Técnicas de Selección y Regularización Ridge y Lasso

Autor:
Grevy Stiben Rápalo García

Asesorado por:
Msc. Roberto Duarte

Resumen

La regresión lineal es el método de aprendizaje estadístico más utilizado en la actualidad, sin embargo, el método estandar de mínimos cuadrados ordinarios no es del todo preciso al aplicarlo a un conjunto de datos de la vida real, ya que hace varias suposiciones de los mismos que, a menudo no son ciertas, los problemas comunes son un modelo demasiado ajustado a los datos y esto sucede por estimadores con alta variabilidad e insesgados, en el presente trabajo de investigación se hace un estudio de técnicas de regularización tales como Ridge y Lasso las cuales son técnicas utilizadas para brindarnos un mejor modelo y a su vez más preciso.
Para lograr el objetivo, se presentan los contenidos teóricos necesarios para la comprensión de dichas técnicas y un posterior desarrollo en un caso práctico con el fin de mostrar su utilidad.

Palabras clave: Regresión, Regularización, Ridge, Lasso

Abstract

Linear regression is the most widely used statistical learning method today, however, the standard método of ordinary square mínimos is not entirely accurate when applied to a real-life dataset, as it makes several assumptions of them that are often not true, common problems are a model too tight to the data and this happens by estimators with high variability and unbiased, in the present research work a study of regularization techniques such as Ridge and Lasso is made, which are techniques used to provide us with a better and at the same time more accurate model. To achieve the objective, the technical contents necessary for the understanding of these techniques and a subsequent development in a practical case are presented in order to show their usefulness.

Keywords: Regression, Regularization, Ridge, Lasso

Índice

Índice de tablas

Índice de figuras

Capítulo 1

Introducción

Los análisis espectométricos son métodos instrumentales empleados en química analítica basados en la interacción de la radiación electromagnética, u otras partículas, con el fin de identificar su concentración química, estos métodos pueden utilizarse en la arqueología, en el estudio de elementos de interés cultural y es importante muchas veces determinar la cantidad de x compuesto químico. Estas concentraciones se pueden determinar o medir también a partir de propios análisis químicos, pero resulta un aumento considerable en el costo del mismo. Es por eso que, se opta por predecir un compuesto químico a través de la energía correspondiente a la frecuencia (análisis espectométrico), pero considerando estas frecuencias y tomando en cuenta el tamaño de la muestra para un estudio, soprepasa al mismo; es decir, nos encontramos con un número menor de muestra en comparación con predictores ($n > p$).

En nuestro caso, para determinar la composición de un conjunto de vasijas de vidrio de un yacimiento arqueológico podríamos optar por el procedimiento habitual al ajustar un modelo de regresión lineal empleando mínimos cuadrados, pero tomando en cuenta que tenemos un tamaño mucho mayor de predictores que de muestra, Sin embargo, surgen dificultades, ya que al tener muchas variables hace que el modelo sea difícil de interpretar, de igual forma, las estimaciones del modelo tendrán mucha varianza y estará sobreajustado (overfitting). La solución a esto es forzar al modelo a ser menos complejo, para así reducir su varianza, una forma de conseguirlo es mediante la regularización (regularization o shrinkage) de la estimación de los parámetros $\beta_1\beta_2, ..., \beta_p$, que consiste en considerar todas las variables predictoras, pero forzando a que algunos de los parámetros se estimen mediante valores muy próximos a cero, o directamente con ceros, esto se logrará recurriendo a los métodos de regresión regularizada como son Ridge y Lasso, estas técnicas van a provocar un pequeño aumento en el sesgo, pero a cambio una notable reducción en la varianza y una interpretación más sencilla del modelo resultante.

1.1. Objetivos

1.1.1. Objetivo General

Aplicar y comparar métodos de selección y regularización Ridge y Lasso en la determinación y selección de parámetros para predecir la composición química de un elemento en un conjunto de vasijas de un yacimiento arqueológico.

1.1.2. Objetivos Específicos

De acuerdo con el trabajo planteado anteriormente, se presentan los siguientes objetivos específicos aplicados a los datos:

- Describir la teoría detrás de los métodos de selección y regularización y proveer una explicación de estas.

- Identificar ventajas y desventajas de ambos métodos.

- Analizar un conjunto de datos reales y comparar los resultados obtenidos mediante los distintos métodos.

1.2. Estructura del documento

A continuación se presentarán los principios teóricos necesarios para la comprensión de la metodología a utilizar. En primer lugar, las generalidades de un modelo, posteriormente, se introducen los modelos de regresión lineal simple, regresión lineal múltiple, así la evaluación de la exactitud de los mismos.

Por último, se presentan los métodos de selección y regulación, introduciendo la definición matemática de las normas l_p debido a su implicación directa en los métodos de regulariación de Ridge y Lasso, y culminando con la selección del párametro de penalización a través de validación cruzada.

Capítulo 2

Marco Teórico

2.1. Estimación de f

De manera más general, supongamos que observamos una respuesta cuantitativa Y y p diferentes predictores $X_1, X_2, ..., X_p$. De igual forma suponemos que hay alguna relación entre Y y $X = (X_1, X_2, ..., X_p)$ que se puede escribir en la forma muy general

$$Y = f(X) + \epsilon \tag{2.1}$$

Aquí f es alguna función fija, pero desconocida de $X_1, X_2, ..., X_p$ y ϵ es un término de error aleatorio que es independiente de X y tiene media cero. En esta formulación , f representa la información sistemática que X proporciona sobre Y.

Ahora el problema principal es la estimación de dicha f y esto se debe a dos razones muy importantes las cuales son predicción e inferencia, que se definirá a continuación.

2.1.0.1. Predicción

En cuanto a la predicción tenemos en muchas situaciones que un conjunto de entradas X esta fácilmente disponible, pero no se puede obtener la salida Y de forma sencilla. En este contexto, el error en 2.1 se promedia a cero y se puede predecir Y usando

$$\hat{Y} = \hat{f}(X) \tag{2.2}$$

donde $\hat{f}$ representa nuestra estimación para f, e $\hat{Y}$ representa la predicción resultante para Y .

La prediccioń a su vez depende de dos cantidades, las cuales son el error reducible y el error irreducible, esto sucede porque de forma general $\hat{f}$ no será una estimación perfecta para f y esto introducirá un error.
Sin embargo, existen casos en que se puede mejorar potencialmebte la precisión de $\hat{f}$ mediante alguna técnica de aprendizaje estadístico más adecuada para la estimación de f, a este se le conoce como error reducible. El error irreducible es aquel que se introduce por el término ϵ, ya que por mejor que se pueda estimar f se sabe que Y es una función de ϵ y este término tiene una variabilidad asociada que no se puede reducir, y por lo tanto afectará la precisión de la predicción.

2.1.0.2. Inferencia

En cuanto a la inferencia, es la comprensión de la asociación y relación entre Y y $X_1, X_2, ..., X_p$, se desea estimar f con el objetivo de descubrir que predictores están asociados con la respuesta y, que tanto se relaciona cada uno

de estos con la respuesta y por último tratar de resumir adecuadamente todo lo anterior usando una relación adecuada.

2.1.1. Modelos lineales paramétricos

Los modelos utlizados en este trabajo son los llamados medelos lineales paramétricos, los cuales son un enfoque de modelos basados en dos pasos:

i) Haciendo una suposición sobre la forma funcional de f de forma muy simple, que f es lineal en X , es decir

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \tag{2.3}$$

ii) Una vez seleccionado el modelo, se necesita de procedimentos que utilicen datos de entrenamiento para ajustar o entrenar el modelo, es decir, estimar los párametros $\beta_0, \beta_1, ..., \beta_p$.

Con esta descripción en dos pasos se reduce el problema de estimar f, a un problema de estimar un conjunto de parámetros.

2.1.2. Precisión e interpretabilidad del modelo

En este apartado hablaremos sobre la flexibilidad de un modelo, que estará relacionado con el enfoque que elijamos, si estamos interesados en inferencia entonces se necesita un modelo con mayor interpretabilidad y así un modelo más restrictivo, es decir, un modelo inflexible, en cambio sí el objetivo es la predicción y la interpretabilidad no es una preocupación, en este contexto, podríamos esperar que será mejor utilizar el modelo más flexible disponible. (aunque puede suceder que se obtenga predicciones más precisas utilizando un método menos flexible, debido al sobrejuste de datos (overftting) como se verá en el siguiente apartado.)

2.1.3. Evaluación de la precisión del modelo

Es de esperar que para distintos conjuntos de datos no existá un método que domine sobre los demás, ya que, para un conjunto de datos en particular, un método específico puede funcionar mejor, pero este mismo método ó uno similar no puede funcionar en otro conjunto de datos distinto. Por lo tanto, es una tarea importante decidir para cualquier conjunto dado de datos, qué método produce mejores resultados.
A continuación veremos algunos conceptos importantes que ayudaran a selecionar dicho método para un conjunto de datos específico.

2.1.3.1. Medición de calidad de ajuste

Ya que tenemos como fin evaluar el rendimiento de un método de aprendizaje estadístico en un conjunto de datos en esepecífico, se necesita medir de alguna forma qué tan bien son sus predicciones, es decir, si coinciden con los datos observados.

En el ajuste de regresión, la medida más utilizada es el error cuadrático medio (MSE), que está dado por:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \tag{2.4}$$

El MSE es calculado utilizando los datos de entrenamiento que se utilizaron para que se ajuste el modelo, un MSE grande nos dice que para algunas observaciones, lo predicho por el modelo con el valor verdadero difieren sustancialmente, en cambio un MSE pequeño dirá que las respuestas predichas están muy cerca de las respuestas verdaderas.

Es importante mencionar que lo que interesa realmente, es qué información proporciona el MSE en la precisión de las predicciones que se obtiene al aplicar el método a lo nunca antes visto, es decir, en datos de testeo.

De forma matématica se supone que ajustamos nuestro método de aprendizaje estadístico en nuestras observaciones de entrenamiento $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$, y obtenemos una estimación $\hat{f}$, podemos entonces calcular $\hat{f}(x_1), \hat{f}(x_2), ..., \hat{f}(x_n)$, sí estos son aproximadamente cercanos a $y_1, y_2, \cdots, y_n$ entonces el MSE de entrenamiento dado en 2.4, será pequeño, sin embargo, no nos interesa saber si $\hat{f}(x_i) \approx y_i$, en cambio, queremos saber si $\hat{f}(x_0) \approx y_0$ donde (x_0, y_0) es una observación nunca antes vista por el método de aprendizaje estadístico.

2.1.3.2. Subajuste (Underfutting) y Sobreajuste (Overfitting)

• Subajuste (Underfitting):
Este ocurre cuando el modelo tiene demasiados parámetros libres para ajustar correctamente los datos. Debido a que el modelo **NO** se ajusta bien a los datos de entrenamiento y es demasiado simplista para el problema, no se generaliza a nuevos ejemplos.

• Sobreajuste (Overfitting):
Si hay demasiadas características, el modelo puede ajustarse muy bien al conjunto de entrenamiento (el error de entrenamiento es pequeño), ya que el modelo es demasiado flexible y se adapta demasiado y coincide estrechamente con el ruido de los datos de entrenamiento entrenados. (El error de entrenamiento debe ser distinto de cero.)

2.1.3.3. Compensación sesgo-varianza

Asumimos en $Y = f(X) + \epsilon$, que podemos derivar una expresión para el error esperado de un ajuste de regresión $\hat{f}(X)$ en un punto de entrada $X = x_0$, usando la pérdida de error al cuadrado

$$Err(X_0) = E[Y - \hat{f}(x_0))^2 | X = x_0] = Var(\hat{f}(x_0)) + [sesgo(\hat{f}(x_0))]^2 + Var(\epsilon) \tag{2.5}$$

donde $E(Y - \hat{f}(x_0))^2$ define la prueba esperada MSE en x_0, que sería la prueba promedio de MSE que obtendriamos al estimar repetidamente f usando un gran número de conjuntos de entrenamiento.

Definimos ahora cada uno de estos componentes:

Error irreductible: varianza del objetivo (Y) alrededor de su verdadera media $f(X_0)$, esto es inevitable sin importar que tan bien estimemos $f(X_0)$. (a menos que $Var(\epsilon) = 0$, por supuesto) Debido al ruido ϵ.

Sesgo: la cantidad por la cual el promedio de nuestras estimaciones difiere de la media real, principalmente porque su estructura es demasiado rígida (modelo demasiado simple), por lo que nuestro modelo f se desvía sistemáticamente de la verdadera función f.

Varianza: la desviación cuadrada esperada de $\hat{f}(x_0)$ sobre su media.
Principalmente por el muestreo y a la estructura demasiado flexible, el modelo estimado f cambia mucho con diferentes muestras de entrenamiento.

Esperamos en 2.5 minimizar el MSE esperado, podemos ver entonces que necesitamos seleccionar un método de aprendizaje estadístico que simultáneamente logre baja varianza y bajo sesgo.

2.2. Regresión lineal

2.2.1. Regresión lineal simple

Es el enfoque para predecir una respuesta cuantitativa Y donde tenemos una variable predictora X, se asume que hay aproximadamenre una relación lineal entre X e Y , matemáticamente podemos escribir esta relación como:

$$Y \approx \beta_0 + \beta_1 X \tag{2.6}$$

donde β_0 y β_1 son dos constantes desconocidas que representan los términos de intersección y pendiente en el modelo lineal (coeficientes o parámetros del modelo lineal).
Es claro que en la práctica β_0 y β_1 son desconocidos, es por eso que para utilizar 2.6 al hacer predicciones debemos usar datos para estimar dichos coeficientes.
Sean $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ que representan n pares de observacion, donde cada uno consiste en una medición de X y una medida de Y, entonces nuestro objetivo será obtener estimaciones de coeficientes $\hat{\beta_0}$ y $\hat{\beta_1}$ tal que el modelo lineal 2.6 se ajuste bien a los datos disponibles , es decir
$y_i \approx \hat{\beta_0} + \hat{\beta_1} x_i$,para $i = 1, 2, \cdots, n$.
Notemos que si: $\hat{y}_i \approx \hat{\beta_0} + \hat{\beta_1} x_i$ la predicción para Y basada en el valor ith de X, tenemos entonces que $e_i = y_i - \hat{y}_i$, representara el residuo ith que será la diferencia entre el valor de respuesta ith observado y el valor de respuesta ith que predice nuestro modelo lineal.

Con lo anterior podemos definir la Suma residual de cuadrados (RSS) como:

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2 \tag{2.7}$$

o de forma equivalente

$$RSS = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2 \tag{2.8}$$

2.2.1.1. Estimación por mínimos cuadrados

Este enfoque elije $\hat{\beta}_0$ y $\hat{\beta}_1$ para minimizar 2.8 tal que:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{2.9}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{2.10}$$

donde $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ y $\bar{x} = \sum_{i=1}^{n} x_i$ son las medias de las muestras.

2.2.1.2. Evaluación de exactitud del modelo

Anteriormente en 2.1 asumimos una relación entre X y Y de la forma $Y = f(X) + \epsilon$ para una función f desconocida. Si nosotros aproximamos f por una función lineal, podemos escribirla como

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X + \epsilon \tag{2.11}$$

Con ϵ un término de error aleaorio con media cero.

Entonces para cuantificar la medida en que este modelo se ajusta a los datos, evaluaremos la calidad de un ajuste de regresión lineal utilizando las siguientes dos cantidades relacioanadas.

- Error estándar Residual (RSE)

Debido a la presencia de los términos de error en 2.11 aunque conozcamos el verdadero valor de los parámetros β_0 y β_1, no seriamos capaces de predecir perfectamente Y a partir de X.
Es entonces que hacemos uso del RSE, que es una estimación de la norma de la desviación de ϵ lo que nos dará la cantidad media que la respuesta se desviará de la verdadera línea de regresión, y se calcula mediante la formula

$$RSE = \sqrt{\frac{1}{n-2} RSS} \tag{2.12}$$

Aquí el RSE se considera una medida de ajuste del modelo, y lo interpretamos de la siguiente manera:
Si RSE es pequeño, podemos decir que el modelo se ajusta muy bien a los datos.

Si RSE es bastante grande, el modelo no se ajusta a los datos, es decir, $\hat{y}_i \neq y_i$ para una o mas observaciones.

- R^2 Estadístico

Esta es una medida absoluta de la falta de ajuste del modelo a los datos, toma forma de una proporción, es la proporción de varianza explicada, tal que

$0 < R^2 < 1$ y es calculado mediante la formula:

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} \tag{2.13}$$

donde $TSS = \sum(y_i - \bar{y})^2$ es la suma total de cuadrados.

Podemos interpretar el R^2 de la siguiente forma:
Si R^2 es cercano a 1 : Gran proporción de la variabilidad en la respuesta será explicada en la regresión.

Si R^2 es cercano a 0 : Indica que la regresión no explica gran parte de la variabilidad de la respuesta, esto se deberá a que el modelo lineal es oncorrecto ó la varianza del error es alta, o ambas.

2.2.2. Regresión lineal multiple

La regresión lineal simple es un enfoque muy útil para predecir respuestas en base a una única variable, sin embargo, en la práctica por lo general, tenemos más de un predictor, entonces un mejor enfoque es extender el modelo de regresión lineal simple 2.6, para trabajar con múltiples predictores, dando a cada predictor un coeficiente dependiente separado en un solo modelo.
Supongamos que tenemos p predictores distintos, entonces la regresión lineal múltiple toma la forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon \tag{2.14}$$

donde X_j representa el predictor jht y β_j cuantifica la asociación entre esa variable y la respuesta, interpretamos β_j como el promedio efecto sobre Y de un aumento de una unidad en X_j, manteniendo fijos todos los demás predictores.

2.2.2.1. Estimación de los coeficientes de regresión

Dado que los coeficientes $\beta_0, \beta_1, ..., \beta_p$ en 2.14 son desconocidos y deben estimarse, podemos hacer las predicciones usando la fórmula

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p \tag{2.15}$$

Aquí los parámetros se estiman usando el enfoque de mínimos cuadrados visto en 2.2.1.1 para minimizar la suma de residuos al cuadrado

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip}) \tag{2.16}$$

Donde los valores $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p$ que minimizan 2.16 son los mínimos cuadrados múltiples de estimaciones del coeficiente de regresión.

2.2.2.2. Relación entre respuesta y predictores

Para la regresión lineal múltiple con p predictores , necesitamos saber si todos los coeficientes de regresión son cero, es decir, si $\beta_1 = \beta_2 = \cdots = \beta_p = 0$, aquí

utilizamos una prueba de hipótesis para responder esta duda, donde probamos la hipótesis nula

$$H_o : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \tag{2.17}$$

versus la aternativa

$$H_a : al\, menos\, un\, \beta_j\, es\, no - cero \tag{2.18}$$

Esta prueba de hipótesis se realiza calculando el estadístico F

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} \tag{2.19}$$

donde $TSS = \sum(y_i - \bar{y})^2$ y $RSS = \sum(y_i - \hat{y})^2$.

En cuanto al ajuste del modelo dos de las medidas numéricas más comunes de ajuste del modelo son el RSE y R^2 la fracción de varianza explicada. Estas cantidades se calculan e interpretan de la misma manera que en la regresión lineal simple 2.12 y 2.13 respectivamente.

2.3. Técnicas de Regularización

Podemos hacer una formulación de las Técnicas de Regularización en el contexo de modelos lineales de la siguiente manera:

$$\hat{\beta} = \underset{\beta}{arg\,min} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2 + \phi_\lambda(|\beta|) \right\} \tag{2.20}$$

Donde $\beta = (\beta_1, \cdots, \beta_p)$, $\lambda \geq 0$ y $\phi_\lambda(|\beta|) = \lambda \sum_{j=1}^{p} \phi_j(|\beta_j|)$
es la función de penalización sobre el tamaño de β, el cual depende de λ

Una familia de funciones de penalización muy utilizada es la correspondiente a la norma l_q dadas por:

$$\phi_\lambda(\beta) = \lambda\, (\parallel \beta \parallel_q)^q = \lambda \sum_{j=1}^{p} |\beta_j|^q\,,\ q > 0 \tag{2.21}$$

Los estimadores resultantes en estos casos son conocidos como estimadores de Ridge y estimadores Lasso, que abordaremos a continuación. Para regresión Ridge tiene norma l_2 en la función 2.21 y regresión Lasso, con norma $l1$.

2.3.1. Ridge

La regresión Ridge es muy similar a los mínimos cuadrados, excepto que los coeficientes se estiman minimizando una cantidad ligeramente diferente, en particular la regresión Ridge estima los valores que minimizan

$$\hat{\beta}^{Ridge} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{n} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = RSS + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{2.22}$$

Donde el término $\sum_{j=1}^{p} \beta_j^2$ es llamado término de penalización por contracción, el cual será pequeño cuando $\beta_1, \cdots, \beta_p$ estén cerca de cero y entonces tendrá el efecto de reducir las estimaciones β_j hacia cero.
De forma equivalente podemos escribir el problema Ridge como:

$$\hat{\beta}^{Ridge} = \underset{\beta}{arg\,min} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} X_{ij}\beta_j \right)^2 \quad s.a. \quad \sum_{j=1}^{p} \beta_j^2 \leq t, \tag{2.23}$$

donde t es el parámetro de penalización.

Podemos también escribir 2.22 de forma vectorial

$$\hat{\beta}^{ridge} = \left(X^T X + \lambda I \right)^{-1} X^T \mathbf{y} \tag{2.24}$$

Donde I es la matriz identidad de orden pxp.

Para todo lo anterior tenemos que:

$\lambda \geq 0$, es un parámetro de penalización, sirve para controlar el impacto relativo de estos dos términos en las estimaciones del coeficiente de regresión.

Cuando $\lambda = 0$, el término de penalización no tiene efecto y entonces la regresión Ridge producirá estimaciones de mínimos cuadrados.

Cuando $\lambda \to \infty$, el impacto de la penalización por contracción crece y el coeficiente de regresión de Ridge que estima se acercará a cero.
Es importante mencionar que la regulacioń Ridge no selecciona variables.

2.3.2. Lasso

Lasso es una técnica de regresión lineal regularizada, así como la regresión Ridge, pero con una leve diferencia en la penalización que trae consecuencias importantes, ya que a partir de cierto valor del parámetro de penalización, el estimador de Lasso produce estimaciones nulas para algunos coeficientes y no nulas para otros, con lo cual realiza una especie de selección de variables en forma contínua, esto debido a la norma l_1. Lasso reduce la variabilidad de las estimaciones por la reducción de los coeficientes, es decir, reduce algunos coeficientes a cero.
Los coeficientes de Lasso minimizan

$$\hat{\beta}^{lasso} = \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| = RSS + \lambda \sum_{j=1}^{p} |\beta_j| \tag{2.25}$$

donde el término $\sum_{j=1}^{p} |\beta_j|$ es llamado termino de penalización por contracción, aquí la penalización l_1 tiene el efecto de forzar que algunas de las estimaciones del coeficiente sean exactamente iguales a cero, cuando el parámetro de penalización λ es lo suficientemente grande.

De forma equivalente podemos escribir el problema Lasso como:

$$\hat{\beta}^{lasso} = \underset{\beta}{arg\,min} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} X_{ij}\beta_j \right)^2 \quad s.a. \sum_{j=1}^{p} |\beta_j| \leq t, \tag{2.26}$$

donde t es el parámetro de penalización.

Podemos también escribir 2.22 de forma vectorial

$$\hat{\beta}^{lasso} = \underset{\beta}{arg\,min} \left\{ \| \mathbf{y} - X\beta \|_2^2 + \lambda \| (\beta_1, ..., \beta_p) \|_1 \right\} \tag{2.27}$$

Para todo lo anterior tenemos que:

$\lambda \geq 0$ es un parámetro de penalización, sirve para controlar el impacto relativo de estos dos términos en las estimaciones del coeficiente de regresión.

Cuando $\lambda = 0$ el término de penalización no tiene efecto y entonces la regresión Lasso producir;a estimaciones de mínimos cuadrados.

Cuando $\lambda \to \infty$ Lasso da el modelo nulo en el que todas las estimaciones de coeficientes son iguales cero.

2.3.3. Selección de parámetro de penalización

Al implementar Ridge y Lasso como vimos en 2.20 dependen de un parámetro de penalización λ, entonces necesitaremos un método para selecionar el valor de dicho parámetro, ya que la elección de este involucra un balance entre los componentes de sesgo y varianza del ECM al estimar β.
Una opción es utilizar validación cruzada, la cual describimos a continuación.

Validación cruzada
Este es un método que estima la tasa de error de prueba, consiste en la división aleatoria de la muestra en dos partes: entrenamiento y validación, de tal forma que manteniene un subconjunto de las observaciones fuera del entrenamiento del proceso de ajuste y luego aplica el método de aprendizaje estadístico al conjunto de validación.

En nuestro caso utilizaremos validación cruzada de **k-fold (k-pliegues)**, este enfoque implica dividir aleatoriamente el conjunto de observaciones en k grupos, o pliegues, de aproximadamente igual tamaño. El primer pliegue se trata como un conjunto de validación, y el método se ajusta en los $k-1$ pliegues restantes.
El error cuadrático medio MSE_1 es el cálculo para las observaciones de ese primer pliegue, este proceso se repite k veces, cada vez, se trata un grupo diferente de observaciones como un conjunto de validación, este proceso da como resultado k estimaciones de error de prueba; $MSE_1, MSE_2, \cdots, MSE_k$, entonces la estimación de $k-$pliegues calcula el promedio de estos valores:

$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^{k} MSE_i \tag{2.28}$$

Ahora para la selección de nuestro parámetro de penalización λ:
Elegimos una cuadrícula de valores λ y calculamos el error de validación cruzada para cada valor de λ, a continuación, seleccionamos el valor del parámetro de ajuste para el que se produce el error de validación cruzada más pequeño, es decir:

$$\hat{\lambda}_{cross} = arg\min_{\lambda} CV(\lambda)$$

donde $CV(\lambda) = \frac{1}{K}\sum_{i=k}^{K} MSE_k(\lambda)$

Finalmente, el modelo se reajusta utilizando todas las observaciones disponibles y el valor seleccionado del parámetro de ajuste $\lambda = \hat{\lambda}_{cross}$.

2.3.4. Estandarización

Dado que los coeficientes al aplicar los métodos de regulación pueden verse alterado si las variables no se estandarizan antes de llevar a cabo el ajuste, por ello, es mejor realizar esta estandarización antes para que todas estén en la misma escala (desviación estándar = 1, media= 0), de la siguiente forma:

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}} \tag{2.29}$$

Donde el denominador es la desviación estándar estimada del predictor jth.

2.3.5. Comparación Ridge y Lasso

Al hacer algunas comparaciones entre Ridge y Lasso podemos observar lo siguiente:

• Una notable ventaja de regresión Lasso sobre regresión Ridge es la selección de variables, ya que esto dará lugar a modelos más simples e interpretables, esto por la implicación de un pequeño subconjunto de predictores.

• Regresión Ridge obtendrá mejores resultados cuando la respuesta sea una función de muchos factores predictivos, todos estos con coeficientes que serán muy pequeños y cercanos a cero. Sin embargo, el número de predictores que se relaciona con la respuesta no se conoce a priori para los conjuntos de datos reales.

• Cuando en la regresión Ridge tiene excesivamente alta varianza (al igual que en estimaciones de mínimos cuadrados), Lasso puede brindarnos una solución aumentando el sesgo y reduciendo la varianza y asi también generar predicciones más exactas.

• La regresión Ridge superará a regresión Lasso en términos de predicción, esto debido a la inclusión de todas las variables.

• Por ultimo, al tomar en cuenta los puntos anteriores, cabe resaltar que no hay un método que siempre tenga dominio sobre otro.

2.3.5.1. Ejemplo corto

- **regresión Ridge**

Tomando en cuenta los siguientes datos de $n = 50$ jugadores, nosotros deseamos predecir el salario de un jugador de béisbol sobre la base de varias estadísticas asociado con el desempeño en el año anterior, tomando en cuenta las variables:
Hits: Número de hits en 1986
HmRuns: Número de jonrones en 1986
Runs: Número de carreras en 1986

Salario	Hits	HmRuns	Runs
475,000	0,81	7	24
480,000	130	18	66
500,000	141	20	65
91,500	87	10	39
750,000	169	4	74
70,000	37	1	23
⋮	⋮	⋮	⋮
950,000	139	31	93

Tabla 2.1
Fuente: Elaboración propia (subconjunto de Datos HItters libreia ISLR2 en R)

Solución
Tomando en cuenta 2.22: $\hat{\beta}^{ridge} = \left(X^T X + \lambda I\right)^{-1} X^T \mathbf{y}$
tenemos que:
tomando en cuenta los siguientes valores para λ , $2, 5, 10$.

$$X^T X = \begin{pmatrix} 50 & 4918 & 504 & 2474 \\ 4918 & 578166 & 60202 & 293688 \\ 504 & 60202 & 7938 & 31631 \\ 2474 & 293688 & 31631 & 154222 \end{pmatrix}$$

y para $\lambda = 2$

$$\left(X^T X + \lambda I\right)^{-1} = \begin{pmatrix} 0{,}099558068 & -0{,}00108888 & 0{,}000209297 & 0{,}000433551 \\ -0{,}00108888 & 6{,}4823E-05 & -1{,}09584E-06 & -0{,}00010575 \\ 0{,}000209297 & -1{,}09584E-06 & 0{,}00068891 & -0{,}000142565 \\ 0{,}000433551 & -0{,}00010575 & -0{,}000142565 & 0{,}000230148 \end{pmatrix}$$

tenemos ahora que :

$$X^T\mathbf{y} = \begin{pmatrix} 50 \\ 4918 \\ 504 \\ 2474 \end{pmatrix} \quad \text{y finalmente,} \quad \hat{\beta}^{ridge} = \begin{pmatrix} 0{,}800883863 \\ 0{,}002177759 \\ -0{,}000418594 \\ -0{,}000867102 \end{pmatrix}$$

Nuestro modelo sería:

$Salario = 0{,}8008839 + 0{,}0021778 * Hits - 0{,}0004185 * HmRuns - 0{,}0008671 * Runs$

para $\lambda = 5$

$$\left(X^TX + \lambda I\right)^{-1} = \begin{pmatrix} 0{,}076658792 & -0{,}000838194 & 0{,}000160971 & 0{,}000333417 \\ -0{,}000838194 & 6{,}20397E-05 & -6{,}12798E-07 & -0{,}000104568 \\ 0{,}000160971 & -6{,}12798E-07 & 0{,}000687328 & -0{,}000142382 \\ 0{,}000333417 & -0{,}000104568 & -0{,}000142382 & 0{,}000229462 \end{pmatrix}$$

tenemos ahora que :

$$X^T\mathbf{y} = \begin{pmatrix} 50 \\ 4918 \\ 504 \\ 2474 \end{pmatrix} \quad \text{y finalmente,} \quad \hat{\beta}^{ridge} = \begin{pmatrix} 0{,}616706041 \\ 0{,}004190969 \\ -0{,}000804854 \\ -0{,}001667087 \end{pmatrix}$$

Nuestro modelo sería:

$Salario = 0{,}8008839 + 0{,}004190969 * Hits - 0{,}000804854 * HmRuns - 0{,}001667087 * Runs$

para $\lambda = 20$

$$\left(X^TX + \lambda I\right)^{-1} = \begin{pmatrix} 0{,}035654492 & -0{,}000389308 & 7{,}44389E-05 & 0{,}000154118 \\ -0{,}000389308 & 5{,}69317E-05 & 1{,}08389E-07 & -0{,}00010218 \\ 7{,}44389E-05 & 1{,}08389E-07 & 0{,}000679838 & -0{,}000140817 \\ 0{,}000154118 & -0{,}00010218 & -0{,}000140817 & 0{,}000227447 \end{pmatrix}$$

tenemos ahora que :

$$X^T\mathbf{y} = \begin{pmatrix} 50 \\ 4918 \\ 504 \\ 2474 \end{pmatrix} \quad \text{y finalmente,} \quad \hat{\beta}^{ridge} = \begin{pmatrix} 0{,}286910157 \\ 0{,}00778617 \\ -0{,}001488778 \\ -0{,}003082356 \end{pmatrix}$$

Nuestro modelo sería:

$Salario = 0{,}286910157 + 0{,}00778617 * Hits - 0{,}001488778 * HmRuns - 0{,}003082356 * Runs$

Conclusión
Como podemos observar, al aumentar el valor de penalización λ, el valor de los coeficientes disminuye y se acerca cada vez más a cero como se habia mencionado en 2.3.1

• **regresión Lasso**
Dado que Lasso es un problema convexo, esto específicamente un problema cuadrático con una restricción convexa. Como tal, hay muchos métodos sofisticados para resolver Lasso. Sin embargo, aunque hay particularmente un simple y efectivo algoritmo computacional, es decir, no tenemos una forma matricial como el método Ridge para aplicar dicho ejemplo

Capítulo 3

Desarrollo del proyecto

Una vez presentado en el capítulo 2 la base teórica necesaria, se procede a presentar una aplicación en el ámbito ambiental, tratándose de un análisis espectométrico para determinar un compuesto químico de un conjunto de vasijas de vidrio de un yacimiento arqueológico, respecto a la variable a modelar , será el compuesto Oxido de Hierro.

El lenguaje de programación ultilizado para esto es R, que cuenta con funciones desarrolladas que permiten el cálculo de cada uno de los elementos necesarios para dicho análisis.

3.1. Descripción de Datos

Los siguientes datos corresponden a un trabajo para determinar la composición de un conjunto de vasijas de vidrio de un yacimiento arqueológico. Dado que el análisis espectrométrico es más barato que el análisis químico, se procuró calibrar el primero para que reemplace al segundo.

Los datos utilizados están conformados por un marco de datos con 180 observaciones cada uno:

• Vessel_X : Cada fila es el espectro de una vasija a la que se le realizó espectometría de rayos X , es decir, la energía corrrespondiente a cada frecuencia j , $j = 1, 2, \cdots, 301$:

$$V1, V2, \cdots, V301$$

• Vessel_Y : Cada fila es un análisis de laboratorio de 13 compuestos qui]ímicos en cada vasija:

$$Na_2O, MgO, AL_2O_3, SiO_2, P_2O_5, SO_3, Cl, K_2O, CaO, MnO, Fe_2O_3, BaO, PbO$$

Vamos a comparar distintos métodos para predecir el compuesto Fe_2O_3 (Oxido de hierro).

3.2. Paquete glmnet en R

Se utilizará el paquete *glmnet* de R, desarrollado por Friedman e al. (2010). el paquete *glmnet* ajusta un modelo lineal generalizado a través de la máxima verosimilitud penalizada.

La regularización se calcula para la penalización *Ridge* y *Lasso* en una cuadrícula de valores para el parámetro de regularización lambda, utilizando validación cruzada.

La función nos permite utilizar los dos casos mencionados anteriormente con la diferencia de que para aplicar *Ridge* ha de usarse el argumento $alpha = 0$,

$alpha = 1$ si se quiere aplicar Lasso.
Es importante mencionar que en el argumento de la función *glmnet* contiene un indicador lógico predeterminado para la estandarización de la variable X antes de ajustar la secuencia del modelo. Los coeficientes siempre se devuelven en la escala original, el valor predeterminado es estandarizar = VERDADERO.

3.3. Ridge, Lasso

En el presente apartado estudiaremos los métodos de regularización *Ridge* y *Lasso*.
Intetaremos seleccionar el modelo más adecuado para la predicción de nuestro compuesto químico seleccionando entre los diferentes métodos basándonos en el MSE sobre nuestros datos de entrenamiento, discutido 2.1.3.1, el valor óptimo para el parámetro λ se estimará según el método especificado en 2.3.3, (validación cruzada).
Además, es importante mencionar que al utilizar validación cruzada se obtienen diferentes λ, pero tomaremos λ_{min} el cual es el valor de λ que minimiza la medida de error utilizada para el cálculo.

3.3.1. Ridge

Hemos dividido los datos $n = 180$,de forma aleatoria en datos de entrenamiento y datos de test, de la siguiente manera:
Entrenamiento = 120
Test = 60
A continuación en la figura 3.1 mostramos la curva generada por la aplicación de validacion cruzada para distintos valores de λ, a partir del cambio del logaritmo de λ. Notamos que $\lambda_{min} = 0{,}8894$, con 301 variables

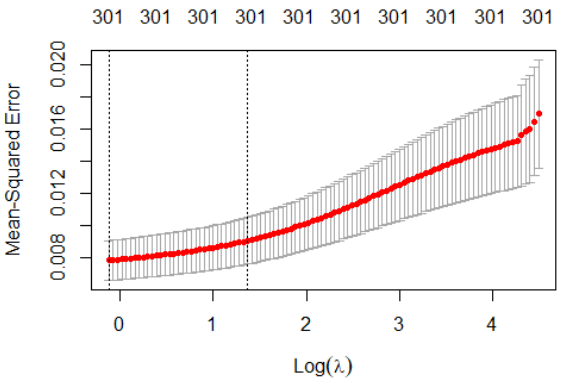

Figura 3.1: Valores de λ Regresión Ridge
Fuente: Elaboración propia

En la figura 3.2 podemos ver como disminuye el número de variables a medida el valor de λ aumenta y dado que la regulación Ridge no selecciona variables (como se había mencionado en 2.3.1), podemos notar que en λ_{min} seleccionado por validación cruzada tendremos todos los 301 coeficientes para el modelo.

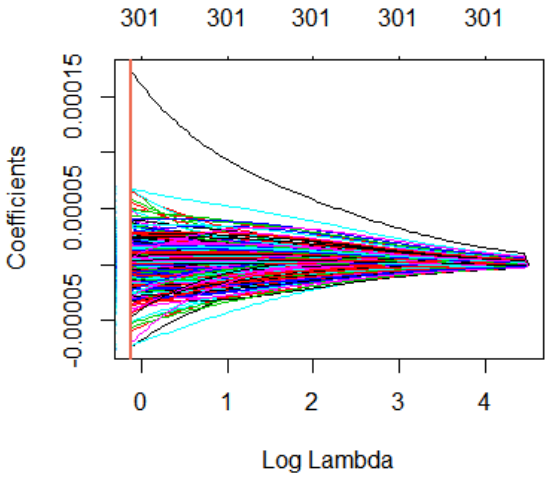

Figura 3.2: Coeficientes para distintos λ
Fuente: Elaboración propia

En la figura 3.3 están representados tanto los valores que va adquiriendo los coeficientes a medida varía el valor de λ, así como el recorrido de los coeficientes al variar la norma de penalización l_2. Notamos que el efecto de penalización a medida aumenta el valor de λ, hay una disminución en la cantidad de coeficientes cercanos a cero.

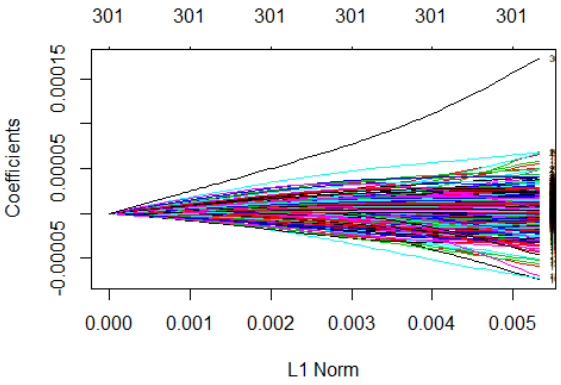

Figura 3.3: Comportamiento de los coeficientes según la norma l_2 y $log(\lambda)$
Fuente: Elaboración propia

Tenemos a continuación la tabla 3.1 de coeficientes Ridge ajustados al valor de $\lambda_{min} = 0{,}8894$

Coeficientes	$\lambda_{min} = 0{,}8894$
(intercep)	$3{,}477026e-01$
$V1$	$-5{,}292886e-06$
$V2$	$-3{,}791804e-06$
$V3$	$-4{,}306726e-06$
$V4$	$-5{,}561770e-06$
$V5$	$-7{,}298931e-06$
⋮	⋮
$V300$	$1{,}130898e-05$
$V301$	$1{,}779409e-04$

Tabla 3.1: Coeficientes Ridge
Fuente: Elaboración propia
La tablas correspondientes a los otros valores de los coeficientes se presentan en el Anexo A (.1.1) debido a su extensa longitud.

3.3.2. Lasso

Hemos dividido los datos $n = 180$, de forma aleatoria en datos de entrenamiento y datos de test, de la siguiente manera:Entrenamiento = 120, Test = 60
A continuación en la figura 3.4 mostramos la curva generada por la aplicación de validacion cruzada para distintos valores de λ, a partir del cambio del logaritmo de λ. Notamos que $\lambda_{min} = 0{,}00089$, con 67 variables.

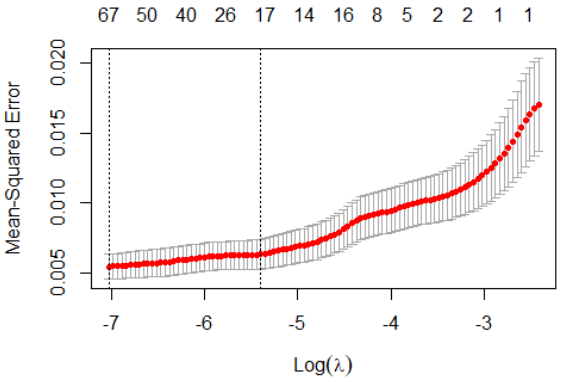

Figura 3.4: Valores de λ Regresión Lasso
Fuente: Elaboración propia

En la figura 3.5 podemos ver como disminuye el número de variables a medida el valor de $log(\lambda)$ aumenta, podemos notar que en λ_{min} tendremos 67 coeficientes para el modelo como se había mencionado anteriormente.

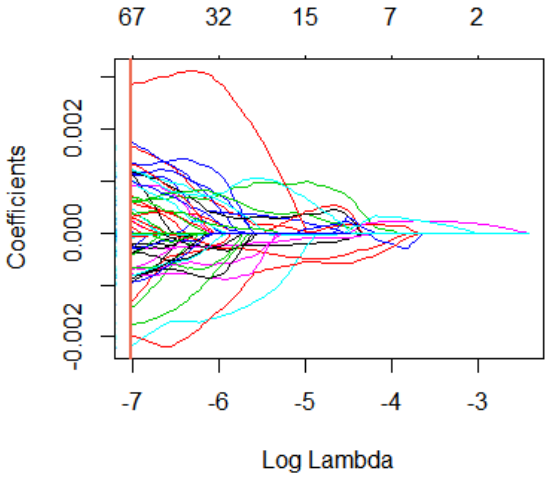

Figura 3.5: Coeficientes para distintos λ
Fuente: Elaboración propia

En la figura 3.6 están represnatados tanto los valores que va adquiriendo los coeficientes a medida varía el valor de λ, así como el recorrido de los coeficientes al variar la norma de penalización l_1. Notamos que el efecto de penalización, a medida aumenta el valor de λ, hay una disminución en la cantidad de coeficientes distintos a cero.

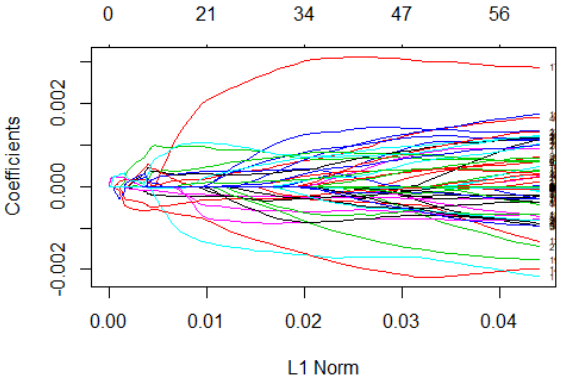

Figura 3.6: Comportamiento de los coeficientes según la norma l_1 y $log(\lambda)$
Fuente: Elaboración propia

Tenemos a continuación la tabla 3.2 de coeficietes Lasso ajustados al valor de $\lambda_{min} = 0{,}00089$

Coeficientes	$\lambda_{min} = 0{,}00089$
(intercep)	$3{,}220296e-01$
$V26$	$6{,}885800e-04$
$V27$	$-1{,}428970e-03$
$V31$	$1{,}362337e-03$
$V32$	$-2{,}055747e-04$
$V34$	$7{,}924745e-05$
$\vdots$	$\vdots$
$V281$	$-4{,}356981e-05$
$\vdots$	$\vdots$
$V300$	$-9{,}383597e-04$
$V301$	$1{,}183660e-03$

Tabla 3.2: Coeficientes Lasso
Fuente: Elaboración propia
La tablas correspondientes a los otros valores de los coeficientes se presentan en el Anexo B (.1.2) debido a su extensa longitud.

3.3.3. MSE Ridge y Lasso

En la siguiente tabla 3.3 mostramos el Error Cuadrático medio MSE, al aplicar las predicciones de los modelos con los datos de test y el mejor λ, es decir, λ_{min}

Modelo	λ_{min}	MSE
Ridge	0,88942	0,0135971
Lasso	0,00089	0,0117466

Tabla 3.3: Comparación MSE Ridge, Lasso
Fuente: Elaboración propia

3.3.4. Ajuste de nuevo modelo con coeficientes seleccionados por Regresión Lasso

Lasso nos selecccionó 67 variables :
$V26, V27, V31, V32, V34, ..., V180, V183, ..., V300, V301$

Seleccionamos los datos de entrenamiento y test para esas variables. Luego ajustamos un nuevo modelo (Modelo 1) y aplicamos lm a dichos datos de entrenamiento, dándonos para una significancia de $p < 0{,}05$ un total de 24 variables las cuales se muestran en la siguiente tabla 3.4

Coeficientes	Estimaciones
(intercep)	$2{,}011e-01$
$V26$	$1{,}708e-03$
$V27$	$-2{,}595e-03$
$V32$	$-2{,}200e-03$
$V34$	$2{,}452e-03$
40	0.011116
$V54$	$-1{,}184e-03$
⋮	⋮
V174	0.010299
V194	0.000440
⋮	⋮
V271	0.002728
V287	0.001751

Tabla 3.4: Coeficientes Modelo 1: Variables seleccionadas por Lasso
Fuente: Elaboración propia
La tablas correspondientes a los otros valores de los coeficientes se presentan en el Anexo C (.1.3) debido a su extensa longitud.

Ahora ajustamos un nuevo modelo (Modelo 2) aplicamos lm a las variables seleccionadas por el Modelo 1, dándonos para una significancia de $p < 0{,}05$ un total de 19 variables las cuales se muestran en la siguiente tabla 3.5

Coeficientes	Estimaciones
(intercep)	0,3627373
$V26$	0,0033389
$V27$	−0,0045747
$V32$	−0,0044642
$V40$	0,0034428
$V54$	−0,0022116
⋮	⋮
$V149$	−0,0018147
⋮	⋮
$V272$	0,0026383
$V287$	−0,0019498

Tabla 3.5: Coeficientes Modelo 2: Variables seleccionadas por Modelo 1
Fuente: Elaboración propia
La tablas correspondientes a los otros valores de los coeficientes se presentan en el Anexo C (.1.3) debido a su extensa longitud

3.3.4.1. Modelo1 y Modelo2

En la siguiente tabla 3.6 mostramos el R^2 ajustado, el Error estándar Residual y Cuadrático medio MSE al aplicar las predicciones de el Modelo 1 y Modelo 2 con los datos de test

Modelo	R^2	RSE	MSE
Modelo 1	0,945	0,03026	0,0137844
Modelo 2	0,8355	0,05235	0,0161221

Tabla 3.6: Comparación Modelo 1 y Modelo 2
Fuente: Elaboración propia

Finalmente realizamos un Anova entre el Modelo 1 y Modelo 2, obteniendo el resultado en tabla 3.7

Modelo	Res.Df	RSS	Df	Sum of Sq	F	$Pr(>F)$	
1	53	0,048526					
2	95	0,260330	-42	$-0{,}2118$	5,5079	$6{,}652e-09$	***

Tabla 3.7: Anova Modelo 1 y Modelo 2
Fuente: Elaboración propia

3.3.5. MSE: Ridge, Lasso, Modelo 1 y Modelo 2

En la siguiente tabla 3.8 mostramos el Error Cuadrático medio MSE al aplicar las predicciones de los modelos con los datos de test con λ_{min}. En regresiones Ridge Y Lasso así como también en los modelos de regresión Modelo 1 y Modelo 2

Modelo	MSE
Ridge	0,0135971
Lasso	0,0117466
Modelo1	0,0137844
Modelo2	0,0161221

Tabla 3.8: Comparación MSE Ridge , Lasso, Modelo 1 y Modelo 2
Fuente: Elaboración propia

Capítulo 4

Análisis de Resultados

Ya que tenemos los resultados de los modelos estudiados, podemos notar que :
En cada uno de los modelos en cuanto a los coeficientes, podemos observar que los interceptos son considerados, así como también la medida de analísis espectométrico que corresponde a los predictores:

$$V26, V27, V32, V40, V54, V92, V115, V118, V124,$$

$$V134, V149, V168, V173, V174, V194, V243, V271, V272, V287$$

Observamos en cuanto a los método de regularización:

Ridge aproxima todos los coeficientes a cero, en cuanto a la regularización Lasso tenemos una seleccción de 67 predictores el cual produce un modelo más simple e interpretable, ya que seleccionó un subconjunto de los 301 predictores, podemos notar que después de la selección de Lasso ajustamos otro modelo (Modelo 1) el cual según la tabla 3.4 nos da un total de 24 predictores de los 67 considerados anteriormente por Lasso, posteriormente ajustamos un nuevo modelo (Modelo 2), dándonos este un total de 19 predictores significativos.

Luego al hacer la comparación entre los 4 modelos anteriores del error de testeo MSE en la tabla 3.8 podemos ver que el menor MSE es la regularización Lasso, seguido regularización Ridge, posteriormente el Modelo 1 y por último el Modelo 2.

Cabe destacar que al relizar el Anova entre el Modelo 1 y Modelo 2, con el Modelo 2 un submodelo del Modelo 1 y tomando en cuenta las siguientes hipótesis:

$$H_0 : \text{Los dos modelos ajustan los datos igualmente bien}$$
$$H_a : \text{El modelo 1 ajusta mejor los datos}$$

Y según los resultados en la tabla 3.7, el estadístico F es 5,5079 y el valor p asociado es prácticamente cero.
Podemos rechazar H_0 y por lo tanto el Modelo 1 ajusta mejor los datos.

Ahora bien, tomando en cuenta lo anterior y los errores de testeo MSE en la tabla 3.8 recomendariamos utilizar en primer lugar el modelo de regularización Lasso que a su vez da una mejor interpretación, ya que tenemos menos predictores, seguidamente con un pequeño aumento en el MSE de testeo en comparación a Ridge recomendariamos el Modelo 1.

Capítulo 5

Conclusiones y trabajos futuros

5.1. Conclusiones

A lo largo de este trabajo se hizo énfasis en la selección de técnicas regularizadas para seleccionar el mejor modelo, esto con el objetivo de disminuir costos en cuanto a la predicción de un compuesto químico en un análisis espectrométrico sustituyendo al análisis químico, a continuación, mostramos las principales conclusiones obtenidas en el desarrollo del mismo.

Las técnicas estudiadas nos permiten perspectivas diferentes en cuanto al modelado de la variable estudiada en términos de los distintos predictores.

Se ha mostrado que la regresión Ridge permite regularizar las estimaciones de los coeficientes acercándolos a cero, el cual nos dio un modelo con todos los predictores, por lo cual se nos dificulta al momento de decidir cuáles de ellos tienen mayor influencia en el análisis espectométrico. Cabe destacar que a diferencia de la estimación de los modelos por mínimos cuadrados ajustados la regresión Ridge nos puede brindar un MSE menor cuando se encuentra el λ adecuado.

Con la regresión Lasso a diferencia de la regresión Ridge si contrae coeficientes a exactamente cero, lo cual nos da un modelo con una cantidad menor de predictores haciendo de este un modelo sea más interpretable, podemos darnos cuenta entonces de las variables que podrían aportar mas a la predicción de nuestro compuesto químico.

Es importante mencionar que al reajustar modelos por mínimos cuadrados con los coeficientes seleccionados por Lasso obtenemos en uno de ellos (Modelo 1), un MSE cercano al arrojado por Ridge, el cual de igual forma nos brinda una cantidad aún menor de predictores y ganamos más interpretabilidad y podemos tener una mejor información de las variables mas influyentes en el análisis químico.

Por último, es importante mencionar la importancia de validación cruzada en las técnicas de regularización, ya que el parámetro de penalización es fundamental para la obtención del menor MSE.

5.2. Lineas futuras

Tras la realización de este trabajo se motiva a diversas líneas de investigaciones futuras, como ser:

• La 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.

•. La técnica de Garrote o estimador de Garrote , siendo una alternativa a la regresión Lasso, ya que también puede dar lugar a modelos más simples, esta es una técnica curiosamente nombrada e introducida por Breman [1995]. En garrote, los coeficientes de mínimos cuadrados individuales β_j son reducido por una cantidad no negativa c_j. Garrote contrae más los coeficientes, se hacen más pequeños y algunos son incluso forzados a cero.

• Otro enfoque sería la aplicación Group Lasso propuesto por Yuan y Lin (2006), esto es de gran ayuda en muchas aplicaciones, donde el vector de parámetros de alta dimensión lleva una estructura, por lo que permite estimar que ciertas covariables agrupadas tengan valor cero. El objetivo es la estimación de alta dimensión en modelos lineales o lineales generalizados siendo escasa con respecto a grupos enteros, entonces Group Lasso, logra tal paridad grupal.

Bibliografía

[1] Peter Bühlmann , Sara van de Geer, *Statistics for High-dimensional Data Methods, Theory and Applications*,
Springer Heidelberg Dordrecht London, New York (2012),

[2] Gareth James, Daniela Witten, Robert Tibshirani, Trevor Hastie, *An Introduction to Statistical Learning with Applications in R*, 2ed
Springer Verlag, New York (2013),

[3] López Cruz, M. A., *Aplicación del Elastic Net LASSO y modelos relacionados en selección genómica basados en marcadores moleculares (Master's thesis).* (2012),

[4] García S. Jasmin, *Aplicaciones del modelo LASSO bayesiano en finanzass,* (2011),

[5] Lücken Giménez, José Ignacio von, *Métodos de Regularización Lasso, Ridge y Elastic Net: Una aplicación a los seguros de no vida. (2021),*

[6] Trevor Hastie, Robert Tibshirani, Martin Wainwrightn, *Statistical Learning with Sparsity The Lasso and Generalizations. Taylor & Francis Group, LLC CRC Press is an imprint of Taylor & Francis Group, an Informa business, Sound Parkway NW (2015)*

[7] George A. F. Seber Alan J. Lee , *An Introduction to Statistical Learning with Applications in R*, 2ed
John Wiley & Sons, Inc., Hoboken, New Jersey (2003)

[8] Brian S. Everitt, Torten Hothorn *A Handbook Statistical Analyses Using R*, 2ed
kTaylor & Francis Group, LLC CRC Press is an imprint of Taylor & Francis Group, an Informa business, Sound Parkway NW (2010)

[9] Marvin H.J. Gruber, *Improving Efficiency by Shrinkage The James-Stein and Ridge Regression Estimators*,
MARCEL DEKKER. INC , New York (1998)

[10] Andrey Thikhonov, *Solution of Incorrectly Formulated Problems and the Regularization Method*,
Soviet Mathematics Doklady (1963)

Anexos

.1. Anexos

.1.1. Anexo A
Coeficientes Ridge

Coeficientes	$\lambda_{min} = 0{,}8894$
(Intercept)	3.477026e-01
V1	-5.292886e-06
V2	-3.791804e-06
V3	-4.306726e-06
V4	-5.561770e-06
V5	-7.298931e-06
V6	-7.620651e-06
V7	-8.005651e-06
V8	-1.105525e-05
V9	-9.563542e-06
V10	-2.290400e-05
V11	-2.109916e-06
V12	-1.554865e-05
V13	-3.033696e-06
V14	3.866279e-06
V15	1.159352e-05
V16	3.533547e-05
V17	1.327023e-05
V18	2.568487e-05
V19	2.044272e-05
V20	1.346833e-05
V21	1.807217e-05
V22	1.970869e-05
V23	1.970285e-05
V24	1.952358e-05
V25	2.045297e-05
V26	3.070277e-05
V27	-1.126637e-06
V28	6.362804e-07
V29	1.123595e-05
V30	-4.830105e-06

Coeficientes	$\lambda_{min} = 0{,}8894$
V31	3.686339e-05
V32	9.886825e-06
V33	7.228648e-07
V34	2.431012e-05
V35	1.560097e-05
V36	4.226519e-05
V37	3.516493e-05
V38	3.582299e-05
V39	3.590546e-05
V40	3.587230e-05
V41	3.458962e-05
V42	2.628258e-05
V43	2.936948e-05
V44	2.552450e-05
V45	2.403916e-05
V46	2.350505e-05
V47	2.472842e-05
V48	2.260052e-05
V49	1.797873e-05
V50	1.629415e-05
V51	6.785012e-06
V52	9.933571e-06
V53	1.585848e-08
V54	-1.572936e-05
V55	-5.634693e-06
V56	-8.354593e-06
V57	-1.250912e-05
V58	-5.020717e-06
V59	-3.962290e-06
V60	-3.694439e-06

Coeficientes	$\lambda_{min} = 0{,}8894$
V61	-2.610168e-06
V62	-2.154400e-06
V63	-1.590025e-06
V64	-1.196992e-06
V65	-9.965747e-07
V66	-8.281983e-07
V67	-7.462155e-07
V68	-6.925209e-07
V69	-6.435594e-07
V70	-6.539858e-07
V71	-7.341954e-07
V72	-7.541868e-07
V73	-8.814915e-07
V74	-1.104702e-06
V75	-1.302522e-06
V76	-1.838732e-06
V77	-2.353570e-06
V78	-3.014590e-06
V79	-4.523173e-06
V80	-5.440781e-06
V81	-7.673350e-06
V82	-9.237827e-06
V83	-1.374153e-05
V84	-1.218694e-05
V85	-1.797475e-05
V86	2.515067e-07
V87	1.147783e-05
V88	1.498306e-05
V89	1.215753e-05
V90	1.936539e-05
V91	2.229456e-05
V92	2.423360e-05

Coeficientes	$\lambda_{min} = 0{,}8894$
V93	1.998716e-05
V94	1.805478e-05
V95	1.885783e-05
V96	1.916887e-05
V97	1.628742e-05
V98	2.620223e-05
V99	2.574982e-05
V100	2.412520e-05
V101	3.150956e-05
V102	2.265090e-05
V103	3.285829e-05
V104	2.481512e-06
V105	1.936995e-05
V106	-1.703593e-05
V107	2.315793e-05
V108	2.264669e-05
V109	-1.197686e-05
V110	-1.102870e-05
V111	-3.043426e-06
V112	-1.635228e-05
V113	2.414053e-05
V114	-2.306061e-05
V115	-7.698524e-05
V116	-6.162368e-05
V117	-2.621506e-05
V118	-3.360780e-05
V119	-5.355703e-05
V120	-3.660883e-05
V121	-3.330862e-05
V122	-3.483091e-05
V123	-3.054914e-05
V124	-3.039723e-05
V125	-1.777673e-05

Coeficientes	$\lambda_{min} = 0{,}8894$
V126	-1.635938e-05
V127	-1.780403e-05
V128	-1.629593e-05
V129	-1.462987e-05
V130	-1.903477e-05
V131	-1.586703e-05
V132	-1.437004e-05
V133	-1.674811e-05
V134	-8.769851e-06
V135	-2.007178e-05
V136	-1.972343e-05
V137	-1.637252e-05
V138	-1.537068e-05
V139	-2.621463e-05
V140	-3.686255e-05
V141	-5.237889e-05
V142	-2.985489e-05
V143	-7.314965e-05
V144	-2.352855e-05
V145	-2.585736e-05
V146	-2.933186e-05
V147	-5.517879e-05
V148	-2.120778e-05
V149	-4.972249e-05
V150	-2.902296e-05
V151	-1.353160e-05
V152	-1.808551e-05
V153	-1.242626e-05
V154	-1.574556e-05
V155	-1.140655e-05
V156	-1.885592e-05
V157	-1.988422e-05
V158	-8.904441e-06
V159	-1.243237e-06

Coeficientes	$\lambda_{min} = 0{,}8894$
V160	-2.154409e-05
V161	-2.618522e-05
V162	-4.014845e-05
V163	-2.940018e-05
V164	-2.147222e-05
V165	-1.509262e-05
V166	-3.423344e-05
V167	-1.993266e-05
V168	3.060964e-06
V169	-3.907695e-05
V170	-4.682867e-05
V171	6.453322e-05
V172	3.639241e-05
V173	6.111024e-07
V174	4.412185e-05
V175	2.425358e-05
V176	-3.876307e-05
V177	2.022362e-06
V178	-1.043565e-05
V179	-2.852943e-05
V180	-4.595055e-05
V181	2.001820e-05
V182	-1.691460e-08
V183	-3.152369e-05
V184	-2.558147e-05
V185	3.792327e-06
V186	-7.300116e-05
V187	-4.898066e-05
V188	6.279263e-05
V189	1.739593e-05
V190	1.999388e-05
V191	4.673709e-05
V192	-2.974373e-05
V193	1.269433e-05

Coeficientes	$\lambda_{min} = 0{,}8894$
V194	5.389703e-05
V195	5.712304e-05
V196	4.265059e-06
V197	6.638996e-05
V198	1.135230e-05
V199	2.430733e-05
V200	-2.027956e-06
V201	-2.523611e-06
V202	3.786507e-05
V203	-6.041660e-07
V204	2.972328e-05
V205	2.165837e-05
V206	8.649715e-06
V207	3.702340e-06
V208	5.032621e-06
V209	3.224066e-06
V210	2.535178e-06
V211	1.445083e-06
V212	7.545233e-07
V213	7.158130e-07
V214	6.300351e-07
V215	4.291881e-07
V216	1.000046e-07
V217	1.670913e-07
V218	2.464675e-07
V219	1.071967e-07
V220	1.064616e-07
V221	9.934336e-08
V222	9.849788e-08
V223	7.619717e-08
V224	8.414609e-08
V225	2.054694e-07
V226	6.049974e-08
V227	3.177326e-07

Coeficientes	$\lambda_{min} = 0{,}8894$
V228	8.013462e-07
V229	8.500724e-07
V230	1.805853e-06
V231	4.459015e-06
V232	9.520990e-06
V233	1.768366e-05
V234	2.097616e-05
V235	3.802925e-05
V236	2.916315e-05
V237	3.694873e-05
V238	1.950775e-05
V239	2.382639e-05
V240	9.732751e-06
V241	8.697756e-06
V242	8.887736e-06
V243	5.033652e-06
V244	5.285178e-06
V245	4.384321e-06
V246	3.804092e-06
V247	3.726986e-06
V248	2.335931e-06
V249	2.106130e-06
V250	1.607433e-06
V251	1.340550e-06
V252	1.197972e-06
V253	1.094071e-06
V254	1.163499e-06
V255	1.090126e-06
V256	1.069618e-06
V257	9.944286e-07
V258	9.987140e-07
V259	1.074034e-06
V260	1.132906e-06

Coeficientes	$\lambda_{min} = 0{,}8894$
V261	1.535822e-06
V262	1.847008e-06
V263	2.174455e-06
V264	2.550367e-06
V265	3.691862e-06
V266	5.393599e-06
V267	8.436069e-06
V268	8.966034e-06
V269	1.467049e-05
V270	1.744560e-05
V271	4.036369e-05
V272	5.014092e-05
V273	4.975751e-05
V274	2.958287e-05
V275	3.615853e-05
V276	2.760231e-05
V277	2.748305e-05
V278	2.806119e-05
V279	1.805052e-05
V280	1.940484e-05
V281	1.065875e-05
V282	1.204232e-05
V283	9.371906e-06
V284	1.060340e-05
V285	8.419205e-06
V286	7.830021e-06
V287	5.502430e-06
V288	6.671523e-06
V289	8.123333e-06
V290	1.271625e-05
V291	1.527806e-05
V292	1.347728e-05
V293	1.545896e-05
V294	1.768504e-05

Coeficientes	$\lambda_{min} = 0{,}8894$
V295	2.572820e-05
V296	3.126051e-05
V297	4.636047e-05
V298	4.469134e-05
V299	7.188400e-05
$V300$	$1{,}130898e-05$
$V301$	$1{,}779409e-04$

Tabla 1: Coeficientes Ridge
Fuente: Elaboración propia

.1.2. Anexo B
Coeficientes Lasso

Coeficientes	$\lambda_{min} = 0{,}00089$
(intercep)	$3{,}220296e-01$
$V26$	$6{,}885800e-04$
$V27$	$-1{,}428970e-03$
$V31$	$1{,}362337e-03$
$V32$	$-2{,}055747e-04$
$V34$	$7{,}924745e-05$
V40	1.707493e-03
V41	5.649239e-04
V48	1.929423e-05
V54	-7.469045e-04
V57	-1.812221e-04
V76	-2.281763e-05
V79	-3.112054e-04
V85	-4.539159e-04
V92	6.371618e-04
V104	-1.568399e-04
V113	1.330848e-03
V115	-2.164076e-03
V116	-6.671020e-04
V118	1.086032e-03
V124	-1.298785e-03
V134	4.113816e-04
V144	2.404530e-04
V147	-2.524058e-04
V149	-2.038722e-03
V150	-3.700231e-04
V159	1.153850e-03
V167	4.220423e-04
V168	1.983434e-04
V170	-4.130083e-04
V171	2.864550e-03
V172	4.274513e-04

Coeficientes	$\lambda_{min} = 0{,}00089$
V173	-3.301915e-04
V174	1.213224e-03
V175	9.299215e-04
V176	-1.470610e-04
V178	1.210711e-04
V180	-5.649020e-04
V183	-1.709555e-04
V185	-1.322240e-04
V186	-7.890162e-04
V187	-5.499555e-05
V188	3.365338e-04
V192	-1.715700e-03
V194	1.695488e-03
V197	1.123270e-03
V199	-1.627968e-05
V200	-7.779666e-04
V201	-4.050386e-0
V216	-1.554361e-05
V224	-2.210156e-05
V226	-3.022900e-05
V233	3.488219e-04
V235	5.452901e-04
V237	9.093122e-04
V243	-7.488761e-04
V271	1.228355e-03
V272	1.285474e-03
V273	6.951782e-04
V278	2.941870e-05
V281	-4.356981e-05
V287	-9.101000e-04
V298	2.011723e-04
V299	5.792137e-05
$V300$	$-9{,}383597e-04$
$V301$	$1{,}183660e-03$

Tabla 2: Coeficientes Lasso
Fuente: Elaboración propia

.1.3. Anexo C
Coeficientes Modelo 1 , Modelo 2

Coeficientes	Estimaciones
(intercep)	$2{,}011e-01$
$V26$	$1{,}708e-03$
$V27$	$-2{,}595e-03$
$V32$	$-2{,}200e-03$
$V34$	$2{,}452e-03$
V40	0.011116
$V54$	$-1{,}184e-03$
V92	4.49e-05
V115	0.000236
V118	0.000121
V124	4.27e-06
V134	0.005751
V149	0.000704
V150	0.007288
V159	0.030952
V167	0.033693
V168	0.030189
V171	6.88e-05
V173	0.010229
V174	0.010299
V194	0.000440
V243	0.032070
V271	0.002728
V287	0.001751

Tabla 3: Coeficientes Modelo 1: Variables seleccionadas por Lasso
Fuente: Elaboración propia

Coeficientes	Estimaciones
(intercep)	0,3627373
$V26$	0,0033389
$V27$	−0,0045747
$V32$	−0,0044642
$V40$	0,0034428
$V54$	−0,0022116
V92	0.003504
V115	0.000309
V118	0.002122
V124	5.02e-05
V134	0.010578
V149	0.006045
V168	0.017828
V173	0.011684
V174	0.018768
V194	0.005938
V243	0.011434
V271	0.000293
$V272$	0,0026383
$V287$	−0,0019498

Tabla 4: Coeficientes Modelo 2: Variables seleccionadas por Modelo 1
Fuente: Elaboración propia