UNIVERSIDAD
DE GRANADA

Facultad de Ciencias

Grado en Matemáticas

TRABAJO DE FIN DE GRADO

Modelos de Regresión Lineal en un contexto de Big Data

Presentado por:
José Javier Miñano Ramos

Curso académico 2023-2024

Modelos de Regresión Lineal en un contexto de Big Data

José Javier Miñano Ramos

José Javier Miñano Ramos *Modelos de Regresión Lineal en un contexto de Big Data*.
Trabajo de fin de Grado. Curso académico 2023-2024.

Responsable de tutorización

María Dolores Martínez Miranda
Estadística e Investigación Operativa

Grado en Matemáticas
Facultad de Ciencias
Universidad de Granada

Declaración de originalidad

D. José Javier Miñano Ramos

Declaro explícitamente que el trabajo presentado como Trabajo de Fin de Grado (TFG), correspondiente al curso académico 2023-2024, es original, entendido esto en el sentido de que no he utilizado para la elaboración del trabajo fuentes sin citarlas debidamente.

En Granada a 28 de noviembre de 2023

Fdo: José Javier Miñano Ramos

Índice general

Summary v

Introducción vii

Notación ix

1 Métodos de regresión lineal multivariante 1
- 1.1 Breve introducción a los modelos de regresión . . . 1
- 1.2 El modelo de regresión lineal múltiple . . . 2
- 1.3 El modelo lineal de Gauss-Markov . . . 3
- 1.4 Estimación de los coeficientes del modelo . . . 4
  - 1.4.1 Estimación por mínimos cuadrados . . . 4
  - 1.4.2 Estimación por máxima verosimilitud . . . 7
- 1.5 Varianza residual . . . 8
- 1.6 Distribución de los estimadores . . . 10
  - 1.6.1 Distribución de los estimadores de los coeficientes $\hat{\beta}_1, ..., \hat{\beta}_m$ . . . 10
  - 1.6.2 Distribución de la varianza residual . . . 11
- 1.7 Inferencia sobre los parámetros del modelo . . . 12
- 1.8 Descomposición de la variabilidad . . . 13
- 1.9 Bondad de ajuste del modelo . . . 14
- 1.10 Problemas debido a la alta dimensionalidad . . . 14
  - 1.10.1 Inconsistencias en la estimación del modelo de regresión lineal . . . 15
  - 1.10.2 Colinealidad . . . 15
  - 1.10.3 Reducción de dimensiones: Métodos de selección de variables independientes . . . 16
- 1.11 Selección de covariables . . . 16
  - 1.11.1 Método de eliminación hacia atrás . . . 17
  - 1.11.2 Método de introducción progresiva . . . 17
  - 1.11.3 Método *stepwise* . . . 18
- 1.12 Criterios para la selección de modelos . . . 18
  - 1.12.1 Criterios basados en la bondad del ajuste . . . 18
  - 1.12.2 Criterio AIC . . . 19
  - 1.12.3 Criterio BIC . . . 19

2 Métodos de regularización 21
- 2.1 Introducción a los métodos de regularización . . . 21
- 2.2 Bias-Variance Trade-off . . . 22
- 2.3 Regresión Ridge . . . 23
  - 2.3.1 Coeficientes de la regresión Ridge . . . 24
  - 2.3.2 Significado de λ . . . 25
- 2.4 Propiedades de la regresión Ridge . . . 25
  - 2.4.1 Sesgo de la regresión Ridge . . . 26

2.4.2 Varianza de la regresión Ridge . . . 26
2.4.3 Error cuadrático medio de la regresión Ridge . . . 28
2.4.4 El estimador Ridge no es invariante a la escala . . . 29
2.4.5 Cómo escoger el parámetro de penalización . . . 30
2.4.6 Regresión Ridge: Conclusiones . . . 31
2.5 Regresión Lasso . . . 31
2.5.1 Unicidad del estimador Lasso . . . 33
2.5.2 Variantes de la regresión Lasso . . . 34
2.5.3 Regresión Lasso: Conclusiones . . . 35
2.6 Elastic net . . . 36
2.6.1 Naïve elastic net . . . 36
2.6.2 Elastic net . . . 39
2.6.3 Deficiencias del método naïve elastic net . . . 39
2.6.4 El estimador elastic net . . . 39

3 Aplicación con datos reales **41**
3.1 Breve descripción del conjunto de datos . . . 41
3.2 Metodología del análisis . . . 42
3.3 Análisis exploratorio . . . 42
3.4 Separar los datos en datos de entrenamiento y de testeo . . . 44
3.5 Entrenamiento de modelos . . . 44
3.5.1 Modelo de regresión lineal simple . . . 44
3.5.2 Modelo de regresión lineal múltiple . . . 45
3.5.3 Método de selección de variables *stepwise* . . . 46
3.5.4 Regresión Ridge . . . 47
3.5.5 Regresión Lasso . . . 49

Bibliografía **53**

Summary

In the era of big data and machine learning, the ability to draw conclusions from large data sets is key in different areas of science and business. However, this task is not necessarily easy because as the number of predictors increases, the difficulty of the problem increases too.

One way to deal with all this information would be through the use of simple models such as the multiple linear regression model. In this paper, we will study this model in depth and in particular the linear Gauss-Markov model. We will deal with the estimation of all the parameters that define the model such as the coefficients and the residual variance. Subsequently, we will study the variability decomposition and the goodness of fit of the model according to different criteria such as the coefficient of determination and the coefficient of determination adjusted by the degrees of freedom.

However, even the linear Gauss-Markov model, which we consider simple, has difficulties in its explainability and interpretability when the number of predictors is large. Particularly when the number of predictors exceeds the sample size (a common situation, for example when studying rare diseases) the linear regression model relating predictors to responses will be poorly conditioned or may not even exist. Therefore, we will introduce two of the problems associated with this situation, namely inconsistency in the estimation of the model and collinearity.

To solve the problems derived from high dimensionality such as collinearity and multicollinearity, several variable selection techniques have been developed such as forward selection, backward elimination or stepwise methods. With the use of these techniques we generate from the same set of predictors different models that offer us the possibility of giving more emphasis to simplicity and interpretability or accuracy in prediction. We will review different model selection criteria such as goodness-of-fit criteria, the Akaike information criterion or the Bayesian information criterion.

In addition to variable selection techniques, regularisation methods have been developed, among which we can highlight the Elastic Net method and its particular cases: the Ridge and Lasso methods. The objective of these techniques is not only to select predictors but also to change the value of the coefficients of the predictors in order to give more "importance" to those variables that have a greater influence on the response. That is, it is a continuous variable selection, as opposed to the dichotomous variable selection of forward selection, backward elimination and stepwise methods where variables can only be selected or discarded.

We will study in depth the theoretical properties of these regularisation methods and how each of them deals with the balance between bias and variance to obtain the best possible model.

Finally, we will apply the methods we have studied throughout the paper to a real dataset and compare their performance.

Introducción

La cantidad de datos generados en el mundo se ha multiplicado en la última década por treinta. De hecho, existen estimaciones que indican que, de media, cada habitante del planeta genera unos 2 MB de datos por segundo cada día. Esta cantidad de información generándose continuamente plantea, al menos, dos retos: su almacenamiento y su análisis. Ambos problemas están íntimamente relacionados pues una gran parte de los datos generados se desechan automáticamente, entre otras cosas, por la imposibilidad de analizarlos a la velocidad necesaria.

Toda esta información permite alimentar distintos algoritmos de aprendizaje estadístico que aúnan áreas como la estadística, el álgebra o la computación para desarrollar tareas tan diversas como la predicción de series temporales, la simulación de ensayos clínicos muy costosos o incluso el modelado del lenguaje natural y la visión automática.

En este trabajo, estudiaremos como los modelos de regresión lineal trabajan en un contexto de *Big Data*. Veremos cuáles son sus principales propiedades y sus limitaciones frente a esta situación, ya sean limitaciones en su capacidad predictiva o en su interpretabilidad.

En particular, nos centraremos en los problemas que tienen los modelos de regresión lineal cuando el número de variables independientes supera al número de observaciones sobre las que pretendemos definir un modelo. Estudiaremos distintos enfoques desarrollados desde finales del siglo XX hasta hoy, como la regresión Ridge, la regresión Lasso o el método Elastic Net, entre otros. Finalmente, aplicaremos estas técnicas a datos reales y las compararemos entre sí.

A lo largo del trabajo, las principales fuentes consultadas han sido el libro *The Elements of Statistical Learning* ([TH09]), los apuntes de la asignatura *Predictive Modeling* del Máster en *Big Data Analytics* de la Universidad Carlos III de Madrid ([GP23]) y el artículo *Regularization and variable selection via the elastic net* ([HZ05]).

Notación

A lo largo del documento notaremos con letras mayúsculas las variables aleatorias: X,Y, etc. Denotaremos en minúsculas las observaciones de esas variables aleatorias.

Las variables serán 'explicativas' $(X_1, ..., X_n)$ o 'explicadas' (Y). Aunque también podremos usar otras denominaciones como 'independientes' / 'dependientes' o la terminología típicamente usada en machine learning como 'predictor' y 'respuesta'.

Usaremos letras resaltadas en negrita para notar vectores $\mathbf{X} = (X_1, ..., X_n)$ y matrices ($\mathbf{A}$). La matriz traspuesta de $\mathbf{A}$ se denotará como $\mathbf{A^t}$.

1 Métodos de regresión lineal multivariante

En este capítulo, realizaremos una breve descripción de los métodos de regresión lineal multivariante, una herramienta esencial en estadística para modelar la relación linear entre las variables independientes, o predictores y las variables dependientes, o respuestas. Comenzamos con una introducción a los modelos de regresión en la sección 1.1 y presentamos el modelo de regresión lineal múltiple y en particular el modelo de Gauss-Markov en las secciones 1.2 y 1.3, respectivamente. En las secciones 1.4 y 1.5 estimamos los parámetros clave del modelo que son los coeficientes y la varianza residual. En la sección 1.6 estudiaremos la distribución de los estimadores previamente mencionados y en la sección 1.7 realizaremos inferencia sobre los parámetros del modelo. La descomposición de la variabilidad y la bondad de ajuste del modelo los trataremos en las seccon 1.8 y 1.9. En la sección 1.10 veremos los problemas que presentan los modelos de regresión lineal múltiple en situaciones de alta dimensionalidad y la necesidad de buscar nuevos modelos que superen esas dificultades. En la sección 1.11 propondremos las primeras alternativas para conseguir mejores modelos en alta dimensionalidad: los métodos de selección de variables *backward elimination, forward selection* y *stepwise*. Finalmente, en la sección 1.12 estudiaremos posibles criterios de selección de modelos.

1.1. Breve introducción a los modelos de regresión

En un modelo de regresión, la variable de interés o respuesta $\mathbf{Y}$ se explica en términos de $m \geq 1$ variables independientes $\mathbf{X} = (X_1, ..., X_m)^t$. Para ello, admitimos como cierto que $\mathbf{Y}$ se relaciona con $\mathbf{X}$ a partir de una función de regresión $m(\cdot)$, que es, en general, desconocida. Así, obtenemos el siguiente modelo de regresión:

$$\mathbf{Y} = m(\mathbf{X}) + \boldsymbol{\varepsilon}$$

donde $\boldsymbol{\varepsilon}$ es el error del modelo que se suele asumir como condicionalmente independiente de $\mathbf{X}$ en términos de $m(\cdot)$.

A la hora de elegir la forma de $m(\cdot)$, existen dos posibles opciones: asumir que el modelo sigue alguna estructura, como por ejemplo lineal ($m(\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$) o aditiva ($m(\mathbf{X}) = \sum_{i=1}^{m} F_i(X_i)$), o elegir técnicas no parámetricas sin ninguna hipótesis sobre la forma del modelo. La primera opción es más restrictiva, pero nos permite medir el impacto de cada predictor en la respuesta. Como punto en contra, asumir que el modelo sigue alguna estructura preliminar implica limitaciones en el caso $m > n$, siendo n el número de observaciones de $\mathbf{Y}$, relacionadas con la estimación de los parámetros que definen dicha estructura.

El modelo lineal corresponde con la estructura $m(\mathbf{X}) = \mathbf{X}^t\boldsymbol{\beta}$ con $\boldsymbol{\beta} \in \mathbb{R}^{m+1}$. Esta es una formulación parámetrica que asume una estructura lineal en $m(\cdot)$. Por tanto, la estimación del modelo equivale a la estimación de los coeficientes $\boldsymbol{\beta}$.

1.2. El modelo de regresión lineal múltiple

Dadas n observaciones independientes de la variable Y, $\{Y_1, ..., Y_n\}$ diremos que siguen un modelo lineal si:

$$Y_i = X_{i1}\beta_1 + X_{i2}\beta_2 + ... + X_{im}\beta_m + \varepsilon_i, \quad i = 1, ..., n$$

El modelo lineal en las observaciones se representa matricialmente como:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} X_{11} & X_{12} & ... & X_{1m} \\ X_{21} & X_{22} & ... & X_{2m} \\ ... & ... & ... & ... \\ X_{n1} & X_{n2} & ... & X_{nm} \end{pmatrix} \cdot \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Que se escribe de forma resumida

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Los elementos del modelo lineal en las observaciones son:

- El vector de observaciones $\mathbf{Y} = (Y_1, ..., Y_n)^t$.
- La matriz del modelo
$$\mathbf{X} = \begin{pmatrix} X_{11} & X_{12} & ... & X_{1m} \\ X_{21} & X_{22} & ... & X_{2m} \\ ... & ... & ... & ... \\ X_{n1} & X_{n2} & ... & X_{nm} \end{pmatrix}$$
cuyos elementos son conocidos.
- El vector de parámetros $\boldsymbol{\beta} = (\beta_1, ..., \beta_m)$.
- El vector de desviaciones aleatorias $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_2)^t$, donde
$$\varepsilon_i = Y_i - X_{i1}\beta_1 - X_{i2}\beta_2 - ... - X_{im}\beta_m$$
es la desviación aleatoria de Y_i.

Definición 1.1. Suponiendo m variables explicativas para describir la variable de respuesta, el modelo de regresión lineal múltiple se formula como:

$$Y_i = \beta_0 + \beta_1 X_{i1} + ... + \beta_m X_{im} + \varepsilon_i, \quad i = 1, ..., n$$

siendo X_{ij} las observaciones de la j-ésima variable explicativa y β_j el efecto marginal de esa variable sobre la variable de respuesta. Usando la notación matricial previa, se puede escribir el modelo de regresión lineal como múltiple como

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

donde:

- $\mathbf{Y} = (Y_1, ..., Y_n)^t$ es el vector de observaciones de la variable de respuesta.

- $\mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{12} & ... & X_{1m} \\ 1 & X_{21} & X_{22} & ... & X_{2m} \\ ... & ... & ... & ... & \\ 1 & X_{n1} & X_{n2} & ... & X_{nm} \end{pmatrix}$ es la matriz de regresión de dimensión $n \times (m+1)$.
- $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_m)$ es el vector de coeficientes de dimensión $m+1$, también llamado vector de efectos.
- $\boldsymbol{\varepsilon} = (\varepsilon_1, ..., \varepsilon_2)^t$ es el vector de errores del modelo que representan las desviaciones aleatorias de Y_i respecto de la función lineal:

$$\varepsilon_i = Y_i - \beta_0 - \sum_{j=1}^{m} \beta_j X_{ij}$$

Una muestra de $(X_1, ..., X_m, Y)$ se denota como $\{(X_{i1}, ..., X_{im}, Y_i)\}_{i=1}^n$ siendo X_{ij} es la i-ésima observación de la variable explicativa X_j. Denotamos como $\mathbf{X}_i := \left(X_{i1}, ..., X_{ip}\right)$ a la i-ésima observación de $(X_1, ..., X_m)$. Así, la notación de dicha muestra se simplifica en $\{(\mathbf{X}_i, Y_i)\}_{i=1}^n$.

1.3. El modelo lineal de Gauss-Markov

Dentro de los modelos lineales, en este primer capítulo nos vamos a centrar en el modelo de Gauss-Markov. Este es un modelo lineal de efectos ($\boldsymbol{\beta}$) fijos en el que se consideran las siguientes tres hipótesis:

1. El vector de errores está centrado ($\mathbb{E}[\varepsilon_i] = 0, \forall i = 1...n$), o equivalentemente

$$\mathbb{E}[Y_i] = \beta_0 + \beta_1 X_{i1} + ... + \beta_m X_{im}$$

2. $Var(\varepsilon_i) = \mathbb{E}[\varepsilon_i^2] = \sigma^2, \forall i = 1, ..., n$, o equivalentemente $Var(Y_i) = \sigma^2$.
3. $Cov\left(\varepsilon_i, \varepsilon_j\right) = \mathbb{E}\left[\varepsilon_i, \varepsilon_j\right] = 0, \forall i \neq j \quad i, j \in \{1, ..., n\}$, o equivalentemente $Cov(Y_i, Y_j) = 0$.

Las tres condiciones impuestas al vector $\boldsymbol{\varepsilon}$ se pueden resumir en las dos siguientes: $\mathbb{E}[\boldsymbol{\varepsilon}] = 0$ y $\mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^t] = \sigma^2 \mathbf{I_n}$.

Las tres condiciones anteriores son las hipótesis básicas que impondremos a un modelo lineal y al que denominaremos **modelo lineal de Gauss-Markov**. Si además suponemos que los errores se distribuyen normalmente, el modelo se denomina **modelo lineal normal de Gauss-Markov** y podemos resumir estas hipótesis como:

$$\boldsymbol{\varepsilon} \sim N_n\left(\mathbf{0}, \sigma^2 \mathbf{I_n}\right)$$

o equivalentemente

$$\mathbf{Y} \sim N_n\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{I_n}\right)$$

La normalidad es una hipótesis adicional a las hipótesis de Gauss-Markov. Asumir la normalidad es importante para ciertos resultados teóricos que veremos más adelante como la estimación por máxima verosimilitud de los coeficientes $\boldsymbol{\beta}$ o para la inferencia sobre los

parámetros del modelo. Sin embargo, no es una hipótesis necesaria para la estimación por mínimos cuadrados o para los métodos de regularización.

En este primer capítulo asumiremos que el número de observaciones n es mayor que el número de variables m. También establecemos que $\mathbf{X}$ es invertible, o dicho de otra manera, que ninguna de las variables explicativas es una combinación lineal de las demás.

1.4. Estimación de los coeficientes del modelo

Esta sección se basa principalmente en el punto 2.2.3 de la fuente [GP23]

Los coeficientes teóricos del modelo $\beta_0, \beta_1, ..., \beta_m$ son no observables, por lo tanto debemos obtener estimadores $\hat{\beta}_0, \hat{\beta}_1, ..., \hat{\beta}_m$ que aproximen de la mejor forma posible a estos coeficientes teóricos. Sin embargo, antes de estudiar los estimadores de los coeficientes del modelo, conviene que entendamos el significado de tales coeficientes. Como ya hemos introducido, el modelo lineal múltiple representa la relación lineal entre varias variables explicativas y la variable explicada.

$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_m X_m + \varepsilon$$

Donde β_0 es la ordenada del modelo, $\beta_1, ..., \beta_m$ son las pendientes respecto de cada variable y ε cumple las hipótesis enumeradas previamente. Por tanto

$$\mathbb{E}\left[Y | X_1 = x_1, ..., X_m = x_m\right] = \beta_0 + \beta_1 x_1 + ... + \beta_m x_m$$

pues

$$\mathbb{E}\left[\varepsilon | X_1 = x_1, ..., X_m = x_m\right] = 0$$

Equivalentemente, podemos interpretar los coeficientes del modelo de la siguiente forma:

- β_0 es la media de Y cuando $X_1 = X_2 = ... = X_m = 0$
- $\beta_i, 1 \leq i \leq m$ es el incremento en la media de Y por cada incremento unitario en $X_i = x_i$, suponiendo que el resto de las variables explicativas se mantengan constantes.

Geométricamente, los coeficientes definen un hiperplano en el espacio vectorial $\mathbb{R}^{m+1}$. Por tanto, si $m = 1$ el hiperplano es una recta y si $m = 2$ es un plano que puede ser representado en un espacio vectorial tridimensional.

Ahora vamos a estimar los coeficientes que determinan el modelo, para ello, vamos a desarrollar dos métodos de estimación. El primer método es a través de la minimización de la función *Suma de Residuos Cuadráticos (SRC)* y el segundo apoyándonos en la *función de máxima verosimilitud*.

1.4.1. Estimación por mínimos cuadrados

Para esta subsección, nos apoyamos en el punto 7.3 de la fuente [ACR07] y en la fuente [Wika].

Definimos la función $SRC(\boldsymbol{\beta})$ como:

$$SRC(\beta_0, \beta_1, ..., \beta_m) := \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 =$$

$$= \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{m} X_{ij}\beta_j \right)^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - ... - \beta_m X_{im})^2 =$$

$$= (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^t (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{Y}^t\mathbf{Y} - \mathbf{Y}^t\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^t\mathbf{X}^t\mathbf{Y} + \boldsymbol{\beta}^t\mathbf{X}^t\mathbf{X}\boldsymbol{\beta}$$

La función *SRC* suma las distancias verticales al cuadrado de los puntos de la muestra al hiperplano de regresión definido por el vector de coeficientes $\boldsymbol{\beta}$. Por lo tanto, pretendemos encontrar un estimador $\hat{\boldsymbol{\beta}}$ tal que:

$$\hat{\boldsymbol{\beta}} := \underset{\boldsymbol{\beta} \in \mathbb{R}^{m+1}}{\arg\min} SRC(\boldsymbol{\beta})$$

Teorema 1.1. *Si* $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ *donde* $\mathbf{X}$ *es una matriz de dimensión* $n \times (m+1)$ *de rango* $m+1 < n$*, entonces el vector* $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, ..., \hat{\beta}_m)$ *que minimiza la función SRC es*

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{Y}$$

Demostración
Deduzcamos explítamente estos estimadores.

$$SRC(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^t(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{Y}^t\mathbf{Y} - \mathbf{Y}^t\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^t\mathbf{X}^t\mathbf{Y} + \boldsymbol{\beta}^t\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} \Rightarrow$$

$$\Rightarrow \frac{\delta SRC}{\delta}(\boldsymbol{\beta}) = -2\mathbf{X}^t\mathbf{Y} + 2\mathbf{X}^t\mathbf{X}\boldsymbol{\beta}$$

Igualamos a cero la derivada y obtenemos las *ecuaciones normales*

$$\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^t\mathbf{Y}$$

Resolvemos para $\boldsymbol{\beta}$ y obtenemos:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\mathbf{Y}$$

Siempre y cuando $\mathbf{X}^t\mathbf{X}$ sea invertible. □

Introduzcamos ahora un resultado teórico clave sobre los estimadores de mínimos cuadrados:

Teorema 1.2. ***(de Gauss-Markov)*** *Bajo las hipótesis de Gauss-Markov, los estimadores de mínimos cuadrados ordinarios* $\hat{\beta}_i, i = 0, 1, ..., m$ *tienen mínima varianza entre los estimadores lineales insesgados.*

Demostración
Empecemos viendo que el estimador es lineal:
En efecto, sea $\mathbf{C} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m1} & c_{m2} & \cdots & c_{mn} \end{pmatrix}$

se tiene que $\hat{\boldsymbol{\beta}}$ se expresa como combinación lineal del vector $\mathbf{Y}$:

$$\hat{\boldsymbol{\beta}} = \mathbf{CY} = \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ c_{m1} & c_{m2} & \cdots & c_{mn} \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} c_{11}Y_1 + c_{12}Y_2 + ... + c_{1n}Y_n \\ c_{21}Y_1 + c_{22}Y_2 + ... + c_{2n}Y_n \\ \vdots \\ c_{m1}Y_1 + c_{m2}Y_2 + ... + c_{mn}Y_n \end{pmatrix}$$

Ahora veamos que el estimador $\hat{\boldsymbol{\beta}}$ de $\boldsymbol{\beta}$ es insesgado:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t(\boldsymbol{\beta} + \boldsymbol{\varepsilon}) =$$
$$= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\boldsymbol{\beta} + (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\boldsymbol{\varepsilon} = \boldsymbol{\beta} + (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\boldsymbol{\varepsilon} \Rightarrow$$
$$\Rightarrow \hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\boldsymbol{\varepsilon}$$

Aplicando la linealidad de la esperanza matemática:

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] = \mathbb{E}[\boldsymbol{\beta} + (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\boldsymbol{\varepsilon}] = \mathbb{E}[\boldsymbol{\beta}] + \mathbb{E}[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\boldsymbol{\varepsilon}]$$
$$= \mathbb{E}[\boldsymbol{\beta}] + (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbb{E}[\boldsymbol{\varepsilon}]$$

Y finalmente, como $\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$:

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] = \mathbb{E}[\boldsymbol{\beta}] = \boldsymbol{\beta}$$

Por último, veamos que tiene mínima varianza en la clase de estimadores lineales e insesgados. Calculemos $Var(\hat{\boldsymbol{\beta}})$.

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \mathbb{E}[(\hat{\boldsymbol{\beta}} - \mathbb{E}[\hat{\boldsymbol{\beta}}])(\hat{\boldsymbol{\beta}} - \mathbb{E}[\hat{\boldsymbol{\beta}}^t] = \mathbb{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^t] =$$
$$= \mathbb{E}[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\boldsymbol{\varepsilon} \cdot \boldsymbol{\varepsilon}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}] = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbb{E}[\boldsymbol{\varepsilon} \cdot \boldsymbol{\varepsilon}^t]\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1} =$$
$$= \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1} = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}$$

Siendo σ^2 la varianza del error. Donde se ha usado que $\hat{\boldsymbol{\beta}}$ es insesgado, $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\boldsymbol{\varepsilon}$ y que $\text{Var}(\boldsymbol{\varepsilon}) = \mathbb{E}[\boldsymbol{\varepsilon} \cdot \boldsymbol{\varepsilon}^t] = \sigma^2 \cdot \mathbf{I_n}$.

Ahora, tomemos un estimador lineal e insesgado cualquiera $\tilde{\mathbf{a}}$ y comprobemos que $Var(\hat{\boldsymbol{\beta}}) < Var(\tilde{\mathbf{a}})$. Entonces, sea $\tilde{\mathbf{a}} = \mathbf{CY}$ con $\mathbf{C} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t + \mathbf{D}$ siendo $\mathbf{D}$ una matriz de dimensión $m \times n$ no nula. Calculemos:

$$\tilde{\mathbf{a}} = \mathbb{E}[\mathbf{CY}] = \mathbb{E}[((\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t + \mathbf{D})(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})] =$$
$$= \mathbb{E}[((\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t + \mathbf{D})(\mathbf{X}\boldsymbol{\beta}) + ((\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t + \mathbf{D})\boldsymbol{\varepsilon}] =$$
$$= ((\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t + \mathbf{D})(\mathbf{X}\boldsymbol{\beta}) + ((\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t + \mathbf{D})\mathbb{E}[\boldsymbol{\varepsilon}] =$$
$$= ((\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t + \mathbf{D})(\mathbf{X}\boldsymbol{\beta}) \quad \text{(puesto que } \mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}\text{)}$$
$$= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} + \mathbf{DX}\boldsymbol{\beta} = (\mathbf{I_n} + \mathbf{DX})\boldsymbol{\beta}$$

Por tanto, $\tilde{\mathbf{a}}$ es insesgado si, y solo sí, $\mathbf{DX} = \mathbf{0}$. Entonces:

$$\text{Var}(\tilde{\mathbf{a}}) = \text{Var}(\mathbf{CY}) = \mathbf{C}\text{Var}(\mathbf{Y})\mathbf{C}^t = \sigma^2\mathbf{CC}^t =$$

$$= \sigma^2((\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t + \mathbf{D})((\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t + \mathbf{D}^t) =$$

$$= \sigma^2\left((\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1} + (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{D}^t + \mathbf{DX}(\mathbf{X}^t\mathbf{X})^{-1} + \mathbf{DD}^t\right) =$$

$$= \sigma^2\left(\mathbf{X}^t\mathbf{X}\right)^{-1} + \sigma^2\left(\mathbf{X}^t\mathbf{X}\right)^{-1}(\mathbf{DX})^t + \sigma^2\mathbf{DX}\left(\mathbf{X}^t\mathbf{X}\right)^{-1} + \sigma^2\mathbf{DD}^t$$

$$= \sigma^2\left(\mathbf{X}^t\mathbf{X}\right)^{-1} + \sigma^2\mathbf{DD}^t \quad \text{(puesto que } \mathbf{DX} = \mathbf{0}) =$$

$$= \text{Var}(\hat{\boldsymbol{\beta}}) + \sigma^2\mathbf{DD}^t \quad \text{(puesto que } \sigma^2\left(\mathbf{X}^t\mathbf{X}\right)^{-1} = \text{Var}(\hat{\boldsymbol{\beta}}))$$

Como $\mathbf{DD}^t$ es una matriz semidefinida positiva, $\text{Var}(\tilde{\mathbf{a}}) > \text{Var}(\hat{\boldsymbol{\beta}})$ probando el teorema. □

1.4.2. Estimación por máxima verosimilitud

Por otra parte, podemos también calcular los coeficientes del modelo usando la función de verosimilitud. Para ello, nos valdremos de la hipótesis de normalidad del modelo:

$$\mathbf{Y} \sim N_n\left(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I_n}\right)$$

Por ende, $Y_1, Y_2, ..., Y_n$ son variables aleatorias, independientes e idénticamente distribuidas con distribución común $N(\beta_0 + \beta_1 X_{i1} + ... + \beta_m X_{im}, \sigma^2)$. Por tanto, la función de densidad de Y_i, $\quad i = 1, 2..., n$ es

$$\varphi_{\beta_0+\beta_1 X_{i1}+...+\beta_m X_{im},\sigma^2}(Y_i) = \frac{1}{\sqrt{2\pi\sigma^2}}\text{exp}\left(-\frac{(Y_i - \beta_0 + \beta_1 X_{i1} + ... + \beta_m X_{im})^2}{2\sigma^2}\right)$$

Y por tanto la estimación por máxima verosimilitud se obtiene de maximizar la función de verosimilitud:

$$L\left(\mathbf{Y}; \boldsymbol{\beta}, \sigma^2\right) = (2\pi\sigma^2)^{-n/2}\text{exp}\left(-\sum_{i=1}^{n}\frac{(Y_i - \beta_0 - \beta_1 x_{i1} - ... - \beta_m x_{im})^2}{2\sigma^2}\right) =$$

$$= (2\pi\sigma^2)^{-n/2}\text{exp}\left(-\frac{(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^t(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2}\right)$$

Que equivale a minimizar la función g:

$$g(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^t(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$$

Ahora, repitiendo el mismo proceso, derivamos respecto a $\boldsymbol{\beta}$ e igualamos a cero para obtener el estimador de $\boldsymbol{\beta}$:

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

Llegando al mismo resultado que obtuvimos en la estimación por mínimos cuadrados. Es decir, el estimador de mínimos cuadrados es también el estimador de máxima verosimilitud

cuando se cumplen las hipótesis de Gauss-Markov.

Una vez que hemos obtenido la expresión de $\hat{\boldsymbol{\beta}}$ con ambos métodos, podemos definir $\hat{Y}_1, ..., \hat{Y}_n$ y $\mathbf{e}$ como:

- Los valores ajustados $\hat{Y}_1, ..., \hat{Y}_n$:

  $$\hat{Y}_i := \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + ... + \hat{\beta}_m X_{im}$$

  son las proyecciones verticales de $Y_1, ..., Y_n$ en el plano. Podemos expresarlas matricialmente como:

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

  siendo $\mathbf{H} := \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t$ la matriz que proyecta $\mathbf{Y}$ en el hiperplano de regresión. A esta matriz la llamaremos *matriz hat*.

- El vector de residuos $\mathbf{e}$

  $$\mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} - \begin{pmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{pmatrix}$$

  Representa la distancia entre los datos reales y los ajustados en cada coordenada.

1.5. Varianza residual

Esta sección se ha escrito teniendo en cuenta la fuente [Jor22].

Ahora pretendemos estimar σ^2. Esto nos dará una medida sobre la bondad del ajuste y será usado posteriormente para estudiar la inferencia sobre el modelo.

El método de los mínimos cuadrados no produce una función de los valores Y y X en la muestra que podamos minimizar para obtener un estimador de σ^2. Vamos a intentar obtener este estimador usando la función de verosimilitud:

Teorema 1.3. *Considérese el modelo de regresión lineal con $\mathbf{X}$ conocida, $\boldsymbol{\beta}$ desconocido y σ^2 desconocida:*

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I_n})$$

Entonces:

1. *El estimador de máxima verosimilitud de σ^2 es:*

   $$\hat{\sigma}^2 = \frac{1}{n}\left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)^t\left(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) = \frac{\mathbf{e}^t\mathbf{e}}{n}$$

   donde

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

2. *$\hat{\sigma}^2$ es un estimador sesgado de σ^2:*

   $$\mathbb{E}[\hat{\sigma}^2] \neq \sigma^2$$

De hecho

$$\mathbb{E}[\hat{\sigma}^2] = \frac{n-m-1}{n}\sigma^2$$

Demostración

1.
Usando la función de verosimilitud:

$$L\left(\mathbf{Y};\boldsymbol{\beta},\sigma^2\right) = (2\pi\sigma^2)^{-n/2}\exp\left(-\frac{(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^t(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})}{2\sigma^2}\right) =$$

$$= \left(2\pi\sigma^2\right)^{-n/2}\exp\left(-\frac{\boldsymbol{\varepsilon}^t\boldsymbol{\varepsilon}}{2\sigma^2}\right)$$

En este caso derivamos respecto a σ^2, igualamos a cero y sustituimos $\boldsymbol{\beta}$ por $\hat{\boldsymbol{\beta}}$

$$\frac{\delta L\left(\mathbf{Y};\boldsymbol{\beta},\sigma^2\right)}{\delta\sigma} = 0 \Rightarrow$$

$$\Rightarrow \frac{-n}{2}\left(2\pi\sigma^2\right)^{\frac{-n}{2}-1}\cdot 4\pi\sigma\exp\left(\frac{-\boldsymbol{\varepsilon}^t\boldsymbol{\varepsilon}}{2\sigma^2}\right) + \left(2\pi\sigma^2\right)^{\frac{-n}{2}}\exp\left(-\frac{\boldsymbol{\varepsilon}^t\boldsymbol{\varepsilon}}{2\sigma^2}\right)\left(\frac{4\sigma\boldsymbol{\varepsilon}^t\boldsymbol{\varepsilon}}{4\sigma^4}\right) = 0$$

$$\Rightarrow -2n\pi\sigma\left(2\pi\sigma^2\right)^{\frac{-n}{2}-1} + \left(2\pi\sigma^2\right)^{\frac{-n}{2}}\left(\frac{\sigma\boldsymbol{\varepsilon}^t\boldsymbol{\varepsilon}}{\sigma^4}\right) = 0 \Rightarrow \frac{-2n\pi\sigma}{2\pi\sigma^2} + \frac{\boldsymbol{\varepsilon}^t\boldsymbol{\varepsilon}}{\sigma^3} = 0$$

$$\Rightarrow \hat{\sigma}^2 = \frac{\mathbf{e}^t\mathbf{e}}{n} = \frac{1}{n}\left(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)^t\left(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}}\right)$$

2.

Sabemos que la suma de residuos cuadráticos dividida entre σ^2 sigue una distribución chi-cuadrado:

$$\frac{\mathbf{e}^t\mathbf{e}}{\sigma^2} \sim \chi^2_{n-m-1}$$

Por tanto

$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-m-1}$$

Ahora, usamos la relación entre la distribución chi cuadrado y la distribución gamma:

$$\mathbf{X} \sim \chi^2_k \Rightarrow c\mathbf{X} \sim \text{Gam}\left(\frac{k}{2}, \frac{1}{2c}\right)$$

De donde deducimos que

$$\hat{\sigma}^2 = \frac{\sigma^2}{n}\frac{n\hat{\sigma}^2}{\sigma^2} \sim \text{Gam}\left(\frac{n-m-1}{2}, \frac{n}{2\sigma^2}\right)$$

Usando la esperanza de la distribución gamma:

$$\mathbf{X} \sim \text{Gam}\,(a,b) \Rightarrow \mathbb{E}\,(\mathbf{X}) = \frac{a}{b}$$

Obtenemos finalmente:

$$\mathbb{E}[\hat{\sigma}^2] = \frac{\frac{n-m-1}{2}}{\frac{n}{2\sigma^2}} = \frac{n-m-1}{n}\sigma^2$$

□

Por tanto, necesitamos encontrar un estimador insesgado de σ^2. Para ello, definimos la varianza residual:

Definición 1.2. Se define la varianza residual como el cociente entre la suma de cuadrados de residuos y el número de residuos linealmente independientes.

$$\hat{\sigma}_R^2 = \frac{\mathbf{e}^t\mathbf{e}}{n-m-1} = \frac{\sum_{i=1}^n e_i^2}{n-m-1}$$

El número de residuos linealmente independientes viene determinado por el rango de la matriz $\mathbf{X}$.

Proposición 1.1. *La varianza residual $\hat{\sigma}_R^2$ es un estimador insesgado de σ^2.*

Demostración

Aplicamos que la esperanza de una distribución χ^2 es igual a sus grados de libertad

$$\mathbb{E}\left[\frac{(n-m-1)\,\hat{\sigma}_R^2}{\sigma^2}\right] = n-m-1 \Rightarrow \mathbb{E}\left[\frac{(n-m-1)\,\hat{\sigma}_R^2}{(n-m-1)}\right] = \sigma^2 \Rightarrow$$

$$\Rightarrow \mathbb{E}\left[\hat{\sigma}_R^2\right] = \sigma^2$$

Aplicando la linealidad de la esperanza matemática, pues tanto $n-m-1$ como σ^2 son constantes. □

1.6. Distribución de los estimadores

1.6.1. Distribución de los estimadores de los coeficientes $\hat{\beta}_1, ..., \hat{\beta}_m$

Hemos visto que $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, es decir, los coeficientes $(\hat{\beta}_0, ..., \hat{\beta}_m)$ se escriben como combinaciones lineales de las variables explicativas las cuales se distribuyen normalmente. Por ende, cada uno de los estimadores $(\hat{\beta}_0, ..., \hat{\beta}_m)$ sigue una distribución normal. Así, para conocer unívocamente la distribución necesitamos calcular la media y la varianza:

$$\mathbb{E}\left[\hat{\boldsymbol{\beta}}\right] = \mathbb{E}\left[\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}\right] = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbb{E}\left[\mathbf{Y}\right] = \boldsymbol{\beta}$$

$$Var\left[\hat{\boldsymbol{\beta}}\right] = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'Var\left[\mathbf{Y}\right]\left(\left(\mathbf{X}\mathbf{X}'\right)^{-1}\mathbf{X}'\right) = \sigma^2\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

Por tanto

$$\hat{\boldsymbol{\beta}} = N_n(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$$

1.6.2. Distribución de la varianza residual

El problema con la distribución de los coeficientes recién obtenida es que σ^2 es desconocida en la práctica, por tanto, necesitamos estimarla. Previamente, hemos definido la varianza residual $\hat{\sigma}_R^2$ como:

$$\hat{\sigma}_R^2 := \frac{1}{n-m-1}\sum_{i=1}^{n} e_i^2$$

El denominador $n-m-1$ representa los grados de libertad: el número de observaciones menos el número de parámetros. Para interpretar $\hat{\sigma}_R^2$, es importante observar que la media de los residuos $e_1, ..., e_n$ es cero. Por tanto $\hat{\sigma}_R^2$ es una varianza muestral reescalada de los residuos que estima la varianza de $\boldsymbol{\varepsilon}$. Para obtener su distribución operamos matricialmente:

$$\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{HY} = (\mathbf{I_n} - \mathbf{H})\mathbf{Y}$$

Siendo $\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{HY}$ con $\mathbf{H}$ la matriz *hat* previamente mencionada. Como $\mathbf{I_n} - \mathbf{H}$ cumple que $(\mathbf{I_n} - \mathbf{H})^2 = (\mathbf{I_n} - \mathbf{H})$ y usando que

$$(\mathbf{I_n} - \mathbf{H})\,\mathbf{X}\boldsymbol{\beta} = \mathbf{0}$$

se tiene:

$$\mathbf{e}^t\mathbf{e} = \mathbf{Y}^t\,(\mathbf{I_n} - \mathbf{H})\,\mathbf{Y} = (\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})^t\,(\mathbf{I_n} - \mathbf{H})\,(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) = \boldsymbol{\varepsilon}^t\,(\mathbf{I_n} - \mathbf{H})\,\boldsymbol{\varepsilon}$$

Habiendo expresado $\mathbf{e}^t\mathbf{e}$ como una forma cuadrática $\boldsymbol{\varepsilon}^t\,(\mathbf{I_n} - \mathbf{H})\,\boldsymbol{\varepsilon}$, de variables distribuidas normalmente con media $\mathbf{0}$ y varianza σ^2, obtenemos:

$$\frac{\mathbf{e}^t\mathbf{e}}{\sigma^2} = \frac{\boldsymbol{\varepsilon}^t\,(\mathbf{I_n} - \mathbf{H})\,\boldsymbol{\varepsilon}}{\sigma^2} \sim \chi^2_{n-m-1}$$

Los grados de libertad se obtienen a partir del rango de la matriz $(\mathbf{I_n} - \mathbf{H})$:

$$Rango\,(\mathbf{I_n} - \mathbf{H}) = Traza\,(\mathbf{I_n} - \mathbf{H}) = n - Traza\,(\mathbf{H})$$

Sustituyendo en $\mathbf{H}$ tenemos que

$$Traza\,(\mathbf{H}) = Traza\left(\mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\mathbf{X}^t\right) = Traza\left(\mathbf{X}^t\mathbf{X}\left(\mathbf{X}^t\mathbf{X}\right)^{-1}\right) = m+1$$

Y finalmente, deducimos que

$$Rango\,(\mathbf{I_n} - \mathbf{H}) = n - Traza\,(\mathbf{H}) = n - m - 1$$

Con los cálculos anteriores se deduce que la distribución de la varianza residual viene dada por:

$$\frac{(n-m-1)\,\hat{\sigma}_R^2}{\sigma^2} = \frac{\mathbf{e}^t\mathbf{e}}{\sigma^2} \sim \chi^2_{n-m-1}$$

1.7. Inferencia sobre los parámetros del modelo

Para esta sección se ha tenido en cuenta el punto 2.4 de la fuente [GP23].

Las hipótesis y los resultados que hemos desarrollado previamente nos ayudarán ahora a aplicar inferencia sobre la distribución del vector aleatorio $\hat{\boldsymbol{\beta}}$. Como hemos visto, $\hat{\boldsymbol{\beta}} \sim N_n(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$. Particularmente, los coeficientes estimados $\hat{\beta}_j, j = 1, ..., m$ siguen la siguiente distribución:

$$\hat{\beta}_j \sim N_n\left(\beta_j, \sigma^2 P_{j+1,j+1}\right), \quad j = 1, ..., m$$

siendo P_{jj} el elemento j-ésimo de la diagonal de $P = (\mathbf{X}'\mathbf{X})^{-1}$. Tipificando obtenemos

$$\frac{\hat{\beta}_j - \beta_j}{\sigma\sqrt{P_{j+1,j+1}}} \sim N(0,1) \quad j = 1, ..., m$$

Ahora, usamos la distribución de la varianza residual que acabamos de obtener

$$\frac{(n-m-1)\hat{\sigma}_R^2}{\sigma^2} \sim \chi^2_{n-m-1}$$

Y aplicamos que la distribución t de Student con n grados de libertad se obtiene a partir de una distribución $X \sim N(0,1)$ y de otra distribución $Y \sim \chi^2_n$ tal que

$$T = \frac{X}{\sqrt{Y/n}}$$

Nos queda entonces que

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_R\sqrt{C_{j+1,j+1}}} \sim t_{n-m-1}$$

Para obtener los intervalos y los contrastes de hipótesis sobre los coeficientes del modelo, usamos este último resultado. Así, un intervalo de confianza de 1-α para β_j es:

$$(\hat{\beta}_j - t_{n-m-1;\alpha/2} \cdot std(\hat{\beta}_j), \hat{\beta}_j + t_{n-m-1;\alpha/2} \cdot std(\hat{\beta}_j))$$

siendo $std(\hat{\beta}_j) = \hat{\sigma}_R\sqrt{C_{j+1,j+1}}$ la desviación típica de $\hat{\beta}_j$.

Y definimos el problema de contraste:

$$\begin{cases} H_0 : \beta_j = \omega_j \\ H_1 : \beta_j \neq \omega_j \end{cases}$$

Donde ω_j es algún valor de interés en la práctica. El estadístico de contraste sería:

$$t = \frac{\hat{\beta}_j - \beta_j}{std(\hat{\beta}_j)}$$

que sigue la distribución t de Student t_{n-k-1} si H_0 es cierta. Así, rechazará la hipótesis nula al nivel de significación α cuando $|t| > t_{n-m-1;\alpha/2}$.

1.8. Descomposición de la variabilidad

En esta sección vamos a estudiar la varianza y su descomposición en varianza explicada y varianza no explicada. Esta descomposición se denomina ANOVA, del inglés *Analysis of Variance*. Sabemos que:

$$VT = VE + VNE \tag{1.1}$$

Donde:

- Varianza Total: $VT = \sum_{i=1}^{n}(Y_i - \bar{\mathbf{Y}})^2 = (\mathbf{Y} - \mathbf{I_n}\bar{\mathbf{Y}})^t(\mathbf{Y} - \mathbf{I_n}\bar{\mathbf{Y}})$
- Varianza Explicada: $VE = \sum_{i=1}^{n}(Y_i - \hat{\mathbf{Y}})^2 = (\mathbf{Y} - \mathbf{I_n}\hat{\mathbf{Y}})^t(\mathbf{Y} - \mathbf{I_n}\hat{\mathbf{Y}})$
- Varianza No Explicada: $VNE = \sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2 = (\mathbf{Y} - \hat{\mathbf{Y}})^t(\mathbf{Y} - \hat{\mathbf{Y}})$

Formulamos el contraste de regresión como:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \\ H_1 : \beta_j \neq 0 \quad \text{para algún j} \quad 1 \leq j \leq k \end{cases}$$

Donde la hipótesis nula indica que las variables independientes de forma conjunta no pueden explicar la respuesta. Podemos deducir que debemos rechazar la hipótesis nula si la Varianza Explicada es suficientemente mayor que la Varianza No Explicada. En este caso, el estadístico de contraste es el cociente entre VE y VNE normalizadas por sus grados de libertad:

$$\left.\begin{array}{c} \frac{VNE}{\sigma^2} = \frac{(n-m-1)\hat{\sigma}_R}{\sigma^2} \sim \chi^2_{n-m-1} \\ \frac{VE}{\sigma^2} \sim \chi^2_m \quad \text{bajo} \quad H_0 \end{array}\right\} \Rightarrow$$

$$\Rightarrow \frac{VE/m}{VNE/(n-m-1)} \sim F_{m,n-m-1} \quad \text{bajo} \quad H_0$$

donde se ha usado que las formas cuadráticas VE y VNE son independientes. Por tanto, rechazaremos la hipótesis nula al nivel de significacion α si el valor del estadístico de contraste F es mayor que $F_{m,n-m-1;\alpha}$. La tabla ANOVA en el caso de la regresión lineal múltiple tiene el formato siguiente:

Fuente	**Grados de libertad**	**SC**	**MC**	**Cociente F**
Regresión	m	VE	VE/m	$\frac{VE/m}{VNE/(n-m-1)}$
Error	n-m-1	VNE	VNE/(n-m-1)	
Total	n-1	VT		

Donde 'Fuente' se refiere a las fuentes de variación del modelo, 'SC' se refiere a la suma de cuadrados y 'MC' la media de cuadrados, es decir, la suma de cuadrados dividida entre los grados de libertad correspondientes. El 'Cociente F' es la relación entre la varianza explicada por la regresión y la varianza no explicada.

1.9. Bondad de ajuste del modelo

A partir de la descomposición de la variabilidad de la ecuación (1.1), introducimos el *coefiente de determinación*:

Definición 1.3. Definimos el coeficiente de determinación R^2 como:

$$R^2 := \frac{VE}{VT} = 1 - \frac{VNE}{VT}$$

Este coeficiente mide la proporción de varianza de la respuesta $\mathbf{Y}$ que es explicada por las variables independientes $X_1, ..., X_m$ a partir del modelo de regresión.

Como claramente $0 \leq VE \leq VT$, tenemos que $0 \leq R^2 \leq 1$. Si $R^2 = 1$ el modelo explica completamente la varianza de la respuesta. En este sentido, la relación entre las variables dependientes y las variables independientes es mayor cuanto más se acerca el valor de R^2 a 1.

El coeficiente R^2 tiene como principal inconveniente que aumenta siempre a medida que se añaden variables al modelo (incluso si esas variables no son significativas), es decir, favorece sistemáticamente a los modelos más complejos.

Para superar esta desventaja, definimos el *coeficiente de determinación corregido por los grados de libertad*, $\bar{R}^2$:

Definición 1.4. Definimos el coeficiente de determinación corregido por los grados de libertad como:

$$\bar{R}^2 = 1 - \frac{VNE/(n-m-1)}{VT/n-1}$$

Y se cumple que:

$$\bar{R}^2 = R^2 - (1-R^2)\frac{m}{n-m-1}$$

Entonces, si $k \geq 1$ se cumple que $R^2 > \bar{R}^2$. De hecho, $\bar{R}^2$ puede ser negativo.

1.10. Problemas debido a la alta dimensionalidad

En un contexto de *Big Data*, son usuales las situaciones de alta dimensionalidad en las variables explicativas $X_1, ..., X_m$. Este número podría ser mucho mayor que el tamaño de la muestra $(m >> n)$. En este tipo de circunstancias los estimadores mínimo cuadráticos pueden no existir o estar mal condicionados.

Ahora, introduciremos dos de los principales problemas que enfrentan los modelos de regresión lineal en esta situación:

- La aparición de inconsistencias en la estimación del modelo.
- El aumento en la probabilidad de colinealidad.

1.10.1. Inconsistencias en la estimación del modelo de regresión lineal

En un contexto de alta dimensionalidad donde m es mayor que n, denotado como $m > n$, algunas técnicas típicas de regresión no se comportan adecuadamente. Aparecen inconsistencias en el proceso de estimación. En particular, veámoslo para el caso de la regresión lineal.

Cuando $m > n$, no es posible obtener el estimador de mínimos cuadrados, $\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}$. Esto ocurre debido a que $\mathbf{X}$ es una matriz de dimensión $n \times (m+1)$, $\mathbf{X}^t\mathbf{X}$ es una matriz de dimensión $(m+1) \times (m+1)$ y el siguiente lema garantiza que $Rango(\mathbf{X}^t\mathbf{X}) \leq n < m+1$. Usando que

$$\exists \left(\mathbf{X}^t\mathbf{X}\right)^{-1} \Leftrightarrow det\left(\mathbf{X}^t\mathbf{X}\right) \neq 0 \Leftrightarrow Rango(\mathbf{X}^t\mathbf{X}) = m+1$$

la inversa no tiene por qué existir. Por tanto, no hay unicidad del estimador de mínimos cuadrados.

Lema 1.1. *Sea A una matriz de dimensión $c \times k$ y B una matriz de dimensión $k \times c$ con $c > k$, se cumple que*

$$\left.\begin{array}{c} Rango(\mathbf{A}\cdot\mathbf{B}) \leq Rango(\mathbf{A}) \\ \wedge \\ Rango(\mathbf{A}\cdot\mathbf{B}) \leq Rango(B) \end{array}\right\} \Rightarrow Rango(\mathbf{A}\cdot\mathbf{B}) \leq k$$

dado que $Rango(\mathbf{A}) \leq k$ y $Rango(\mathbf{B}) \leq k$ porque $c > k$.

Hay varios métodos para resolver este inconveniente que se basan en imponer una penalización a los coeficientes $\boldsymbol{\beta}$. En el siguiente capítulo estudiaremos tres de estos métodos: Ridge, LASSO y Elastic Net.

1.10.2. Colinealidad

La *colinealidad* se produce cuando alguno de los predictores del modelo puede ser expresado en función de los demás predictores a partir de una relación lineal. Si una de las variables explicativas X_j se expresa de forma exacta como combinación lineal de las restantes, la matriz $\mathbf{X}$ tendrá un rango menor que $m+1$ y entonces $\mathbf{X}^t\mathbf{X}$ será no invertible. En este caso se dice que X_j es colineal al resto y el sistema de ecuaciones que determina las estimaciones de los parámetros no tiene solución única.

No es común llegar a la situación extrema en que una variable explicativa es exactamente colineal al resto, pero es posible que algunas variables explicativas estén altamente correladas entre sí, esto es lo que llamamos *multicolinealidad*. La multicolinealidad no perjudica las predicciones obtenidas desde el modelo de regresión, ni en principio a la estimación de σ^2, pero sí afecta a la desviación estándar de los coeficientes estimados $\hat{\beta}_j$. Estos errores se incrementan considerablemente, lo que lleva a conclusiones falsas en los contrastes de significación individual.

Por lo tanto, considerar un número grande de covariables, m, aumenta la probabilidad de que el modelo lineal sufra colinealidad. Debemos buscar métodos para la reducción de dimensiones y así evitar estos inconvenientes.

1.10.3. Reducción de dimensiones: Métodos de selección de variables independientes

En general, el estimador de mínimos cuadrados ofrece dos desventajas:

- El método de los mínimos cuadrados obtiene estimadores con un un sesgo bajo, como contraparte obtiene una varianza alta. Es decir, existe un problema con la precisión de las predicciones.
- Cuando el número de predictores es alto se obtiene un modelo complicado de interpretar. Sería lógico intentar obtener un subconjunto de predictores que explicaran la mayoría de los efectos del modelo, aún sacrificando parte de la información.

Particularmente cuando el número de covariables supera al tamaño de la muestra, la alta dimensionalidad en los problemas de regresión es un inconveniente difícil de superar.

Vistos todos estos inconvenientes, se vuelve evidente la necesidad de obtener modelos simples cuando $m > n$. En los modelos de regresión lineal, una de las formas de obtener una reducción en la dimensión es a través de técnicas de selección de variables explicativas, imponiendo penalizaciones en el proceso de estimación para evitar inconsistencias.

Con esta clase de métodos seleccionamos cuáles de las m variables independientes son importantes para la regresión y descartamos las que no lo son. Es decir, se pretende que toda la información incluida en el modelo sea relevante y que el ruido y las dificultades computaciones que plantea el exceso de variables no relevantes queden eliminados. No obstante, en el caso $m > n$, las técnicas clasicas de selección de variables explicativas, como el *Principal Component Analisis* (PCA) no se comportan adecuadamente. En consecuencia, en este caso se necesitan técnicas específicas de reducción de covariables. Esto es lo que estudiaremos en el próximo capítulo.

Ahora, introduciremos algunos métodos de selección de covariables que nos permitirán obtener modelos más sencillos e interpretables.

1.11. Selección de covariables

Esta sección está inspirada en el punto 3.2 de la fuente [GP23].

Como hemos visto previamente, al añadir más predictores pagamos una penalización en forma de mayor variabilidad de los coeficientes estimados, mayores dificultades para la interpretación o el aumento en la probabilidad de predictores mutuamente dependientes linealmente, lo que se ha definido como colinealidad. En general, el número de predictores m que se pueden considerar en un modelo lineal multivariante está acotado por:

$$m \leq n - 2$$

Equivalentemente, el tamaño mínimo de una muestra para ajustar cierto modelo lineal con m variables independientes es $n \geq m + 2$. Esto se debe a que al ajustar un modelo lineal con m variables explicativas tenemos que estimar $m + 2$ parámetros: $(\beta_0, \beta_1, ..., \beta_m, \sigma^2)$ a partir de n datos. Por tanto, cuanto más se acerque $m + 2$ a n, menos precisa será la estimación obtenida.

Si $n = m + 2$ cada elemento de la muestra determina un parámetro estimado. Como vimos previamente

$$(\hat{\beta}_j - t_{n-m-1;\alpha/2} \cdot std(\hat{\beta}_j), \hat{\beta}_j + t_{n-m-1;\alpha/2} \cdot std(\hat{\beta}_j))$$

Donde al disminuir el número de grados de libertad, el factor $t_{n-m-1;\alpha/2}$ aumenta, ensanchando el intervalo y empeorando la precisión de la predicción. La selección de variables explicativas en la regresión multivariante trata de elegir el mejor subconjunto de estas variables que describa la variable explicada.

Nuestro objetivo con este tipo de métodos es simplificar el modelo eliminando predictores innecesarios, reduciendo la varianza y obteniendo un modelo más sencillo con el que trabajar computacionalmente.

Para ello, estudiaremos tres procedimientos de selección de variables: el método de eliminación hacia atrás, el método de selección hacia delante y la regresión *stepwise*.

1.11.1. Método de eliminación hacia atrás

El método de eliminación hacia atrás (*backward elimination*) comienza con el modelo completo obtenido previamente y va eliminando en cada paso la variable explicativa que menos afecta al ajuste del modelo. El algoritmo tiene los siguientes pasos:

- *Paso 1:* Ajustar el modelo lineal multivariante con todas las variables explicativas $X_i, i \in 1, ..., m$.

- *Paso 2:* Calcular los contrastes para cada X_i, basados en la distribución t de Student y localizar el mayor p-valor que supere α_{salir}. Entonces, eliminamos la variable asociada a dicho p-valor.

- *Paso 3:* Volver a ajustar el modelo, volver a calcular los contrastes y detener el proceso cuando todos los p-valores sean menores que α_{salir}.

El problema del método de *backward elimination* es su coste computacional. Se requiren muchas operaciones y se vuelve más ineficiente cuántos más predictores estemos considerando desde el inicio.

1.11.2. Método de introducción progresiva

El método de introducción progresiva, también llamado *forward selection* sigue un proceso inverso al del *backward elimination*. En este caso comenzamos con un modelo sin variables y vamos añadiendo progresivamente las de mayor significación. Podemos estructurarlo en los siguientes pasos:

- *Paso 1:* Definimos un nivel de significación arbitrario α_{entrar} y tomamos un modelo sin predictores.

- *Paso 2:* Se les realiza el contraste a todos los predictores que no están en el modelo y añadimos el predictor de mayor significación, esto es, el predictor asociado al menor p-valor menor que α_{entrar}.

- *Paso 3:* Repetir iterativamente el *Paso 2* hasta que no reste ningún predictor que cumpla los requisitos del segundo paso.

La ventaja de este método es que requiere menos capacidad computacional que el método *backward elimination*. La desventaja es que en este proceso se añaden variables que a posteriori se demuestran innecesarias cuando se añaden otras nuevas.

1.11.3. Método *stepwise*

El método de regresión *stepwise* se diferencia del método de *forward selection* en que cada vez que se añade un nuevo predictor al modelo se realiza el contraste de todas las variables presentes. Esto permite descartar las variables no significativas que se han añadido en los pasos previos, resolviendo así la principal desventaja del método de introducción progresiva. El algoritmo consta de las siguientes reglas, que se aplican en cada paso:

- *Regla de entrada* de variables: cada una de las variables entrará en el modelo si supone una contribución significativa al nivel α_{entrar}. Para determinar la contribución de una variable al modelo se evalúa el estadístico de contraste t asociado, siendo significativa si p-valor$< \alpha_{entrar}$. En caso de que el número de variables cumpliendo esta condición sea mayor que uno, solo entrará en el modelo la que más contribuya.
- *Regla de salida* de variables: Una variable previamente incluida en el modelo saldrá de la selección si no supone una contribución significativa al nivel α_{salir}. Es decir, saldrá la variable que tenga el mayor p-valor por encima de α_{salir}.

1.11.3.1. Notas sobre la regresión *stepwise*

Estos tres métodos son muy populares en situaciones de alta dimensionalidad, incluso si $m < n$, dadas las dificultades en la interpretación y en la computación de un modelo con un número excesivo de covariantes. Sin embargo, hay que tener prudencia al utilizarlos. Por ejemplo, en caso de multicolinealidad en las variables estos métodos no deberían aplicarse. Como vimos previamente, la multicolinealidad afecta a los contrastes de significación individual conduciendo a errores en muchos casos. Además, eliminar variables a partir de los contrastes no implica que las variables no estén relacionadas con la respuesta sino que su efecto es menor. Por tanto, si la finalidad del modelo es hacer predicciones puede ser recomendable mantener variables aunque tengan una influencia pequeña en la respuesta.

1.12. Criterios para la selección de modelos

La regresión stepwise necesita medir continuamente el rendimiento de los diferentes modelos que se van generando para así poder compararlos. Para ello se utiliza un *criterio de información*. Dicho criterio debe tener en cuenta tanto la calidad de ajuste del modelo como el número de predictores empleados. Veamos ahora algunos criterios de información usados en los métodos de selección de modelos.

1.12.1. Criterios basados en la bondad del ajuste

Estos criterios son en general aquellos que se basan en la varianza residual, como por ejemplo comparar modelos a partir de su coeficiente de determinación R^2. Este criterio

no es recomendable porque como vimos en la sección 1.8 este coeficiente crece al añadir predictores, favoreciendo la selección de modelos más complejos o con variables innecesarias. El coeficiente corregido $\bar{R}^2$ evita este problema pero aún así supone una regla demasiado permisiva a la hora de introducir variables.

1.12.2. Criterio AIC

EL Criterio de Información de Akaike (AIC), usa la función de verosimilitud y castiga a los modelos al aumentar la cantidad de parámetros estimados. Al valorar un modelo con m parámetros ($m-1$ variables independientes), el AIC se define como

$$AIC = n\log(\hat{\sigma}_m^2) + 2m$$

siendo $\hat{\sigma}_m^2$ el estimador de máxima verosimilitud de la varianza del modelo con m parámetros. Es evidente que al crecer m, disminuye $\hat{\sigma}_m^2$ y aumenta $2m$. Dado un conjunto de modelos, el mejor modelo es el que tiene el valor mínimo en el AIC.

1.12.3. Criterio BIC

E Criterio Bayesiano de Schwarz (BIC) que se define como

$$BIC = n\log(\hat{\sigma}_m^2) + m\log(n)$$

Al igual que con el AIC, el BIC identifica los mejores modelos con los valores pequeños que obtienen al ser evaluados por el criterio. Como podemos ver, se penalizan los modelos con un número grande de parámetros a medida que crecen las observaciones n.

Tanto el criterio AIC como el BIC pueden ser empleados en el método stepwise. Los aplicaremos tras cada etapa de entrada y salida de variables para evaluar cada posible modelo y compararlo con los obtenidos previamente

2 Métodos de regularización

En este capítulo, abordamos los métodos de regularización, herramientas esenciales en estadística para resolver los problemas de alta dimensionalidad planteados en el capítulo 1. Comenzamos introduciendo estos métodos en la sección 2.1. En la sección 2.2 trataremos el concepto de *Bias-Variance Trade-off* y su relación con los métodos de regularización. En las secciones 2.3 y 2.4 veremos en particular la regresión Ridge y algunos resultados teóricos de esta. En las secciones 2.5 y 2.6 haremos lo propio con las regresiones Lasso y Elastic Net, respectivamente.

2.1. Introducción a los métodos de regularización

Para realizar esta sección nos hemos basado en la fuente [Neu14].

Los estimadores de mínimos cuadrados tienen mínima varianza entre los estimadores lineales insesgados siempre y cuando se cumplan las hipótesis del teorema de Gauss-Markov, pero en ocasiones estas hipótesis no se cumplen. Otras veces, preferimos minimizar aún más la varianza aún a riesgo de obtener un estimador sesgado.

En condiciones de gran dimensionalidad, los estimadores de mínimos cuadrados se ven gravemente afectados por la multicolinealidad y el modelo se vuelve difícil de interpretar.

Al usar métodos de selección de variables explicativas como *backward elimination*, *forward selection y stepwise*, obtenemos modelos más fácilmente interpretables y más sencillos de tratar computacionalmente. Además, si hemos realizado correctamente la selección de estos predictores obtendremos errores de predicción menores que los obtenidos con el modelo completo. A pesar de estas ventajas, por ser métodos discretos y dicotómicos donde las variables se escogen o se descartan sin considerar ninguna opción intermedia, pueden ofrecer una variabilidad excesivamente alta.

Estas desventajas nos muestran la necesidad de buscar alternativas. Necesitamos modelos tales que:

- Seleccionen solo las variables relevantes para la regresión
- Reduzcan la varianza de tal forma que mejore la precisión de las predicciones del modelo.
- Sean sencillos de entender e interpretar.

En este capítulo estudiaremos tres modelos de regularización: Ridge, Lasso (*least absolute shrinkage and selection operator*) y el método *Elastic Net* que combina los dos primeros métodos. Como veremos, estos modelos ofrecen mayor continuidad, consiguiendo reducir la variabilidad.

Estos métodos regularizan los coeficientes, reducen la varianza y en el caso de Lasso y Elastic Net, seleccionan los coeficientes que anular completamente.

Como vimos en el anterior capítulo, los estimadores de mínimos cuadrados $\hat{\boldsymbol{\beta}}$ del modelo lineal multivariante

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

son los que minimizan la función SRC definida como

$$SRC(\boldsymbol{\beta}) =:= \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - ... - \beta_m X_{im})^2$$

Siempre y cuando se cumplan las hipótesis establecidas previamente. En ese caso, vimos que

$$\hat{\boldsymbol{\beta}} = \mathcal{N}_n(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$$

Y podemos afirmar que $\hat{\boldsymbol{\beta}}$ es un estimador no insesgado de $\boldsymbol{\beta}$. Sin embargo, el sesgo de una estimación no es el único parámetro que determina la calidad de las predicciones de un modelo, también es necesario tener en cuenta su varianza. De la idea de optimizar ambos parámetros surge el concepto de *Bias-Variance Tradeoff*, basado en la descomposición del Error Cuadrático Medio (ECM) en sesgo y varianza. Por ejemplo, para el estimador $\hat{\beta}_i$ de β_i, se tiene:

$$ECM[\hat{\beta}_j] := \mathbb{E}[(\hat{\beta}_i - \beta_i)^2] = B^2(\hat{\beta}_j) + Var(\hat{\beta}_j) = (\mathbb{E}[\hat{\beta}_i] - \beta_i)^2 + Var[\hat{\beta}_i]$$

siendo $B(\hat{\beta}_j)$ la función sesgo asociada el estimador $\hat{\beta}_j$.

Los métodos de regularización pretenden reducir la varianza a cambio de añadir sesgo a $\hat{\boldsymbol{\beta}}$.

2.2. Bias-Variance Trade-off

Esta sección está basada en el punto 3.1 de la fuente [Neu14].

Para entender el avance que representan estos métodos de regresión respecto a los mínimos cuadrados clásicos tenemos que entender la relación entre sesgo y varianza. Especialmente tenemos que relacionarlas con las interpretaciones más recientes asociadas a campos como el *machine learning*.

Como ya sabemos, el sesgo representa la diferencia entre la predicción esperada y el valor que estamos prediciendo en realidad. Por otra parte la varianza mide cuánto varían las predicciones para cada observación de los predictores alrededor de su media. El sesgo y la varianza también afectan a otro de los factores que hemos explicado antes que pretendemos optimizar: la complejidad y explicabilidad del modelo.

En el campo del *machine learning*, para tratar con modelos con muchos predictores se dividen las observaciones de los predictores y las respuestas en conjuntos de entrenamiento y conjuntos de testeo. En este caso, imponer insesgadez en el modelo nos lleva a lo que se

denomina como *overfitting* o 'sobre-entrenamiento'. Esto significa que el modelo obtenido se comporta muy bien en el conjunto sobre el que hemos estimado sus parámetros (el conjunto de entrenamiento) y obtiene una varianza excesivamente alta en el conjunto de testeo que hace que el modelo tenga poca precisión en la predicción. El fenómeno del sobre-entrenamieto suele ser consecuencia de la complejidad del modelo.

Cuando el modelo es muy complejo, observamos una varianza alta en las predicciones y por tanto un error alto en el conjunto de testeo. Cuando la complejidad del modelo es baja, puede ocurrir el efecto contrario, el *under-fitting*, donde obtenemos un modelo altamente sesgado. En general, el modelo con las mejores capacidades predictivas es aquel que encuentra un equilibrio entre el sesgo y la varianza. Esto es lo que pretendemos obtener con los siguientes métodos.

2.3. Regresión Ridge

Esta sección se ha escrito teniendo en cuenta el punto 3.4.1 de la fuente [TH09].

La regresión Ridge es un método para estimar los coeficientes de modelos lineales múltiples cuando los predictores están altamente correlacionados. Por tanto, es útil en los casos $m >> n$ que nos preocupan, donde la cantidad de covariables hace aumentar la probabilidad de tener multicolinealidad. Como hemos explicado previamente, la regresión Ridge pretende reducir la variabilidad en la estimación de los parámetros a partir de admitir cierta cantidad de sesgo.

La regresión Ridge reduce los coeficientes de la regresión lineal al imponer una penalización a su tamaño. Los coeficientes obtenidos con este método minimizan una función SRC penalizada. En el caso de la regresión Ridge, se impone una penalización cuadrática a los coeficientes que determinan la *pendiente* del modelo, es decir, a los coeficientes $\boldsymbol{\beta}_r = (\beta_1, ..., \beta_m)^t$. Se pretende encontrar $\hat{\boldsymbol{\beta}}^{Ridge}$ tal que

$$\hat{\boldsymbol{\beta}}^{Ridge} = \underset{\boldsymbol{\beta}}{\arg\min} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{m} X_{ij}\beta_j \right)^2$$
$$\text{sujeto a } \sum_{j=1}^{m} \beta_j^2 \leq t$$

para cierto $t > 0$. Esta expresión muestra explícitamente la restricción de tamaño a los coeficientes. Equivalentemente, podemos expresar este problema en su forma lagrangiana:

$$\hat{\boldsymbol{\beta}}^{Ridge} = \underset{\boldsymbol{\beta}}{\arg\min} \left\{ \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{m} X_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{m} \beta_j^2 \right\}$$

con $\lambda \geq 0$ siendo el parámetro de penalización que controla la cantidad de regularización. Cuánto mayor sea λ, mayor regularización.

2.3.1. Coeficientes de la regresión Ridge

Teorema 2.1. *Los coeficientes de la regresión Ridge se expresan como*

$$\hat{\boldsymbol{\beta}}^{Ridge} = \left(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m}\right)^{-1}\mathbf{X}^t\mathbf{Y}$$

Demostración
Podemos escribir la función a minimizar como:

$$SRC_r(\lambda) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^t (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^t\boldsymbol{\beta}$$

Primero, igualamos el gradiente de la función $SRC_r(\lambda)$ respecto a $\boldsymbol{\beta}$ a cero:

$$\frac{\delta SRC(\beta)}{\delta\beta} = 0$$

Es decir,

$$\frac{\delta((\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^t (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^t\boldsymbol{\beta})}{\delta\beta} = 0 \Rightarrow$$

$$\Rightarrow -2\mathbf{X}^t (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + 2\lambda\boldsymbol{\beta} = 0$$

$$(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m})\boldsymbol{\beta} = \mathbf{X}^t\mathbf{Y}$$

La matriz $(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m})$ es definida positiva para todo $\lambda > 0$ porque para todo vector $\mathbf{a} \neq 0$ de dimensión $m \times 1$ tenemos:

$$\mathbf{a}^t(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m})\mathbf{a} = (\mathbf{X}\mathbf{a})^t(\mathbf{X}\mathbf{a}) + \lambda\mathbf{a}^t\mathbf{a} =$$

$$\sum_{i=1}^{n} (\mathbf{X}_i\mathbf{a})^2 + \lambda \sum_{k=1}^{m} a_k^2 > 0$$

Por tanto, la matriz es invertible. Obteniendo finalmente:

$$\hat{\boldsymbol{\beta}}^{Ridge} = \left(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m}\right)^{-1}\mathbf{X}^t\mathbf{Y}$$

Ahora, verifiquemos que esta expresión obtenida representa efectivamente un mínimo. Para ello, usando la matriz Hessiana de la función SRC_λ:

$$\frac{\delta^2 SRC_\lambda}{\delta\boldsymbol{\beta}^2} = Hess(SRC_\lambda) = 2(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m})$$

Que es definida positiva por serlo $(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m})$. Por tanto, SRC_λ es estrictamente convexa en $\hat{\boldsymbol{\beta}}^{Ridge}$ siendo este un mínimo □

En esta expresión podemos observar dos puntos claves sobre los que se basa la regresión Ridge:

1. La solución de la regresión Ridge es, al igual que la regresión de mínimos cuadrados

clásica, una función lineal de $\mathbf{Y}$.

2. Al sumar $\lambda \mathbf{I_m}$ a $\mathbf{X}^t\mathbf{X}$ estamos sumando una constante positiva a la diagonal de $\mathbf{X}^t\mathbf{X}$, obteniendo una matriz invertible incluso si $\mathbf{X}^t\mathbf{X}$ no lo es. En realidad, este era el principal objetivo de la regresión Ridge cuando fue diseñada por Arthur Hoerl y Robert Kennard en 1970.

Como demostraremos en la siguiente sección, los coeficientes obtenidos con la regresión de Ridge dependen de la escala de las observaciones con las que se estiman. Por tanto, antes de calcular los coeficientes es recomendable normalizar los datos.

2.3.2. Significado de λ

En esta subsección nos basaremos en el punto 3.3 de la fuente [Neu14].

Antes de estudiar las propiedades de la regresión Ridge, analicemos la importancia de λ y cómo influye en la complejidad del modelo o en el equilibrio varianza-sesgo.

El parámetro λ controla la cantidad de regularización en los estimadores Ridge y el tamaño de sus coeficientes. Al crecer λ, $\hat{\boldsymbol{\beta}}^{Ridge}$ va tendiendo a $\mathbf{0}$. Por otra parte, cuando $\lambda = 0$, obtenemos el estimador de mínimos cuadrados clásico.

Como veremos en los siguientes resultados teóricos, el valor de este parámetro es clave para la capacidad predictiva del modelo: a medida que λ aumenta se reduce la varianza, pero aumenta el sesgo en el modelo y viceversa.

Otro concepto relacionado con λ son los grados de libertad. Para el método de mínimos cuadrados, los grados de libertad son iguales al número de parámetros libres, que denotamos por $m+1$. Para la regresión Ridge, los grados de libertad se definen como una función de λ. En particular:

$$GL(\lambda) = Traza[\mathbf{X}(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^t] =$$

$$= \sum_{i=1}^{m+1} \frac{d_i^2}{d_i^2 + \lambda}$$

donde d_i son los valores propios de la matriz $\mathbf{X}^t\mathbf{X}$. Es evidente que esta es una función decreciente de λ. Así, como cabría esperar, cuando $\lambda = 0$, $GL(\lambda) = m+1$ pues estaríamos en el caso de mínimos cuadrados clásicos. Por otra parte, si

$$\lambda \to \infty \Rightarrow GL(\lambda) \to 0$$

Es decir, a mayor nivel de regularización, menores grados de libertad tiene el modelo.

2.4. Propiedades de la regresión Ridge

Esta sección está basada en la fuente [Tab21].

En esta sección obtenemos el sesgo y la varianza del estimador de Ridge bajo las hipótesis de Gauss-Markov introducidas en la sección 1.3.

2.4.1. Sesgo de la regresión Ridge

Teorema 2.2. *El sesgo del estimador Ridge es*

$$\mathbb{E}\left[\hat{\boldsymbol{\beta}}^{Ridge}\right] - \boldsymbol{\beta} = [(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m})^{-1} - (\mathbf{X}^t\mathbf{X})^{-1}]\mathbf{X}^t\mathbf{X}\boldsymbol{\beta}$$

Demostración
El estimador $\hat{\boldsymbol{\beta}}^{Ridge}$ se puede expresar como

$$\hat{\boldsymbol{\beta}}^{Ridge} = \left(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m}\right)^{-1}\mathbf{X}^t\mathbf{Y} =$$

$$= \left(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m}\right)^{-1}\mathbf{X}^t(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) =$$

$$= \left(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m}\right)^{-1}\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} + \left(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m}\right)^{-1}\mathbf{X}^t\boldsymbol{\varepsilon} =$$

Y por tanto

$$\mathbb{E}\left[\hat{\boldsymbol{\beta}}^{Ridge}\right] = \mathbb{E}\left[\left(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m}\right)^{-1}\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} + \left(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m}\right)^{-1}\mathbf{X}^t\boldsymbol{\varepsilon}\right] =$$

$$= \left(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m}\right)^{-1}\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} + \left(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m}\right)^{-1}\mathbf{X}^t\mathbb{E}[\boldsymbol{\varepsilon}] =$$

$$= \left(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m}\right)^{-1}\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} + \left(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m}\right)^{-1}\mathbf{X}^t\mathbb{E}\cdot\mathbf{0} =$$

$$= \left(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m}\right)^{-1}\mathbf{X}^t\mathbf{X}\boldsymbol{\beta}$$

Sabemos que el estimador será insesgado si y solamente si $\mathbb{E}\left[\hat{\boldsymbol{\beta}}^{Ridge}\right] = \boldsymbol{\beta}$. Cosa que sólo puedo ocurrir si

$$\left(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m}\right)^{-1}\mathbf{X}^t\mathbf{X} = \mathbf{I_m}$$

Que equivale a $\lambda = 0$, lo cuál no es posible pues $\lambda > 0$. De hecho, si $\lambda = 0$ estaríamos en el caso de los coeficientes obtenidos por el método clásico de la minimización del error cuadrático. Por tanto, el estimador Ridge es no insesgado y podemos calcular su sesgo:

$$\mathbb{E}\left[\hat{\boldsymbol{\beta}}^{Ridge}\right] - \boldsymbol{\beta} =$$

$$= \left(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m}\right)^{-1}\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta} =$$

$$= \left(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m}\right)^{-1}\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} - (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} =$$

$$[(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m})^{-1} - (\mathbf{X}^t\mathbf{X})^{-1}]\mathbf{X}^t\mathbf{X}\boldsymbol{\beta}$$

□

2.4.2. Varianza de la regresión Ridge

Proposición 2.1. *La varianza del estimador de Ridge es*

$$Var(\hat{\boldsymbol{\beta}}^{Ridge}) = \sigma^2(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m})^{-1}\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m})^{-1}$$

Demostración
Previamente habíamos demostrado que la varianza del estimador $\hat{\boldsymbol{\beta}}$ obtenido a partir de minimizar la función SRC es igual a $\sigma^2(\mathbf{X}^t\mathbf{X})^{-1}$.

Expresando $\hat{\boldsymbol{\beta}}^{Ridge}$ en función del $\boldsymbol{\beta}$ clásico:

$$\hat{\boldsymbol{\beta}}^{Ridge} = \left(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m}\right)^{-1}\mathbf{X}^t\mathbf{Y}$$

$$= \left(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m}\right)^{-1}\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}$$

$$= \left(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m}\right)^{-1}\mathbf{X}^t\mathbf{X}\hat{\boldsymbol{\beta}}$$

Y aplicando la varianza a ambos lados de la igualdad, obtenemos:

$$Var(\hat{\boldsymbol{\beta}}^{Ridge}) = \left(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m}\right)^{-1}\mathbf{X}^t\mathbf{X}Var(\hat{\boldsymbol{\beta}})[\left(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m}\right)^{-1}\mathbf{X}^t\mathbf{X}]^t =$$

$$= \left(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m}\right)^{-1}\mathbf{X}^t\mathbf{X}Var(\hat{\boldsymbol{\beta}})\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m})^{-1} =$$

$$= \left(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m}\right)^{-1}\mathbf{X}^t\mathbf{X}\sigma^2(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m})^{-1} =$$

$$= \sigma^2(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m})^{-1}\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m})^{-1}$$

□

Ya hemos explicado sobre como estos métodos buscan aumentar el sesgo a cambio de reducir la varianza de los estimadores. Para comparar la varianza de dos estimadores tenemos que ver si la matriz que representa la diferencia entre ambas matrices de covarianzas es definida positiva. Así, podemos formular la siguiente proposición:

Proposición 2.2. *La varianza del estimador Ridge es siempre menor que la varianza del estimador de mínimos cuadrados clásico. Es decir,*

$$Var(\hat{\boldsymbol{\beta}}) - Var(\hat{\boldsymbol{\beta}}^{Ridge})$$

es una matriz definida positiva.

Demostración
Empecemos suponiendo que la la matriz $\mathbf{X}$ es invertible y por tanto el estimador de mínimos cuadrados clásicos existe. Dicho estimador tiene la siguiente varianza:

$$Var(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1}$$

Ahora, definamos la matriz

$$Z = \mathbf{X}^t\mathbf{X}(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m})^{-1}$$

que es no singular. Ahora expresamos la varianza del estimador de Ridge como

$$Var[\hat{\boldsymbol{\beta}}^{Ridge}] = \sigma^2(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m})^{-1}\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m})^{-1} =$$

$$= \sigma^2(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m})^{-1}\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{X}(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m})^{-1} =$$

$$= \sigma^2 Z^t (\mathbf{X}^t\mathbf{X})^{-1} Z$$

Ahora, restamos ambas varianzas

$$Var(\hat{\boldsymbol{\beta}}) - Var(\hat{\boldsymbol{\beta}}^{Ridge}) = \sigma^2(\mathbf{X}^t\mathbf{X})^{-1} - \sigma^2 Z^t(\mathbf{X}^t\mathbf{X})^{-1}Z =$$

$$\sigma^2[\mathbf{Z}^t(\mathbf{Z}^t)^{-1}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{Z}^{-1}\mathbf{Z} - \mathbf{Z}^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{Z}] =$$

$$= \sigma^2\mathbf{Z}^t[(\mathbf{Z}^t)^{-1}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{Z}^{-1} - (\mathbf{X}^t\mathbf{X})^{-1}]\mathbf{Z} =$$
$$= \sigma^2\mathbf{Z}^t[(\mathbf{X}^t\mathbf{X})^{-1}(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m})(\mathbf{X}^t\mathbf{X})^{-1}(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m})(\mathbf{X}^t\mathbf{X})^{-1} - (\mathbf{X}^t\mathbf{X})^{-1}]\mathbf{Z} =$$
$$= \sigma^2\mathbf{Z}^t[((\mathbf{X}^t\mathbf{X})^{-1} + \lambda(\mathbf{X}^t\mathbf{X})^{-2})(\mathbf{I_m} + \lambda(\mathbf{X}^t\mathbf{X})^{-1}) - (\mathbf{X}^t\mathbf{X})^{-1}]\mathbf{Z} =$$
$$= \sigma^2\mathbf{Z}^t[2\lambda(\mathbf{X}^t\mathbf{X})^{-2} + \lambda^2(\mathbf{X}^t\mathbf{X})^{-3}]\mathbf{Z} =$$
$$= \sigma^2(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m})^{-1}[2\lambda\mathbf{I_m} + \lambda^2(\mathbf{X}^t\mathbf{X})^{-1}](\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m})^{-1}$$

Como $\lambda > 0$, la última matriz obtenida es definida positiva puesto que para todo $\mathbf{u} \neq \mathbf{0}$ se cumple que

$$\mathbf{z} = (\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m})^{-1}\mathbf{u} \neq 0$$

y que

$$\mathbf{u}^t[Var(\hat{\boldsymbol{\beta}}) - Var(\hat{\boldsymbol{\beta}}^{Ridge})]\mathbf{u} =$$
$$= \sigma^2\mathbf{u}^t[2\lambda\mathbf{I_m} + \lambda^2(\mathbf{X}^t\mathbf{X})^{-1}]\mathbf{u}$$
$$\geq \sigma^2\lambda\mathbf{u}^t\mathbf{u} + \sigma^2\lambda^2\mathbf{u}^t(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{u} > 0$$

usando que $(\mathbf{X}^t\mathbf{X})$ y $(\mathbf{X}^t\mathbf{X})^{-1}$ son definidas positivas. □

2.4.3. Error cuadrático medio de la regresión Ridge

El Error Cuadrático Medio (ECM) de la regresión Ridge es igual a la traza de su varianza más la norma $\|\cdot\|_2$ de su sesgo. Es decir:

$$ECM(\hat{\boldsymbol{\beta}}^{Ridge}) =$$

$$= \mathbb{E}[\left\|\hat{\boldsymbol{\beta}}^{Ridge} - \boldsymbol{\beta}\right\|^2] =$$

$$= Traza(Var(\hat{\boldsymbol{\beta}}^{Ridge})) + \left\|B(\hat{\boldsymbol{\beta}}^{Ridge})\right\|^2$$

Previamente habíamos probado que el estimador de mínimos cuadrados clásico es insesgado, por tanto su error cuadrático medio cumple:

$$ECM(\hat{\boldsymbol{\beta}}) = Traza(Var(\hat{\boldsymbol{\beta}}))$$

Por tanto, restando ambos ECM

$$ECM(\hat{\boldsymbol{\beta}}) - ECM(\hat{\boldsymbol{\beta}}^{Ridge}) =$$

$$= Traza(Var(\hat{\boldsymbol{\beta}}) - Var(\hat{\boldsymbol{\beta}}^{Ridge})) - \left\| B(\hat{\boldsymbol{\beta}}^{Ridge}) \right\|^2$$

Usando que la matriz $Var(\hat{\boldsymbol{\beta}}) - Var(\hat{\boldsymbol{\beta}}^{Ridge})$ es definida positiva podemos afirmar que $Traza(Var(\hat{\boldsymbol{\beta}}) - Var(\hat{\boldsymbol{\beta}}^{Ridge}))$ es estrictamente positiva. Por tanto en principio no podríamos asegurar si la diferencia $Traza(Var(\hat{\boldsymbol{\beta}}) - Var(\hat{\boldsymbol{\beta}}^{Ridge})) - \left\| B(\hat{\boldsymbol{\beta}}^{Ridge}) \right\|^2$ es positiva o negativa.

Se puede probar (Theobald 1974 [The74]) que el signo de dicha diferencia depende del λ escogido y que siempre es posible encontrar un λ que haga positiva esta diferencia. De aquí, deducimos la siguiente proposición:

Proposición 2.3. *Para todo $\boldsymbol{\beta} \in \mathbb{R}^{m+1}$ existe λ tal que*

$$ECM(\hat{\boldsymbol{\beta}}^{Ridge}) < ECM(\hat{\boldsymbol{\beta}})$$

Este resultado es muy importante desde un punto de vista práctico, pero también teórico. Previamente, habíamos enunciado y demostrado el teorema de Gauss-Markov que afirma que el estimador de mínimos cuadrados clásico tiene la menor varianza y el menor error cuadrático medio entre los estimadores insesgados, pero no entre todos los estimadores. Aquí vemos como existe al menos un λ tal que el estimador Ridge asociado (que no es insesgado) tiene un error cuadrático medio menor que el estimador dado por el teorema de Gauss-Markov.

2.4.4. El estimador Ridge no es invariante a la escala

El estimador de mínimos cuadrados tiene la ventaja de que es invariante a las escalas. Es decir, que si multiplicamos la matriz de diseño por una matriz no singular $\mathbf{A}$, el estimador de mínimos cuadrados que obtenemos es igual al anterior multiplicado por $\mathbf{A}^{-1}$. Consideremos el estimador de mínimos cuadrados $\hat{\boldsymbol{\beta}}$:

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

Y sea $\mathbf{Q}$ la matriz de diseño reescalada:

$$\mathbf{Q} = \mathbf{X}\mathbf{A}$$

Calculemos el estimador de mínimos cuadrados que se obtiene con la nueva matriz de diseño:

$$\begin{aligned}
\tilde{\boldsymbol{\beta}} &= (\mathbf{Q}^t\mathbf{Q})^{-1}\mathbf{Q}^t\mathbf{Y} = \\
&= (\mathbf{A}^t\mathbf{X}^t\mathbf{X}\mathbf{A})^{-1}\mathbf{A}^t\mathbf{X}^t\mathbf{Y} = \\
&= \mathbf{A}^{-1}(\mathbf{X}^t\mathbf{X})^{-1}(\mathbf{A}^t)^{-1}\mathbf{A}^t\mathbf{X}^t\mathbf{Y} = \\
&= \mathbf{A}^{-1}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y} = \\
&= \mathbf{A}^{-1}\hat{\boldsymbol{\beta}}
\end{aligned}$$

El estimador Rigde no conserva esta propiedad. De hecho, podemos formular la siguiente proposición:

Proposición 2.4. *El estimador Ridge es invariante a escalas si y solo si la matriz de cambio de escala es ortonormal.*

Demostración
Sea $\hat{\boldsymbol{\beta}}^{Ridge}$ el estimador de Rigde obtenido como

$$\hat{\boldsymbol{\beta}}^{Ridge} = (\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m})^{-1}\mathbf{X}^t\mathbf{Y}$$

Ahora, el estimador de Ridge asociado a la matriz reescalada

$$\mathbf{Q} = \mathbf{XA}$$

se expresa como

$$\begin{aligned}\tilde{\boldsymbol{\beta}} = (\mathbf{Q}^t\mathbf{Q} + \lambda\mathbf{I_m})^{-1}\mathbf{Q}^t\mathbf{Y} = \\ = (\mathbf{A}^t\mathbf{X}^t\mathbf{XA} + \lambda\mathbf{I_m})^{-1}\mathbf{A}^t\mathbf{X}^t\mathbf{Y} = \\ = (\mathbf{A}^t\mathbf{X}^t\mathbf{XA} + \lambda\mathbf{A}^t(\mathbf{A}^t)^{-1}\mathbf{A}^{-1}\mathbf{A})^{-1}\mathbf{A}^t\mathbf{X}^t\mathbf{Y} = \\ = (\mathbf{A}^t(\mathbf{X}^t\mathbf{X} + \lambda(\mathbf{A}^t)^{-1}\mathbf{A}^{-1})\mathbf{A})^{-1}\mathbf{A}^t\mathbf{X}^t\mathbf{Y} = \\ = \mathbf{A}^{-1}(\mathbf{X}^t\mathbf{X} + \lambda(\mathbf{A}^t)^{-1}\mathbf{A}^{-1})^{-1}(\mathbf{A}^t)^{-1}\mathbf{A}^t\mathbf{X}^t\mathbf{Y} = \\ = \mathbf{A}^{-1}(\mathbf{X}^t\mathbf{X} + \lambda(\mathbf{A}^t)^{-1}\mathbf{A}^{-1})^{-1}\mathbf{X}^t\mathbf{Y}\end{aligned}$$

Que es igual a $\mathbf{A}^{-1}\hat{\boldsymbol{\beta}}^{Ridge}$ si y solo si

$$(\mathbf{A}^t)^{-1}\mathbf{A}^{-1} = \mathbf{I_m} \Leftrightarrow \mathbf{AA}^t = \mathbf{I_m}$$

□

Al no poder apoyarnos en la invariabilidad frente a la escala de los estimadores debemos tener cuidado con la elección de la escala de las variables pues afectará a los estimadores de los coeficientes. Para evitar estas situaciones, normalizamos las variables antes de realizar el ajuste del modelo.

2.4.5. Cómo escoger el parámetro de penalización

Para finalizar con el estudio de la regresión Ridge nos queda por responder una pregunta natural: ¿Cómo podemos elegir λ?

Uno de los métodos más comunes es el denominado como *leave-one-out cross-validation* (LOOCV). El algoritmo es el siguiente:

1. Elegimos un conjunto de k posibles parámetros $\lambda_1, ..., \lambda_k$.
2. Para cada $i = 1, ..., n$ descartamos la i-ésima observación (X_i, Y_i) de la muestra aleatoria. Entonces:
   2.1 Usamos los $n-1$ observaciones restantes para obtener k estimadores Ridge de $\boldsymbol{\beta}$, que en este caso denotaremos como $\hat{\boldsymbol{\beta}}^{Ridge}_{\lambda_p,i}$, donde el subíndice λ_p, i denota que el parámetro de penalización es igual a λ_p y la i-ésima observación se ha excluido.
   2.2 Calculamos el error cuadrático medio de las predicciones:

$$ECM_{\lambda_p} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{\mathbf{Y}}_{\lambda_p,i}\right)^2$$

con $p = 1, ..., k$.

2.3 Finalmente escogemos como el parámetro de penalización óptimo (λ^o) el que minimiza el error cuadrático medio:

$$\lambda^o = \arg\min_{\lambda_p} ECM(\lambda_p)$$

Es decir, imponemos como parámetro de minimización aquel que genera el menor error cuadrático medio en la validación cruzada.

2.4.6. Regresión Ridge: Conclusiones

En general, la regresión Ridge ofrece dos ventajas respecto al método de mínimos cuadrados y al método de selección de variables *stepwise*:

1. Mejoramos la precisión en la predicción del modelo al reducir la varianza (aún a costa de sacrificar la insesgadez).
2. El modelo se vuelve más 'estable´ al tener la capacidad de calibrar la importancia de las variables explicativas, mejorando al método *stepwise* que simplemente las introducía o las rechazaba según un test.

Y a pesar de estos avances, el método Ridge no resuelve completamente nuestros problemas en casos de alta dimensionalidad pues no realiza selección de predictores, debido a que ningún coeficiente se llega a anular. Así, mantiene la difícil interpretabilidad del modelo como en el caso de los mínimos cuadrados. Además, Ridge puede volverse computacionalmente muy costoso, dada la expresión de sus coeficientes $\hat{\boldsymbol{\beta}}^{Ridge} = \left(\mathbf{X}^t\mathbf{X} + \lambda\mathbf{I_m}\right)^{-1}\mathbf{X}^t\mathbf{Y}$, que implica calcular la inversa de una matriz. En la próxima sección, estudiaremos un método que mejora en muchos aspectos a la regresión Ridge: la regresión Lasso.

2.5. Regresión Lasso

En esta sección nos hemos basado en los punto 3.4.2 y 3.4.3 de la fuente [TH09] y en el punto 5.1 de la fuente [Neu14].

En esta sección continuamos continuamos con los métodos de regularización. Como hemos visto, la regresión Ridge obtiene un estimador de $\boldsymbol{\beta}$ a partir de la minimización de:

$$\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_2^2$$

La elección de la penalización $\lambda \|\boldsymbol{\beta}\|_2^2$ se originó para resolver la posible no invertibilidad de la matriz $\mathbf{X}^t\mathbf{X}$ en codiciones de alta dimensionalidad. Este motivo nos sugiere que la elección de la norma euclídea de $\boldsymbol{\beta}$ como término de penalización es arbitraria y que podríamos probar otras opciones. Uno de esos posibles términos de penalización es la penalización Lasso, lo que nos lleva a la regresión Lasso. Al igual que la regresión Ridge, la regresión Lasso es una variación del método de mínimos cuadrados clásicos. La diferencia con la regresión Ridge es que la regresión Lasso usa la norma L_1 en la penalización. Así, el estimador Lasso es aquel que minimiza la función:

$$\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \tag{2.5}$$

Y por tanto, el estimador Lasso ($\hat{\boldsymbol{\beta}}^{lso}$), se define como:

$$\hat{\boldsymbol{\beta}}^{lso} = \underset{\boldsymbol{\beta}}{\arg\min} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{m} X_{ij}\beta_j \right)^2 \quad \text{sujeto a } \sum_{j=1}^{m} |\beta_j| \leq t$$

También podemos escribir el estimador de Lasso de forma equivalente en su forma lagrangiana

$$\hat{\boldsymbol{\beta}}^{lso} = \underset{\boldsymbol{\beta}}{\arg\min} \left\{ \frac{1}{2} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{m} X_{ij}\beta_j \right)^2 + \lambda \sum_{j=1}^{m} |\beta_j| \right\}$$

Esta nueva penalización es no derivable debido al valor absoluto, provocando que los coeficientes $\hat{\boldsymbol{\beta}}^{lso}$ no tengan una expresión equivalente a la que hemos demostrado previamente para los coeficientes Ridge en el Teorema 2.1 .

Al igual que en la regresión Ridge, el parámetro λ (o t) se utiliza para definir la cantidad de regularización de los coeficientes Lasso. Notemos $t_0 = \sum_{j=1}^{n} |\boldsymbol{\beta}_{MCO}|$ a la suma del valor absoluto de los coeficientes obtenidos por el método de mínimos cuadrados ordinarios. Si $t > t_0$, no se produce ninguna regularización y los coeficientes obtenidos por el método Lasso coinciden con los obtenidos a partir de los mínimos cuadrados clásicos. Por otra parte, si $0 < t < t_0$, se produce cierta regularización y los coeficientes obtenidos se van aproximando a cero. Si λ es suficientemente grande (o equivalentemente t suficientemente pequeña) se anularán algunos estimadores, realizando una selección de covariables continua. Así, se mejoran las capacidades del estimador Ridge que no realiza selección de covariables y del método *stepwise* que realiza una selección discreta. Al igual que en la regresión Ridge, λ debe ser elegido tal que minimice el error de predicción. Por el contrario, en esta regresión los grados de libertad no se pueden obtener explícitamente de forma sencilla.

Ilustremos gráficamente la restricción de Ridge y Lasso en dimensión 2. Supongamos el caso más general, donde la matriz de diseño $\mathbf{X}$ es no ortonormal. En este caso la función Suma de Residuos Cuadráticos SRC tiene contornos elípticos. La restricción para la regresión Ridge es el disco $\beta_1^2 + \beta_2^2 \leq t$ mientras que para Lasso es el rombo $|\beta_1| + |\beta_2| \leq t$. Ambos métodos encuentran el punto donde los contornos elípticos de la función SRC alcanzan la región de la restricción. La diferencia entre ambos que provoca que el método Lasso anule coeficientes y el método Ridge no, es que la restricción de Lasso tiene 'picos'. Cuando la solución ocurre en uno de ellos, al menos uno de los estimadores es cero.

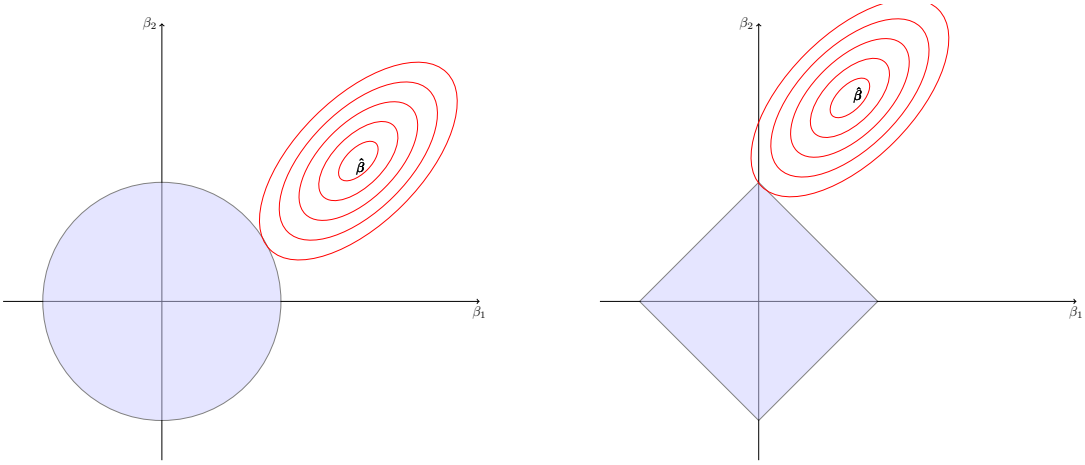


Figura 2.1: A la izquierda, la región delimitada por la condición $\beta_1^2 + \beta_2^2 \leq t$ de la penalización Ridge. A la derecha, la región delimitada por la condición $|\beta_1| + |\beta_2| \leq t$ de la penalización Lasso.

2.5.1. Unicidad del estimador Lasso

Para el desarrollo de esta subsección nos basaremos en el punto 6.1 de la fuente [vW15].

El estimador de la regresión Lasso no tiene por qué ser único. Esto se deduce a partir de la función de pérdida asociada. Esta función es la suma de una función cuadrática ($\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$) y un valor absoluto ($\lambda \|\boldsymbol{\beta}\|_1$). Ambas funciones son convexas en $\boldsymbol{\beta}$, pero en condiciones de alta dimensionalidad ($\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$) no es estrictamente convexa por lo que la función de pérdida asociada a la regresión Lasso tampoco lo es. Por lo tanto, su mínimo no es único. Sin embargo, el conjunto de soluciones de un problema de minimización convexo es convexo, así que en caso de existir varios mínimos de la función de pérdida de Lasso, estos constituyen un conjunto convexo de mínimos. Así, si $\hat{\boldsymbol{\beta}}_a^{lso}$ y $\hat{\boldsymbol{\beta}}_b^{lso}$ son estimadores Lasso, también lo son sus combinaciones convexas:

$$(1-\theta)\,\hat{\boldsymbol{\beta}}_a^{lso} + \theta\hat{\boldsymbol{\beta}}_b^{lso}, \quad \forall\theta \in (0,1)$$

La no unicidad del estimador Lasso es un punto en contra y plantea problemas prácticos. No obstante, el siguiente lema asegura que bajo ciertas condiciones en la matriz de diseño, este estimador puede tener unicidad.

Lema 2.1. *Si los elementos de $\mathbf{X} \in \mathcal{M}^{n,m}$ se obtienen a partir de una distribución continua de probabilidad en $\mathbb{R}^{n\times p}$, entonces para cualquier $\mathbf{Y}$ y $\lambda > 0$, el problema de minimización (2.5) tiene solución única con probabilidad uno. Esta solución tiene, como máximo, $\min\{n, m\}$ componentes no nulos.*

La demostración de este lema la podemos encontrar en [Tib12].

2.5.2. Variantes de la regresión Lasso

Esta subsección se basa en el punto 3.8 de la fuente [TH09] y en la fuente [Wikb]

Existen modificaciones de la regresión Lasso para adaptarlo a problemas particulares. Estas variantes se basan en la dependencia que puedan tener varias covariables entre sí.

El *group Lasso* nos da la posibilidad de seleccionar grupos de predictores de forma conjunta, como una unidad. Esto es útil en situaciones donde no tendría sentido incluir ciertas covariantes y descartar otras con las que están muy correlacionadas.

Por otra parte, el *fused* Lasso es útil en situaciones donde la estructura del problema está relacionada con series temporales, por ejemplo.

Por último, cabe mencionar el método *elastic net*, que estudiaremos más tarde en detalle. Este método está especialmente diseñado para el caso $m > n$, pues ni la regresión Ridge, ni la regresión Lasso clásica se comportan bien aquí.

2.5.2.1. Group Lasso

El método *group* Lasso permite seleccionar o descartar grupos de predictores definidos previamente. Esto es especialmente útil en situaciones donde una variable categórica se representa como una colección de predictores binarios.

Por ejemplo, supongamos que estamos realizando un estudio genómico para identificar genes relacionados con una cierta enfermedad. En este caso, cada gen se puede representar como una covariable. Sin embargo, sabemos que los genes suelen trabajar en grupos o están involucrados en los mismos procesos biológicos.

Usar el *group* Lasso nos permitiría tener en cuenta que la actividad biológica suele ser el resultado de una interacción compleja entre varios genes y así seleccionar grupos de genes que trabajan juntos.

La función objetivo de este método es una generalización de la función objetivo del método Lasso estándar:

$$\hat{\boldsymbol{\beta}}^{glso} = \arg\min_{\boldsymbol{\beta}} \left\{ \left\| Y - \sum_{j=1}^{J} \mathbf{X}_j \beta_j \right\|_2^2 + \lambda \sum_{j=1}^{J} \left\| \beta_j \right\| K_j \right\}, \quad \|z\| K_j = \left(z^t K_j z \right)^{1/2}$$

donde la matriz de diseño $\mathbf{X}$ y el vector de coeficientes $\boldsymbol{\beta}$ se sustituyen por una colección de matrices de diseño $\mathbf{X}_j$ y de vectores de coeficientes $\boldsymbol{\beta}_j$, uno por cada uno de los J grupos. El término que define la penalización, $\lambda \sum_{j=1}^{J} \left\| \beta_j \right\| K_j$, es una suma de las norma euclídeas de los grupos de coeficientes multiplicadas por las matrices definidas positivas K_j. Si cada predictor tiene su propio grupo (es decir, no agrupamos) y $K_j = \mathbf{I}$, entonces tenemos el método Lasso estándar, mientras que si tenemos un único grupo y $K_1 = \mathbf{I}$ estaríamos en el caso de la regresión Ridge. Como el término de penalización se define con la norma euclídea en los subespacios definidos por cada agrupación, no se pueden seleccionar solo algunas de las covariables del grupo, al igual que no se podría en la regresión Ridge.

Como vemos, la penalización es la suma sobre las distintas normas de los subespacios y, como en Lasso, la restricción tiene puntos no diferenciables que corresponden a los conjuntos de predictores que se anularan, por tanto el mecanismo es el mismo que usa Lasso: regulariza algunos coeficentes de los grupos y anula otros.

Este método se puede extender y seleccionar predictores individuales dentro de cada grupo añadiendo una penalización adicional l^1 a cada subespacio.

2.5.2.2. Fused Lasso

El método fused Lasso es una variante de la regresión Lasso utilizada para abordar situaciones en las que se espera que los coeficientes de las variables adyacentes estén relacionados o tengan algún patrón de proximidad.

Un ejemplo concreto podría ser el análisis y predicción de series temporales donde cabría esperar que las observaciones consecutivas muestren una cierta tendencia conjunta. Aquí, el fused Lasso será útil para penalizar diferencias entre coeficientes adyacentes, favoreciendo soluciones donde los cambios en los coeficientes estimados sean suaves.

$$\min_{\beta} \left\{ \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{m} X_{ij}\beta_j \right)^2 + \lambda_1 \sum_{j=1}^{m} |\beta_j| + \lambda_2 \sum_{j=1}^{m-1} |\beta_{j+1} - \beta_j| \right\}$$

Esta función objetivo es estrictamente convexa respecto a $\boldsymbol{\beta}$, por tanto existe una única solución. El primer término de penalización ejecuta la selección de covariables y el segundo penaliza los cambios bruscos en la serie temporal, lo que fuerza a los coeficientes a variar de forma suave para reflejar el comportamiento de la serie.

El término de penalización $\lambda_2 \sum_{j=1}^{m-1} |\beta_{j+1} - \beta_j|$ asume que la variable índice t está uniformemente espaciada. Si eso no ocurriera, podemos usar una generalización de esta penalización basada en las diferencias divididas:

$$\lambda_2 \sum_{j=1}^{m-1} \frac{|\beta_{j+1} - \beta_j|}{|t_{j+1} - tj|}$$

2.5.3. Regresión Lasso: Conclusiones

En esta subsección nos basamos en en el punto 1 de la fuente [HZ05].

Dada la naturaleza de su penalización con norma L^1, Lasso permite la regularización continua y la selección de variables, algo imposible usando la regresión Ridge. Sin embargo, Lasso tiene algunas limitaciones:

1. En el caso $m > n$, el método Lasso selecciona como máximo n variables, debido a la naturaleza de optimización convexa de su función objetivo. Esta es una característica limitante para un método de selección de variables.

2. Si hay un grupo de variables altamente correlacionadas entre sí, Lasso selecciona solo una de las variables.

3. Para las situaciones $n > m$, si existen predictores altamente correlacionados entre sí, se ha demostrado que la capacidad predictiva de Lasso es peor que la de Ridge.

Los puntos 1. y 2. indican que Lasso podría no ser el método de selección de variables más óptimo en algunas situaciones. De hecho, si $m >> n$ y en las situaciones de predictores agrupados donde Lasso selecciona como máximo n variables de m candidatos, su precisión predictiva es mala.

Por tanto, debemos investigar otros métodos que mejoren la capacidad predictiva de Lasso. Es decir, encontrar un nuevo método que funcione bien en las situaciones en las que Lasso funciona bien y además pueda resolver los problemas que se generan en las situaciones 1. y 2. mencionadas previamente. Además, debería tener más precisión en la predicción que Lasso en el tercer escenario.

2.6. Elastic net

Para redactar esta sección nos basamos en el artículo [HZ05].

En esta sección, se propone un nuevo método de regularización que pretende mejorar los dos métodos previamente mencionados, Ridge y Lasso: la regularización *elastic net*. Al igual que Lasso, el método elastic net realiza simultáneamente selección de variables y regularización. Además, tiene la capacidad de seleccionar conjuntos de variables correlacionadas.

Primero, definiremos el *naïve elastic net*, que es un método de mínimos cuadrados penalizados con una nueva penalización. Más tarde, veremos como el método naïve regulariza en exceso e introduciremos el método elastic net que resuelve este problema.

2.6.1. Naïve elastic net

2.6.1.1. Definición de la penalización naïve elastic net

Sea un conjunto de datos de n observaciones y m variables independientes. Sea $\mathbf{Y} = (Y_1, ..., Y_n)^t$ el vector de respuestas y $\mathbf{X}$ la matriz de diseño. Suponiendo que los predictores están estandarizados, podemos definir para cada λ_1 y λ_2 no negativos el criterio naïve elastic net:

$$L(\lambda_1, \lambda_2, \boldsymbol{\beta}) = \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_2 \|\boldsymbol{\beta}\|_2^2 + \lambda_1 \|\boldsymbol{\beta}\|_1$$

Es decir, el estimador naïve elastic net $\hat{\boldsymbol{\beta}}_{nEN}$ cumple

$$\hat{\boldsymbol{\beta}}_{nEN} = \arg\min_{\boldsymbol{\beta}} L(\lambda_1, \lambda_2, \boldsymbol{\beta})$$

Lo que, de nuevo, es un método de mínimos cuadrados penalizado. Siendo

$$\alpha = \frac{\lambda_2}{\lambda_1 + \lambda_2}$$

Obtenemos la formulación equivalente en forma de problema de optimización:

$$\hat{\boldsymbol{\beta}}_{nEN} = \begin{array}{c} \arg\min_{\boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 \\ \text{sujeto a } \alpha \|\boldsymbol{\beta}\|_2^2 + (1-\alpha)\|\boldsymbol{\beta}\|_1 \leq t \quad \text{para cierto } t \end{array}$$

La función $\alpha \|\boldsymbol{\beta}\|_2^2 + (1-\alpha)\|\boldsymbol{\beta}\|_1$ es la penalización elastic net. Esta penalización es una combinación convexa de las penalizaciones de Lasso y Ridge. Si $\alpha = 1$, el naïve elastic net es una regresión Ridge. Para todo $\alpha \in [0,1)$, la penalización elastic net es no derivable en o y estrictamente convexa para $\alpha > 0$. Para el caso $\alpha = 0$, obtenemos la regresión Lasso que es convexa pero no estrictamente convexa.

2.6.1.2. Solución del problema naïve elastic net

Lema 2.2. *Dado un par* $(\mathbf{Y}, \mathbf{X})$ *y* (λ_1, λ_2)*, definimos el par* $(\mathbf{Y}^*, \mathbf{X}^*)$*:*

$$\mathbf{X}^* = (1+\lambda_2)^{-1/2} \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda_2}\mathbf{I} \end{pmatrix}, \quad \mathbf{Y}^* = \begin{pmatrix} \mathbf{Y} \\ 0 \end{pmatrix} \tag{2.6.1.2}$$

Sean

$$\gamma = \frac{\lambda_1}{\sqrt{1+\lambda_2}}, \quad \boldsymbol{\beta}^* = \sqrt{1+\lambda_2}\boldsymbol{\beta}$$

Entonces el criterio naïve elastic net se puede expresar como

$$L(\gamma, \boldsymbol{\beta}^*) = \|\mathbf{Y}^* - \mathbf{X}^*\boldsymbol{\beta}^*\|_2^2 + \gamma \|\boldsymbol{\beta}^*\|_1$$

Obteniendo

$$\hat{\boldsymbol{\beta}}^* = \arg\min_{\boldsymbol{\beta}^*} \quad L\{(\gamma, \boldsymbol{\beta}^*)\}$$

Entonces

$$\hat{\boldsymbol{\beta}}_{nEN} = \frac{1}{\sqrt{1+\lambda_2}}\hat{\boldsymbol{\beta}}^*$$

Este lema nos indica que podemos transformar el problem naïve elastic net en un problema Lasso equivalente. Notemos que el tamaño de muestra del nuevo problema es $n+m$ y que $\mathbf{X}^*$ tiene rango m. Esto significa que el método naïve podría seleccionar las m variables explicativas. Con esta propiedad conseguimos superar la primera de las tres desventajas del método Lasso que hemos mencionado antes. Además, este lema muestra como el método naïve elastic net puede seleccionar variables de forma similiar al método Lasso.

Una vez que hemos superado una de las limitaciones del método Lasso, veamos como el método naïve elastic net tiene la capacidad de seleccionar variables agrupadas, una propiedad que no presenta el método Lasso.

2.6.1.3. Agrupamiento de variables mediante naïve elastic net

En las situaciones $m >> n$, las variables agrupadas son una preocupación recurrente a la que se le han dado distintas posibles soluciones como el análisis de componentes principales

o el *clustering*. Si consideramos los métodos de penalización de forma genérica:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda J(\beta) \tag{2.6.1.3}$$

con $J(\cdot)$ una función positiva para $\boldsymbol{\beta} \neq 0$. Podemos decir que un método de regresión presenta un comportamiento de agrupamiento de variables cuando los coeficientes de regresión de un grupo de variables altamente correlacionas son muy similares o iguales en valor absoluto (si la correlación es negativa los coeficientes serán similares salvo el signo). Particularmente, si dos variables son idénticas, los coefientes de regresión también deberían serlo.

Lema 2.3. *Sean* $\mathbf{X}_i = \mathbf{X}_j, \quad i, j \in \{1, ..., m\}$.

- *Si* $J(\cdot)$ *es estrictamente convexa, entonces* $\hat{\beta}_i = \hat{\beta}_j, \quad \forall \lambda > 0$.

- *Si* $J(\boldsymbol{\beta}) = \|\boldsymbol{\beta}\|_1$, *entonces* $\hat{\beta}_i \hat{\beta}_j \geq 0$ *y* $\hat{\boldsymbol{\beta}}^*$ *es otro mínimo de la ecuación (2.6.1.3), donde*

$$\hat{\beta}_k^* = \begin{cases} \hat{\beta}_k, & si \quad k \neq i \quad y \quad k \neq j, \\ (\hat{\beta}_i + \hat{\beta}_j) \cdot (s) & si \quad k = i \\ (\hat{\beta}_i + \hat{\beta}_j) \cdot (1 - s) & si \quad k = j \end{cases}$$

  Para cualquier $s \in [0, 1]$.

Este lema muestra la distinción entre las funciones de penalización estrictamente convexas y la penalización Lasso. La convexidad estricta garantiza el agrupamiento en caso de existir predictores idénticos. Sin embargo, la regresión Lasso ni siquiera tiene una solución única. La penalización elastic net es estrictamente convexa si $\lambda_2 > 0$, lo que nos da la propiedad deseada.

Teorema 2.3. *Dados los datos* $(\mathbf{Y}, \mathbf{X})$ *y los parámetros* (λ_1, λ_2) *con el vector de respuestas* $\mathbf{Y}$ *centrado y los predictores* $\mathbf{X}$ *estandarizados. Sea* $\hat{\boldsymbol{\beta}}_{nEN}(\lambda_1, \lambda_2)$ *el estimador naïve elastic net. Supongamos que* $\hat{\beta}_{nEN_i}(\lambda_1, \lambda_2)\hat{\beta}_{nEN_j}(\lambda_1, \lambda_2) > 0$. *Se define*

$$D_{\lambda_1,\lambda_2}(i, j) = \frac{1}{\|\mathbf{Y}\|_1} \left| \hat{\beta}_{nEN_i}(\lambda_1, \lambda_2) - \hat{\beta}_{nEN_j}(\lambda_1, \lambda_2) \right|$$

Entonces

$$D_{\lambda_1,\lambda_2}(i, j) \leq \frac{1}{\lambda_2} \sqrt{2(1 - \rho)}$$

Donde $\rho = X_i^t X_j$ *es la correlación muestral.*

$D_{\lambda_1,\lambda_2}(i, j)$ es una cantidad adimensional que mide la distancia entre los predictores estimados $\hat{\beta}_{nEN_i}(\lambda_1, \lambda_2)$ y $\hat{\beta}_{nEN_j}(\lambda_1, \lambda_2)$. Si X_i y X_j están altamente correlacionados (es decir, si $\rho = 1$), la diferencia entre ambos predictores estimados debe ser o. Si $\rho = -1$, se considera $-X_j$. Esta cota superior de $D_{\lambda_1,\lambda_2}(i, j)$ es una representación del efecto de agrupamiento del naïve elastic net.

2.6.2. Elastic net

2.6.3. Deficiencias del método naïve elastic net

Como acabamos de ver, el método naïve elastic net supera las limitaciones del método Lasso en los casos (1.) y (2.) de la sección 2.5.3. No obstante, la evidencia empírica (que no trataremos en este documento) demuestra que el método naïve no se comporta satisfactoriamente a no ser que esté muy cerca de la regresión Ridge o del método Lasso. Es por esto por lo que se le llama 'naïve'.

El estimador naïve elastic net se obtiene en un proceso de dos etapas:

1. Para cada λ_2, se obtienen los coefientes de la regresión Ridge.
2. A estos coeficientes se les aplica la regularización Lasso.

Este proceso nos produce una 'doble regularización' que no regulariza lo suficiente los coeficientes y además añade un sesgo excesivo e innecesario, en comparación con los métodos Ridge y Lasso simples. Veamos como podemos mejorar las predicciones de método naïve corrigiendo esta doble regularización.

2.6.4. El estimador elastic net

Dado el par $(\mathbf{Y}, \mathbf{X})$, los parámetros de penalización (λ_1, λ_2) y el par $(\mathbf{Y}^*, \mathbf{X}^*)$ tal y como lo hemos definido en (2.6.1.2) el naïve elastic net resuelve un problema de tipo Lasso:

$$\hat{\boldsymbol{\beta}}^* = \arg\min_{\boldsymbol{\beta}^*} \|\mathbf{Y}^* - \mathbf{X}^*\boldsymbol{\beta}^*\|_2^2 + \frac{\lambda_1}{\sqrt{1+\lambda_2}} \|\boldsymbol{\beta}^*\|_1$$

El estimador elastic net (corregido) $\hat{\boldsymbol{\beta}}_{EN}$ se define como

$$\hat{\boldsymbol{\beta}}_{EN} = \sqrt{1+\lambda_2}\hat{\boldsymbol{\beta}}^*$$

Si recordamos que $\hat{\boldsymbol{\beta}}_{nEN} = \frac{1}{\sqrt{1+\lambda_2}}\hat{\boldsymbol{\beta}}^*$, entonces obtenemos la relación entre el estimador elastic net y naïve elastic net:

$$\hat{\boldsymbol{\beta}}_{EN} = (1+\lambda_2)\hat{\boldsymbol{\beta}}_{nEN}$$

Por lo tanto, los coeficientes elastic net son coeficientes naïve elastic net reescalados. Dicho cambio de escala preserva la propiedad de selección de variables del naïve elastic net y es el método más sencillo para deshacer la regularización excesiva. Así, todas las propiedades positivas del naïve elastic net que habíamos descrito previamente se mantienen en el método elastic net.

Se ha demostrado empíricamente que el estimador elastic net se comporta muy bien en comparación el estimador obtenido mediante el método Lasso o el método Ridge.

3 Aplicación con datos reales

En este capítulo desarrollaremos las técnicas expuestas a lo largo del documento de forma práctica usando el paquete de software estadístico R. Para ello nos valdremos de dos librerías principalmente: *olsrr* y *glmnet*.

La librería *olsrr* la usaremos para construir modelos de mínimos cuadrados clásicos a partir de las técnicas de selección de variables *forward selection, backward elimination* y regresión *stepwise*.

Por otra parte, la libería *glmnet* la utilizaremos para implementar las regresiones Ridge, Lasso y *Elastic Net*.

3.1. Breve descripción del conjunto de datos

https://www.kaggle.com/datasets/jamieleech/boston-housing-dataset

Los datos que usaremos se han obtenido de *Kaggle*, una sitio web dedicado a la ciencia de datos y el *machine learning*. El *dataset* cuenta con 506 observaciones de 13 predictores y una variable respuesta. Los predictores son los siguientes:

1. CRIM: Crimen per cápita en la localidad.
2. ZN: Fracción de la superficie del terreno en esa zona destinada a construcción residencial en lotes de gran tamaño.
3. INDUS: Proporción de negocios distintos de venta al por menor en la localidad.
4. CHAS: Variable que toma el valor 1 si la propiedad limita con el río Charles y 0 si no lo hace.
5. NOX: concentración de óxido nítrico en la localidad (en partes por cada 10 millones).
6. RM: número medio de habitaciones por vivienda.
7. AGE: Proporción de viviendas de la localidad construidas antes del 1940.
8. DIS: La distancia promedio a cinco centros de trabajo de Boston.
9. RAD: Índice de accesibilidad a las circunvalaciones.
10. TAX: Ratio de impuestos a la propiedad por cada 10.000 dólares.
11. PTRATIO: Ratio del número de alumnos por profesor en cada localidad.
12. B: $1000(Bk - 0.63)^2$ donde Bk es la proporción de población de mínorias sociales de Estados Unidos.

13. LSTAT: Porcentaje de la población por debajo del umbral de la pobreza.

La variable que trataremos de predecir es $MEDV$. Este es el valor mediano de las propiedades en miles de dólares.

3.2. Metodología del análisis

El análisis que vamos a realizar consta de varias etapas que nos permitirán realizar predicciones y comparar entre sí los modelos que obtengamos:

1. **Análisis exploratorio**: Primero, veremos el tipo de datos que estamos usando o la influencia de distintas variables individuales respecto a la respuesta.
2. **Separar el conjunto de datos en datos de entrenamiento y de testeo**: Sobre los datos de entrenamiento ajustaremos los modelos y los compararemos posteriormente usando los datos de testeo.
3. **Aplicación de modelos en el set de entrenamiento y elección de parámetros**: En este paso usaremos los métodos de selección y regularización sobre el set de entrenamiento comparando los resultados obtenidos con distintos parámetros.
4. **Comparación de modelos**: Una vez hemos seleccionado los parámetros que optimizan el comportamiento del modelo, comparemos los modelos optimizados en el set de testeo.

3.3. Análisis exploratorio

Primero, veamos superficialmente la distribución de cada una de las variables. En la imagen de abajo podemos ver el mínimo, el máximo, la media y los cuartiles de todas las variables. Vemos que hay variables muy heterogéneas como el índice de criminalidad de la localidad que va desde 0 hasta 89 con media de 3,61. Por el contrario, podemos ver que la distribución del óxido nítrico es bastante regular.

```
      CRIM                ZN              INDUS            CHAS              NOX             RM           AGE              DIS       
 Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000   Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
 1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000   1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100  
 Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000   Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207  
 Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917   Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795  
 3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000   3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188  
 Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000   Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127  
      RAD              TAX            PTRATIO             B              LSTAT            MEDV      
 Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32   Min.   : 1.73   Min.   : 5.00  
 1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38   1st Qu.: 6.95   1st Qu.:17.02  
 Median : 5.000   Median :330.0   Median :19.05   Median :391.44   Median :11.36   Median :21.20  
 Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67   Mean   :12.65   Mean   :22.53  
 3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23   3rd Qu.:16.95   3rd Qu.:25.00  
 Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90   Max.   :37.97   Max.   :50.00  
```

Figura 3.1: Resumen del conjunto de datos

Antes de aplicar modelos lineales, conviene conocer si el *dataset* tiene valores no disponibles. En caso afirmativo habría que estudiar de que manera manejar este problema, si sustituir estos valores no disponibles (NA, del inglés *Not Available*) por otros como por ejemplo la media de los valores disponibles o simplemente eliminar estas observaciones. En este caso, nada de esto será necesario porque ninguna de las variables contiene valores no disponibles:

CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX	PTRATIO	B	LSTAT	MEDV
0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figura 3.2: Número de NAs por variable

La respuesta sobre la que pretendemos aplicar modelos de regresión es el precio mediano de las viviendas de la zona en miles de dólares. Esta variable se distribuye según nos muestra el siguiente histograma:

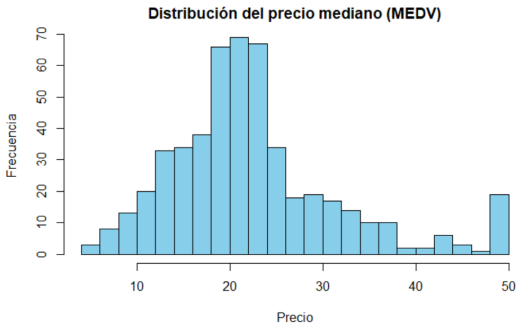


Figura 3.3: Histograma de la variable MEDV

También podemos visualizar fácilmente la correlación entre distintas variables usando la matriz de correlaciones coloreada. Esta matriz toma tonos azul oscuro cuando el coeficiente de correlación entre dos variables es igual a uno y tonos rojos oscuros en el caso de que la correlación sea negativa:

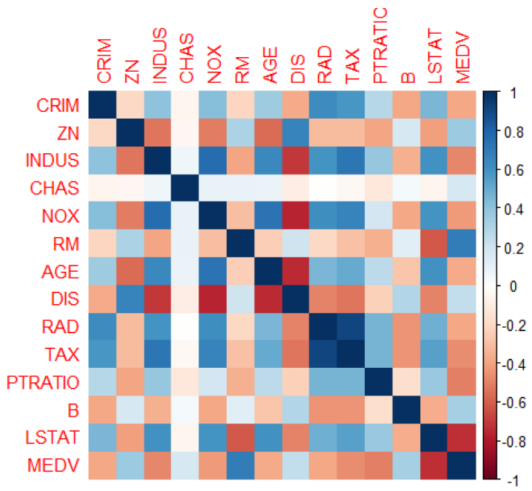


Figura 3.4: Matriz de correlaciones

Como podemos observar en la matriz de correlaciones, la variable con mayor correlación en valor absoluto con el precio mediano es *LSTAT*. Ya dijimos que dicha variable mide el porcentaje de la población de la localidad por debajo del umbral de la pobreza. Intuitivamente, tiene sentido que ambas variables estén negativamente correlacionadas.

3.4. Separar los datos en datos de entrenamiento y de testeo

Este paso es clave para poder comparar posteriormente los modelos obtenidos. Realizaremos un *split* de los datos con el que tomaremos el 80 % de ellos como datos de entrenamiento. Será aquí donde ajustemos los parámetros.

En el caso de los métodos de selección de variables como *backward elimination, forward selection* y *stepwise*, es necesario definir niveles de significación de entrada o salida de cada variable del modelo. Para ello utilizaremos la validación cruzada: definiremos varios subconjuntos de datos de entrenamiento y aplicaremos los métodos de selección de variables a cada subconjunto con un cierto nivel de significación de entrada o salida según corresponda. El modelo que optimice los criterios de selección será el que usaremos en el set de testeo.

En los métodos de regularización, el parámetro a optimizar es la variable de penalización λ. El procedimiento será parecido al que hemos explicado previamente para los métodos de selección de variables: Se dividirá el set de entrenamiento en varios subconjuntos y se probarán distintas penalizaciones.

Primero, estandarizaremos los datos, esto no es estrictamente necesario para el modelo lineal general pero sí lo es para aplicar métodos de regularización. Para evaluar los sucesivos modelos en este análisis utilizaremos tres criterios: el coeficiente de determinación, el criterio AIC y el Error Cuadrático Medio (ECM).

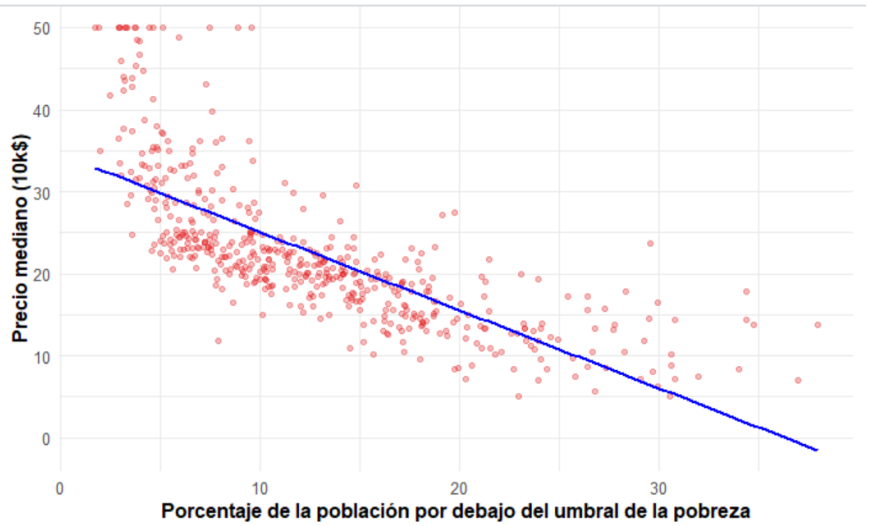
Así, con un *split* del 80 % de los datos obtenemos un conjunto de entrenamiento de 407 observaciones y un set de testeo de 99 observaciones.

3.5. Entrenamiento de modelos

3.5.1. Modelo de regresión lineal simple

Después de observar como la variable con mayor correlación con el precio mediano es *LSTAT*, podemos tomar como modelo de referencia o modelo base el modelo que explica $MEDV$ a partir de *LSTAT* exclusivamente.

Así, ajustamos el modelo usando los datos de entrenamiento y obtenemos el siguiente gráfico:



A primera vista, existe una tendencia clara en las observaciones que el modelo consigue reflejar. Ahora, aplicamos el modelo al set de testeo y obtenemos los criterios que usaremos para compararlo con futuros modelos:

Resultados de Evaluación del Modelo

Error Cuadrático Medio	Coeficiente de Determinación	Criterio AIC
6.503817	0.513163	2636.747

El criterio AIC de momento no es útil, será necesario compararlo con el de otros modelos Recordemos que el mejor modelo según el criterio AIC es el que minimiza dicho criterio.

3.5.2. Modelo de regresión lineal múltiple

Una vez hemos establecido el modelo base, vamos a intentar mejorar nuestra capacidad predictiva con otro modelo. En este caso usaremos todos los predictores que tenemos disponibles. Aquí, ya tenemos uno de los problemas que habíamos mencionado previamente: la difícil explicabilidad del modelo. Es evidente que en este caso no podemos explorar la relación lineal entre los predictores y las respuestas de una forma tan sencilla como en el modelo anterior, donde sólo teníamos un predictor. Repetimos el proceso: ajustamos el modelo en el set de entrenamiento y obtenemos los siguientes coeficientes:

```
Coefficients:
              Estimate Std. Error t value             Pr(>|t|)
(Intercept)  37.733617   5.619935   6.714   0.0000000000663462
CRIM         -0.093857   0.039157  -2.397             0.016999
ZN            0.039436   0.015987   2.467             0.014062
INDUS        -0.012988   0.069595  -0.187             0.852059
CHAS          2.290187   0.940621   2.435             0.015346
NOX         -17.130560   4.342272  -3.945   0.0009944788182833
RM            3.499219   0.451445   7.751   0.0000000000000787
AGE           0.009823   0.015510   0.633             0.526905
DIS          -1.390769   0.230614  -6.031   0.0000000037679571
RAD           0.330939   0.077135   4.290   0.0000224836384194
TAX          -0.012386   0.004342  -2.852             0.004568
PTRATIO      -0.960676   0.150307  -6.391   0.0000000004660371
B             0.009841   0.002935   3.353             0.000877
LSTAT        -0.562095   0.059180  -9.498 < 0.0000000000000002
```

Obteniendo que en el modelo de regresión lineal múltiple mejora al modelo lineal simple en todas las métricas que habíamos definido previamente:

Resultados de Evaluación del Modelo

Error Cuadrático Medio	**Coeficiente de Determinación**	**Criterio AIC**
4.588948	0.761126	2447.721

De hecho, incluso el criterio AIC que tiene en cuenta la complejidad del modelo es menor en este caso que en el anterior.

3.5.3. Método de selección de variables *stepwise*

Para aplicar el método *stepwise*, tenemos que optimizar dos parámetros: α_{salir} y α_{entrar}. Optimizar, por tanto, este modelo puede ser un proceso computacionalmente costoso. En nuestro caso vamos a tomar 5 posibles α_{salir} y otros 5 posibles α_{entrar}. A su vez, usaremos 5 subconjuntos de entrenamiento sobre el que probaremos las distintas combinaciones $(\alpha_{entrar}, \alpha_{salir})$. En particular:

- $\alpha_{salir} = (0.05, 0.01, 0.005, 0.001, 0.0005)$
- $\alpha_{entrar} = (0.05, 0.01, 0.005, 0.001, 0.0005)$

Por ejemplo, para el par $(\alpha_{entrar}, \alpha_{salir}) = (0.001, 0.0005)$ el proceso ha sido el siguiente:

```
[1] "p-value remove: 0.0005"
[1] "p-value enter: 0.001"

                          Stepwise Selection Summary
-----------------------------------------------------------------------------------------
                    Added/                   Adj.
Step    Variable    Removed     R-Square     R-Square      C(p)        AIC         RMSE
-----------------------------------------------------------------------------------------
   1      LSTAT     addition       0.552        0.551    260.2910   2636.7466    6.1435
   2       RM       addition       0.635        0.634    138.9790   2555.0333    5.5500
   3    PTRATIO     addition       0.675        0.673     81.8250   2509.8103    5.2437
   4       DIS      addition       0.688        0.685     64.9400   2495.5022    5.1460
   5       NOX      addition       0.704        0.700     43.9190   2476.6964    5.0224
   6        B       addition       0.711        0.707     34.3870   2467.8601    4.9622
   7        B       removal        0.704        0.700     43.9190   2476.6964    5.0224
   8      CHAS      addition       0.710        0.705     36.6490   2470.0091    4.9753
   9      CHAS      removal        0.704        0.700     43.9190   2476.6964    5.0224
  10        ZN      addition       0.706        0.702     42.3040   2475.3306    5.0079
  11        ZN      removal        0.704        0.700     43.9190   2476.6964    5.0224
  12       RAD      addition       0.706        0.702     42.4540   2475.4714    5.0088
  13       RAD      removal        0.704        0.700     43.9190   2476.6964    5.0224
-----------------------------------------------------------------------------------------
```

Y el proceso ha terminado con la selección de cinco variables. Como vemos, las variables $B, CHAS, ZN$ y RAD entraron al modelo y en la siguiente iteración salieron. Aunque este

ejemplo es representativo para entender como funciona el método *stepwise*, no ha sido en el que se ha alcanzado un menor AIC. El menor AIC, que es uno de los criterios de selección de modelos que mencionamos la sección 1.13, se obtiene con el par $(\alpha_{entrar}, \alpha_{salir}) = (0.001, 0.05)$.

```
[1] "p-value remove: 0.05"
[1] "p-value enter: 0.001"

                        Stepwise Selection Summary
-----------------------------------------------------------------------------------------
                    Added/                   Adj.
Step    Variable    Removed     R-Square     R-Square     C(p)         AIC         RMSE
-----------------------------------------------------------------------------------------
   1     LSTAT      addition       0.552        0.551    260.2910    2636.7466    6.1435
   2      RM        addition       0.635        0.634    138.9790    2555.0333    5.5500
   3    PTRATIO     addition       0.675        0.673     81.8250    2509.8103    5.2437
   4     DIS        addition       0.688        0.685     64.9400    2495.5022    5.1460
   5     NOX        addition       0.704        0.700     43.9190    2476.6964    5.0224
   6      B         addition       0.711        0.707     34.3870    2467.8601    4.9622
   7     CHAS       addition       0.717        0.712     28.2800    2462.0658    4.9211
   8     RAD        addition       0.722        0.717     22.2990    2456.2438    4.8801
   9     TAX        addition       0.727        0.721     17.2530    2451.2110    4.8442
  10      ZN        addition       0.730        0.724     14.2530    2448.1456    4.8203
  11     CRIM       addition       0.734        0.727     10.4360    2444.1728    4.7911
-----------------------------------------------------------------------------------------
```

En este caso, el método de selección ha introducido las once variables que vemos arriba. Si probamos este modelo en el set de testeo, obtenemos:

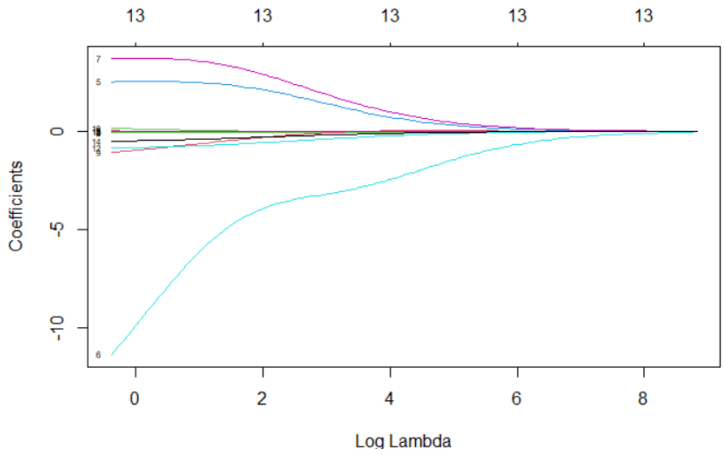
Resultados de Evaluación del Modelo

Error Cuadrático Medio	**Coeficiente de Determinación**	**Criterio AIC**
4.560599	0.764644	2444.173

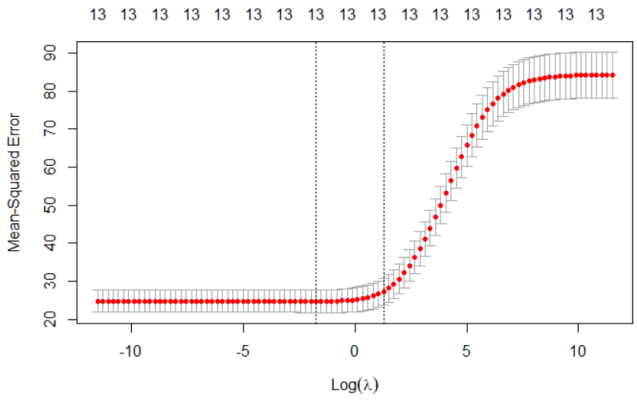
Que mejora ligeramente los resultados del modelo que tenía en cuenta todos los predictores.

3.5.4. Regresión Ridge

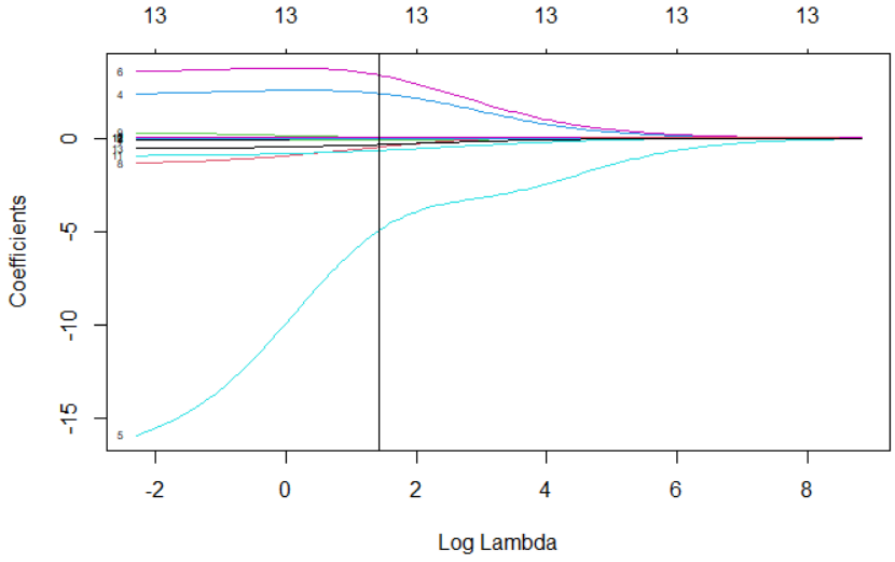
Ahora, usemos el método Ridge para ajustar un nuevo modelo multivariante. Para ello usaremos la librería *glmnet*. En este caso, la normalización se realiza automáticamente usando la función *glmnet* y posteriormente se devuelven los coeficientes en su dimensión original. En el gráfico inferior podemos ver como evolucionan los coeficientes del modelo en función del valor de λ. Cuánto mayor es λ, mayor regularización y por tanto menores coeficientes.



Ahora, seleccionamos λ usando el método que introdujimos en la sección 2.4.5: *leave-one-out cross-validation*. El λ óptimo será aquel que minimice el Error Cuadrático Medio. En el gráfico inferior podemos ver como se comporta el ECM en función del valor de λ. Las dos líneas verticales que se muestran en el gráfico muestran dos posibles valores óptimos para λ. El motivo por el que aparecen dos valores óptimos es debido a que nosotros pretendemos minimizar el ECM, pero también realizar cierto nivel de regularización y obtener menores coeficientes. Podría ocurrir que el valor que minimiza e ECM estuviese en el extremo izquierdo del eje horizontal, es decir, que el valor que optimizase el ECM es un λ excesivamente pequeño que no realiza ninguna regularización. Por tanto, se considera también un valor λ mayor, que nos asegure a la vez minimización del ECM y regularización.



Ambos valores de λ son en este caso 0.06892612 y 3.593814, respectivamente. Para este ejemplo, utilizaremos el mayor de ellos puesto que realizará una mayor regularización y será útil a efectos de visualizarla gráficamente. En el gráfico inferior, podemos ver en particular los valores aproximados que tomarán los coeficientes para dicho valor λ.



En particular, los coeficientes obtenidos con el método Ridge son los siguientes:

```
 (Intercept)         CRIM           ZN        INDUS         CHAS          NOX           RM          AGE          DIS          RAD          TAX
22.021803285 -0.064084939  0.017199999 -0.080356502  2.426803351 -5.345105109  3.453730640 -0.005697804 -0.530496973  0.044467267 -0.002654900
     PTRATIO            B        LSTAT
-0.686849514  0.008188415 -0.366660664
```

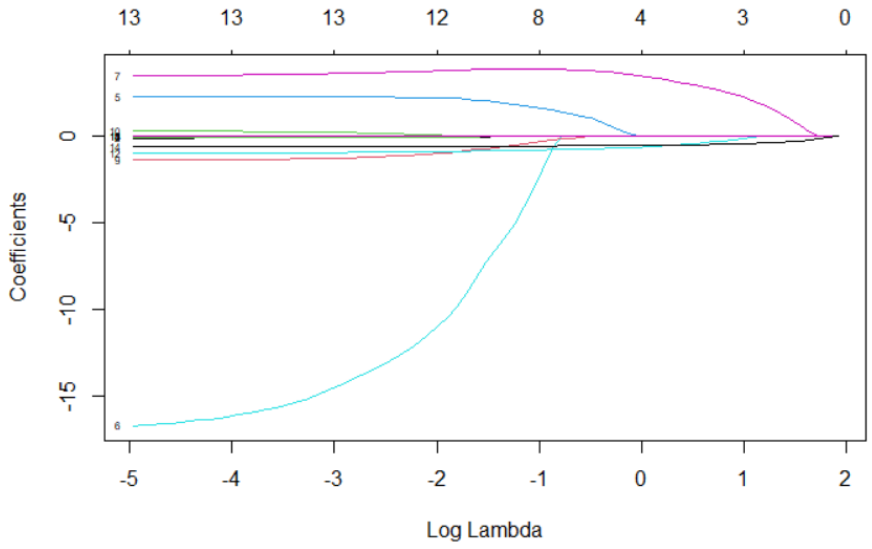
Y finalmente, obtenemos las métricas de este modelo simplemente multiplicando la matriz de diseño del set de testeo por los coeficientes regularizados, obteniendo: **REVISAR LAS METRICAS**

Resultados de Evaluación del Modelo

Error Cuadrático Medio	**Coeficiente de Determinación**	**Criterio AIC**
3.985399	0.739404	1783.982

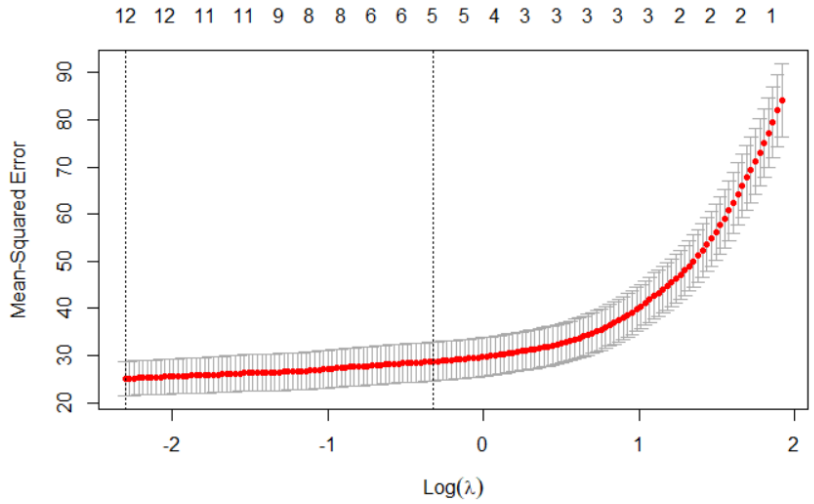
3.5.5. Regresión Lasso

El proceso para la regresión Lasso es equivalente. Sin embargo, el resultado es muy diferente a la regresión Ridge. Como podemos ver en el gráfico inferior, ahora los coeficientes no se mueven de forma suave con el valor λ sino que algunos de ellos acaban tendiendo abruptamente a cero, anulándose y por tanto descartando esas variables.

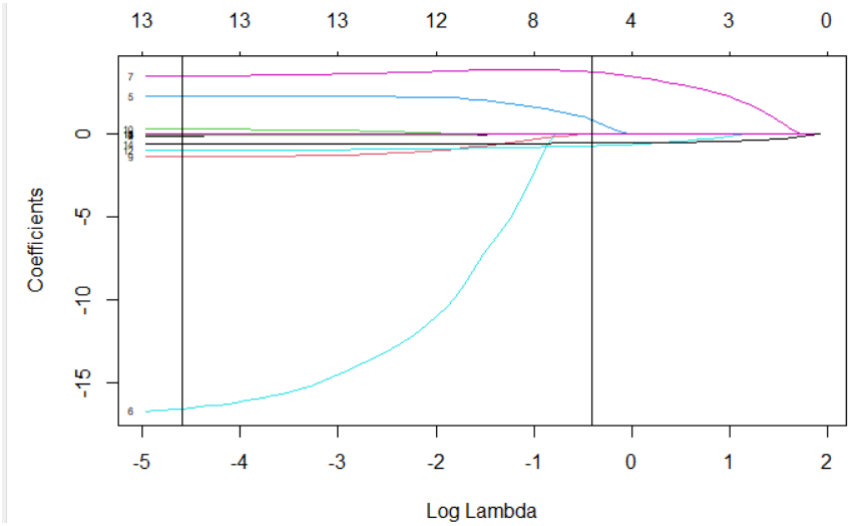


El gráfico inferior ilustra de nuevo la validación cruzada para elegir λ. Sin embargo, en este caso el valor óptimo está en uno de los extremos, concretamente el extremo izquierdo. Es decir, el valor óptimo según el criterio del Error Cuadrático Mínimo implicaría la ausencia de regularización. Para evitarlo, volvemos a tomar un valor que está a una desviación típica

del λ que minimiza el ECM. En este caso, $\lambda = 0.6647195$.



Ahora, en el gráfico inferior podemos ver aproximadamente el valor de los coeficientes para ambos λ. Como explicábamos previamente, el λ menor no realiza ninguna regularización ni selección de variables. Por el contrario, con $\lambda = 0.6647195$, vemos claramente como algunos de los coeficientes incluso se han anulado.



Así, los coeficiente obtenidos son:

Y evaluando el modelo.

(Intercept)	CRIM	ZN	INDUS	CHAS	NOX	RM	AGE	DIS	RAD	TAX
16.545421045	0.000000000	0.000000000	0.000000000	0.868555282	0.000000000	3.764394315	0.000000000	0.000000000	0.000000000	0.000000000
PTRATIO	B	LSTAT								
-0.697942205	0.004795069	-0.516317757								

Resultados de Evaluación del Modelo

Error Cuadrático Medio	**Coeficiente de Determinación**	**Criterio AIC**
4.27344	0.689566	1485.352

Bibliografía

[ACR07] G. Bruce Schaalje Alvin C. Rencher. *LINEAR MODELS IN STATISTICS*. Wiley, 2007.

[GP23] Eduardo García-Portugués. Notes for predictive modeling, oct 2023.

[HZ05] Trevor Hastie Hui Zou. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society.*, 67(2), 2005.

[Jor22] JoramSoch. Proof: Maximum likelihood estimator of variance in multiple linear regression is biased, dec 2022. Recurso online, accedido el 27 de octubre del 2023.

[Neu14] Dirk Neumann. Ridge regression and lasso, 2014. Recurso online, accedido el 14 de Noviembre del 2023.

[Tab21] Marco Taboga. Ridge regression, 2021. Recurso online, accedido el 1 de Noviembre del 2023.

[THo9] Jerome Friedman Trevor Hastie, Robert Tibshirani. *The Elements of Statistical Learning*. Springer, second edición, 2009.

[The74] C. M. Theobald. Generalizations of mean square error applied to ridge regression, 1974.

[Tib12] Ryan J. Tibshirani. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 2012.

[vW15] Wessel N. van Wieringen. Lecture notes on ridge regression. https://arxiv.org/pdf/1509.09169.pdf, 2015.

[Wika] Wikipedia. Gauss–markov theorem. Recurso online, accedido el 13 de octubre del 2023.

[Wikb] Wikipedia. Lasso (statistics). Recurso online, accedido el 27 de Noviembre del 2023.