

# Regresión Lineal, Ridge y Lasso

D. Sierra-Porta

26 de febrero de 2026

## Resumen

Estas notas presentan una introducción teórica y práctica a la regresión lineal ordinaria (OLS), la regresión Ridge y la regresión Lasso, con énfasis en su formulación matemática e interpretación en problemas de aprendizaje supervisado para variables continuas.

El objetivo principal es mostrar, de manera gradual, cómo la regresión lineal surge como un problema de minimización del error cuadrático y cómo Ridge y Lasso extienden este enfoque mediante términos de regularización que controlan la magnitud de los coeficientes. Se desarrollan las funciones objetivo de cada método y se derivan, en el caso de un predictor, expresiones útiles para comprender el papel del parámetro de regularización  $\lambda$ .

Además, se discuten diferencias conceptuales clave entre OLS, Ridge y Lasso (ajuste, encogimiento de coeficientes, estabilidad e interpretabilidad), junto con observaciones prácticas como la estandarización de variables y la no penalización del intercepto. Finalmente, se incluye un ejemplo numérico sencillo, pensado para trabajo en clase y pizarra, que permite comparar los tres métodos de forma transparente.

## 1. Introducción

En problemas de aprendizaje supervisado para variables continuas, una de las primeras preguntas es cómo construir un modelo que permita *predecir* una respuesta  $y$  a partir de una o varias variables explicativas  $x_1, \dots, x_p$ . La familia de los *modelos lineales* ocupa un lugar central en esta tarea porque ofrece una combinación muy valiosa de simplicidad, interpretabilidad y control matemático. En particular, la regresión lineal ordinaria (OLS, *Ordinary Least Squares*) suele ser el punto de partida natural para introducir ideas fundamentales de modelado, estimación y optimización.

La idea básica de la regresión lineal es proponer una relación funcional simple entre variables:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p,$$

donde los parámetros  $\beta_0, \beta_1, \dots, \beta_p$  se estiman a partir de datos. En OLS, esta estimación se obtiene minimizando la suma de errores cuadráticos:

$$\min_{\beta_0, \beta_1, \dots, \beta_p} \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

Esta formulación convierte el problema de estimación en un problema de *minimización* (u optimización), lo cual es una idea transversal en Machine Learning: entrenar un modelo consiste, en gran medida, en definir una función objetivo y encontrar los parámetros que la hacen lo más pequeña posible.

Entonces, ¿por qué existen varios tipos de regresión si ya tenemos OLS? La respuesta corta es que, aunque OLS es un método fundamental, no siempre es suficiente para obtener modelos estables o con buena capacidad de generalización. En la práctica, pueden aparecer situaciones como:

- variables altamente correlacionadas (*multicolinealidad*),
- muchas variables en relación con el número de observaciones,
- coeficientes estimados con gran variabilidad,
- buen ajuste en entrenamiento pero mal desempeño en datos nuevos (*sobreajuste*).

Estas dificultades motivan la introducción de métodos de *regularización*, que modifican la función objetivo original agregando un término de penalización sobre los coeficientes del modelo. La idea general es:

$$\min_{\beta} \underbrace{\text{error de ajuste}}_{\text{qué tan bien explica los datos}} + \underbrace{\text{penalización}}_{\text{qué tan complejo es el modelo}}.$$

De esta forma, el problema ya no consiste únicamente en ajustar lo mejor posible los datos observados, sino también en controlar la complejidad del modelo para favorecer estabilidad e interpretabilidad.

En este contexto aparecen, entre otros, dos métodos muy importantes:

- **Ridge regression:** agrega una penalización cuadrática ( $\ell_2$ ) sobre los coeficientes, lo que tiende a reducir su magnitud de forma continua.
- **Lasso regression:** agrega una penalización absoluta ( $\ell_1$ ), que además de encoger coeficientes puede anular algunos de ellos, produciendo modelos más simples.

Desde el punto de vista conceptual, OLS, Ridge y Lasso no son métodos desconectados, sino variantes de una misma idea: estimar parámetros lineales mediante optimización, con distintos criterios para balancear *ajuste* y *complejidad*. Esta perspectiva unificada es especialmente útil en un curso introductorio de Machine Learning, porque permite relacionar temas de estadística, cálculo y optimización dentro de un mismo marco.

Estas notas de clase se enfocan en el caso univariado (un predictor) para presentar de manera clara la formulación matemática de los tres métodos, sus diferencias principales y el papel del parámetro de regularización  $\lambda$ . Aunque el tratamiento será introductorio, se buscará mantener una justificación matemática suficiente para entender no solo *cómo* se aplican estos métodos, sino también *por qué* funcionan.

## 2. Regresión lineal ordinaria (OLS)

### 2.1. Motivación y formulación del problema

La regresión lineal ordinaria (OLS, *Ordinary Least Squares*) es uno de los modelos más importantes en estadística y aprendizaje automático, no solo por su utilidad práctica, sino también porque introduce con claridad varias ideas fundamentales que reaparecen en métodos más avanzados: formulación de modelos, definición de una función de pérdida y estimación de parámetros mediante optimización.

En el caso univariado, se busca modelar una variable respuesta continua y a partir de un único predictor  $x$  mediante una relación lineal de la forma

$$\hat{y}_i = \beta_0 + \beta_1 x_i,$$

donde  $\beta_0$  representa el intercepto (nivel base del modelo) y  $\beta_1$  representa la pendiente (cambio esperado en la respuesta por unidad de cambio en  $x$ ). Aunque esta forma es sencilla, ya permite discutir de manera rigurosa la lógica de la predicción supervisada: proponer una familia de funciones y luego elegir, dentro de esa familia, los parámetros que mejor se ajustan a los datos observados.

Para cuantificar qué significa “ajustar bien”, se introduce el residuo de la observación  $i$ :

$$e_i = y_i - \hat{y}_i = y_i - (\beta_0 + \beta_1 x_i).$$

La estimación OLS consiste en escoger  $\beta_0$  y  $\beta_1$  minimizando la suma de cuadrados de los residuos:

$$\min_{\beta_0, \beta_1} \sum_{i=1}^n e_i^2 = \min_{\beta_0, \beta_1} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2.$$

Definimos entonces la función objetivo

$$J(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2.$$

Esta formulación muestra una idea clave: en OLS, “entrenar” el modelo equivale a resolver un problema de minimización. En otras palabras, el método no adivina los parámetros, sino que los obtiene como solución de un problema de optimización bien definido.

La elección del error cuadrático tiene además varias ventajas. Penaliza con mayor fuerza errores grandes, produce una función diferenciable (lo que facilita el cálculo) y conduce, en el caso lineal, a expresiones cerradas para los parámetros. Por estas razones, OLS suele ser el punto de partida natural antes de introducir variantes regularizadas como Ridge y Lasso.

## 2.2. Derivación de las ecuaciones normales y solución cerrada (caso univariado)

Para encontrar el mínimo de  $J(\beta_0, \beta_1)$ , derivamos respecto a  $\beta_0$  y  $\beta_1$ , e imponemos las condiciones de primer orden. Este procedimiento produce las llamadas *ecuaciones normales*, que caracterizan el estimador OLS en el caso de un predictor.

Partimos de

$$J(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2.$$

Derivando respecto a  $\beta_0$ :

$$\frac{\partial J}{\partial \beta_0} = -2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0.$$

Al reorganizar términos,

$$\sum_{i=1}^n y_i - n\beta_0 - \beta_1 \sum_{i=1}^n x_i = 0,$$

por lo que se obtiene

$$\boxed{n\beta_0 + \beta_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i.} \quad (E1)$$

Derivando ahora respecto a  $\beta_1$ :

$$\frac{\partial J}{\partial \beta_1} = -2 \sum_{i=1}^n x_i (y_i - \beta_0 - \beta_1 x_i) = 0.$$

Equivalentemente,

$$\sum_{i=1}^n x_i y_i - \beta_0 \sum_{i=1}^n x_i - \beta_1 \sum_{i=1}^n x_i^2 = 0,$$

y por tanto

$$\boxed{\beta_0 \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i.} \quad (E2)$$

Las ecuaciones (E1) y (E2) forman un sistema lineal en  $\beta_0$  y  $\beta_1$ . Al resolverlo (por eliminación o sustitución), se obtiene la solución cerrada del problema OLS:

$$\boxed{\beta_1 = \frac{n \sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i) (\sum_{i=1}^n y_i)}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}}$$

y luego

$$\boxed{\beta_0 = \frac{\sum_{i=1}^n y_i - \beta_1 \sum_{i=1}^n x_i}{n} = \bar{y} - \beta_1 \bar{x}.$$

Estas expresiones son especialmente útiles porque permiten calcular los parámetros directamente a partir de sumas simples de los datos. Además, muestran con claridad que la pendiente  $\beta_1$  depende de cómo varían conjuntamente  $x$  y  $y$ , mientras que el intercepto ajusta el nivel del modelo para que la recta pase por el punto  $(\bar{x}, \bar{y})$ .

Una condición importante para que la fórmula de  $\beta_1$  sea válida es que el denominador no sea cero:

$$n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2 \neq 0.$$

Esto equivale a exigir que los valores de  $x_i$  no sean todos iguales. Si no hay variación en  $x$ , no existe información suficiente para estimar una pendiente.

Como referencia conceptual, la pendiente también puede escribirse en forma centrada:

$$\beta_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \beta_0 = \bar{y} - \beta_1 \bar{x}.$$

Esta forma es algebraicamente equivalente y pone en evidencia la relación entre covariación y pendiente. Sin embargo, en estas notas se mantiene la formulación sin centrar como expresión principal por su continuidad con la derivación directa de las ecuaciones normales.

### 2.3. Interpretación, alcance del método y ejemplo numérico

Una vez obtenidos  $\beta_0$  y  $\beta_1$ , el modelo OLS queda completamente determinado. Su interpretación básica es simple pero muy importante:

- $\beta_0$  representa la predicción del modelo cuando  $x = 0$  (si este valor tiene sentido en el contexto del problema);
- $\beta_1$  representa el cambio promedio estimado en  $y$  por cada unidad adicional de  $x$ .

Por ejemplo, si  $\beta_1 = 0.9$ , el modelo sugiere que, en promedio, un incremento de una unidad en  $x$  se asocia con un incremento de 0.9 unidades en  $y$ . Esta interpretabilidad directa es una de las razones por las que los modelos lineales siguen siendo una herramienta central, incluso cuando se dispone de métodos más complejos.

También es importante precisar qué garantiza OLS y qué no garantiza. OLS minimiza la suma de errores cuadrados *sobre los datos usados para estimar el modelo*. Por ello, dentro de la familia de rectas de la forma  $\beta_0 + \beta_1 x$ , OLS proporciona el mejor ajuste cuadrático en entrenamiento. Sin embargo, ese resultado no implica automáticamente que el modelo sea el mejor posible para datos nuevos. Esta distinción entre ajuste en entrenamiento y capacidad de generalización será una motivación central para introducir regularización en las secciones siguientes.

Consideremos ahora un ejemplo numérico sencillo:

$$x = \{1, 2, 3, 4\}, \quad y = \{2, 3, 3, 5\}.$$

Se calculan las cantidades básicas:

$$n = 4, \quad \sum x_i = 10, \quad \sum y_i = 13, \quad \sum x_i^2 = 30, \quad \sum x_i y_i = 37.$$

Sustituyendo en la fórmula de la pendiente:

$$\beta_1 = \frac{4(37) - 10(13)}{4(30) - 10^2} = \frac{148 - 130}{120 - 100} = \frac{18}{20} = 0.9.$$

Luego, para el intercepto:

$$\beta_0 = \frac{13 - 0.9(10)}{4} = \frac{13 - 9}{4} = 1.$$

Por tanto, el modelo OLS estimado es

$$\hat{y} = 1 + 0.9x.$$

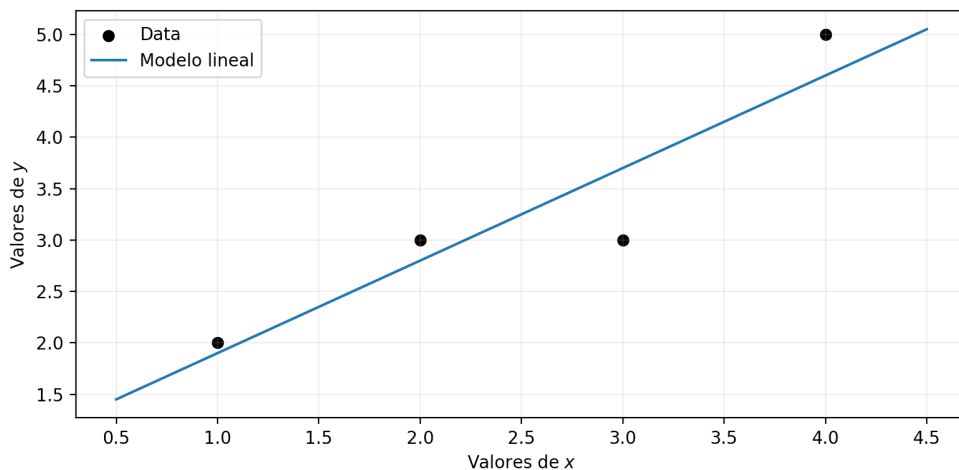


Figura 1: Inspección visual de la regresión simple mediante OLS del ejemplo anterior.

Este ejemplo ilustra la mecánica completa del método: definición del modelo, construcción de la función objetivo, obtención de las ecuaciones normales y cálculo explícito de la solución. A partir de esta base, es posible introducir Ridge y Lasso como modificaciones de la misma estructura de estimación, en las que se agrega una penalización para controlar la complejidad del modelo sin abandonar la formulación lineal.

### 3. Regresión Ridge

#### 3.1. Motivación y formulación del problema

La regresión Ridge puede entenderse como una extensión natural de la regresión lineal ordinaria (OLS) cuando se desea conservar la estructura lineal del modelo, pero incorporando un mecanismo explícito de control sobre la magnitud de los coeficientes. En otras palabras, Ridge mantiene la misma idea de predecir mediante una combinación lineal de variables, pero modifica el criterio de estimación para evitar pendientes excesivamente grandes o inestables.

En el caso univariado, el modelo sigue siendo

$$\hat{y}_i = \beta_0 + \beta_1 x_i,$$

de modo que no cambia la forma funcional del predictor. Lo que cambia es la función objetivo que se minimiza para estimar  $\beta_0$  y  $\beta_1$ . Mientras que OLS minimiza únicamente la suma de errores cuadrados, Ridge agrega un término de penalización sobre la pendiente. Con la convención que usaremos en estas notas, la función objetivo es:

$$J_R(\beta_0, \beta_1) = \frac{1}{2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 + \frac{\lambda}{2} \beta_1^2, \quad \lambda \geq 0.$$

Esta expresión permite visualizar con claridad los dos componentes que Ridge intenta balancear. El primer término mide qué tan bien se ajusta el modelo a los datos observados (error de ajuste), mientras que el segundo término penaliza pendientes de gran magnitud (control de complejidad). El parámetro  $\lambda$  regula la importancia relativa de dicha penalización:

- si  $\lambda = 0$ , el problema coincide con OLS;
- si  $\lambda > 0$ , la penalización comienza a influir y la pendiente tiende a reducirse;
- si  $\lambda$  crece, el modelo se vuelve progresivamente más conservador.

Es importante notar que en estas notas **no penalizamos el intercepto**  $\beta_0$ . Esta decisión es estándar en regresión regularizada, ya que el objetivo principal de la regularización es controlar la contribución de las variables explicativas al modelo, no desplazar artificialmente el nivel base de la predicción.

Desde un punto de vista más general, Ridge introduce una idea central en Machine Learning: el entrenamiento de un modelo no consiste solamente en minimizar error sobre datos observados, sino en definir una función objetivo que combine *ajuste* e *inductiva de complejidad*. En este sentido, Ridge es uno de los ejemplos más claros y accesibles de regularización.

#### 3.2. Derivación de las ecuaciones de Ridge y solución cerrada (caso univariado)

Partimos de la función objetivo

$$J_R(\beta_0, \beta_1) = \frac{1}{2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 + \frac{\lambda}{2} \beta_1^2.$$

La estrategia de estimación es la misma que en OLS: derivar respecto a los parámetros e imponer condiciones de primer orden. La diferencia está en que ahora, al derivar respecto a  $\beta_1$ , aparece el término adicional  $\lambda \beta_1$  proveniente de la penalización cuadrática.

Derivando respecto a  $\beta_0$ , se obtiene:

$$\frac{\partial J_R}{\partial \beta_0} = - \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0,$$

lo cual equivale a

$$\sum_{i=1}^n y_i - n\beta_0 - \beta_1 \sum_{i=1}^n x_i = 0,$$

y por tanto

$$\boxed{n\beta_0 + \beta_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i.} \quad (R1)$$

Esta primera ecuación coincide con la de OLS. La razón es simple: la penalización no depende de  $\beta_0$ , por lo que el comportamiento del intercepto se mantiene formalmente igual en la condición de primer orden.

Derivando ahora respecto a  $\beta_1$ , obtenemos:

$$\frac{\partial J_R}{\partial \beta_1} = - \sum_{i=1}^n x_i (y_i - \beta_0 - \beta_1 x_i) + \lambda \beta_1 = 0.$$

Al expandir y reorganizar términos:

$$\sum_{i=1}^n x_i y_i - \beta_0 \sum_{i=1}^n x_i - \beta_1 \sum_{i=1}^n x_i^2 + \lambda \beta_1 = 0,$$

de donde resulta

$$\beta_0 \sum_{i=1}^n x_i + \beta_1 \left( \sum_{i=1}^n x_i^2 + \lambda \right) = \sum_{i=1}^n x_i y_i. \quad (R2)$$

Las ecuaciones (R1) y (R2) forman el sistema lineal que caracteriza el estimador Ridge en el caso univariado. Comparadas con las ecuaciones normales de OLS, la diferencia estructural está en la aparición del término  $+\lambda$  junto a  $\sum x_i^2$ , lo cual anticipa el efecto de encogimiento sobre la pendiente.

Al resolver este sistema (eliminando  $\beta_0$ ), se obtiene la solución cerrada:

$$\beta_1^{\text{Ridge}} = \frac{n \sum_{i=1}^n x_i y_i - \left( \sum_{i=1}^n x_i \right) \left( \sum_{i=1}^n y_i \right)}{n \sum_{i=1}^n x_i^2 - \left( \sum_{i=1}^n x_i \right)^2 + n\lambda}$$

y luego

$$\beta_0^{\text{Ridge}} = \frac{\sum_{i=1}^n y_i - \beta_1^{\text{Ridge}} \sum_{i=1}^n x_i}{n} = \bar{y} - \beta_1^{\text{Ridge}} \bar{x}.$$

Una forma especialmente útil de interpretar esta expresión consiste en definir

$$A = n \sum x_i y_i - \left( \sum x_i \right) \left( \sum y_i \right), \quad D = n \sum x_i^2 - \left( \sum x_i \right)^2.$$

Con esta notación,

$$\beta_1^{\text{OLS}} = \frac{A}{D}, \quad \beta_1^{\text{Ridge}} = \frac{A}{D + n\lambda}.$$

Esta comparación hace visible el efecto de la regularización de manera inmediata: para  $\lambda > 0$ , el denominador de Ridge es mayor que el de OLS, y por ello la magnitud de la pendiente se reduce. En este sentido, Ridge no cambia la lógica básica de la estimación lineal, sino que la modifica suavemente añadiendo una restricción implícita sobre la escala del coeficiente.

Como referencia (y por conexión con formulaciones más compactas), la pendiente Ridge también puede escribirse en términos de variables centradas:

$$\beta_1^{\text{Ridge}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2 + \lambda}.$$

Sin embargo, en estas notas se prioriza la forma no centrada por su transparencia algebraica y su continuidad con la derivación de OLS presentada previamente.

### 3.3. Interpretación, comportamiento del parámetro $\lambda$ y ejemplo numérico

La característica distintiva de Ridge es que introduce un encogimiento *continuo* de los coeficientes. Esto significa que la pendiente estimada tiende a disminuir en magnitud a medida que crece  $\lambda$ , pero en general no se vuelve exactamente cero (a diferencia de Lasso). Por ello, Ridge suele interpretarse como un método que conserva todas las variables en el modelo, aunque moderando su influencia.

Desde el punto de vista del compromiso entre ajuste e interpretación, Ridge no busca necesariamente mejorar el error sobre los datos de entrenamiento; de hecho, para  $\lambda > 0$ , lo normal es que el ajuste cuadrático puro empeore respecto a OLS. Su valor está en que produce modelos menos sensibles a variaciones en los datos y, en contextos multivariados, más estables frente a colinealidad entre predictores.

También es útil considerar los casos límite. Si  $\lambda = 0$ , se recupera exactamente la solución OLS. Si  $\lambda \rightarrow \infty$ , entonces

$$\beta_1^{\text{Ridge}} \rightarrow 0,$$

y por consiguiente

$$\beta_0^{\text{Ridge}} \rightarrow \bar{y}.$$

En ese régimen, el modelo tiende a predecir una constante (aproximadamente la media de la respuesta), lo cual refleja una regularización extrema: se renuncia casi por completo a usar la variable  $x$ .

Para ilustrar el procedimiento, consideremos el conjunto de datos

$$x = \{1, 2, 3, 4\}, \quad y = \{2, 3, 3, 5\}, \quad \lambda = 1.$$

Se calculan las sumas:

$$n = 4, \quad \sum x_i = 10, \quad \sum y_i = 13, \quad \sum x_i^2 = 30, \quad \sum x_i y_i = 37.$$

Con ello,

$$A = 4(37) - 10(13) = 148 - 130 = 18, \quad D = 4(30) - 10^2 = 120 - 100 = 20.$$

La pendiente Ridge es

$$\beta_1^{\text{Ridge}} = \frac{A}{D + n\lambda} = \frac{18}{20 + 4(1)} = \frac{18}{24} = 0.75,$$

y el intercepto:

$$\beta_0^{\text{Ridge}} = \frac{13 - 0.75(10)}{4} = \frac{13 - 7.5}{4} = 1.375.$$

Por tanto, el modelo estimado es

$$\hat{y} = 1.375 + 0.75x.$$

Si se compara con OLS en el mismo conjunto de datos ( $\hat{y} = 1 + 0.9x$ ), se observa claramente el efecto de Ridge: la pendiente es menor y el modelo resulta más conservador. Esta diferencia resume la idea central de la regularización  $\ell_2$ : aceptar una pequeña pérdida de ajuste en entrenamiento a cambio de un mayor control sobre la complejidad del modelo.

En problemas con varias variables, Ridge penaliza típicamente

$$\sum_{j=1}^p \beta_j^2$$

(excluyendo  $\beta_0$ ). En ese contexto, una recomendación práctica importante es **estandarizar** las variables antes de aplicar Ridge, de modo que la penalización actúe de manera comparable entre predictores que pueden estar en escalas muy distintas.

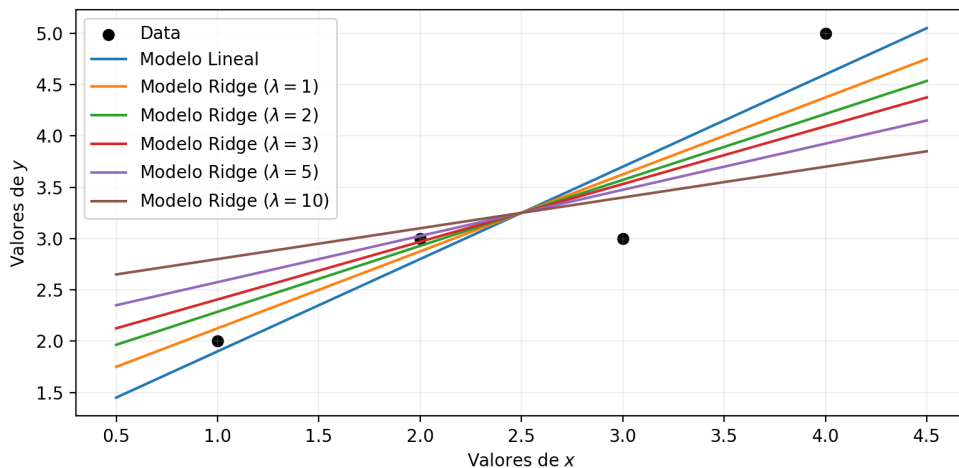


Figura 2: Inspección visual de la regresión Ridge del ejemplo anterior para  $\lambda$  igual a 1, 2, 3, 5 y 10; en comparación con el modelo lineal simple.

## 4. Regresión Lasso

### 4.1. Motivación y formulación del problema

La regresión Lasso (de *Least Absolute Shrinkage and Selection Operator*) es otra extensión de la regresión lineal ordinaria (OLS) dentro de la familia de métodos regularizados. Al igual que Ridge, Lasso mantiene la estructura lineal del modelo y modifica la función objetivo para incorporar un control explícito sobre la magnitud de los coeficientes. La diferencia principal está en el tipo de penalización utilizada: en lugar de una penalización cuadrática ( $\ell_2$ ), Lasso emplea una penalización basada en el valor absoluto ( $\ell_1$ ).

En el caso univariado, el modelo sigue siendo

$$\hat{y}_i = \beta_0 + \beta_1 x_i.$$

Por tanto, la forma del predictor no cambia; lo que cambia es el criterio con el cual se estiman  $\beta_0$  y  $\beta_1$ . Con la convención adoptada en estas notas, la función objetivo de Lasso es:

$$J_L(\beta_0, \beta_1) = \frac{1}{2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 + \lambda |\beta_1|, \quad \lambda \geq 0.$$

Esta formulación conserva el mismo término de error cuadrático que en OLS y Ridge, pero reemplaza la penalización cuadrática por una penalización absoluta. Conceptualmente, esto significa que Lasso también busca balancear ajuste y complejidad, pero lo hace con una geometría distinta de penalización, lo cual tiene consecuencias importantes sobre la solución.

Como en Ridge, en estas notas **no se penaliza el intercepto**  $\beta_0$ . Esta decisión permite que el nivel base del modelo se ajuste libremente a los datos, mientras que la regularización actúa sobre la pendiente  $\beta_1$ , que es el parámetro asociado al uso de la variable explicativa.

El parámetro  $\lambda$  controla la intensidad de la regularización:

- si  $\lambda = 0$ , se recupera OLS;
- si  $\lambda > 0$ , la pendiente comienza a encogerse;
- si  $\lambda$  es suficientemente grande, Lasso puede llevar la pendiente exactamente a cero.

Este último punto es la diferencia conceptual más importante frente a Ridge. Mientras Ridge tiende a reducir coeficientes de forma continua, Lasso puede producir soluciones *dispersas* (sparse), es decir, con algunos coeficientes exactamente nulos. En el caso univariado esto se observa como una pendiente igual a cero; en el caso multivariado, se interpreta como una forma de selección automática de variables.

### 4.2. Derivación de las condiciones de optimalidad y solución (caso univariado)

Partimos de la función objetivo:

$$J_L(\beta_0, \beta_1) = \frac{1}{2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 + \lambda |\beta_1|.$$

La derivación sigue la misma lógica general que en OLS y Ridge, con una diferencia técnica importante: el término  $|\beta_1|$  **no es derivable en**  $\beta_1 = 0$ . Por ello, para Lasso se trabaja con derivación por casos (cuando  $\beta_1 > 0$  o  $\beta_1 < 0$ ) y con la noción de subgradiente en  $\beta_1 = 0$ . Esta es una buena oportunidad para mostrar que no todos los problemas de optimización en Machine Learning se resuelven con derivadas ordinarias en todos los puntos.

Derivando respecto a  $\beta_0$ , la penalización no contribuye (porque no depende de  $\beta_0$ ), de modo que se obtiene la misma condición que en OLS y Ridge:

$$\frac{\partial J_L}{\partial \beta_0} = - \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0,$$

lo que equivale a

$$\boxed{n\beta_0 + \beta_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i.} \quad (L1)$$

Para  $\beta_1$ , la condición de optimalidad se expresa mediante subgradientes:

$$0 \in - \sum_{i=1}^n x_i (y_i - \beta_0 - \beta_1 x_i) + \lambda \partial |\beta_1|,$$

donde

$$\partial|\beta_1| = \begin{cases} \{1\}, & \beta_1 > 0, \\ [-1, 1], & \beta_1 = 0, \\ \{-1\}, & \beta_1 < 0. \end{cases}$$

Esta expresión resume el comportamiento de la penalización absoluta: fuera de cero, el término de regularización aporta una contribución constante ( $+\lambda$  o  $-\lambda$ ); en cero, admite un rango de valores, lo que hace posible que la solución se “pegue” exactamente a  $\beta_1 = 0$ .

Para obtener una expresión cerrada útil, es conveniente eliminar  $\beta_0$  usando (L1) y trabajar con las cantidades

$$A = n \sum x_i y_i - \left( \sum x_i \right) \left( \sum y_i \right), \quad D = n \sum x_i^2 - \left( \sum x_i \right)^2.$$

Con esta notación, la solución de Lasso en el caso univariado puede escribirse de forma compacta como

$$\beta_1^{\text{Lasso}} = \frac{\text{sign}(A) \max(|A| - n\lambda, 0)}{D}$$

y, una vez obtenida la pendiente,

$$\beta_0^{\text{Lasso}} = \frac{\sum_{i=1}^n y_i - \beta_1^{\text{Lasso}} \sum_{i=1}^n x_i}{n} = \bar{y} - \beta_1^{\text{Lasso}} \bar{x}.$$

La fórmula anterior muestra de manera explícita el mecanismo de Lasso. En lugar de dividir simplemente  $A$  por  $D$  (como en OLS) o por  $D + n\lambda$  (como en Ridge), Lasso primero aplica un *umbral* sobre  $A$ : si  $|A|$  no supera  $n\lambda$ , entonces el numerador se anula y la pendiente queda exactamente en cero. Este comportamiento se conoce como *soft-thresholding* (umbral suave).

Como referencia, en variables centradas la expresión equivalente es

$$\beta_1^{\text{Lasso}} = \text{sign}(S_{xy}) \max\left(\frac{|S_{xy}| - \lambda}{S_{xx}}, 0\right),$$

donde

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}), \quad S_{xx} = \sum_{i=1}^n (x_i - \bar{x})^2.$$

No obstante, en estas notas se mantiene la formulación sin centrar como expresión principal para conservar continuidad algebraica con las secciones anteriores.

#### 4.3. Interpretación, comparación con OLS/Ridge y ejemplo numérico

La regresión Lasso comparte con Ridge la idea de regularizar para controlar la complejidad del modelo, pero difiere en el efecto cualitativo sobre los coeficientes. Ridge produce un encogimiento continuo que normalmente no anula la pendiente; Lasso, en cambio, puede encoger y además fijar exactamente  $\beta_1 = 0$  cuando la señal asociada al predictor no es suficientemente fuerte frente a la penalización.

Desde el punto de vista de interpretación, este comportamiento hace de Lasso una herramienta especialmente atractiva cuando se desea favorecer modelos más simples. En el caso univariado, una pendiente nula significa que el modelo decide no usar la variable  $x$ , y la predicción se reduce a una constante. En el caso multivariado, este mismo principio se traduce en selección de variables: algunas quedan activas y otras son descartadas por el propio proceso de optimización.

También conviene destacar, como en Ridge, que Lasso no está diseñado para mejorar necesariamente el error cuadrático sobre el conjunto de entrenamiento. OLS sigue siendo el minimizador del error cuadrático puro en entrenamiento. La contribución de Lasso está en introducir una restricción efectiva sobre la complejidad del modelo que puede mejorar interpretabilidad y, en muchos escenarios, capacidad de generalización.

Para ilustrar el cálculo, consideremos el mismo conjunto de datos usado en las secciones anteriores:

$$x = \{1, 2, 3, 4\}, \quad y = \{2, 3, 3, 5\}, \quad \lambda = 1.$$

Se tiene:

$$n = 4, \quad \sum x_i = 10, \quad \sum y_i = 13, \quad \sum x_i^2 = 30, \quad \sum x_i y_i = 37.$$

Por tanto,

$$A = 4(37) - 10(13) = 148 - 130 = 18, \quad D = 4(30) - 10^2 = 120 - 100 = 20.$$

Aplicando la fórmula de Lasso:

$$\beta_1^{\text{Lasso}} = \frac{\text{sign}(18) \max(18 - 4(1), 0)}{20} = \frac{14}{20} = 0.7.$$

Luego,

$$\beta_0^{\text{Lasso}} = \frac{13 - 0.7(10)}{4} = \frac{13 - 7}{4} = 1.5.$$

Así, el modelo estimado es

$$\hat{y} = 1.5 + 0.7x.$$

Si se compara con los modelos obtenidos previamente,

$$\hat{y}_{\text{OLS}} = 1 + 0.9x, \quad \hat{y}_{\text{Ridge}} = 1.375 + 0.75x, \quad \hat{y}_{\text{Lasso}} = 1.5 + 0.7x,$$

se observa que Lasso produce, en este caso, una pendiente aún más pequeña que Ridge para el valor de  $\lambda$  elegido. Esto ilustra el efecto de encogimiento de la penalización  $\ell_1$ . Sin embargo, la característica más distintiva de Lasso aparece al aumentar la regularización.

Por ejemplo, si en este mismo conjunto de datos se toma  $\lambda = 5$ , entonces

$$|A| - n\lambda = 18 - 4(5) = 18 - 20 = -2 < 0,$$

y por tanto

$$\beta_1^{\text{Lasso}} = 0.$$

En consecuencia,

$$\beta_0^{\text{Lasso}} = \bar{y} = \frac{13}{4} = 3.25,$$

y el modelo se reduce a

$$\hat{y} = 3.25.$$

Este resultado resume de manera muy clara la lógica de Lasso: si la evidencia a favor de una pendiente distinta de cero no supera el umbral impuesto por la penalización, el método prefiere un modelo más simple. Esa propiedad explica su importancia en problemas modernos de aprendizaje automático y análisis de datos, especialmente cuando se trabaja con múltiples predictores y se desea combinar predicción con interpretabilidad.

Comentario práctico (anticipando el caso multivariable). En problemas con varias variables, Lasso penaliza típicamente

$$\sum_{j=1}^p |\beta_j|$$

(excluyendo  $\beta_0$ ). Al igual que en Ridge, es muy recomendable **estandarizar** las variables antes de aplicar Lasso, para que la penalización actúe de manera comparable entre predictores con escalas distintas.

## 5. Comparación entre OLS, Ridge y Lasso

### 5.1. Marco común y diferencias conceptuales

Las tres regresiones estudiadas en estas notas (OLS, Ridge y Lasso) comparten una misma estructura básica: todas parten de un *modelo lineal* para predecir una respuesta continua a partir de uno o más predictores. En el caso univariado, esta estructura común es

$$\hat{y}_i = \beta_0 + \beta_1 x_i.$$

Esto significa que la diferencia entre los métodos no está en la forma del predictor, sino en el *criterio de estimación* usado para obtener los parámetros.

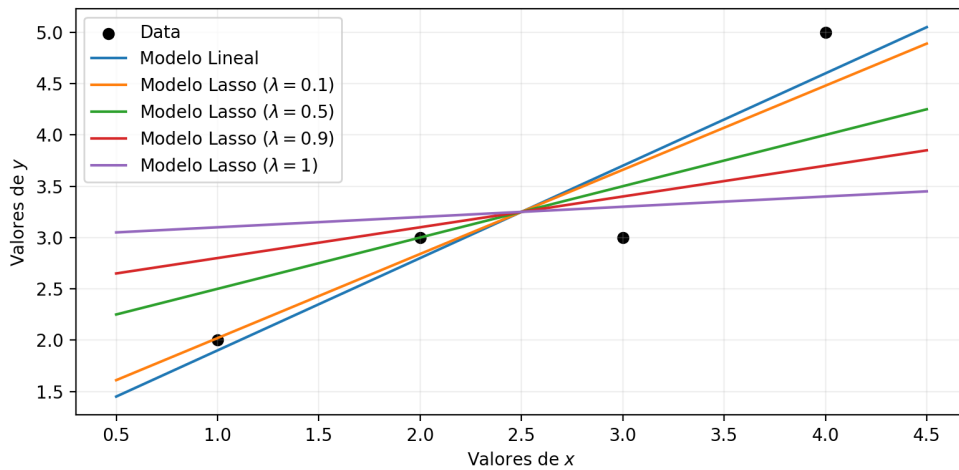


Figura 3: Inspección visual de la regresión Lasso del ejemplo anterior para  $\lambda$  igual a 0.1, 0.5, 0.9 y 1; en comparación con el modelo lineal simple.

Desde esta perspectiva, resulta útil entender OLS, Ridge y Lasso como tres variantes de una misma idea general de aprendizaje:

entrenar un modelo  $\iff$  definir una función objetivo y minimizarla.

Lo que cambia entre los métodos es la función objetivo, es decir, la manera en que se balancean dos aspectos:

- **ajuste a los datos** (qué tan pequeños son los errores de predicción),
- **complejidad del modelo** (qué tan grandes son los coeficientes).

En OLS, solo se minimiza el error cuadrático:

$$J_{\text{OLS}}(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2.$$

En Ridge, se mantiene el error cuadrático pero se añade una penalización  $\ell_2$ :

$$J_{\text{Ridge}}(\beta_0, \beta_1) = \frac{1}{2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 + \frac{\lambda}{2} \beta_1^2.$$

En Lasso, la penalización es  $\ell_1$ :

$$J_{\text{Lasso}}(\beta_0, \beta_1) = \frac{1}{2} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 + \lambda |\beta_1|.$$

Esta comparación permite identificar una diferencia conceptual importante:

- **OLS** prioriza exclusivamente el ajuste en entrenamiento.
- **Ridge** acepta una pequeña pérdida de ajuste para obtener coeficientes más estables y de menor magnitud.
- **Lasso** también regulariza, pero además puede producir coeficientes exactamente nulos, favoreciendo modelos más simples e interpretables.

En términos de comportamiento de la pendiente  $\beta_1$ , la diferencia entre Ridge y Lasso es especialmente relevante:

- Ridge produce un *encogimiento continuo*: la pendiente disminuye en magnitud, pero usualmente no se hace exactamente cero.
- Lasso produce un *encogimiento con umbral*: si la señal es débil frente a la penalización, la pendiente puede anularse.

Esta distinción es la base de una interpretación muy útil en problemas multivariables: Ridge tiende a conservar todas las variables (con pesos moderados), mientras que Lasso puede actuar como un mecanismo de selección de variables.

## 5.2. Comparación matemática y numérica en el caso univariado

Una forma compacta de comparar los tres métodos en el caso univariado (sin centrar) consiste en usar las cantidades

$$A = n \sum x_i y_i - \left( \sum x_i \right) \left( \sum y_i \right), \quad D = n \sum x_i^2 - \left( \sum x_i \right)^2.$$

Con esta notación, las pendientes estimadas se escriben como:

$$\beta_1^{\text{OLS}} = \frac{A}{D}, \quad \beta_1^{\text{Ridge}} = \frac{A}{D + n\lambda}, \quad \beta_1^{\text{Lasso}} = \frac{\text{sign}(A) \max(|A| - n\lambda, 0)}{D}.$$

Estas tres expresiones muestran, de forma muy clara, cómo se modifica la estimación:

- OLS usa directamente la “señal”  $A$  y la variabilidad del predictor  $D$ .
- Ridge aumenta el denominador en  $n\lambda$ , reduciendo la pendiente.
- Lasso aplica primero un umbral a  $A$  y solo después divide por  $D$ .

Por su parte, el intercepto se obtiene en los tres casos (con la convención de no penalizar  $\beta_0$ ) mediante

$$\beta_0 = \bar{y} - \beta_1 \bar{x}.$$

Esto resalta que la diferencia esencial entre los métodos se concentra en la estimación de la pendiente (o, en el caso multivariado, en los coeficientes de las variables explicativas).

Ejemplo comparativo (mismo conjunto de datos). Consideremos nuevamente

$$x = \{1, 2, 3, 4\}, \quad y = \{2, 3, 3, 5\},$$

para el cual se obtuvo:

$$n = 4, \quad \sum x_i = 10, \quad \sum y_i = 13, \quad \sum x_i^2 = 30, \quad \sum x_i y_i = 37,$$

y por tanto

$$A = 18, \quad D = 20.$$

Tomando  $\lambda = 1$ , los modelos estimados fueron:

$$\hat{y}_{\text{OLS}} = 1 + 0.9x, \quad \hat{y}_{\text{Ridge}} = 1.375 + 0.75x, \quad \hat{y}_{\text{Lasso}} = 1.5 + 0.7x.$$

La comparación muestra el patrón esperado:

- OLS produce la pendiente más grande (mejor ajuste cuadrático en entrenamiento).
- Ridge reduce la pendiente por regularización  $\ell_2$ .
- Lasso también reduce la pendiente y, para este valor de  $\lambda$ , la encoge aún más que Ridge.

Si se evalúa el error cuadrático sobre el mismo conjunto de entrenamiento (SSE), se obtiene:

$$\text{SSE}_{\text{OLS}} = 0.70, \quad \text{SSE}_{\text{Ridge}} = 0.8125, \quad \text{SSE}_{\text{Lasso}} = 0.90.$$

Este resultado es consistente con la teoría: OLS debe tener el menor SSE en entrenamiento porque fue definido precisamente para minimizar ese criterio. Ridge y Lasso, en cambio, optimizan funciones objetivo distintas, que incluyen penalización.

Este punto es didácticamente muy importante: **no tiene sentido comparar métodos usando un criterio que no coincide con el que cada método optimiza sin aclarar esa diferencia**. OLS “gana” en SSE de entrenamiento, pero Ridge y Lasso introducen un objetivo más amplio que combina ajuste y control de complejidad.

Para enfatizar esta idea, nótese que con  $\lambda = 1$ :

$$J_{\text{Ridge}}(\beta_0, \beta_1) = \frac{1}{2} \text{SSE} + \frac{1}{2} \lambda \beta_1^2, \quad J_{\text{Lasso}}(\beta_0, \beta_1) = \frac{1}{2} \text{SSE} + \lambda |\beta_1|.$$

Por construcción, la solución Ridge minimiza  $J_{\text{Ridge}}$ , y la solución Lasso minimiza  $J_{\text{Lasso}}$ , aunque sus SSE sean mayores que el de OLS.

Resumen comparativo. La Tabla 1 sintetiza las diferencias principales en el caso univariado.

Cuadro 1: Comparación conceptual y matemática entre OLS, Ridge y Lasso (caso univariado).

Método	Función objetivo	Efecto sobre $\beta_1$	Propiedad distintiva
OLS	$\sum (y_i - \beta_0 - \beta_1 x_i)^2$	Sin penalización	Minimiza SSE en entrenamiento
Ridge	$\frac{1}{2} \sum (y_i - \beta_0 - \beta_1 x_i)^2 + \frac{\lambda}{2} \beta_1^2$	Encogimiento continuo	Mayor estabilidad (regularización $\ell_2$ )
Lasso	$\frac{1}{2} \sum (y_i - \beta_0 - \beta_1 x_i)^2 + \lambda  \beta_1 $	Encogimiento con umbral	Puede anular coeficientes (regularización $\ell_1$ )

### 5.3. Criterios de uso, ventajas y consideraciones prácticas

La pregunta “¿qué regresión es mejor?” no tiene una respuesta única fuera de contexto. La elección entre OLS, Ridge y Lasso depende del objetivo del análisis, del tamaño y calidad de los datos, del número de predictores y del balance deseado entre ajuste, estabilidad e interpretabilidad.

Si el interés principal es construir una línea base simple y entender con claridad la relación entre variables, OLS es una excelente primera opción. Su formulación es directa, sus parámetros tienen interpretación inmediata y constituye el punto de partida natural para introducir ideas de estimación y optimización.

Ridge resulta particularmente útil cuando se sospecha que las estimaciones pueden ser inestables, especialmente en problemas con varios predictores y colinealidad. En esos escenarios, la penalización  $\ell_2$  ayuda a reducir la varianza de los coeficientes y suele producir modelos más robustos. Aunque en entrenamiento puede tener un SSE mayor que OLS, su valor suele apreciarse al evaluar desempeño en datos no vistos.

Lasso es especialmente atractivo cuando, además de predecir, se desea simplificar el modelo. Su capacidad para llevar coeficientes exactamente a cero lo convierte en una herramienta útil para obtener modelos más parsimoniosos y potencialmente más interpretables. En contextos con muchos predictores, esta propiedad puede ayudar a identificar subconjuntos relevantes de variables, aunque la selección final siempre debe interpretarse con cuidado y en diálogo con el conocimiento del problema.

Una consideración práctica crucial para Ridge y Lasso (y, en general, para modelos regularizados) es la **estandarización de variables**. Cuando los predictores están en escalas diferentes, la penalización no actúa de manera comparable sobre todos los coeficientes, lo que puede distorsionar la regularización. Por esta razón, en aplicaciones reales con múltiples variables, suele ser recomendable centrar y escalar los predictores antes del ajuste.

Finalmente, el parámetro  $\lambda$  no debe verse como una constante arbitraria, sino como un hiperparámetro que controla el compromiso entre ajuste y complejidad. En estas notas se ha fijado  $\lambda$  para mostrar el efecto de la regularización de forma transparente; en aplicaciones reales, su elección suele apoyarse en procedimientos de validación (por ejemplo, validación cruzada), precisamente porque el objetivo final no es solo ajustar el conjunto de entrenamiento, sino generalizar adecuadamente a nuevos datos.

Cierre de la comparación. OLS, Ridge y Lasso no compiten como métodos completamente independientes; más bien, forman una secuencia conceptual muy útil para aprender Machine Learning con base matemática. OLS introduce la lógica de la estimación por minimización; Ridge añade regularización continua; y Lasso extiende esta idea con regularización capaz de inducir sparsidad. Comprender esta progresión permite interpretar mejor no solo estos métodos, sino una parte importante de los modelos supervisados modernos.

## 6. Ejercicios propuestos

### 6.1. Ejercicios numéricos y de comparación

6. **Cálculo completo con OLS, Ridge y Lasso.** Use el conjunto de datos

$$x = \{1, 2, 3, 4\}, \quad y = \{2, 3, 3, 5\}.$$

- a) Calcule  $n$ ,  $\sum x_i$ ,  $\sum y_i$ ,  $\sum x_i^2$ ,  $\sum x_i y_i$ .  
b) Calcule  $A$  y  $D$ :

$$A = n \sum x_i y_i - (\sum x_i)(\sum y_i), \quad D = n \sum x_i^2 - (\sum x_i)^2.$$

- c) Obtenga el modelo OLS.

- d) Obtenga el modelo Ridge con  $\lambda = 1$ .  
 e) Obtenga el modelo Lasso con  $\lambda = 1$ .

7. **Comparación de SSE en entrenamiento.** Para los tres modelos obtenidos en el ejercicio anterior, calcule:

$$SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

Verifique que OLS tiene el menor SSE en entrenamiento y explique por qué este resultado es esperable desde la definición del método.

8. **Efecto de  $\lambda$  en Ridge.** Usando el mismo conjunto de datos del ejercicio 6, calcule  $\beta_0^{\text{Ridge}}$  y  $\beta_1^{\text{Ridge}}$  para

$$\lambda \in \{0, 0.5, 1, 2, 10\}.$$

Describa cómo cambia la pendiente y qué ocurre con el modelo cuando  $\lambda$  aumenta.

9. **Efecto de  $\lambda$  en Lasso y umbral.** Con el mismo conjunto de datos, calcule  $\beta_1^{\text{Lasso}}$  para

$$\lambda \in \{0, 1, 2, 4, 5, 6\}.$$

Identifique el valor a partir del cual la pendiente se vuelve cero. Interprete el resultado en términos de complejidad del modelo.

10. **Comparación de funciones objetivo.** Para el conjunto de datos del ejercicio 6 y  $\lambda = 1$ , evalúe:

$$J_{\text{Ridge}}(\beta_0, \beta_1) = \frac{1}{2}SSE + \frac{1}{2}\lambda\beta_1^2$$

en los coeficientes de OLS y en los coeficientes de Ridge. Repita el mismo análisis con

$$J_{\text{Lasso}}(\beta_0, \beta_1) = \frac{1}{2}SSE + \lambda|\beta_1|$$

evaluando en los coeficientes de OLS y en los coeficientes de Lasso. Discuta por qué cada método minimiza su propia función objetivo, aunque no minimice necesariamente el SSE.

## 6.2. Ejercicios conceptuales y de reflexión

11. **Interpretación de coeficientes.** En un problema aplicado, ¿qué significa interpretar  $\beta_0$  como el valor esperado de  $y$  cuando  $x = 0$ ? Dé un ejemplo donde esta interpretación tenga sentido y otro donde no sea razonable.
12. **¿Qué significa regularizar?** Explique con sus palabras la diferencia entre:
  - minimizar solo error de ajuste;
  - minimizar error de ajuste + penalización.
 Relacione su respuesta con la idea de compromiso entre ajuste y complejidad.
13. **Ridge vs. Lasso.** Compare Ridge y Lasso en términos de:
  - tipo de penalización,
  - efecto sobre la magnitud de los coeficientes,
  - posibilidad de anular coeficientes,
  - interpretabilidad del modelo.
14. **¿Cuál método “es mejor”?** Discuta por qué la pregunta “¿qué método es mejor?” no tiene una respuesta única. Proponga al menos tres criterios distintos de comparación (por ejemplo: SSE en entrenamiento, interpretabilidad, generalización esperada, estabilidad, simplicidad).
15. **Importancia de la estandarización.** Explique por qué en problemas con múltiples predictores es recomendable estandarizar variables antes de aplicar Ridge o Lasso. ¿Qué podría ocurrir si una variable está medida en miles y otra en decimales?
16. **Extensión al caso multivariable (lectura guiada).** Sin realizar derivaciones completas, escriba cómo cambiarían las funciones objetivo de OLS, Ridge y Lasso al pasar del caso univariado al caso con  $p$  predictores. Indique explícitamente qué coeficientes se penalizan y cuál no.

Sugerencia de trabajo. Una forma útil de resolver estos ejercicios es separar el proceso en tres etapas: (i) cálculo de sumas básicas de los datos, (ii) obtención de coeficientes, y (iii) interpretación del resultado. Esta secuencia ayuda a conectar la parte algebraica con la parte conceptual del modelado.

## 7. Comentarios finales

Estas notas han presentado una progresión conceptual y matemática desde la regresión lineal ordinaria (OLS) hasta dos de sus extensiones regularizadas más importantes: Ridge y Lasso. A lo largo del desarrollo, el objetivo no ha sido solo mostrar fórmulas de estimación, sino resaltar una idea central del aprendizaje automático: la construcción de modelos puede entenderse como un problema de optimización en el que se define una función objetivo y se buscan los parámetros que la minimizan.

En OLS, esa función objetivo está dada únicamente por el error cuadrático, lo que produce un método sencillo, interpretable y con solución cerrada en el caso lineal. Ridge y Lasso conservan esta misma estructura de modelado, pero incorporan penalizaciones que permiten controlar la complejidad del modelo. Esta modificación, aunque algebraicamente simple, introduce una diferencia conceptual profunda: el mejor modelo ya no es necesariamente el que ajusta más al conjunto de entrenamiento, sino aquel que logra un equilibrio adecuado entre ajuste, estabilidad e interpretabilidad.

Desde el punto de vista pedagógico, OLS, Ridge y Lasso forman una secuencia especialmente valiosa para un curso introductorio de Machine Learning. OLS permite establecer el lenguaje de modelos, residuos, funciones de pérdida y minimización. Ridge introduce la regularización continua y la idea de encogimiento de coeficientes. Lasso, por su parte, amplía este panorama mostrando que la regularización también puede inducir sparsidad y simplificación estructural del modelo. Comprender esta secuencia facilita el estudio posterior de métodos más generales y de algoritmos de entrenamiento más complejos.

Aunque en estas notas se trabajó principalmente el caso univariado para mantener transparencia algebraica, las ideas centrales se extienden al caso multivariable, donde la regularización adquiere aún mayor relevancia. En aplicaciones reales, aspectos como la estandarización de variables, la elección del parámetro  $\lambda$  y la evaluación en datos no vistos se vuelven fundamentales. Por ello, estas notas deben entenderse como una base conceptual y matemática inicial sobre la cual se pueden construir desarrollos posteriores más orientados a práctica computacional y validación de modelos.

Finalmente, más allá de los detalles técnicos de cada método, una conclusión importante es que el modelado estadístico y el aprendizaje automático no consisten únicamente en “ajustar una fórmula”, sino en tomar decisiones explícitas sobre qué se optimiza, qué se penaliza y qué tipo de modelo se considera razonable para un problema dado. Esa perspectiva crítica y estructurada será útil en cualquier tema posterior del curso.

## Referencias

- [1] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., Springer, New York, 2009. <https://doi.org/10.1007/978-0-387-84858-7>.
- [2] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: with Applications in R*, 2nd ed., Springer, New York, 2021.
- [3] T. Hastie, R. Tibshirani, and M. Wainwright, *Statistical Learning with Sparsity: The Lasso and Generalizations*, CRC Press, Boca Raton, 2015. <https://doi.org/10.1080/24754269.2021.1980261>.
- [4] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis*, 5th ed., Wiley, Hoboken, NJ, 2012. <https://lccn.loc.gov/91021553>.
- [5] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, 2004. <https://web.stanford.edu/~boyd/cvxbook/>.
- [6] A. E. Hoerl and R. W. Kennard, Ridge Regression: Biased Estimation for Nonorthogonal Problems, *Technometrics*, vol. 12, no. 1, pp. 55–67, 1970. <https://doi.org/10.1080/00401706.1970.10488634>.
- [7] R. Tibshirani, Regression Shrinkage and Selection via the Lasso, *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996. <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>.
- [8] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, Least Angle Regression, *The Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004. <https://doi.org/10.1214/009053604000000067>.
- [9] H. Zou and T. Hastie, Regularization and Variable Selection via the Elastic Net, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005. <https://doi.org/10.1111/j.1467-9868.2005.00503.x>.
- [10] J. Friedman, T. Hastie, and R. Tibshirani, Regularization Paths for Generalized Linear Models via Coordinate Descent, *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–22, 2010. <https://doi.org/10.18637/jss.v033.i01>

- [11] G. H. Golub, M. Heath, and G. Wahba, Generalized Cross-Validation as a Method for Choosing a Good Ridge Parameter, *Technometrics*, vol. 21, no. 2, pp. 215–223, 1979. <https://doi.org/10.1080/00401706.1979.10489751>.