UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

Técnicas de Regularización en el marco del Aprendizaje de Máquina: Regresiones Ridge y Lasso

MONOGRAFÍA DE TRABAJO DE GRADO PARA OPTAR POR EL TÍTULO DE MATEMÁTICO
PROYECTO CURRICULAR DE MATEMÁTICAS

Emanuelle Moreno Quintero y Gisell Cristiano Muñoz
Dirigido por: Luis Alejandro Masmela Caita

Bogotá DC
Octubre de 2021

Resumen

La regresión lineal es uno de los métodos de aprendizaje de maquina más utilizados en la actualidad. Sin embargo, en el método estándar de mínimos cuadrados ordinarios se hacen varias suposiciones sobre los datos que a menudo no son ciertas en los conjuntos de datos de la vida real. Esto puede causar numerosos problemas cuando el modelo se ajusta mediante mínimos cuadrados. Uno de los problemas más comunes es que el modelo se ajuste demasiado a los datos, esto sucede cuando el estimador es insesgado, pero tiene alta variabilidad. Las regresiones Ridge y Lasso son dos técnicas de regularización utilizadas para crear un modelo mejor y más preciso. En este trabajo se explica cómo se produce la alta variabilidad en el estimador de mínimos cuadrados. Se incluye un ejemplo con un conjunto de datos de la vida real y se comparan estos métodos con el estimador de mínimos cuadrados para inferir los beneficios e inconvenientes de cada método.

Palabras clave: Aprendizaje de máquina, regresión Ridge, regresión Lasso, regularización
Clasificación AMS: 62Jxx, 62J07.

Agradecimientos: Agradecemos a la Universidad Distrital Francisco José de Caldas y en especial al Proyecto Curricular de Matemáticas por permitirnos explorar un poco del infinito mundo de las matemáticas. También a todos nuestros compañeros y profesores por compartir sus amplios conocimientos con nosotros. Queremos agradecer también a nuestros familiares por el apoyo incondicional.

1. Introducción

Diversas técnicas de aprendizaje de máquina consisten en encontrar los coeficientes que minimizan una función de coste. Las técnicas de regularización se basan en añadir una penalización a dicha función de coste lo cual produce modelos más simples, con mejores ajustes y predicciones. En el caso de la regresión lineal la función de coste que se desea minimizar es la suma de los errores al cuadrado, dicho procedimiento se conoce como mínimos cuadrados, sin embargo esta estimación posee problemas de variabilidad si algunos predictores están correlacionados. La regresión Ridge y Lasso son las dos técnicas de regularización más conocidas que solucionan este problema cuando el número de predictores es menor que el número de observaciones o datos.

Parece difícil datar el origen de las técnicas de regularización, pero actualmente es común identificarlo con los trabajos pioneros de Tikhonov [6]. La motivación de partida proviene del concepto de problema mal planteado, propuesto por Hadamard. Este último, que consiste en la existencia, la unicidad y la dependencia estable de los datos de entrada, en el contexto de la ecuaciones diferenciales parciales. En un entorno discreto de regresión estadística, una idea similar para tratar problemas mal condicionados se desarrolló bajo el término de la regresión Ridge [3] que abrió paso a los estimadores de penalización. Esta metodología utiliza la misma función de costo de regresión más un termino de penalización también conocido como penalización ℓ_2, la cual contrae hacia cero los coeficientes de regresión. Sin embargo, este procedimiento no contrae los coeficientes a exactamente cero, es decir no elimina predictores, lo cual se conoce como una solución no dispersa. Con este fin, Tibshirani [7] propuso el operador de selección y contracción mínima absoluta (Lasso), también conocido como penalización ℓ_1, que dada su formulación tiene la capacidad de producir una solución dispersa y por lo tanto mejorar la interpretabilidad del modelo.

El presente documento pretende ilustrar el desarrollo de estos dos tipos de regresión, Ridge y Lasso, que en su momento dieron solución a los problemas mencionados con anterioridad. Con este fin, el documento se divide en 7 secciones. En la sección 2 se considera el modelo de regresión lineal usual y algunos resultados sobre el estimador de mínimos cuadrados, así como los efectos negativos sobre este en presencia de multicolinealidad. En la sección 3 se presenta la regresión Ridge y sus propiedades. En la sección 4 se introduce la regresión Lasso y una solución particular que existe para este. En la sección 5 se presenta formulaciones equivalentes a la regresiones Ridge y Lasso. En la sección 6 se realiza un ejemplo en donde se ilustra la comparación de la regresión lineal usual con respecto a Ridge y Lasso en un conjunto de datos. En la sección 7 se presentan las conclusiones del trabajo y por el último se encuentra las bibliografías usadas para el desarrollo del mismo.

2. Preliminares

2.1. Modelo de Regresión Lineal

Consideremos un modelo de regresión lineal usual expresado de la siguiente forma:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

En (1) consideramos los siguientes supuestos. La matriz $\boldsymbol{X}$ es no estocástica de tamaño $n \times p$ y de rango columna completo $p < n$, esta matriz contiene las p variables predictoras en cada uno de los n puntos de datos; $\boldsymbol{\beta}$ es el vector de parámetros a estimar de tamaño $p \times 1$; $\boldsymbol{\epsilon}$ es un vector estocástico no observable de tamaño $n \times 1$, con $E(\boldsymbol{\epsilon}) = 0$ y $Cov(\boldsymbol{\epsilon}) = \sigma^2 \boldsymbol{I}_n$ e $\boldsymbol{y}$ es el vector estocástico de salidas de tamaño $n \times 1$.

Uno de los métodos mas conocidos para estimar el vector de parámetros $\boldsymbol{\beta}$ es el de mínimos cuadrados ordinarios. El cual encuentra el vector de parametros $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ que minimiza la función de costo

$$f(\boldsymbol{\beta}_*) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_*\|_2^2 = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_*)^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_*) \tag{2}$$

De ahora en adelante nuestro estimador de mínimos cuadrados será notado como $\hat{\boldsymbol{\beta}}$. Dos propiedades importantes en un estimador es su esperanza y varianza, en el caso de $\hat{\boldsymbol{\beta}}$ tenemos que $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ y $Cov(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

Siempre que nuestro modelo satisfaga los supuestos para (1), podemos estar tranquilos sabiendo que estamos obteniendo las mejores estimaciones posibles. Esto es conocido como el Teorema de Gauss Markov, el cual garantiza que $\hat{\boldsymbol{\beta}}$ es el mejor estimador lineal insesgado o tambien se conoce como un estimador BLUE, por sus siglas en ingles Best Lineal Unbiased Estimator.

Lo que hace que $\hat{\boldsymbol{\beta}}$ sea el mejor estimador lineal insesgado es que este tiene la menor varianza entre todos los demás estimadores lineales insesgados. Como $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$ entonces $\hat{\boldsymbol{\beta}}$ es insesgado, sin embargo, en presencia de multicolinealidad los valores de $\hat{\boldsymbol{\beta}}$ varían mucho de una muestra a otra, lo que hace que se aleje bastante del verdadero valor del vector de parámetros $\boldsymbol{\beta}$. En algunas ocasiones es mejor considerar otros estimadores que aunque sean sesgados ofrecen mejores aproximaciones al verdadero valor de $\boldsymbol{\beta}$, debemos tener en cuenta que sacrificamos la propiedad de ser insesgado por una varianza más baja.

2.2. Error Cuadrático Medio (ECM)

Cuando estimamos el vector de parámetros $\hat{\boldsymbol{\beta}}$ lo hacemos con el fin de predecir nuevos valores de la variable de respuesta o salida y por tanto estaríamos interesados en medir su bondad de ajuste.

En el análisis de regresión una de la formas más conocidas de medir esto se conoce como el error cuadrático medio, el cual lo definimos así

$$ECM = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y)$$

De manera equivalente podemos calcular el ECM relacionando la varianza y el sesgo de un estimador de la siguiente forma

$$ECM[\hat{\boldsymbol{\theta}}] = Var[\hat{\boldsymbol{\theta}}] + [Sesgo[\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}]]^2$$

Notemos que como $\hat{\boldsymbol{\beta}}$ es insesgado se tiene que $ECM[\hat{\boldsymbol{\beta}}] = Var[\hat{\boldsymbol{\beta}}]$.

2.3. Multicolinealidad

La multicolinealidad o colinealidad es la existencia de relaciones casi lineales entre los regresores, predictores o variables de entrada. Se sabe que en presencia de multicolinealidad exacta entre las columnas de la matriz de diseño $\boldsymbol{X}$, la matriz no será de rango columna completo y por tanto $\boldsymbol{X}^T\boldsymbol{X}$ no es invertible, aunque podemos encontrar una pseudo-inversa mediante la inversa generalizada de Moore-Penrose esta no proporciona una única estimación de $\hat{\boldsymbol{\beta}}$.

Lo que hace que la multicolinealidad sea un problema particular es el comportamiento de la inversa de la matriz $\boldsymbol{X}^T\boldsymbol{X}$. Si la multicolinealidad está presente, entonces pequeños cambios relativos en la matriz $\boldsymbol{X}^T\boldsymbol{X}$ producirá grandes cambios relativos en la matriz $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, a esto se le conoce como matriz mal condicionada. Además, algunos elementos de la diagonal principal de $Cov(\hat{\boldsymbol{\beta}}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ serán bastante grandes, lo que significa que algunos elementos del estimador de mínimos cuadrados tendrán una gran varianza.

Ejemplo 1: Consideremos el modelo de regresión lineal $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ con

$$\boldsymbol{y} = \begin{pmatrix} 6.0521 \\ 7.0280 \\ 7.1230 \\ 4.4441 \\ 5.0813 \end{pmatrix}, \quad \boldsymbol{X} = \begin{pmatrix} 1 & 1.9 \\ 1 & 2.1 \\ 1 & 2 \\ 1 & 2 \\ 1 & 1.8 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}$$

Así la matriz $\boldsymbol{X}^T\boldsymbol{X}$ y su inversa están dadas por:

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{pmatrix} 5 & 9.8 \\ 9.8 & 19.26 \end{pmatrix}, \quad (\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{pmatrix} 74.0769 & -37.6923 \\ -37.6923 & 19.2308 \end{pmatrix}$$

Supongamos ahora que la matriz $\boldsymbol{X}$ está dado por:

$$\boldsymbol{X} = \begin{pmatrix} 1 & 1.9 \\ 1 & 2.05 \\ 1 & 2 \\ 1 & 2 \\ 1 & 1.85 \end{pmatrix}$$

Entonces:

$$\boldsymbol{X}^T\boldsymbol{X} = \begin{pmatrix} 5 & 9.8 \\ 9.8 & 19.235 \end{pmatrix}, \quad (\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{pmatrix} 142.4815 & -72.5926 \\ -72.5926 & 37.0370 \end{pmatrix}$$

mostrando que un pequeño cambio relativo en $\boldsymbol{X}^T\boldsymbol{X}$ produce un gran cambio relativo en $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$. Esto se traslada a la estimación por mínimos cuadrados ordinarios ya que la nueva estimación $\hat{\boldsymbol{\beta}} = (-6.6226, 6.4124)^T$ es muy diferente de la anterior $\hat{\boldsymbol{\beta}} = (-4.2489, 5.2013)^T$. Esto da como resultado una alta varianza de nuestro estimador $\hat{\boldsymbol{\beta}}$.

3. Regresión Ridge

El procedimiento de estimación Ridge proporciona una forma de abordar el problema de la multicolinealidad. Aquí, la matriz del modelo se altera para evitar las implicaciones derivadas de su mal condicionamiento.

Hoerl y Kennard [3] propusieron optimizar la siguiente expresión, donde su solución es en efecto el estimador Ridge:

$$f(\boldsymbol{\beta}_*) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2 \tag{3}$$

Notemos que en la parte izquierda de la suma tenemos la expresión (2) que se optimiza para obtener el estimador por mínimos cuadrados $\hat{\boldsymbol{\beta}}$. Ahora a esta misma expresión le sumamos el termino $\lambda\|\boldsymbol{\beta}\|_2^2$ que penaliza la norma al cuadrado del vector de parámetros mediante el hiperparametro $\lambda > 0$ que será el encargado de generar la regularización. Cabe aclarar que si $\lambda = 0$ entonces la penalidad no tiene efecto y volveríamos a (2).

Sin embargo, a medida que $\lambda \to \infty$, el impacto de la penalización aumenta y las estimaciones del coeficiente de regresión Ridge se acercarán a cero. La selección de un buen valor para λ es fundamental, usualmente en el ámbito del aprendizaje de máquina la estimación de λ se hace mediante validación cruzada.

Acudiendo a un método semejante al utilizado en la regresión lineal tradicional el siguiente teorema nos presenta la estimación para $\boldsymbol{\beta}$ en la regresión Ridge.

Teorema 1. *Bajo el modelo de regresión lineal usual, la función*

$$f(\boldsymbol{\beta}_*) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_*\|_2^2 + \lambda\|\boldsymbol{\beta}_*\|_2^2$$

es minimizada por $\boldsymbol{\beta}_* = \hat{\boldsymbol{\beta}}_{\boldsymbol{R}} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{y}$.

Demostración. Primero notemos que

$$\begin{aligned} f(\boldsymbol{\beta}_*) &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_*)^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}_*) + \lambda\|\boldsymbol{\beta}_*\|_2^2 \\ &= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta}_* - \boldsymbol{\beta}_*{}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{\beta}_*{}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}_* + \lambda\boldsymbol{\beta}_*{}^T\boldsymbol{\beta}_* \end{aligned}$$

Derivando con respecto a $\boldsymbol{\beta}_*$ tenemos que:

$$\nabla_{\boldsymbol{\beta}_*} f = \frac{\partial}{\partial \boldsymbol{\beta}_*} f(\boldsymbol{\beta}_*) = -2\boldsymbol{X}^T\boldsymbol{y} + 2\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}_* + 2\lambda\boldsymbol{\beta}_*$$

Igualando a 0:

$$\boldsymbol{\beta}_* = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Notemos que $(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_p)$ es definida positiva, ya que para todo $\boldsymbol{v} \in \mathbb{R}^p$, $\boldsymbol{v} \neq \boldsymbol{0}$

$$\boldsymbol{v}^T(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_p)\boldsymbol{v} = (\boldsymbol{X}\boldsymbol{v})^T(\boldsymbol{X}\boldsymbol{v}) + \lambda\boldsymbol{v}^T\boldsymbol{v} = \|\boldsymbol{X}\boldsymbol{v}\|^2 + \lambda\|\boldsymbol{v}\|^2 > 0$$

Por tanto $(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_p)$ es invertible.

Ahora veamos que es un mínimo:

$$H_{\boldsymbol{\beta}_*} f = \frac{\partial^2}{\partial \boldsymbol{\beta}_* \partial \boldsymbol{\beta}_*{}^T} f(\boldsymbol{\beta}_*) = 2(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_p)$$

Como $H_{\boldsymbol{\beta}_*} f$ es definida positiva entonces $\boldsymbol{\beta}_*$ es el vector que minimiza a $f(\boldsymbol{\beta}_*)$. □

En el teorema anterior cabe advertir que aún si la matriz de diseño $\boldsymbol{X}$ no es de rango columna completo el estimador Ridge es bien definido, ya que $(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_p)$ sigue siendo invertible.

El siguiente teorema nos proporciona la esperanza y matriz de covarianza para el estimador Ridge $\hat{\boldsymbol{\beta}}_{\boldsymbol{R}}$.

Teorema 2. *Bajo el modelo de regresión lineal usual se tiene que*

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}_{\boldsymbol{R}}) &= (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \\ Cov(\hat{\boldsymbol{\beta}}_{\boldsymbol{R}}) &= \sigma^2(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_p)^{-1}(\boldsymbol{X}^T\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_p)^{-1} \end{aligned}$$

Demostración. Para calcular el valor esperado notemos que:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\boldsymbol{R}} &= (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{y} \\ &= (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) \\ &= (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}) + (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}(\boldsymbol{X}^T\boldsymbol{\epsilon})\end{aligned}$$

Por lo tanto:

$$\begin{aligned}E(\hat{\boldsymbol{\beta}}_{\boldsymbol{R}}) &= E[(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}) + (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}(\boldsymbol{X}^T\boldsymbol{\epsilon})] \\ &= (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}) + (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{X}^T E(\boldsymbol{\epsilon}) \\ &= (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}(\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta})\end{aligned}$$

Para calcular la matriz de covarianza notemos que:

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\boldsymbol{R}} &= (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{y} \\ &= (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \\ &= (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}}\end{aligned}$$

Por lo tanto:

$$\begin{aligned}Cov(\hat{\boldsymbol{\beta}}_{\boldsymbol{R}}) &= Cov[(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}}] \\ &= (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}(\boldsymbol{X}^T\boldsymbol{X})Cov(\hat{\boldsymbol{\beta}})[(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}(\boldsymbol{X}^T\boldsymbol{X})]^T \\ &= (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}(\boldsymbol{X}^T\boldsymbol{X})(\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})(\boldsymbol{X}^T\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \\ &= \sigma^2(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}\end{aligned}$$

□

Del teorema anterior podemos concluir que $\hat{\boldsymbol{\beta}}_{\boldsymbol{R}}$ es un estimador lineal sesgado, además hay que aclarar que para el cálculo de la matriz de covarianzas es necesario que la matriz de diseño $\boldsymbol{X}$ sea de rango columna completo, de lo contrario obtendremos una aproximación mediante la inversa generalizada de $\boldsymbol{X}^T\boldsymbol{X}$.

En el siguiente teorema se muestra que el estimador Ridge tiene menor varianza que el estimador de mínimos cuadrados, lo cual lo hace muy atractivo en caso de multicolinealidad.

Teorema 3. *Bajo el modelo de regresión lineal usual el estimador Ridge $\hat{\boldsymbol{\beta}}_{\boldsymbol{R}}$ es de menor varianza que el estimador de mínimos cuadrados $\hat{\boldsymbol{\beta}}$.*

Demostración. Definamos la matriz:

$$\boldsymbol{W} = \boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}$$

la cual es invertible. Entonces, podemos reescribir la matriz de covarianza del estimador Ridge de la siguiente manera

$$\begin{aligned} Cov(\hat{\boldsymbol{\beta}}_{\boldsymbol{R}}) &= \sigma^2(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \\ &= \sigma^2(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \\ &= \sigma^2\boldsymbol{W}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{W} \end{aligned}$$

Por lo tanto la diferencia entre sus matrices de covarianzas es:

$$Cov(\hat{\boldsymbol{\beta}}) - Cov(\hat{\boldsymbol{\beta}}_{\boldsymbol{R}}) = \quad = \quad \sigma^2(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_P)^{-1}[2\lambda \boldsymbol{I}_P + \lambda^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}](\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_P)^{-1}$$

Tomemos $\boldsymbol{z}$ como:

$$\boldsymbol{z} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_P)^{-1}\boldsymbol{u}$$

Para cualquier vector $\boldsymbol{u} \in \mathbb{R}^p$, $\boldsymbol{u} \neq 0$. Luego, tenemos que:

$$\begin{aligned} \boldsymbol{u}^T[Cov(\hat{\boldsymbol{\beta}}) - Cov(\hat{\boldsymbol{\beta}}_{\boldsymbol{R}})]\boldsymbol{u} &= \sigma^2\boldsymbol{z}^T[2\lambda \boldsymbol{I}_p + \lambda^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}]\boldsymbol{z} \\ &= 2\sigma^2\lambda\boldsymbol{z}^T\boldsymbol{z} + \sigma^2\lambda^2\boldsymbol{z}^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{z} > 0 \end{aligned}$$

Por lo tanto el estimador Ridge $\hat{\boldsymbol{\beta}}_{\boldsymbol{R}}$ tiene menor varianza que el estimador de mínimos cuadrados $\hat{\boldsymbol{\beta}}$. □

¿Por qué el estimador $\hat{\boldsymbol{\beta}}_{\boldsymbol{R}}$ es mejor que el estimador $\hat{\boldsymbol{\beta}}$?
La ventaja de la regresión Ridge sobre los mínimos cuadrados tiene su origen en la compensación de sesgo-varianza. A medida que λ aumenta, la varianza de la regresión Ridge disminuye, pero el sesgo aumenta. Esto se ilustra en la Figura 1, utilizando un conjunto de datos simulado que contiene $p = 30$ predictores y $n = 50$ observaciones. En las estimaciones del coeficiente de mínimos cuadrados, que corresponden a la regresión Ridge con $\lambda = 0$, la varianza es alta pero no hay sesgo. Pero a medida que λ aumenta, la penalización de las estimaciones de los coeficientes Ridge conduce a una reducción sustancial de la varianza de las predicciones, a expensas de un ligero aumento del sesgo. Recordemos que el ECM, es una función de la varianza más el sesgo al cuadrado. Para valores de λ hasta aproximadamente 10, la varianza disminuye rápidamente, con muy poco aumento del sesgo. En consecuencia, el ECM disminuye considerablemente a medida que λ aumenta de 0 a 10. Más allá de este punto, la disminución de la varianza debida al aumento de λ se ralentiza, y la contracción de los coeficientes hace que se subestimen significativamente, lo que da lugar a un gran aumento del sesgo. El ECM mínimo se alcanza aproximadamente en $\lambda = 4.80$. Curiosamente, debido a su alta varianza,

el ECM asociado al ajuste por mínimos cuadrados, cuando $\lambda = 0$, es casi tan alto como el del modelo nulo para el que todas las estimaciones de los coeficientes son cero, cuando $\lambda = \infty$. Sin embargo, para un valor intermedio de λ, el ECM es considerablemente menor.

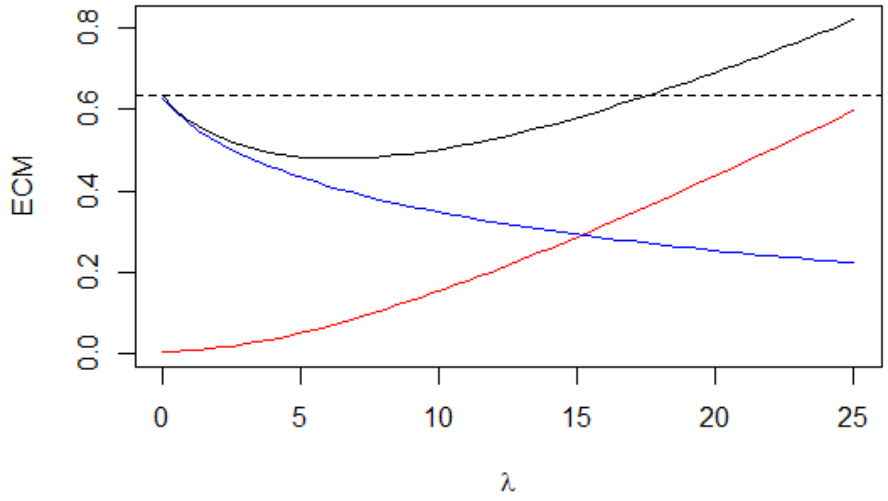

Figura 1: ECM de color negro, varianza de color rojo y sesgo de color azul, para las predicciones de la regresión Ridge en un conjunto de datos simulados, en función de λ. La linea negra punteada es el menor ECM posible.

4. Regresión Lasso

La regresión Ridge tiene una desventaja al no producir una solución dispersa, pues la penalización ℓ_2 hace tender los coeficiente de la regresión a cero, pero en general no los hará exactamente cero. Puede que este no sea un problema para la precisión de la predicción, pero puede crear un desafio en la interpretación del modelo en un entorno en el que número de variables p es bastante grande.

Tibshirani [7] propuso un estimador alternativo que supera esta desventaja. Este estimador es conocido como Lasso (least absolute shrinkage and selection operator), el cual a diferencia del estimador Ridge añade una penalización ℓ_1 a la suma de los errores al cuadrado. Tibshirani propuso optimizar la siguiente expresión:

$$f(\boldsymbol{\beta}_*) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \tag{4}$$

Al igual que en el estimador Ridge, el hiperparametro $\lambda > 0$ se encargará de generar la regularización,

adicionalmente si $\lambda = 0$ no habrá penalización y volveríamos al estimador de mínimos cuadrados. También de manera general el hiperparámetro λ en el caso de Lasso se calcula mediante validación cruzada.

Podemos ver que la función (4) es convexa al ser una suma de funciones convexas, como ocurre con la función (3), sin embargo $\|\boldsymbol{\beta}\|_1$ no es estrictamente convexa y por tanto no es posible garantizar que el estimador Lasso sea único. Una consecuencia de esto es que en general no existe una solución analítica para Lasso. Sin embargo existen diferentes formas de hallar las soluciones a Lasso, que se pueden encontrar en [7], [1] y [2] mediante diversos algoritmos.

El siguiente teorema muestra un caso particular bajo el cual es posible obtener una solución analítica para Lasso.

Teorema 4. *Bajo el modelo de Regresión Lineal usual y la ortogonalidad de la matriz de diseño* $\boldsymbol{X}$ *el* $i-$*ésimo estimador Lasso* $\hat{\beta}_{L_i}$ *viene dado por:*

$$\hat{\beta}_{L_i} = sgn(\hat{\beta}_i)(|\hat{\beta}_i| - \gamma)^+$$

Demostración.

$$\begin{aligned} f(\boldsymbol{\beta_*}) &= (\boldsymbol{y} - \boldsymbol{X\beta_*})^T(\boldsymbol{y} - \boldsymbol{X\beta_*}) + \lambda\|\boldsymbol{\beta_*}\|_1 \\ &= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X\beta_*} - \boldsymbol{\beta_*}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{\beta_*}^T\boldsymbol{X}^T\boldsymbol{X\beta_*} + \lambda\|\boldsymbol{\beta_*}\|_1 \\ &= \boldsymbol{y}^T\boldsymbol{y} - \hat{\boldsymbol{\beta}}^T\boldsymbol{\beta_*} - \boldsymbol{\beta_*}^T\hat{\boldsymbol{\beta}} + \boldsymbol{\beta_*}^T\boldsymbol{\beta_*} + \lambda\|\boldsymbol{\beta_*}\|_1 \\ &= \boldsymbol{y}^T\boldsymbol{y} - 2\hat{\boldsymbol{\beta}}^T\boldsymbol{\beta_*} + \boldsymbol{\beta_*}^T\boldsymbol{\beta_*} + \lambda\|\boldsymbol{\beta_*}\|_1 \end{aligned}$$

Consideremos los siguientes casos para optimizar la función objetivo, minimizando por separado su i-ésimo elemento.

Si $\hat{\beta}_i > 0$ entonces $\beta_{*i} \geq 0$. Lo anterior lo podemos ver de manera gráfica como en la figura 2. Si por ejemplo el estimador de mínimos cuadrados se encuentra en el primer cuadrante entonces la solución estará en el primer o cuarto cuadrante, ya que los contornos son verticales, esto ultimo derivado de que la matriz de diseño sea ortogonal, entonces la correlación entre los coeficientes de los predictores es cero. Derivando con respecto a β_{*i} tenemos que

$$\frac{\partial}{\partial\beta_{*_i}}(y_i^2 - 2\hat{\beta}_i\beta_{*_i} + \beta_{*_i}^2 + \lambda\beta_{*_i}) = -2\hat{\beta}_i + 2\beta_{*_i} + \lambda$$

Igualando a 0 tenemos que:

$$\begin{aligned} -2\hat{\beta}_i + 2\beta_{*_i} + \lambda &= 0 \\ \beta_{*_i} &= \hat{\beta}_i - \frac{\lambda}{2} \\ \beta_{*_i} &= \hat{\beta}_i - \gamma, \quad \gamma = \frac{\lambda}{2} \\ \beta_{*_i} &= \hat{\beta}_i - \gamma \\ \beta_{*_i} &= sgn(\hat{\beta}_i)(|\hat{\beta}_i| - \gamma)^+ \end{aligned}$$

Si $\hat{\beta}_i < 0$ entonces $\beta_{*i} \leq 0$, de manera análoga tenemos que

$$\beta_{*_i} = (\hat{\beta}_i + \gamma)^- = -(-\hat{\beta} - \gamma)^+ = sgn(\hat{\beta})(|\hat{\beta}| - \gamma)^+$$

□

En el teorema anterior utilizamos la hipótesis que nuestra matriz de diseño $\boldsymbol{X}$ es ortogonal, esto lo necesitamos para asegurar que las estimaciones de β_i son independientes. Lo bueno de la ortogonalidad es que asegura que $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ es diagonal, por lo tanto $Cov(\hat{\beta}_i, \hat{\beta}_j) = 0$ para $i \neq j$, de manera que la estimación de β_i no cambiara dependiendo de si β_j está o no en el modelo.

Formulas explicitas para la esperanza y matriz de covarianza del estimador Lasso no existen, sin embargo algunas aproximaciones las podemos encontrar en [7] y [5]. No obstante, se sabe que a medida que $\lambda \to \infty$ el sesgo crece y la varianza disminuye.

5. Otra Formulación para la Regresión Ridge y Lasso

Minimizar las funciones (3) y (4) resulta equivalente a los dos problemas

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|, \quad \text{sujeto a} \quad \|\boldsymbol{\beta}\|_2^2 \leq s, \tag{5}$$

y

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|, \quad \text{sujeto a} \quad \|\boldsymbol{\beta}\|_1 \leq s \tag{6}$$

respectivamente. En otras palabras, para cada valor de λ, existe algún s tal que (3) y (5) permitirá obtener las mismas estimaciones de Ridge. De manera similar, para cada valor de λ existe algún s tal

que (4) y (6) permitirá obtener la misma estimación Lasso. Cuando $p = 2$, entonces (5) indica que el estimador Ridge posee la menor suma de los residuos al cuadrado (RSS) de todos los puntos dentro del circulo definido por $\beta_1^2 + \beta_2^2 \leq s$. De igual forma, el estimador Lasso tiene el menor RSS de todos los puntos dentro del diamante definido por $|\beta_1| + |\beta_2| \leq s$, ver figura 2.

Podemos pensar en (5) de la siguiente forma. Cuando aplicamos regresión Ridge intentamos encontrar el estimador que posea el menor RSS, sujeto a la restricción de que existe un límite s para lo grande que puede ser $\|\boldsymbol{\beta}\|_2^2$. Cuando s es muy grande entonces este límite no es muy restrictivo, por lo que el estimador Ridge puede ser grande. De hecho, si s es lo suficientemente grande como para que la solución de mínimos cuadrados este dentro del límite, entonces (5) simplemente obtendrá la solución de mínimos cuadrados. Por el contrario, si s es pequeño, entonces $\|\boldsymbol{\beta}\|_2^2$ debe ser pequeña para no sobrepasar el límite. De manera similar (6) indica que cuando aplicamos regresión Lasso, buscamos el estimador tal que el RSS sea lo más pequeño posible, sujeto a $\|\boldsymbol{\beta}\|_1$ no exceda el limite s.

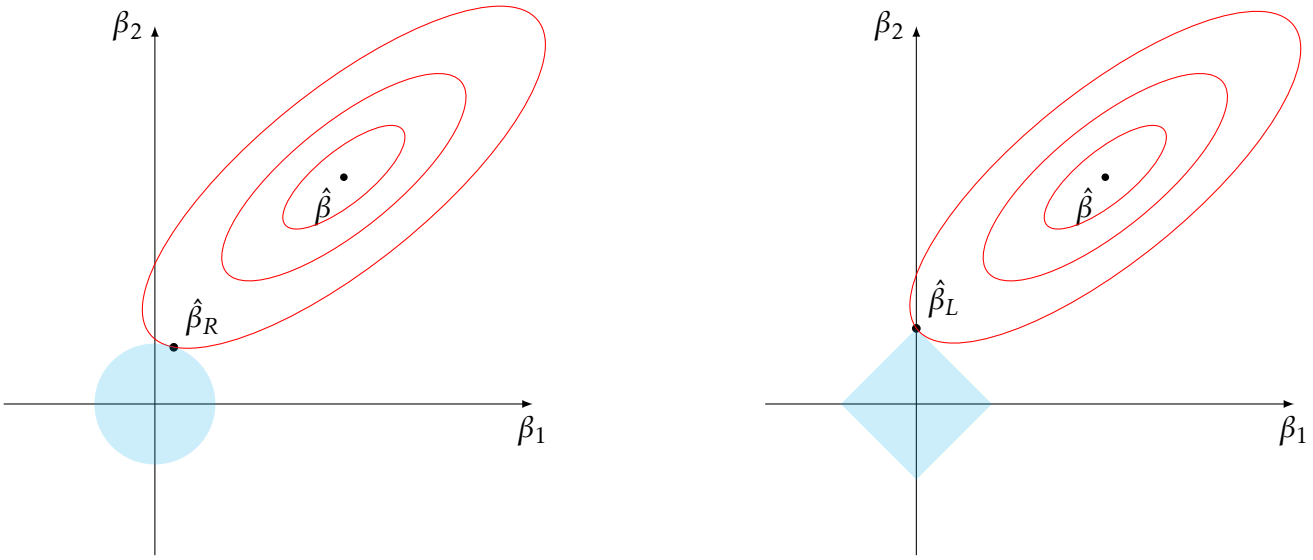

Figura 2: Las áreas de color azul son las regiones de restricción, $\beta_1^2 + \beta_2^2 \leq s$ y $|\beta_1| + |\beta_2| \leq s$, mientras que las elipses rojas son los contornos del RSS.

6. Ejemplo

En este ejemplo se analiza las bondades de la Regresión Ridge y Lasso en la estimación del vector de parámetros de la regresión lineal, con respecto a lo obtenido mediante el método de los mínimos cuadrados ordinarios en presencia de colinealidad. Todos los cálculos se realizaron con el software

estadístico R utilizando el paquete `glmnet`.Tomamos el conjunto de datos llamado **Hitters** que hace parte de la libreria `ISLR` y es usado en [4]. El conjunto de datos describe las estadísticas de 322 jugadores participante en las grandes ligas de béisbol de la temporada 1986. El objetivo es predecir el salario de los jugadores en miles de dolares teniendo en cuenta sus estadísticas respecto al número de veces que el jugador ha bateado en la temporada (AtBat) y durante su carrera (CAtBat); número de Hits o llegadas a primera base en la temporada (Hits) y durante su carrera (CHits);número de Home Run's hecho en la temporada (HmRun) y durante su carrera (HmRun);número de carreras o puntos hechos en la temporada (Runs) y durante su carrera (Runs).

Como la estimación del parámetro λ puede verse afectado por la presencia de valores atípicos en los datos, estos se redujeron a aquellos que estuvieran entre el primer y tercer cuartil, obteniendo un conjunto de 187 observaciones.

Primero partimos nuestro conjunto de datos de manera aleatoria en un conjunto de entrenamiento con 130 observaciones (70%) y un conjunto de prueba con 57 observaciones (30%). En la siguiente gráfica se puede apreciar el comportamiento de los coeficientes de las regresiones Ridge y Lasso para un amplio rango de valores de λ.

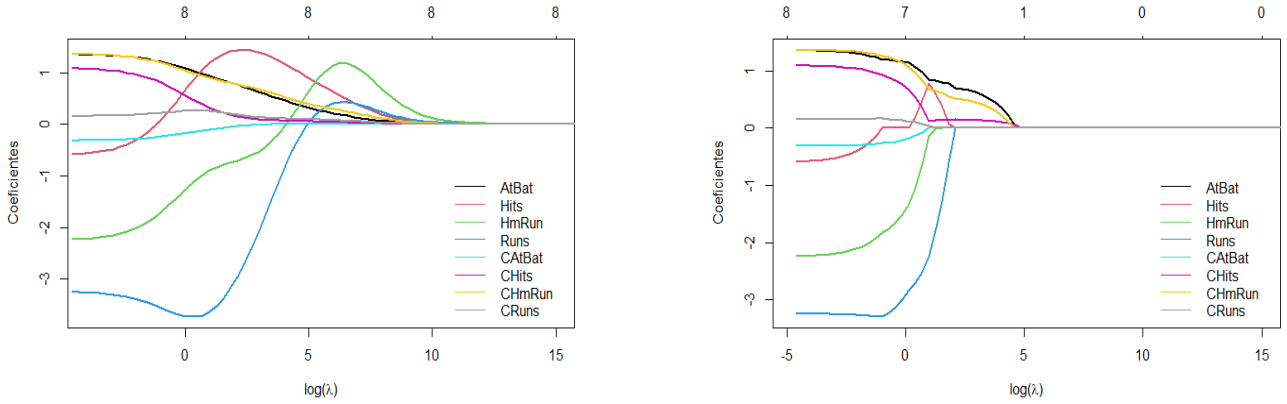

Figura 3: Al lado izquierdo se muestran las estimaciones de los coeficientes de la regresión Ridge para diferentes valores de $\log(\lambda)$, de manera análoga al derecho para la regresión Lasso. Los números que se encuentran en la parte superior de cada gráfica corresponde al número de predictores presentes en el modelo para cada valor de $\log(\lambda)$.

En en lado izquierdo de la Figura 3, cada curva corresponde a la estimación del coeficiente de regresión Ridge para las ocho variables en función de $\log(\lambda)$. Por ejemplo, la curva de color verde representa la estimación para el coeficiente de HmRun, al variar λ. En el extremo izquierdo de la gráfica, λ es esencialmente cero, por lo que las estimaciones del coeficiente Ridge correspondientes

son las mismas que las estimaciones de mínimos cuadrados. Pero a medida que λ aumenta, las estimaciones de los coeficientes se aproximan a cero. Cuando λ es muy grande las estimaciones de los coeficientes son prácticamente cero. Mientras las estimaciones de los coeficientes Ridge tienden a disminuir en conjunto a medida que λ aumenta. Los coeficientes individuales, como Hits, HmRun y Runs, pueden aumentar ocasionalmente a medida que aumenta λ. De igual forma en el lado derecho de la figura se observan las curvas que corresponden a la estimaciones de los coeficientes de regresión Lasso para las ocho variables en función de λ. A diferencia de Ridge cuando λ aumenta, Lasso puede hacer exactamente cero algunos predictores. En cambio la regresión Ridge casi siempre incluirá todos los coeficientes en el modelo.

El mejor valor λ del modelo se podría determinar a partir de la Figura 4. En este gráfico vemos el valor del error cuadrático medio en el eje vertical, mientras que en el eje horizontal los valores de $\log(\lambda)$. La línea del lado izquierdo punteada muestra el valor $\log(\lambda)$ que produce el menor error cuadrático medio. El error se determina utilizando validación cruzada, que es el método más común para los modelos de regresión Ridge y Lasso. `Glmnet` realiza una validación cruzada de diez iteraciones en el conjunto de entrenamiento. Esto significa que los datos se dividen en diez particiones, luego se elige una partición como conjunto de validación. A continuación, el método se ejecuta en las nueve particiones restantes y se compara con el conjunto de validación para determina el error cuadrático medio. Este proceso se repite para que cada partición se utilice como conjunto de validación y luego se promedian los errores cuadráticos medios, que se representan en este gráfico. Este proceso se utiliza tanto para la regresión Ridge como para Lasso.

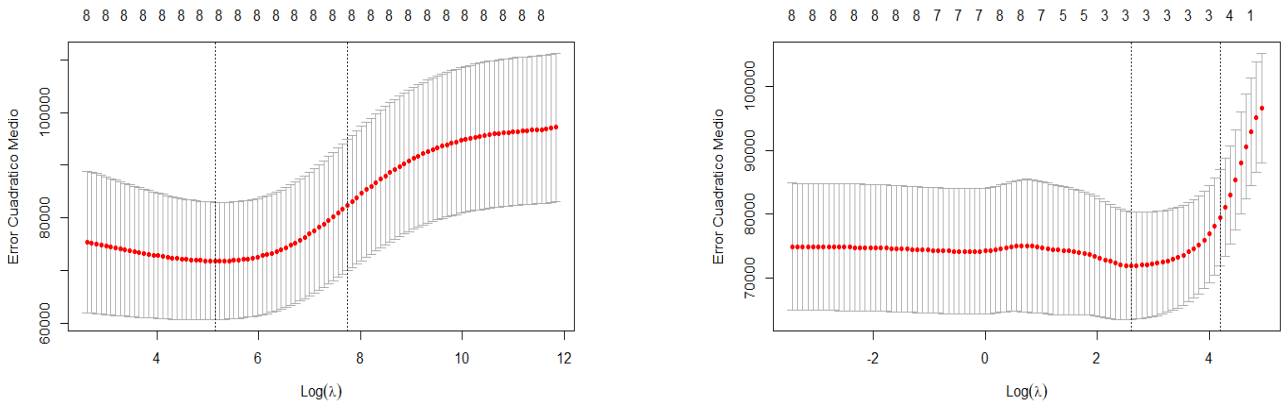

Figura 4: Los puntos rojos representan el ECM para cada valor de $\log(\lambda)$, la barra superior e inferior de cada punto denotan el ECM más o menos su desviación estándar. La gráfica del lado izquierdo repesenta estos valores para la regresión Ridge, de manera análoga se encuentra para la regresión Lasso en el lado derecho.

Después de realizar validación cruzada usando la función `cv.glmnet` y considerando el λ con menor ECM en el conjunto de entrenamiento obtuvimos $\lambda = 155.579$ y $\lambda = 16.298$ para la regresión Ridge y Lasso respectivamente. Estos valores se representan en la linea punteada del lado de izquierdo de cada una de las gráficas de la Figura 4. Así, estimando el vector de parámetros para cada regresión obtenemos la Tabla 1. Note que aunque ningún coeficiente de la regresión Lasso es cero, esta posee menor error cuadrático medio que la regresión por mínimos cuadrados, en este caso para el λ óptimo encontrado todos los coeficientes son significativos para el modelo. Por ejemplo si tomamos $\lambda = 54.626$ obtenemos $\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}{}^{T} = (185.202, 1.116, 0, -0.246, -3.151, -0.295, 1.020, 1.176, 0.173)$, con $ECM = 69629.69$, el cual es mayor que el ECM de la estimación por mínimos cuadrados, apesar de haber eliminado la variable Hit.

Coeficiente	$\hat{\boldsymbol{\beta}}$	$\hat{\boldsymbol{\beta}}_{\boldsymbol{R}}$	$\hat{\boldsymbol{\beta}}_{\boldsymbol{L}}$
Intercepto	184.915	183.305	184.155
AtBat	1.368	1.355	1.257
Hit	−0.657	−0.617	−0.333
HmRun	−2.191	−2.229	−1.516
Runs	−3.227	−3.220	−3.247
CAtBat	−0.324	−0.318	−0.309
CHit	1.111	1.093	1.061
CHmRun	1.377	1.363	1.297
CRuns	0.154	0.153	0.164
ECM	65 449.03	63 737.45	64 842.62

Tabla 1: Estimaciones de las regresiones con sus errores cuadráticos medios

En la Tabla 2 encontramos algunas predicciones a partir de los modelos de regresión considerados y para el conjunto de datos de prueba. Se logra evidenciar que la regresión Ridge y Lasso tienen predicciones que se logran aproximar más a los verdaderos valores de los salarios reales.

Salario Real	Mínimos Cuadrados	Ridge	Lasso
431.500	444.864	430.557	427.305
625.000	566.507	618.158	610.603
450.000	407.335	431.639	431.788
612.500	628.590	613.076	613.151
850.000	751.166	772.707	779.270

Tabla 2: Predicciones

Aunque en este ejemplo la regresión Lasso no hizo cero ningún coeficiente con el λ que minimiza el ECM, es posible obtener otros conjuntos de datos con un número mayor de predictores en donde se

logre llevar a cero varios de los coeficientes sin aumentar su error cuadrático medio y así encontrar un modelo que atenúa el efecto de la correlación entre predictores y que reduzca la influencia de los predictores menos relevantes, como se ilustra en [7].

7. Conclusiones

Hemos mostrado que la regresión Ridge permite regularizar las estimaciones de los coeficientes realizadas por mínimos cuadrados en presencia de multicolinealidad, incluso en ausencia de esta $\hat{\boldsymbol{\beta}}_{\boldsymbol{R}}$ es de menor varianza que $\hat{\boldsymbol{\beta}}$. Adicionalmente, el estimador Ridge puede brindar un ECM menor que el estimador de mínimos cuadrados, cuando se encuentra el λ adecuado. Sin embargo la regresión Ridge no contrae a exactamente cero los coeficientes de regresión, lo que genera un modelo con igual cantidad de predictores que con la regresión por mínimos cuadrados.

Para el caso de la regresión Lasso pudimos ver que de manera general no existe una solución cerrada, sin embargo bajo la hipótesis de ortogonalidad es posible obtenerla. En la vida real es muy difícil obtener matrices de diseño que sean ortogonales, esto impulso el desarrollo de algoritmos para poder estimar su solución. La regresión Lasso a diferencia de Ridge si puede contraer coeficientes a exactamente cero, lo cual genera un modelo con una cantidad menor de predictores y más interpretable.

La escogencia del hiperparámetro λ es fundamental en ambas regresiones, ya que como nos pudimos dar cuenta en el caso de Lasso podríamos obtener un menor ECM pero no contraer algún coeficiente a exactamente cero. Aunque en este documento se estimo el valor de λ mediante validación cruzada, el mejor método para estimarlo aún es tema de investigación.

Existe otra técnica de regularización llamada ElacticNet que combina la regresión Ridge y Lasso en una misma función de coste, la cual tiene la capacidad de eliminar algunos y no todos los predictores.

Aunque las regresiones Ridge y Lasso solucionan el problema de multicolinealidad en un modelo de regresión lineal, este no es el único incoveniente que el modelo podría presentar. Por ejemplo ¿ que pasa con las estimaciones de Ridge y Lasso cuando $p > n$?

Referencias

[1] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani. Least angle regression. *The Annal of statistics*, 32(2):407–451, 2004.

[2] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Regularization paths for generalized linear modelos via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

[3] Arthur Hoerl and Robert Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 42(1):80–87, 1970.

[4] Gareth James, Daniela Witten, Trevor Hatie, and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Springer Texts in Statistics, 2021.

[5] Michael Osborne, Brett Presnell, and Berwin Turlach. On the lasso and its dual. *Journal of Computational and Graphica Statistics*, 9(2):319–337, 2000.

[6] Andrey Thikhonov. *Solution of Incorrectly Formulated Problems and the Regularization Method*. Soviet Mathematics Doklady, 1963.

[7] Robert Tibshirani. Regresion shrinkage and selection via the lasso. *Royal Statictical Society*, 58(1):267–288, 1995.