Técnicas de selección y Regularización Ridge y Lasso

Grevy Stiben Rápalo
Asesorado por Msc. Roberto Duarte del departamento de matemáticas
UNAH

Marzo, 2023

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.
En este trabajo se trata de determinar la composición de un conjunto de vasijas de vidrio de un yacimiento arqueológico, en particular la concentración de Oxido de Hierro. Dado que el análisis espectrométrico es más barato que el análisis químico, se procuró calibrar el primero para que reemplace al segundo.
Para la modelación dicho trabajo se utilizaron tecnicas de selección y regularización Ridge y Lasso.

Estimación de f

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

$$Y = f(X) + \epsilon \tag{1}$$

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

¿Por qué estimar f?

1. Predicción:
   El error en (1) se promedia a cero y se puede predecir Y usando

$$\hat{Y} = \hat{f}(X) \tag{2}$$

   donde $\hat{f}$ representa nuestra estimación para f, e $\hat{Y}$ representa la predicción resultante para Y.
   - Error reducible
   - Error irreducible
2. Inferencia:
   Es la comprensión de la asociación y relación entre Y y $X_1, X_2, ..., X_p$

¿ Cómo estimar f ?

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

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

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \tag{3}$$

• 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 $\beta_0, \beta_1, ..., \beta_p$.

Precisión e interpretabilidad del modelo

• En la Precisión e interpretabilidad nos interesa la flexibilidad de un modelo, y esto estará relacionado con el enfoque que elijamos.

- Flexibles: interés es la predicción, son modelos con menor interpretabilidad y se ajustan demasiado a los datos
- Inflexibles: interés en la inferencia, son modelos con mayor interpretabilidad y más restrictivos.

Medición de calidad de ajuste

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

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \tag{4}$$

- Un MSE grande nos dice que para algunas observaciones, lo predicho por el modelo con el valor verdadero difieren sustancialmente.

- Un MSE pequeño dirá que las respuestas predichas están muy cerca de las respuestas verdaderas

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

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

Error irreducible: varianza del objetivo (Y) alrededor de su verdadera media $f(X_0)$.

Sesgo: la cantidad por la cual el promedio de nuestras estimaciones difiere de la media real.

Varianza: la desviación cuadrada esperada de $\hat{f}(x_0)$ sobre su media.

Underfitting y Overfitting

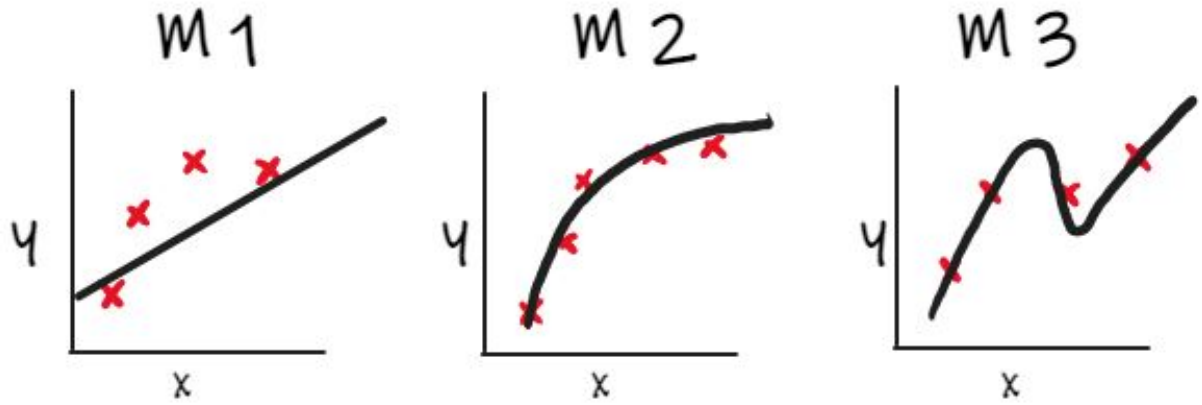
- Underfitting (Subajuste) :

Este ocurre cuando el modelo tiene demasiados parámetros libres para ajustar correctamente los datos.

- Overfitting: (Sobreajuste)

El modelo puede ajustarse muy bien al conjunto de entrenamiento, es demasiado flexible, se adapta demasiado y coincide estrechamente con el error de los datos de entrenamiento.

Explicacion gráfica



Fuente: Elaboración propia

$m_1 : \beta_0 + \beta_1 x$
Underffiting:
Baja flexibilidad, sesgo alto, baja varianza, alta interpretabilidad.

$m_2 : \beta_0 + \beta_1 x + \beta_2 x^2$
Intermedio:
flexibilidad, Sesgo, varianza, interpretabilidad.

$m_3 : \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 x^4$
Overffiting:
Alta flexibilidad, Sesgo bajo, alta varianza, baja interpretabilidad.

Regresión Lineal Múltiple

En la regresión múltiple, intentamos predecir una variable dependiente o de respuesta y en la base de una relación lineal asumida con varias variables independientes o predictoras $X_1, X_2, ..., X_p$. Además de construir un modelo para la predicción, es posible que deseemos evaluar el alcance de la relación entre Y y las variables X.
En este modelo se supone que la función de regresión que relaciona la variable dependiente con las variables independientes es lineal, es decir:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon \tag{5}$$

donde X_j representa el predictor *jht* y β_j cuantifica la asociación entre esa variable y la respuesta.

Regresión Lineal Múltiple

Dado que los coeficientes $\beta_0, \beta_1, ..., \beta_p$ en (5) son desconocidos y deben estimarse, podemos hacer las predicciones usando la fórmula

$$\hat{y} \approx \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p \tag{6}$$

Notemos que en (6), 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*

Ahora podemos definir la Suma residual de cuadrados (RSS) como:

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2 \tag{7}$$

o de forma equivalente

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip}) \tag{8}$$

Mínimos cuadrados

Podemos plantear el modelo en forma matricial de la siguiente manera:
$\mathbf{Y} = \mathbf{X}\beta + \epsilon$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_p \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{p1} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_p \end{pmatrix}$$

Aquí el vector de parámetros β que minimiza *RSS*, se estima usando el enfoque de mínimos cuadrados:

$$\hat{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y} \tag{9}$$

Donde $\mathbf{X}^t$ es la matriz transpuesta de $\mathbf{X}$

¿Por qué usar regularización?

Dado que en nuestro conjunto de datos consta de una muestra de 180 vasijas y tenemos para cada una 301 medidas de espectometría (las que utilizaremos para predecir el compuesto químico). Sabemos que la regresión lineal es el método de aprendizaje estadístico más utilizado en la actualidad, sin embargo, el método estandar de mínimos cuadrados ordinarios, que aún y cuando tienen poco sesgo, no es del todo preciso cuando $p > n$, ya que se vuelven extremandamente variable. Es por eso que consideramos técnicas de selección y regularización, ya que podremos reducir considerablemente la varianza a expensas de un pequeño aumento en el sesgo.

Técnicas de Selección y Regularización

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

$$\hat{\beta} = \underset{\beta}{arg\ min} \left\{ \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2 + \phi_\lambda(|\beta|) \right\} \tag{10}$$

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:

$$\phi_\lambda(\beta) = \lambda \left(\| \beta \|_q \right)^q = \lambda \sum_{j=1}^{p} |\beta_j|^q, \ q > 0 \tag{11}$$

Los estimadores resultantes en estos casos son conocidos como estimadores de Ridge con tiene norma l_2 y estimadores Lasso con norma $l1$.

Regresión 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

$$\hat{\beta}^{Ridge} = \sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{n}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = RSS + \lambda\sum_{j=1}^{p}\beta_j^2 \quad (12)$$

Donde el término $\lambda\sum_{j=1}^{p}\beta_j^2$ 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.

Podemos también escribir (12) de forma matricial

$$\hat{\beta}^{ridge} = \left(X^T X + \lambda I\right)^{-1} X^T \mathbf{y} \tag{13}$$

Donde I es la matriz identidad de orden pxp.

* Estandarización

$$\tilde{x}_{ij} = \frac{x_{ij}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}} \tag{14}$$

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

* Es importante mencionar que la regulacioń Ridge no selecciona variables.

Regresión Lasso

Lasso posee una leve diferencia en la penalización con respecto a Ridge, ya que a partir de cierto valor del parámetro de penalización, el estimador de Lasso produce estimaciones nulas para algunos coeficientes cuando el parámetro de penalización λ es lo suficientemente grande 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.
Los coeficientes de Lasso minimizan

$$\hat{\beta}^{lasso} = \sum_{i=1}^{n}\left(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j| \quad (15)$$

donde el término $\lambda\sum_{j=1}^{p}|\beta_j|$ es llamado termino de penalización por contracción.

λ en Ridge y Lasso

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

Cuando $\lambda = 0$ el término de penalización no tiene efecto y entonces las regresiónes Ridge y Lasso producirán estimaciones de mínimos cuadrados.

Cuando $\lambda \to \infty$

- En Ridge :el impacto de la penalización por contracción crece y el coeficiente de regresión de Ridge que estima se acercará a cero.

- En Lasso dará el modelo nulo en el que todas las estimaciones de coeficientes son iguales cero.

Validación Cruzada para elegir λ

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 utilizamos validación cruzada de $k-$folds.
El algoritmo es el siguiente:

Algoritmo validación cruzada $k-$folds para elegir λ

i) Dividimos la muestra en K partes o folds de igual tamaño, elegidas de forma aleatoria , $\cup_{k=1}^{K} B_K = \{1, 2, ..., n\}$. Generalmente se toma $K = 5$ ó $K = 10$
Fijamos un conjunto de valores λ

ii) Para cada fold, $k = 1, ...K$.
• **Entrenamiento** Para cada λ en el conjunto, calculamos los estimadores a partir de las observaciones de la muestra de entrenamiento, es decir, las que **no** pertenecen a B_K
• **Validación o testeo** Para cada observación (x_j, y_j) tal que $j \in B_K$, predecimos el valor de y_j con los estimadores calculados. Y luego calculamos el error cuadrático medio de predicción para cada λ en en el conjunto dado, $MSE_K(\lambda)$

Algoritmo validación cruzada $k-$folds para elegir λ

iii) Para cada λ en el conjunto de lambdas promediamos estas K estimaciones del error de predicción, lo que produce una curva de error de validación cruzada.

$$CV(\lambda) = \frac{1}{K}\sum_{i=k}^{K} MSE_k(\lambda) \tag{16}$$

Luego, seleccionamos el valor del parámetro de ajuste para el que se produce el menor error de validación cruzada , es decir:

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

Finalmente el modelo se vuelve a ajustar usando todas las observaciones y el valor seleccionado para el parámetro de penalización $\lambda = \hat{\lambda}_{cross}$

Validación cruzada $k = 10$

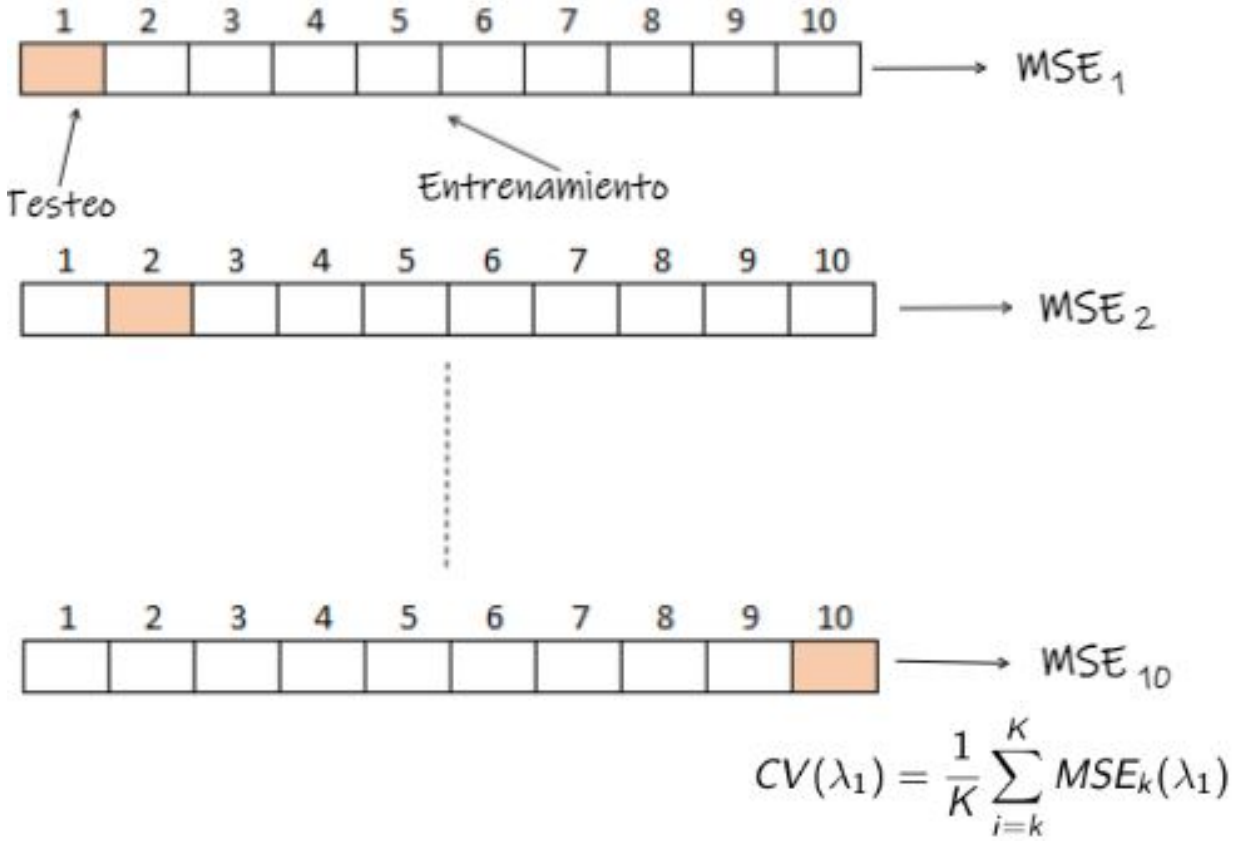


Figure: Validación cruzada: datos entrenamiento y testeo

Fuente: Elaboración propia

Aplicación al problema

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

Descripción de los datos

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

• Variable a predecir : Análisis de laboratorio del compuesto químico Fe_2O_3 en cada vasija

• 301 variables predictoras, las cuales son el espectro de una vasija a la que se le realizó espectometría de rayos X , es decir, la energía corrrespondiente a cada frecuencia j , $j = 1, 2, \cdots, 301$:

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

Para realizar el trabajo hemos dividido los datos $n = 180$,de forma aleatoria en datos de entrenamiento y datos de testeo, de la siguiente manera:

Entrenamiento $= 120$

Testeo $= 60$

Regresión Ridge

En la siguiete gráfica mostramos la curva generada por la aplicación de validacion cruzada para distintos valores de λ, notamos que $\lambda_{min} = 0.8894$, $(log(0.8894) = -0.1172)$ da 301 variables.

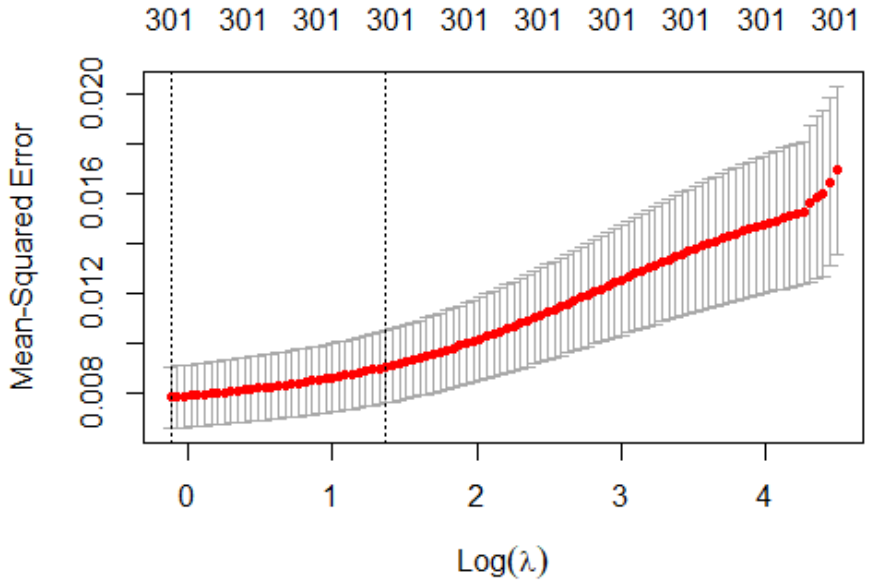


Figure: Valores de λ Regresión Ridge

Fuente: Elaboración propia

En el siguiente gráfico podemos ver los valores que van adquiriendo los coeficientes a medida que varía el valor de λ. Dado que la regulación Ridge no selecciona variables, podemos notar que en λ_{min} seleccionado por validación cruzada tendremos todos los 301 coeficientes para el modelo.

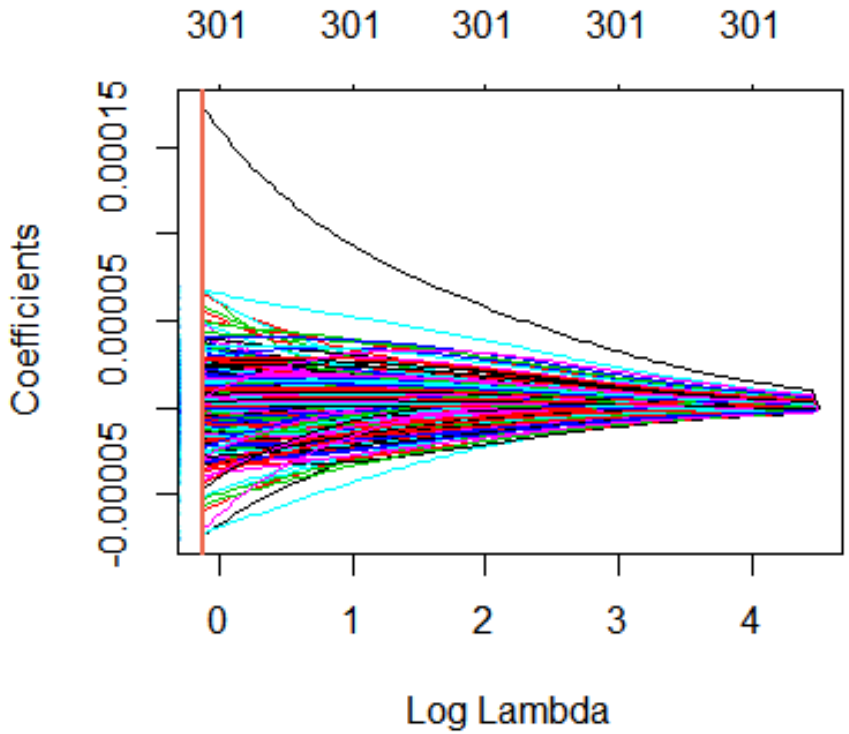


Figure: Coeficientes para distintos λ

En el siguiente gráfico está representado el recorrido de los coeficientes al variar la norma de penalización l_2. Notamos el efecto de penalización, a medida disminuye el valor de la norma, el valor que toman los coeficientes se acerca a cero.

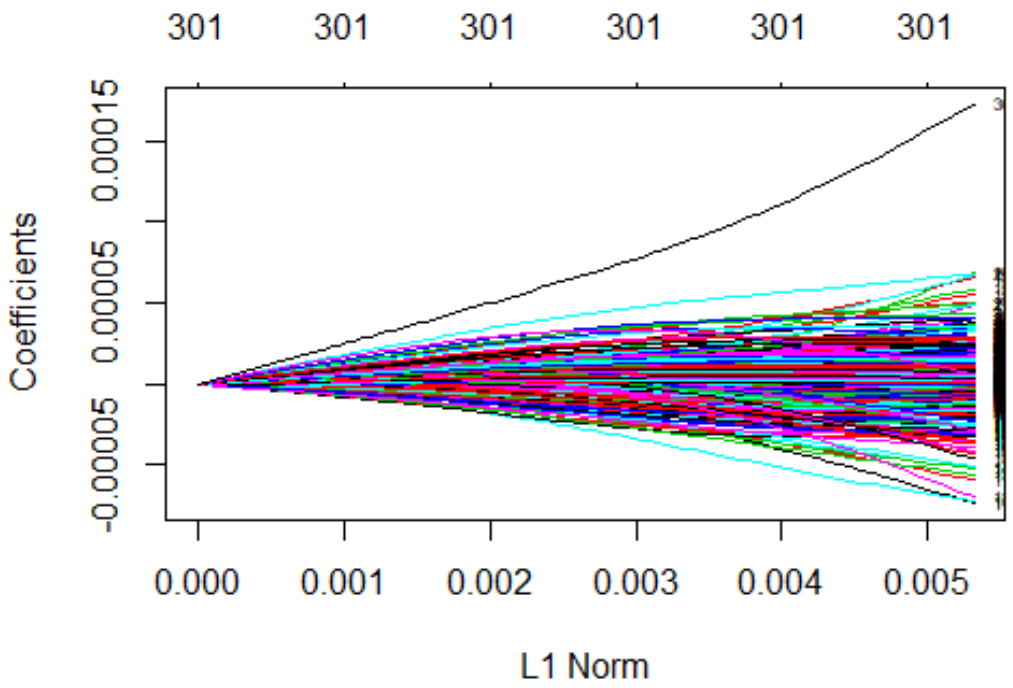


Figure: Comportamiento de los coeficientes según la norma l_2 y $log(\lambda)$

Fuente: Elaboración propia

Tenemos en la siguiente tabla los coeficientes Ridge ajustados al valor de $\lambda_{min} = 0.8894$

Coeficientes	$\lambda_{min} = 0.8894$
(intercep)	$3.477026e-01$
$V1$	$-5.292886e-06$
$V2$	$-3.791804e-06$
$V3$	$-4.306726e-06$
$V4$	$-5.561770e-06$
$V5$	$-7.298931e-06$
$\vdots$	$\vdots$
$V300$	$1.130898e-05$
$V301$	$1.779409e-04$

Table: Coeficientes Ridge

Fuente: Elaboración propia

Regresión Lasso

En el siguiente gráfico mostramos la curva generada por la aplicación de validacion cruzada para distintos valores de λ, notamos que $\lambda_{min} = 0.00089$, $(log(0.00089) = -7.025)$ selecciona 67 variables.

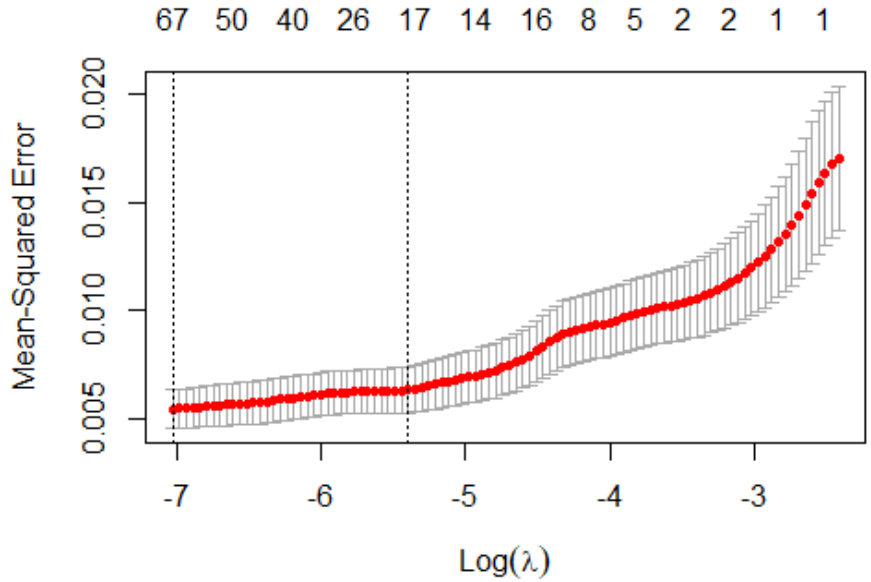


Figure: Valores de λ Regresión Lasso

Fuente: Elaboración propia

En el siguiente gráfico podemos ver los valores que van adquiriendo los coeficientes a medida que varía el valor de λ, a medida incrementa el valor de λ disminuye la cantidad de coeficientes distintos a cero.

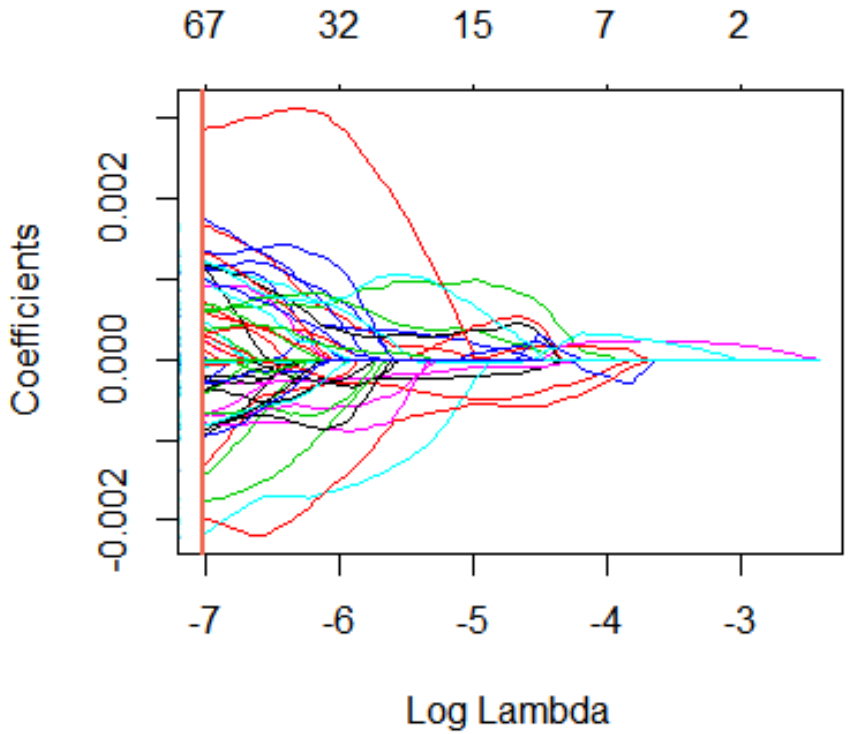


Figure: Coeficientes para distintos λ

Fuente: Elaboración propia

En el siguiente gráfico está representado el recorrido de los coeficientes al variar la norma de penalización l_1. Notamos el efecto de penalización, a medida disminuye el valor de la norma, el valor que toman los coeficientes se acerca a cero.

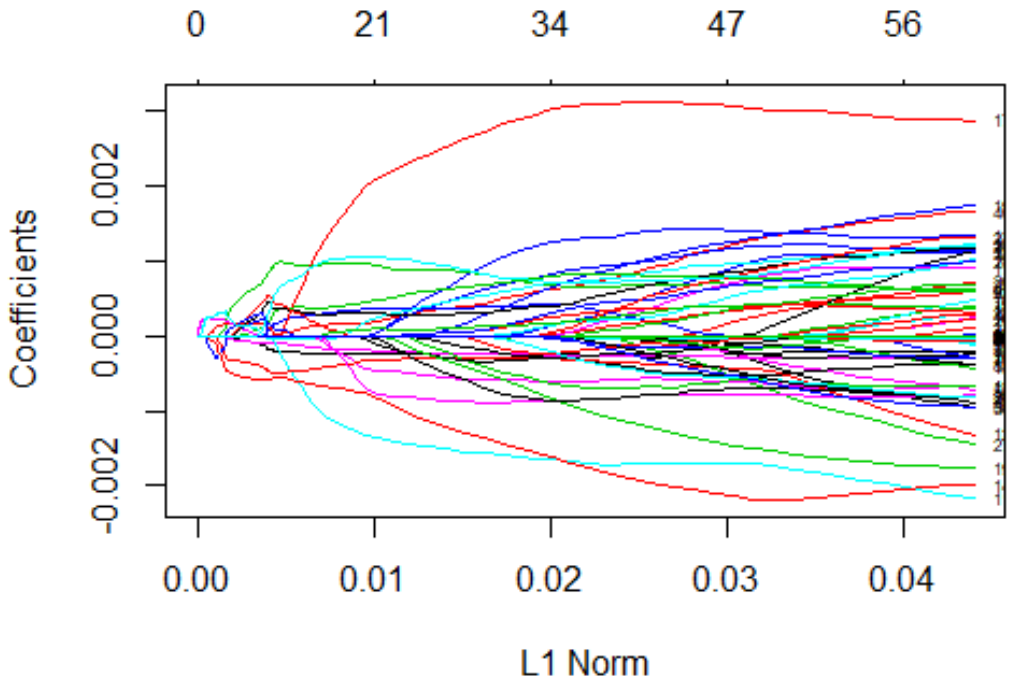


Figure: Comportamiento de los coeficientes según la norma l_1 y $log(\lambda)$

Fuente: Elaboración propia

Tenemos a continuación la tabla de coeficientes Lasso ajustados al valor de $\lambda_{min} = 0.00089$

Coeficientes	$\lambda_{min} = 0.00089$
(intercep)	$3.220296e-01$
$V26$	$6.885800e-04$
$V27$	$-1.428970e-03$
$V31$	$1.362337e-03$
$V32$	$-2.055747e-04$
$\vdots$	$\vdots$
$V281$	$-4.356981e-05$
$\vdots$	$\vdots$
$V300$	$-9.383597e-04$
$V301$	$1.183660e-03$

Table: Coeficientes Lasso

Fuente: Elaboración propia

MSE Ridge y Lasso

En la siguiente tabla 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

Table: Comparación MSE Ridge, Lasso

Fuente: Elaboración propia

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 Regresión lineal a dichos datos de entrenamiento.

dándonos para una significancia de $p < 0.05$ un total de 24 variables las cuales algunas se muestran en la siguiente tabla

Ajuste de nuevo modelo con coeficientes seleccionados por Regresión Lasso

Coeficientes	Estimaciones
(intercep)	$2.011e-01$
$V26$	$1.708e-03$
$V27$	$-2.595e-03$
⋮	⋮
V174	0.010299
V194	0.000440
⋮	⋮
V271	0.002728
V287	0.001751

Table: Coeficientes Modelo 1: Variables seleccionadas por Lasso

Fuente: Elaboración

Ajuste modelo 2

Ahora ajustamos un nuevo modelo (Modelo 2) aplicamos regresión lineal 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 algunas en la siguiente tabla

Ajuste modelo 2

Coeficientes	Estimaciones
(intercep)	0.3627373
$V26$	0.0033389
$V27$	-0.0045747
$V32$	-0.0044642
$V54$	-0.0022116
$\vdots$	$\vdots$
$V149$	-0.0018147
$\vdots$	$\vdots$
$V272$	0.0026383
$V287$	-0.0019498

Table: Coeficientes Modelo 2: Variables seleccionadas por Modelo 1

Fuente: Elaboración propia

Anova Modelo 1 y Modelo 2

Finalmente realizamos un 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 siguiente tabla , 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.

Modelo	Res.Df	RSS	Df	Sum of Sq	F	$Pr(>F)$	
1	53	0.0485					
2	95	0.2603	-42	-0.212	5.508	$6.6e-11$	***

Table: Anova Modelo 1 y Modelo 2

Fuente: Elaboración propia

MSE entre modelos

MSE: Ridge, Lasso, Modelo 1 y Modelo 2

En la siguiente tabla 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 lineal Modelo 1 y Modelo 2

Modelo	*MSE*
Ridge	0.0135971
Lasso	0.0117466
Modelo1	0.0137844
Modelo2	0.0161221

Table: Comparación MSE: Ridge , Lasso, Modelo 1 y Modelo 2

Fuente: Elaboración propia

Anális 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 análisis espectométrico que corresponde a los predictores:

$$V26, V27, V32, V40, V54, V92, V115, V118, V124, V134, V149, V150,$$

$$V159, V167V168, V171, V173, V174, V194, V243, V271, V272, V287$$

• Ridge: aproxima todos los coeficientes a cero, como se había mencionado, no selecciona predictores.

• Lasso: hace una seleccción de 67 predictores el cual produce un modelo más simple e interpretable

• Modelo 1: podemos notar que después de la selección de Lasso ajustamos otro modelo el cual nos da un total de 24 predictores de los 67 considerados anteriormente por Lasso.

• Modelo 2: Podemos descartarlo, esto debido el ANOVA realizado.

• Luego al hacer la comparación entre los 4 modelos anteriores del error de testeo *MSE* podemos notar que el menor *MSE* es la regularización Lasso, seguido regularización Ridge, posteriormente el Modelo 1.

Conclusiones

• Tomando en cuenta lo anterior y los errores de testeo *MSE*, para tener una mejor predicción del compuesto químico Oxido de Hierro recomendaríamos 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 recomendaríamos el Modelo 1.

• 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 más a la predicción de nuestro compuesto químico.

- Notamos que las técnicas estudiadas nos permiten perspectivas diferentes en cuanto al modelado de la variable estudiada en términos de los distintos predictores.

- 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.

Bibliografía

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

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

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

García S. Jasmin,
Aplicaciones del modelo LASSO bayesiano en finanzass,
(2011),


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


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)


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)

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)

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

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

¡Muchas Gracias!