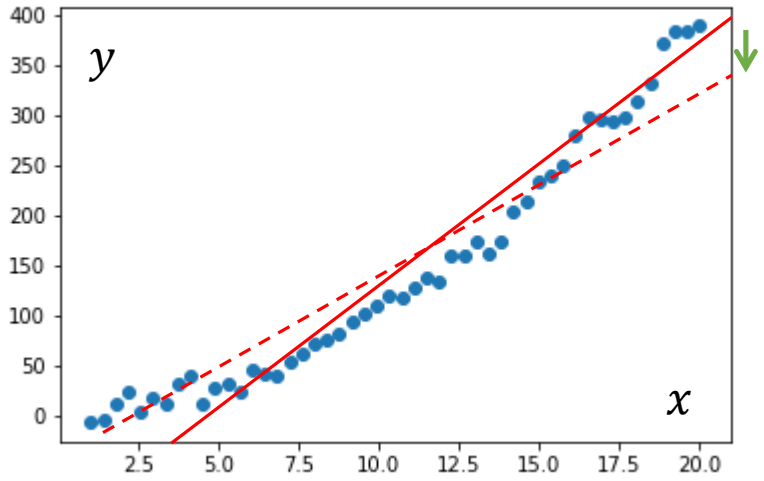
- Insesgado

$$E\left(\widehat{\beta_j}/x_{11}, x_{21}, \dots, x_{k-1N}, x_{kn}\right) = \beta_j, j = 0{,}1, \dots, k$$

- Eficiente

$$\text{Var}\left(\hat{\beta}/x_{kn}\right) = (x'x)^{-1}x'y \rightarrow$$ La mínima dispersión posible respecto a otros estimadores lineales e insesgados.

Regresión Ridge



$$\hat{y}_{n1} = x_{nk}\hat{\beta}_{k1}$$

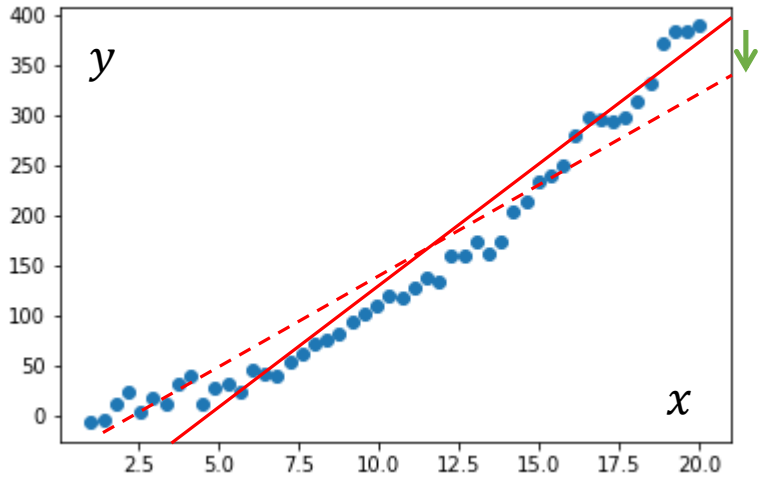
Ridge:

$$Argmín\ L\colon\ (y_{n1} - \hat{y}_{n1})'(y_{n1} - \hat{y}_{n1})$$

$$s.a.:\ \|\beta\|'_2\|\beta\|_2 < c$$

Norma L2: distancia euclidiana entre vectores

Regresión Ridge



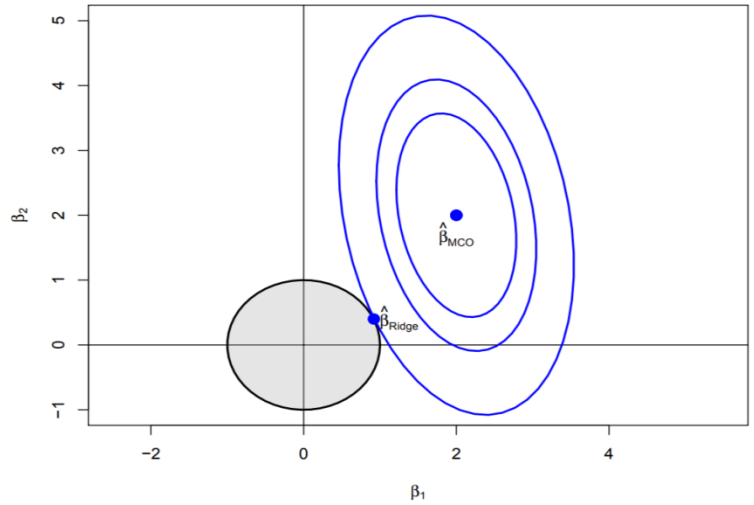
$$\hat{y}_{n1} = x_{nk}\hat{\beta}_{k1}$$

Ridge:

$$Argmín\ L:\ (y_{n1} - \hat{y}_{n1})'(y_{n1} - \hat{y}_{n1})$$

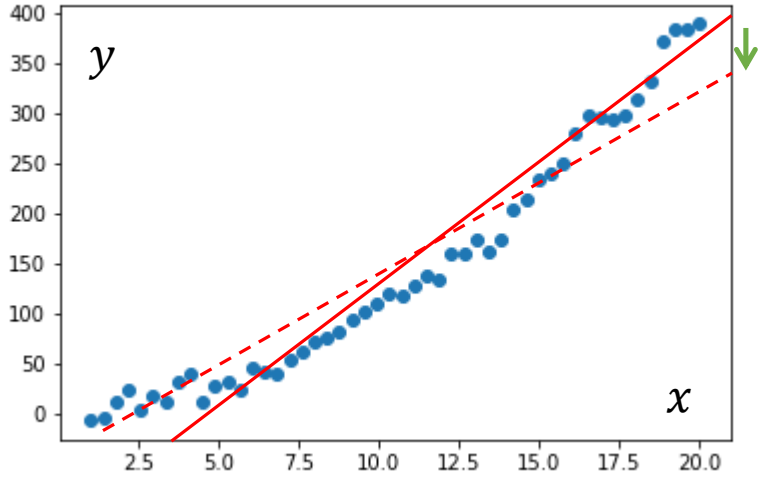
$$s.a.:\ \|\beta\|'_2\|\beta\|_2 < c \longrightarrow$$

Caso bivariado: $(\hat{\beta}_1)^2 + (\hat{\beta}_2)^2 < c$



- Esta restricción al conjunto de posibles valores obliga a disminuir la relevancia de las variables.
- Gráficamente, disminuye la pendiente de la variable.
- En consecuencia, disminuye el overfitting.

Regresión Ridge



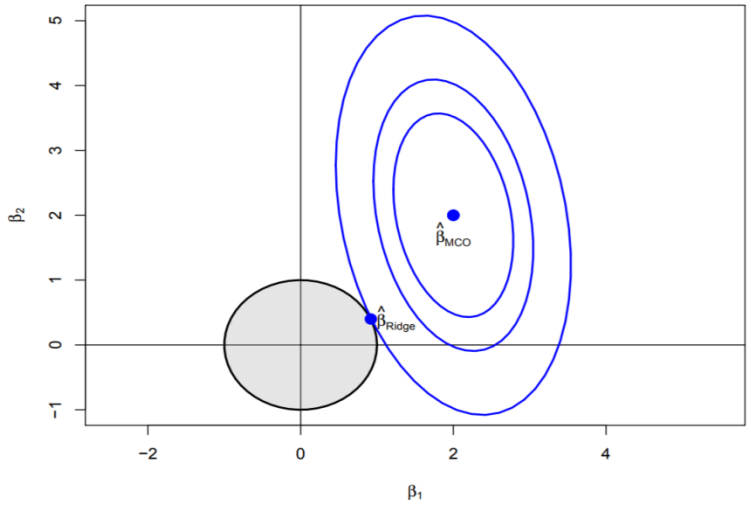
$$\hat{y}_{n1} = x_{nk}\hat{\beta}_{k1}$$

Ridge:

$$Argmín\ L\colon\ (y_{n1} - \hat{y}_{n1})'(y_{n1} - \hat{y}_{n1})$$

$$s.a.:\ \|\beta\|'_2\|\beta\|_2 < c \longrightarrow$$

Caso bivariado: $(\hat{\beta}_1)^2 + (\hat{\beta}_2)^2 < c$

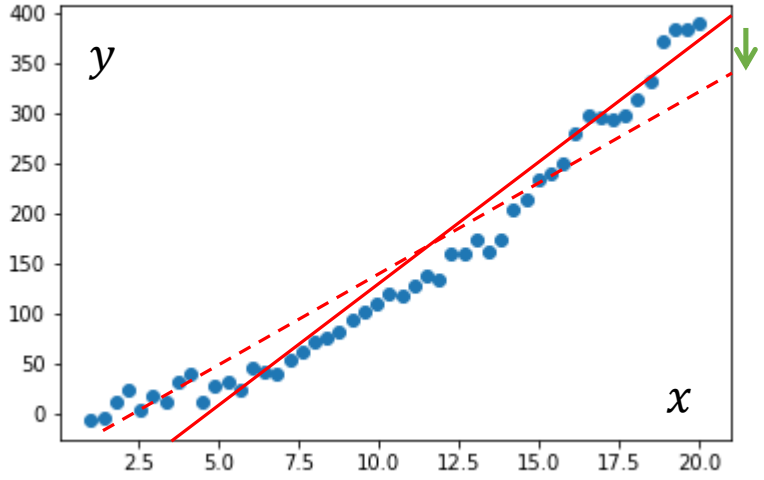


$$\hat{\beta}_{ridge} = (x_{nk}{}' x_{nk} - \lambda A)^{-1} x'_{nk} y_{n1}$$

A: *matriz kxk, parecida a* I.

λ: *hiperparámetro de penalización* L_2

Regresión Ridge



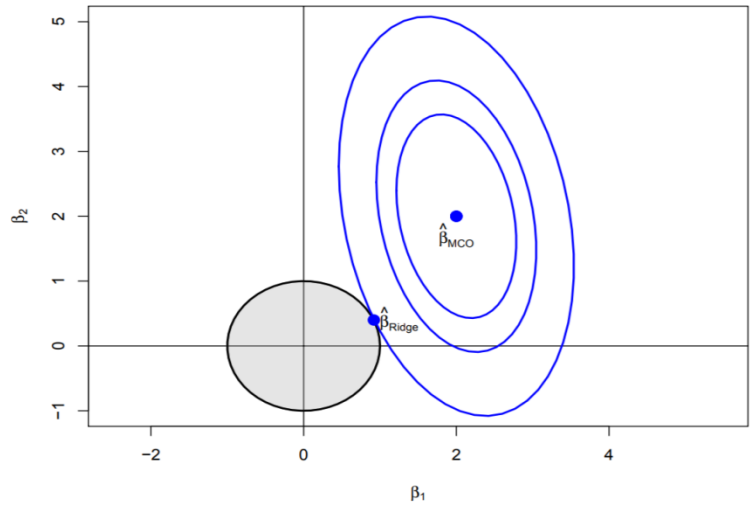
$$\hat{y}_{n1} = x_{nk}\hat{\beta}_{k1}$$

Ridge:

$$Argmín\ L:\ (y_{n1} - \hat{y}_{n1})'(y_{n1} - \hat{y}_{n1})$$

$$s.a.:\ \|\beta\|'_2\|\beta\|_2 < c \longrightarrow$$

Caso bivariado: $(\hat{\beta}_1)^2 + (\hat{\beta}_2)^2 < c$



$$\hat{\beta}_{ridge} = (x_{nk}{}' x_{nk} - \lambda A)^{-1} x'_{nk} y_{n1}$$

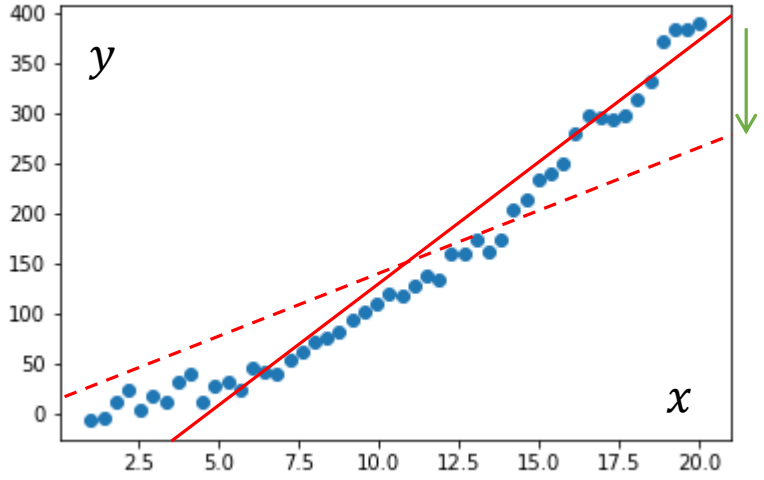
A: $matriz\ kxk, parecida\ a$ I.

λ: $hiperparámetro\ de\ penalización\ L_2$

- ¿Es necesario el escalamiento de variables?
- ¿Qué pasa con el intercepto?
- ¿La función de pérdida es mayor que en el caso anterior?

- ¿Es MELI?
- ¿Qué pasa con la varianza del estimador?
- ¿Los estimadores pueden llegar a ser cero?

Regression LASSO



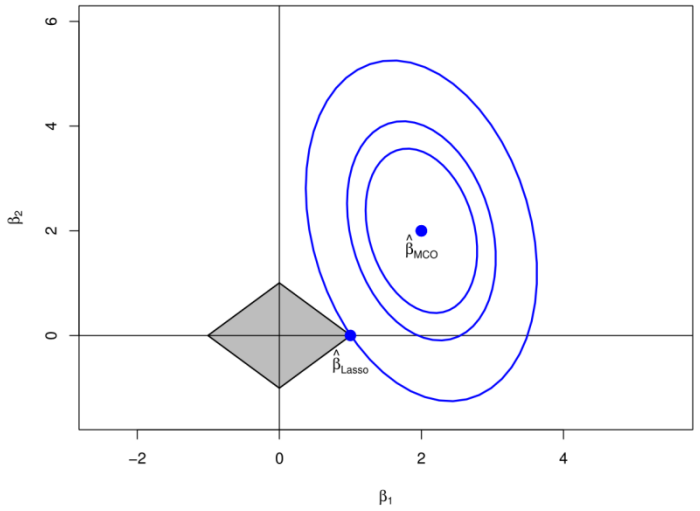
$$\hat{y}_{n1} = x_{nk}\hat{\beta}_{k1}$$

Lasso:

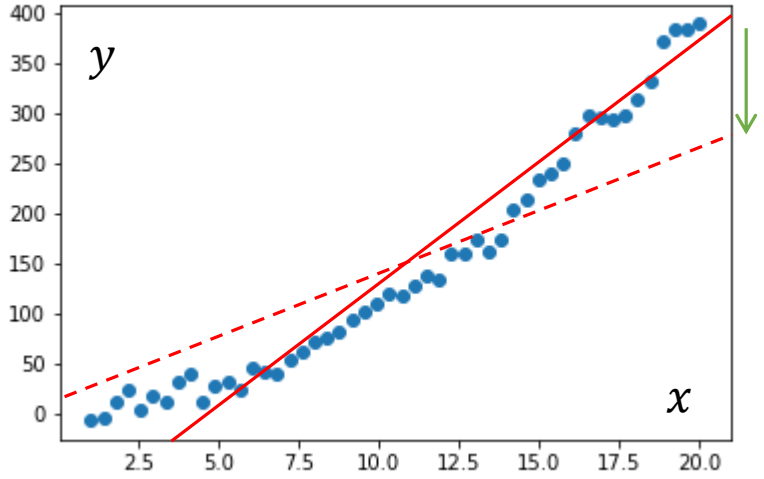
$$Argmín\ L\colon\ (y_{n1} - \hat{y}_{n1})'(y_{n1} - \hat{y}_{n1})$$

$$s.a.:\ \|\beta\|'_1\|\beta\|_1 < c \longrightarrow$$

Caso bivariado: $|\hat{\beta}_1| + |\hat{\beta}_2| < c$



Regression LASSO



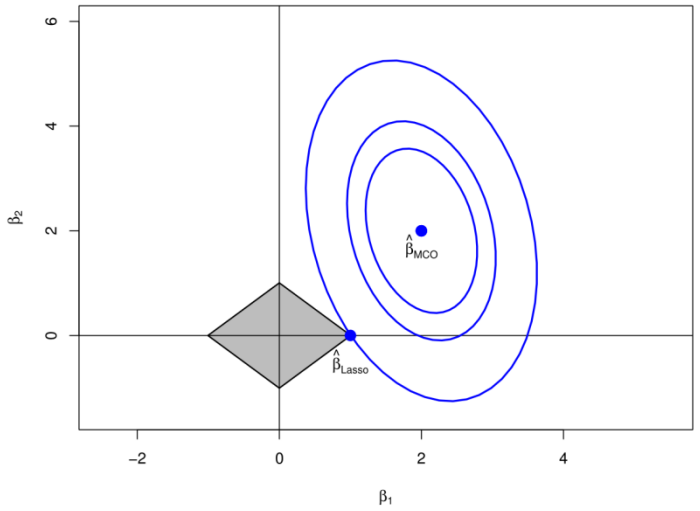
$$\hat{y}_{n1} = x_{nk}\hat{\beta}_{k1}$$

Lasso:

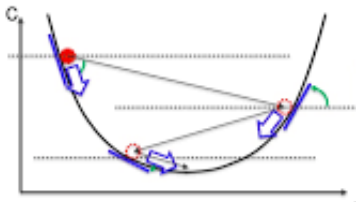
$$Argmín\ L\colon\ (y_{n1} - \hat{y}_{n1})'(y_{n1} - \hat{y}_{n1})$$

$$s.a.:\ \|\beta\|'_1\|\beta\|_1 < c \longrightarrow$$

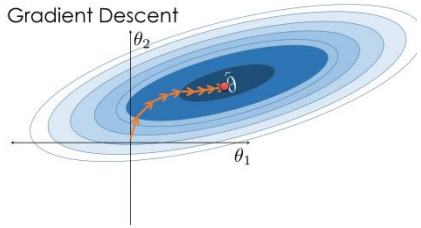
Caso bivariado: $|\hat{\beta}_1| + |\hat{\beta}_2| < c$



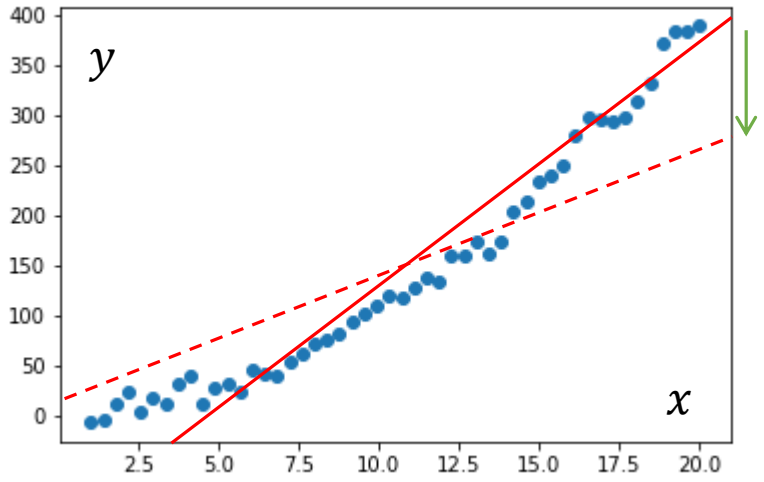
- Python Sklearn-Learn:
  Optimización mediante SGD:



≡



Regression LASSO



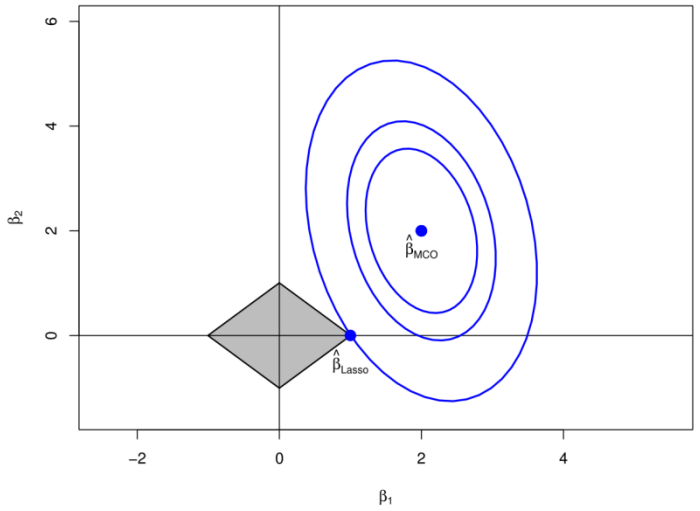
$$\hat{y}_{n1} = x_{nk}\hat{\beta}_{k1}$$

Lasso:

$$Argmín\ L\colon\ (y_{n1} - \hat{y}_{n1})'(y_{n1} - \hat{y}_{n1})$$

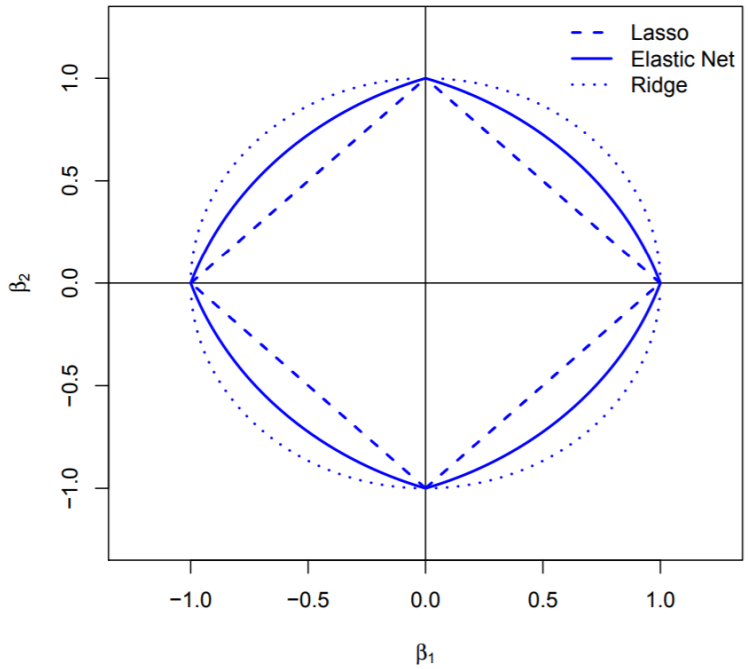
$$s.a.:\ \|\beta\|'_1\|\beta\|_1 < c \longrightarrow$$

Caso bivariado: $|\hat{\beta}_1| + |\hat{\beta}_2| < c$

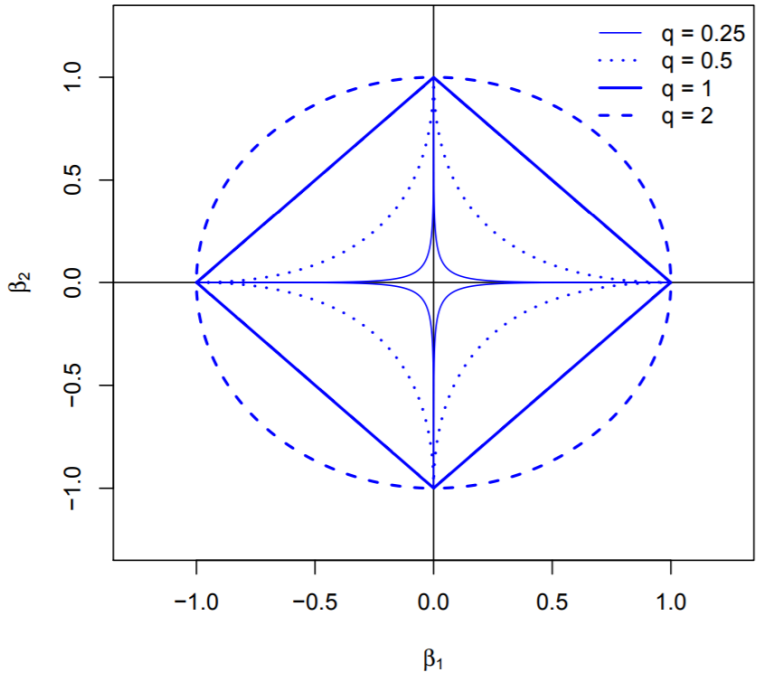


- La función de pérdida no es diferenciable en todos los puntos.
- Más estricto que Ridge
- Selección de variables
- Se evita en XGBoost
- Alasso

Diversas normas



- Elastic net es un caso intermedio entre LASSO y Ridge.
- Asigna pesos para L1 y L2.



- A menor orden de norma, tendremos un proceso más penalizador.

Fuentes

Castro, Sebastián. «Estimación y selección de variables en grandes dimensiones»

Geron, Aurelien (2017). «Hands-On Machine Learning with Scikit Learn»

Hui Zou (2006) «The adaptive Lasso and Its Oracle Properties»

James & Witten. «Introduction to Statistical Learning with Applications in R»

Maqbool, Faisal (2011). «Ordinal Ridge Regression with Categorical Predictors»

Penn State – Elberly College of Science «Applied Data Mining and Statistical Learning»

Tutz, Gerhard (2016).«Regularized regression for categorical data»

Agradecimientos

Alexis Coronado Oritz
Data Scientist Senior

Fabrizio Chavez Anampa
Data Scientist

Luis Chacón Montalván
Sub gerente - BCP

Luis Gavidia Msc (c) Statistics
Data Scientist

Erick Saavedra Palacios
Sub gerente - BCP