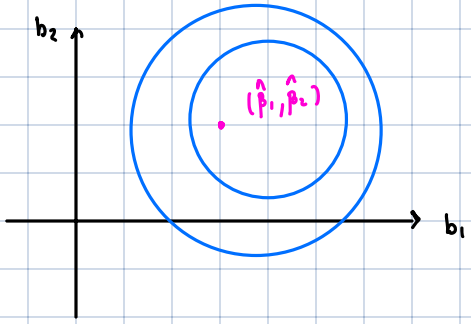
Recuerden nuestro modelo lineal

$$Y_i = X_i'\beta + u_i \quad , \quad E[u_i X_i] = 0.$$

Tenemos que $\hat{\beta}_n$ es la solución a

$$\hat{\beta}_n = \underset{b}{\operatorname{argmin}} \sum_{i=1}^{n} \frac{(Y_i - X_i' b)^2}{2n}$$



Las curvas de nivel se minimizan en $(\hat{\beta}_1, \hat{\beta}_2)$

1. Ridge

La regresión Ridge propone estimar

$$\hat{\beta}_n^R = \underset{b}{\operatorname{argmin}} \sum_{i=1}^{n} \frac{(Y_i - X_i' b)^2}{2n} + \underbrace{\lambda \, b^T b / 2}_{\text{Penalizamos } \|b\|_2^2 := \sum_{j=1}^{n} b_j^2} ,$$

donde λ es un parámetro externo que elegimos (no estimamos directamente). A esto se le conoce como un hiperparámetro.

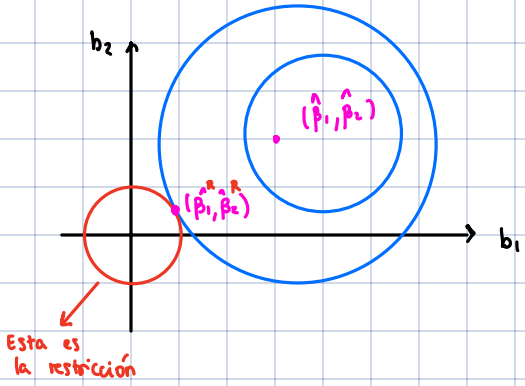
Derivando $\quad Q_n(b;\lambda) = \frac{1}{2n}(Y - Xb)^T(Y - Xb) + \frac{\lambda}{2} b^T b$

$$- X^T(Y - X\hat{\beta}_n^R) + \lambda \hat{\beta}_n^R = 0$$

$$\Rightarrow \quad X^T X \hat{\beta}_n^R + \lambda \hat{\beta}_n^R = X^T Y$$

$$\Rightarrow \quad \hat{\beta}_n^R = \underbrace{(X^T X + \lambda I_n)}^{-1} X^T Y$$

Incluso cuando $X^T X$ es cercano a ser singular, al sumar este último término hacemos que una solución siempre exista.



* La regresión Ridge acerca los coeficientes hacia cero (shrinkage)

Sin embargo, noten que hemos introducido sesgo incluso en muestras finitas. Suponemos $E[u_i | X_i] = 0$.

$$\hat{\beta}_n^R = (X^T X + \lambda I_k)^{-1} X^T X \beta + (X^T X + \lambda I_k)^{-1} X^T u$$

$$\Rightarrow \quad E[\hat{\beta}_n | X] = (X^T X + \lambda I_k)^{-1} X^T X \beta \neq \beta.$$

Es decir, Ridge introduce sesgo a costa de reducir la varianza de las predicciones. Noten como estabiliza la varianza en el caso $k=1$ y homocedástico:

$$\hat{\beta}_{in} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} (x_i^2 + \lambda)}, \quad Var(\hat{\beta}_1 | X) = \frac{\sigma^2}{\sum_{i=1}^{n} (x_i^2 + \lambda)}$$

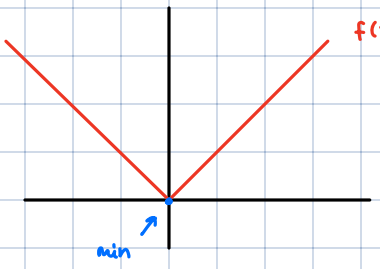
Este denominador está más lejos de 0.

Finalmente, ya que Ridge NO impone una solución de esquina (i.e. que algún $\hat{\beta}_j$ sea 0), no sirve para hacer selección de regresores. Por ello, estudiaremos un siguiente estimador.

2. Minimización Convexa

Existen casos donde queremos minimizar una función objetivo que es convexa. Por tanto, definitivamente podemos encontrar un óptimo. Sin embargo, no siempre son objetos diferenciables.

Se halla como $\partial f(x) = 2x = 0$



¿cómo se halla?

2.1. Subgradiente

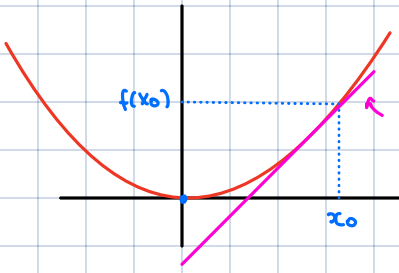
Def.: g es subgradiente de f en x_0 si $\forall x \in \mathbb{R}^K$ tenemos

$$f(x) - f(x_0) \geq g^T(y - x_0)$$

Para ganar intuición consideremos $K = 1$. La definición implica que

$$f(x) \geq f(x_0) - g(x_0) + gx$$
$$= f(x_0) + g(x - x_0)$$

Ejemplo:



$\tilde{f}(x) = f(x_0) + g(x - x_0)$ [linealización alrededor de x_0]

Es decir, la subgradiente es una de las posibles pendientes de las líneas que pasan por el punto $(x_0, f(x_0))$, y que siempre estén por debajo de $f(x)$. Vemos que en el caso de $f(x) = x^2$ solo hay una posible línea.

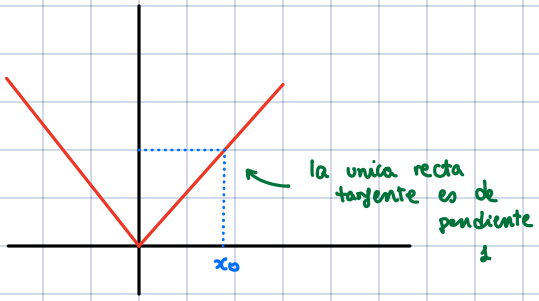
Def.: El set de todos las subgradientes en x_0 se conoce como subdiferencial en x_0.

$$\partial f(x_0) = \{ g \in \mathbb{R}^K : f(x) - f(x_0) \geq g^T(x - x_0) \quad \forall x \in \mathbb{R}^K \}.$$

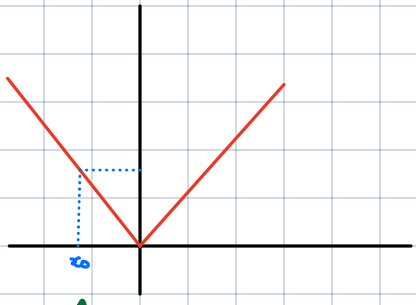
Decimos que una función es diferenciable en x_0 si su subdiferencial en x_0 solo incluye un único elemento.

$$\partial f(x_0) = \{g\} \iff g = \nabla f(x).$$

Veamos ahora el valor absoluto:

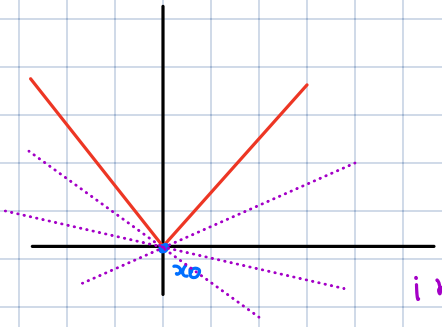


$\partial f(x_0) = 1$ si $x_0 > 0$



la única recta tangente es de pendiente -1

$\partial f(x_0) = -1$ si $x_0 < 0$



¡Hay varias posibles tangentes!

$\partial f(x_0) = [-1, 1]$ si $x_0 = 0$

Es decir, cualquier pendiente entre -1 y 1 es un subgradiente.

Por ello, decimos que:

$$\partial |x| = \begin{cases} -1 & , \quad x < 0 \\ [-1, 1] & , \quad x = 0 \\ 1 & , \quad x > 0 \end{cases}.$$

Proposición.- Sea M el set de puntos mínimos de una función convexa f:

$$M = \{x^* \in \mathbb{R}^K : \ f(x^*) \leq f(x) \ \ \forall x \in \mathbb{R}^K\}$$

Entonces

$$x^* \in M \iff 0 \in \partial f(x^*).$$

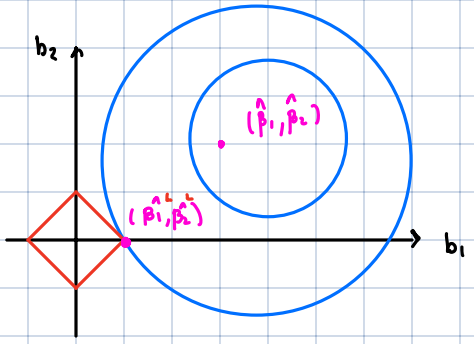
Por ello, podemos decir que $x_0 = 0$ minimiza $f(x) = |x|$:

$$0 \in \partial |x| \quad \text{cuando} \quad x = x_0 \quad \Rightarrow \quad x^* \in M.$$

3. LASSO

La regresión Ridge propone estimar

$$\hat{\beta}_n^L = \underset{b}{\operatorname{argmin}} \sum_{i=1}^{n} \frac{(y_i - x_i' b)^2}{2n} + \lambda \sum_{j=1}^{k} |b_j| = \frac{\| Y - Xb \|_2^2}{2n} + \lambda \| b \|_1$$



Está diseñado para hallar soluciones de esquina donde algún $\hat{\beta}_j$ será igual a 0. Es decir, selecciona regresores.

Importantemente, este es un problema de minimización convexa.

3.1. Solución Analítica (caso especial)

Consideremos

$$\sum_{i=1}^{n} \frac{x_{ij}\, x_{il}}{n} = \begin{cases} 0 & j \neq l \\ 1 & j = l \end{cases}$$

} regresores ortogonales y segundo momento igual a 1.

En ese caso

$$\frac{X'X}{n} = I_k$$

$$\Rightarrow \quad \hat{\beta}^{MCO} = \frac{X'Y}{n},$$

por lo que

$$\hat{\beta}_j^{MCO} = \sum_{i=1}^{n} \frac{x_i y_i}{n}.$$

Vamos a definir $(x)^+ = \max\{x, 0\}$

$$\operatorname{sign}(x) = \begin{cases} -1 & , x < 0 \\ 0 & , x = 0 \\ 1 & , x > 0 \end{cases}$$

Proposición :- Suponer $X'X/n = I_K$. Entonces el estimador LASSO resuelve la minimización de

$$Q_{n,\lambda}(b) = \left\{ \frac{1}{2n} \|Y - Xb\|_2^2 + \lambda \|b\|_1 \right\}$$

y satisface

$$\hat{\beta}_j = \text{sign}(\hat{\beta}_j^{OLS}) \left(|\hat{\beta}_j^{OLS}| - \lambda \right)^+ \qquad \forall j = 1, 2, \dots, K.$$

Achica los coeficientes en una magnitud λ.
Si $\lambda > |\hat{\beta}_j^{OLS}|$ lo reemplaza $\hat{\beta}_j^{OLS}$ por $\hat{\beta}_j = 0$.

Shrinkage + selection

proof :

Definimos

$$\underbrace{v}_{K\times 1} + \underbrace{c}_{1\times 1} \underbrace{S}_{\text{set de vectores } K\times 1} = \{ v + cg : g \in S \}.$$

Usando esa notación definimos

$$\partial Q_{n,\lambda}(b) = -\frac{X'(Y - Xb)}{n} + \lambda \partial \|b\|_1$$

$$= -\frac{X'Y}{n} + \frac{X'X}{n} b + \lambda \partial \|b\|_1$$

$$= -\hat{\beta}^{OLS} + I_K b + \lambda \partial \|b\|_1$$

$$= -(\hat{\beta}^{OLS} - b) + \lambda \partial \|b\|_1$$

$$= -\begin{pmatrix} \hat{\beta}_1^{OLS} - b_1 - \lambda \partial |b_1| \\ \vdots \\ \hat{\beta}_K^{OLS} - b_K - \lambda \partial |b_K| \end{pmatrix}.$$

Noten que podemos resolver elemento por elemento. En concreto, el estimador LASSO satisface

$$0 \in \hat{\beta}_j^{OLS} - \hat{\beta}_j - \lambda \partial |\hat{\beta}_j|$$

Entonces $\hat{\beta}_j = 0$ si y solo si

$$0 \in \hat{\beta}_j^{OLS} - 0 - \lambda \partial |0| = \beta_j^{OLS} - \lambda [-1, 1] = [\hat{\beta}^{OLS} - \lambda, \hat{\beta}^{OLS} + \lambda]$$

Esto es equivalente a decir

$$\hat{\beta}^{ols} - \lambda \leq 0 \leq \hat{\beta}^{ols} + \lambda$$

$$\Rightarrow \quad -\lambda \leq \hat{\beta}^{ols} \leq \lambda$$

$$\Rightarrow \quad |\hat{\beta}^{ols}| < \lambda$$

$$\Rightarrow \quad (|\hat{\beta}^{ols}| - \lambda)^{+} = 0 .$$

El estimador LASSO $\hat{\beta}_j$ será $\neq 0$ si y solo si $|\hat{\beta}^{ols}| > \lambda$, que ocurre si

$$\hat{\beta}^{ols} > \lambda \quad o \quad \hat{\beta}^{ols} < -\lambda .$$

En esos casos tenemos:

$$\underbrace{\hat{\beta}_j > \lambda}_{\hat{\beta}_j > 0} \quad \Rightarrow \quad \hat{\beta}_j^{ols} - \hat{\beta}_j - \lambda\, \partial|\hat{\beta}_j| = \hat{\beta}_j^{ols} - \hat{\beta}_j - \lambda \times 1 = 0$$

$$\Rightarrow \quad \hat{\beta}_j = \hat{\beta}_j^{ols} - \lambda = +(|\hat{\beta}_j^{ols}| - \lambda)^{+} .$$

$$\underbrace{\hat{\beta}_j < -\lambda}_{\hat{\beta}_j < 0} \quad \Rightarrow \quad \hat{\beta}_j^{ols} - \hat{\beta}_j - \lambda\, \partial|\hat{\beta}_j| = \hat{\beta}_j^{ols} - \hat{\beta}_j - \lambda \times (-1) = 0$$

$$\Rightarrow \quad \hat{\beta}_j = \hat{\beta}_j^{ols} + \lambda = -(|\hat{\beta}_j^{ols}| - \lambda)^{+} .$$

■

Ahora, recuerden que $\hat{\beta}_j^{ols} = \beta_j + O_p(n^{-1/2})$, que para $\beta_j = 0$ implica $\hat{\beta}_j^{ols} = O_p(n^{-1/2})$.
Para que LASSO seleccione correctamente, λ_n debe dominar ese componente de ruido.
Sin embargo λ_n tampoco puede ser muy grande o enviará todo a 0. Esto empezará a ocurrir si $\lambda > \min_{j \in A_0} |\beta_j|$.

Supongamos que $\min_j |\beta_j| \geq \delta > 0$, $\lambda_n = \sqrt{\frac{\log n}{n}}$.

- $Pr(\hat{\beta}_j = 0 \mid \beta_j = 0) = P(|\hat{\beta}_j^{ols}| < \lambda_n \mid \beta_j = 0)$

$$= P\left(O_p(1/\sqrt{n}) < \sqrt{\tfrac{\log n}{n}}\right) \to 1 .$$

- $Pr(\hat{\beta}_j \neq 0 \mid \beta_j \neq 0) = P(|\hat{\beta}_j^{ols}| > \lambda_n \mid \beta_j \neq 0)$

$$= P\left(\left|\beta_j + O_p(n^{-1/2})\right| > \sqrt{\tfrac{\log n}{n}} \;\middle|\; \beta_j \neq 0\right)$$

$$= P\left(\delta + O_p(n^{-1/2}) > \sqrt{\tfrac{\log n}{n}}\right) \to 1 .$$

3.2. Caso General

NO EXISTE una fórmula cerrada de la solución, pero la vamos a caracterizar. En concreto LASSO resuelve

$$-\frac{1}{n}\sum_{i=1}^{n} X_i (Y_i - X_i'\hat{\beta}) + \lambda_n \hat{g} = 0 \quad ,$$

donde $\hat{g} \in \partial \|\hat{\beta}\|_1$: $\hat{g}_j = \begin{cases} \text{sign}(\hat{\beta}_j) & \hat{\beta}_j \neq 0 \\ \in [-1, 1] & \hat{\beta}_j = 0 \end{cases}$

Denotamos $\hat{A_n}$ como los regresores seleccionados por Lasso

$$\hat{A_n} = \{j : \hat{\beta_j} \neq 0\}$$

Podemos escribir la condición de primer orden así

$$\frac{1}{n}\sum_{i=1}^{n} X_{i,\hat{A_n}} (Y_i - X_{i,\hat{A_n}}' \hat{\beta}_{\hat{A_n}}) - \lambda_n \underbrace{\text{sign}(\hat{\beta}_{\hat{A_n}})}_{\text{sign}(\cdot) \text{ de cada uno de sus elementos}} = 0$$

Tenemos que Lasso excluye regresores ($j \in \hat{A_n^c}$ o $\hat{\beta_j} = 0$) cuando

$$0 = -\frac{1}{n}\sum_{i=1}^{n} X_{ij} (Y_i - X_{i,\hat{A_n}}' \hat{\beta}_{\hat{A_n}}) + \lambda_n \cdot \hat{g_j} \quad , \text{ donde } |\hat{g_j}| \leq 1$$

o equivalentemente

$$\left| \frac{1}{n}\sum_{i=1}^{n} X_{ij} (Y_i - X_{i,\hat{A_n}}' \hat{\beta}_{\hat{A_n}}) \right| \leq \lambda_n$$

Denotamos $A_0 = \{j : \beta_j \neq 0\}$ como el set de regresores relevantes, tal que el verdadero modelo es

$$Y_i = X_{i,A_0}' \beta_{A_0} + U_i$$

Finalmente, fíjense que $\text{sign}(\hat{\beta}) = \text{sign}(\beta)$ implica que seleccionamos los regresores correctos

$\beta_j = 0 \Rightarrow \hat{\beta_j} = 0$ (ocurre si $j \in A_0^c$)

$\beta_j \neq 0 \Rightarrow \hat{\beta_j} \neq 0$ (ocurre si $j \in A_0$)

$\Rightarrow \hat{A_n} = A_0$

Vamos a demostrar cuando ocurre esto.

Proposición - Supongamos que $E X_{i,A_0} X_{i,A_0}'$ es p.d. Entonces $\text{sign}(\hat{\beta}) = \text{sign}(\beta)$ si y solo si

$$\text{①} \quad \text{sign}(\beta_{A_0}) = \text{sign}\left(\beta_{A_0} + \left(\sum \frac{X_{i A_0} X_{i A_0}'}{n} \right)^{-1} \left(\frac{1}{n} \hat{\sum} X_{i,A_0} u_i - \lambda_n \text{sign}(\beta_{A_0}) \right) \right)$$

y para todo $j \in A_0^c$

$$\text{②} \quad \left| \frac{1}{n} \hat{\sum_{i=1}} X_{ij} X_{i,A_0}' \left(\frac{\sum X_{i A_0} X_{i A_0}}{n} \right)^{-1} \left(\frac{\sum X_{i A_0} u_i}{n} - \lambda_n \text{sign}(\beta_{A_0}) \right) - \frac{\sum X_{ij} u_i}{n} \right| \leq \lambda_n$$

proof: ($\Rightarrow$) Supongamos $\text{sign}(\hat{\beta}) = \text{sign}(\beta)$. Entonces se cumple

$$0 = \frac{1}{n} \sum X_{i,A_0} (y_i - X_{i,A_0}' \hat{\beta}_{A_0}) - \lambda_n \text{sign}(\beta_{A_0})$$

$$\Rightarrow \quad 0 = \frac{1}{n} \hat{\sum} X_{i,A_0} (u_i - X_{i,A_0}' (\hat{\beta}_{A_0} - \beta_{A_0})) - \lambda_n \text{sign}(\beta_{A_0})$$

$$\Rightarrow \quad \hat{\beta_{A_0}} = \beta_{A_0} + \left(\sum \frac{X_{i A_0} X_{i A_0}'}{n} \right)^{-1} \left(\frac{1}{n} \hat{\sum} X_{i,A_0} u_i - \lambda_n \text{sign}(\beta_{A_0}) \right)$$

lo que significa que

$$\text{sign}(\beta_{A_0}) = \text{sign}(\hat{\beta_{A_0}})$$

$$= \text{sign}\left(\beta_{A_0} + \left(\sum \frac{X_{i A_0} X_{i A_0}'}{n} \right)^{-1} \left(\frac{1}{n} \hat{\sum} X_{i,A_0} u_i - \lambda_n \text{sign}(\beta_{A_0}) \right) \right).$$

(obtenemos ①)

La ecuación ② viene de la cond. de primer orden, ya que $\hat{A_n} = A_0$ excluye a j si:

$$\left| \frac{1}{n} \hat{\sum_{i=1}^{n}} X_{ij} (y_i - X_{i,A_0}' \hat{\beta}_{A_0}) \right| \leq \lambda_n$$

↑ reemplazamos la solución de arriba y nos da ②

($\Leftarrow$) revisarla en las notas, pero sigue técnicas que ya usamos.

Podemos re-escribir la proposición

① $$\text{sign}(\beta_{A_0}) = \text{sign}\left(\beta_{A_0} + \left(\sum \frac{x_{i A_0} x_{i A_0}'}{n}\right)^{-1}\left(\frac{1}{n}\sum x_{i,A_0} u_i - \lambda_n \text{sign}(\beta_{A_0})\right)\right)$$

para todo $j \in A_0^c$

② $$\left|\frac{1}{n}\sum_{i=1} x_{ij}\, x_{i,A_0}'\left(\frac{\sum x_{iA_0} x_{iA_0}}{n}\right)^{-1}\left(\frac{\sum x_{iA_0} u_i}{n} - \lambda_n \text{sign}(\beta_{A_0})\right) - \frac{\sum x_{ij} u_i}{n}\right| \leq \lambda_n$$

Como

$$\left|\beta_j + \left[\left(\sum \frac{x_{iA_0} x_{iA_0}'}{n}\right)^{-1}\left(\frac{1}{n}\sum x_{i,A_0} u_i - \lambda_n \text{sign}(\beta_{A_0})\right)\right]_j\right| > 0 \quad , j \in A_0$$

$$\left|\frac{1}{n}\sum_{i=1} x_{ij}\, x_{i,A_0}'\left(\frac{\sum x_{iA_0} x_{iA_0}}{n}\right)^{-1}\left(\frac{\sum x_{iA_0} u_i}{n} - \lambda_n \text{sign}(\beta_{A_0})\right) - \frac{\sum x_{ij} u_i}{n}\right| \leq \lambda_n \quad , j \in A_0^c$$

Supongamos que $\lambda_n = \sqrt{\frac{\log n}{n}}$.

Entonces, para todo $j \in A_0$

$$\left|\beta_j + \left[\left(\sum \frac{x_{iA_0} x_{iA_0}'}{n}\right)^{-1}\left(\frac{1}{n}\sum x_{i,A_0} u_i - \lambda_n \text{sign}(\beta_{A_0})\right)\right]_j\right|$$

$$= \left|\beta_j + O_p(1)\left(O_p\left(\frac{1}{\sqrt{n}}\right) - O\left(\sqrt{\frac{\log n}{n}}\right)\right)\right|$$

$$= \left|\beta_j + o_p(1)\right| > 0$$ con probabilidad aprox. a 1 siempre que $\min_{j \in A_0} |\beta_j| \geq \delta > 0$.

Ahora, para todo $j \in A_0^c$

$$\left|\frac{1}{n}\sum_{i=1} x_{ij}\, x_{i,A_0}'\left(\frac{\sum x_{iA_0} x_{iA_0}}{n}\right)^{-1}\left(\underbrace{\frac{\sum x_{iA_0} u_i}{n}}_{O_p(1/\sqrt{n})} - \lambda_n \text{sign}(\beta_{A_0})\right) - \underbrace{\frac{\sum x_{ij} u_i}{n}}_{O_p(1/\sqrt{n})}\right| \leq \lambda_n$$

$$\Rightarrow \left|O_p(1/\sqrt{n}) + \lambda_n \sum_{i=1} x_{ij}\, x_{i,A_0}'\left(\frac{\sum x_{iA_0} x_{iA_0}}{n}\right)^{-1} \text{sign}(\beta_{A_0})\right| \leq \lambda_n$$

Fíjense que debe cumplirse que en muestras grandes

$$\left| E\, x_{ij}\, x_{i,A_0}' \left(E x_{i,A_0}\, x_{i,A_0}'\right)^{-1} \operatorname{sign}(\beta_{A_0}) \right| \leq 1$$

para $j \in A_0^c$. Esta condición se llama **irrepresentabilidad**. Lo que dice es que los regresores irrelevantes no deben estar fuertemente correlacionados con los relevantes. De lo contrario, Lasso mantendrá regresores irrelevantes.