INFOX
educación sin fronteras

Clase 6 : Modelos de regresión

Mg. Gloria Rivas

Agenda

1. **Random Forest**
2. **Lasso Regression**
3. **Ridge Regression**

Agenda

1. Random Forest

INTRODUCCIÓN

- En esta clase, veremos los métodos basados en árboles para clasificación y regresión.
- Esto involucra segmentar nuestro espacio de predictores en un número de regiones simples. Para hacer una predicción para una observación x, típicamente usamos la media o la moda de las observaciones en el training de la región a la que pertenece.
- Dado que las reglas para partir o segmentar el espacio predictor pueden ser resumidas en un árbol, este tipo de modelos son conocidos como "árboles de decisión (Decision Trees)".
- Estos métodos son simples y fáciles de interpretar. Sin embargo, muchas veces estos no compiten con otros métodos de supervised Learning en términos de predicción. Es por eso que también aprenderemos Bagging y Random Forest.
- Cada una de esas aproximaciones involucra muchos árboles que combinados llevan a una sola predicción.
- Vamos a ver también que combinar muchos árboles puede ayudar en la predicción pero va afectar la interpretación del modelo.

Regression Trees

- Los árboles de decisión pueden ser aplicados tanto para modelos de regresión como para modelos de clasificación. Primero veremos los modelos de regresión y luego los de clasificación.

- Vamos a empezar con un ejemplo para motivar regression trees.

- Usamos Hitters data para predecir el salario de un jugador de baseball basado en los **años** (el número de años que ha jugado en las grandes ligas) y los **hits** (el número de hits del año anterior).

- Primero removemos las observaciones que tienen missing values y transformamos en logaritmo al salario para que tenga una mejor forma.

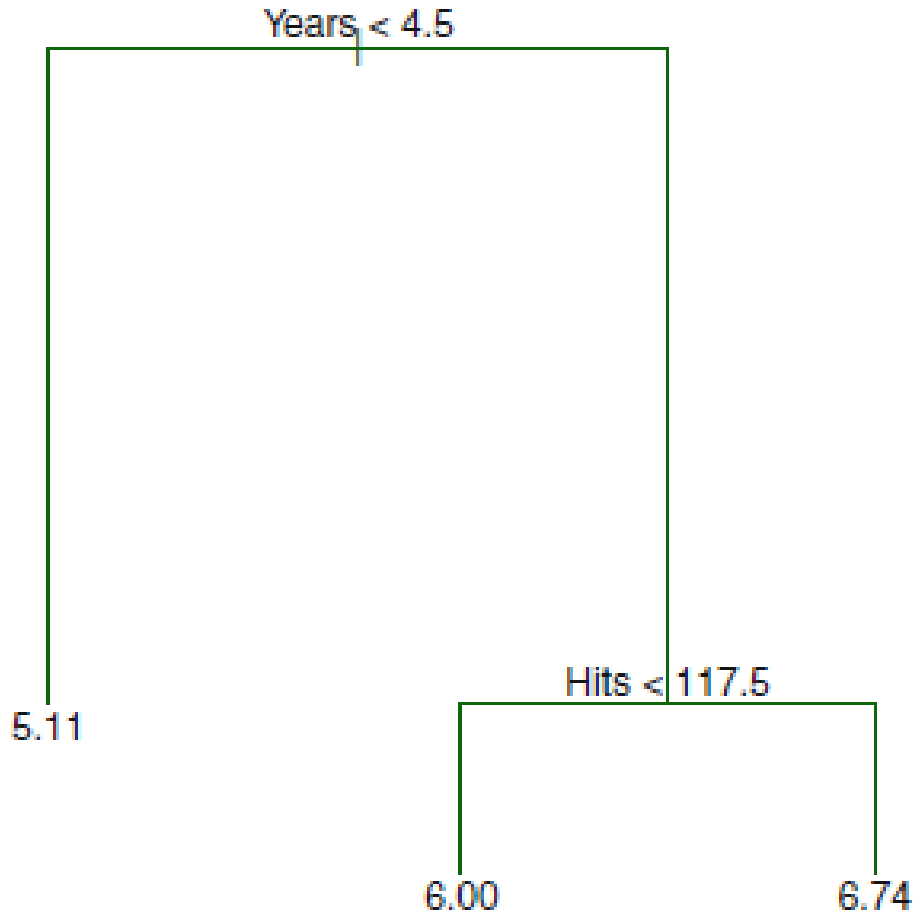
- En la siguiente figura vamos un regression tree ajustarse a la data.

Regression Trees

- Este árbol consiste en una serie de reglas para partirlo. En el tope del árbol se asignan observaciones que tienen Years <4.5 a la parte izquierda del árbol.

- La predicción de salario que tienen es el promedio de esa región.

- Para esos jugadores, el salario promedio en logaritmos es 5.11, entonces hacemos la siguiente predicción de $e^{5.107}$ que lleva a $165 174 dólares.



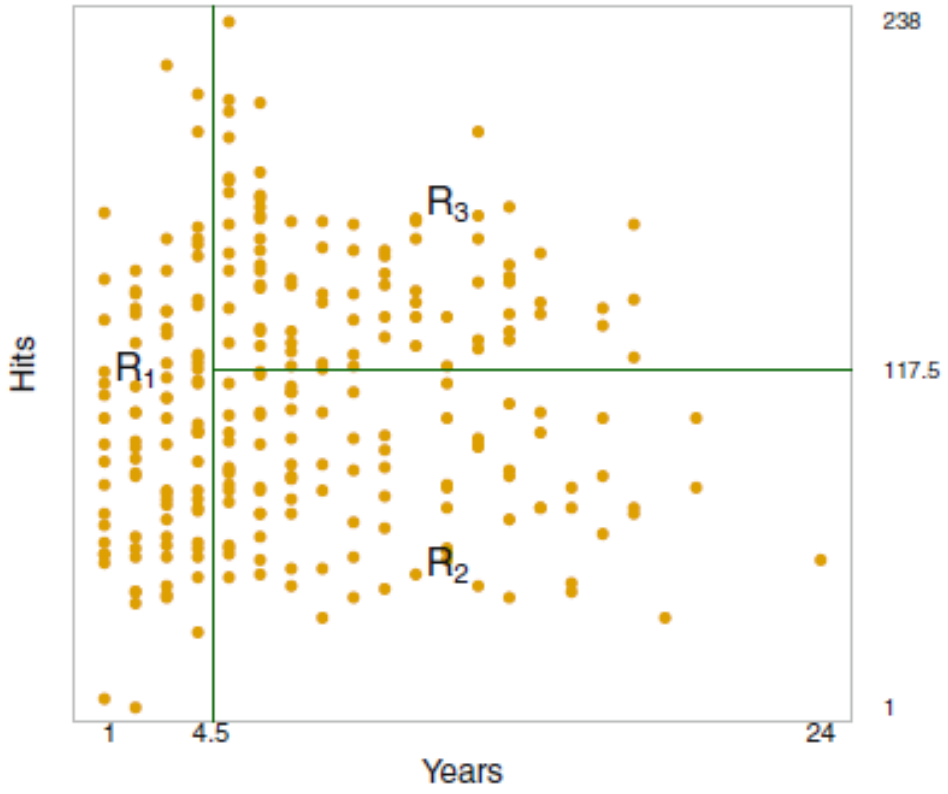
INFOX
educación sin fronteras

Regression Trees

- En general, el árbol segmenta a los jugadores en 3 regiones del espacio de predicción. Jugadores que han jugado 4 años o menos, jugadores que han jugado por más de 5 años y aquellos que han hecho menos de 118 hits el año pasado; y jugadores que han jugado por más de 5 años pero que tienen al menos 118 hits.

- Esas 3 regiones pueden ser escritas como $R_1 = \{X \mid \texttt{Years<4.5}\}$ $R_2 = \{X \mid \texttt{Years>=4.5}, \texttt{Hits<117.5}\}$, $R_3 = \{X \mid \texttt{Years>=4.5}, \texttt{Hits>=117.5}\}$.

- Continuando con la analogía del árbol, las regiones R1, R2,y R3 son conocidas como los nodos terminales o las hojas. Y los intermedios son los nodos intermedios.



INFOX
educación sin fronteras

Regression Trees

- Si interpretáramos la primera figura, los años es el factor más importante al momento de explicar el salario de los jugadores de baseball. Y jugadores con menos experiencia ganan un menor sueldo que aquellos jugadores más experimentados.

- El árbol mostrado en la figura 1 es una simplificación sobre la verdadera relación entre los hits, años y salario. Sin embargo, sobre otro tipo de modelos es fácil de interpretar y tiene una bonita representación gráfica.

Regression Trees

- Ahora vamos a discutir el proceso de construir un Regression tree. Hay 2 etapas:

- 1. Dividimos el espacio predictor- esto es el set de posibles valores para X1,X2..Xp- en J distintos y no entre cruzadas regiones R1, R2,.. Rj.

- 2. Para cada observación que cae en la región Rj, hacemos la misma predicción, que es simplemente la media de los valores para las observaciones en el training en Rj.

- Por ejemplo, pongamos como supuesto que en la etapa 1 obtenemos 2 regiones, R1 y R2, y que el promedio en las observaciones del training de la primera región (R1) sea 10, mientras que el promedio en las observaciones del training de la segunda región (R2) es 20.

- Entonces, para cada observación $X = x,$ si $x \in R_1$ vamos a predecir un valor de 10 y si $x \in R_2$ predecimos un valor de 20.

Regression Trees

- Ahora vamos a elaborar en la etapa 1. ¿Cómo construimos las regiones R1, … Rj? En teoría las regiones podrían tener cualquier forma. Sin embargo, escogemos dividir el espacio predictor en rectángulos dimensionales o cajas, por simplicidad y para una mayor facilidad en la interpretación de los modelos.

- El objetivo es encontrar cajas R1, ..., Rj que minimice la suma de residuos al cuadrado (RSS) dada por

$$\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2,$$

- Donde $\hat{y}_{R_j}$ es la respuesta promedio para las observaciones dentro del training set de la región Rj.

- Desafortunadamente, computacionalmente no es viable considerar cada posible partición del espacio en J cajas. Por esta razón, tomamos una aproximación llamada “recursive binanry splitting”.

- Este método va de arriba hacia abajo porque empieza en el tope del árbol (punto en el que todas las observaciones pertenecen a una misma región) y luego sucesivamente va partiendo el espacio predictor.

Regression Trees

- Para realizar el “recursive binary splitting”, primero debemos seleccionar el predictor Xi y el punto de corte s tal que al cortar el espacio predictor en regiones $\{X|X_j < s\}$ and $\{X|X_j \geq s\}$ lleve a la mayor posible reducción en RSS.

- Esto es, consideremos todos los predictores X1,X2,…, Xp, y todos los posibles valores del punto de corte s para cada uno de los predictores y luego escogemos al predictor y el punto de corte tal que el resultado del árbol tenga el menor RSS.

- En mayor detalle, para cualquier j y s, definimos

$$R_1(j,s) = \{X|X_j < s\} \text{ and } R_2(j,s) = \{X|X_j \geq s\},$$

- Y buscamos el valor de j y s que minimiza la ecuación

$$\sum_{i:\, x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:\, x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2,$$

- Donde $\hat{y}_{R_1}$ es la respuesta promedio para las observaciones del training set en $R_1(j,s)$

Regression Trees

- Encontrar los valores j y s que minimizan la previa ecuación puede hacerse de forma rápida, especialmente cuando el número de predictores p no es muy grande.

- Después, repetimos el proceso, buscando el mejor predictor y el mejor punto de corte para partir la data y así minimizar el RSS dentro de cada región. Sin embargo, esta vez, en vez de usar todo el espacio de predictores, vamos a partir una de las 2 regiones identificadas previamente.

- Ahora tenemos 3 regiones. De nuevo, buscamos partir una de esas 3 regiones y así minimizar el RSS, este proceso continua hasta que el criterio de parar se alcance. Por ejemplo, podemos continuar hasta que ninguna región contenga más de 5 observaciones.

- Una vez que las regiones R1,…, Rj, se hayan creado, vamos a predecir la respuesta para una observaciones del test set usando el promedio del training set de la región a la que pertenece esa observación.

Regression Trees – Tree prunning

- El proceso descrito anteriormente nos puede llevar a buenas predicciones en el training set, pero es muy probable que haga un overfit a la data, llevando a un pobre desempeño en el test set.

- Esto resulta porque el árbol final puede ser muy complejo. Un árbol pequeño con pocos cortes (esto es pocas regiones) puede llevar a una menor varianza y a una mejor interpretación con un costo de un pequeño sesgo.

- Una alternativa es crear un árbol bastante grande T0, y luego podarlo (prunning) para obtener un subtree. Cómo determinamos la mejor manera de podar un árbol?

- Intuitivamente, nuestro objetivo es seleccionar un subtree que nos lleve al menor test error rate. Dado un subtree podemos estimar este test error usando cross validation.

- Sin embargo, estimar el cross validation error para cada posible subtree sería muy trabajoso, dado que hay un gran número de posibilidades. En vez de eso, debemos seleccionar un pequeño set de subtrees.

Regression Trees – Tree prunning

- Cost complexity prunning, también conocido como weakest link prunning, nos da una manera de hacer eso. Antes de considerar todos los posibles subtrees, vamos a considerar una secuencia de árboles indexados por parámetro no negativo α.

- Para cada valor de α corresponde un subtree $T \subset T_0$ tal que

$$\sum_{m=1}^{|T|} \sum_{i:\, x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T|$$

- Es lo más pequeño posible. Aquí $|T|$ indicar el número de nodos terminales en el árbol T, R_m es el rectángulo correspondiente al nodo terminal $m\text{th}$ y $\hat{y}_{R_m}$ es la predicción asociada con Rm, esto es el promedio de las observaciones en el training set en Rm.

Regression Trees – Tree prunning

- El parámetro controla el trade off entre la complexidad del subtree y el fit del modelo en el training set. Cuando $\alpha = 0$, el subtree T va a ser igual a T0 porque la previa ecuación sólo mide el training error.

- Sin embargo, cuando α aumenta, hay un precio que hay que pagar para tener un árbol con varios nodos terminales y así la cantidad tenderá a ser minimizada para un menor subtree.

- En el siguiente cuadro podemos ver el proceso resumido.

Regression Trees – Tree prunning

1. Use recursive binary splitting to grow a large tree on the training data, stopping only when each terminal node has fewer than some minimum number of observations.

2. Apply cost complexity pruning to the large tree in order to obtain a sequence of best subtrees, as a function of α.

3. Use K-fold cross-validation to choose α. That is, divide the training observations into K folds. For each $k = 1, \ldots, K$:

   (a) Repeat Steps 1 and 2 on all but the kth fold of the training data.

   (b) Evaluate the mean squared prediction error on the data in the left-out kth fold, as a function of α.

   Average the results for each value of α, and pick α to minimize the average error.

4. Return the subtree from Step 2 that corresponds to the chosen value of α.

Classification Trees

- Un classification tree es muy similar a un decisión tree, excepto que en este buscamos predecir una respuesta cualitativa y no una cuantitativa.

- Por ejemplo, en el Regression Tree la predicción se daba por el promedio de la región a la que pertenece la observación que queremos predecir. En un classification Tree, predecimos la observación que pertenece a la clase común de observaciones en el training a las que pertenece.

- Cuando predecimos en un clasification tree no solo estamos interesados en saber la clase a la que corresponde sino también la proporción dentro de la clase.

- La forma es como se forma estos árboles de clasificación es muy similar a los árboles de regresión. Ya que así como en árboles de regresión usamos el “recursive binanry splitting” para crear un árbol de clasificación.

- Sin embargo, aquí no podemos usar el RSS como criterio de partición , se usa el classification error rate.

Classification Trees

- Dado que buscamos un plan para asignar una observación a la que mayormente pertenezca en el training set. El “classification error rate” es una simple fracción de las observaciones en el training set en la región que no corresponden a la gran mayoría de la clase.

$$E = 1 - \max_k(\hat{p}_{mk}).$$

- Donde $\hat{p}_{mk}$ representa la proporción de las observaciones del training set en la mth región que son parte de la clase kth. Sin embargo, este classification error rate no es muy sensible cuando se podan los árboles, es por eso que como medida usamos el Gini Index.

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}),$$

- Que es una medida de la varianza total entre las K clases.

Agenda

2. Lasso Regression

INTRODUCCIÓN

- En cuanto a regresiones, el modelo lineal estándar es el siguiente:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

- Se utiliza comúnmente para describir la relación entre una variable Y y un conjunto de variables X1, X2,. . ., Xp; y normalmente este modelo se ajusta utilizando mínimos cuadrados

- El modelo lineal tiene claras ventajas en términos de inferencia y, en ayudar a resolver problemas del mundo real. Asimismo, es a menudo sorprendentemente competitivo en relación con los métodos no lineales.

- En esta clase se discutirán algunas formas en las que se puede mejorar el modelo lineal simple, reemplazando el ajuste de mínimos cuadrados por algunos procedimientos de ajuste alternativos.

- Los procedimientos de **ajuste alternativos pueden producir una i) mejor precisión de predicción, e ii) interpretación de modelos.**

Precisión de la predicción:

- Siempre que la verdadera relación entre la respuesta y los predictores sea aproximadamente lineal, las estimaciones de mínimos cuadrados tendrán un sesgo bajo.

- Si N>>P ,el número de observaciones, es mucho mayor que p, entonces las estimaciones de mínimos cuadrados tiende a tener una baja varianza, y por lo tanto a tener un buen rendimiento en las observaciones de la prueba.

- Sin embargo, si N no es mucho mayor que P, entonces puede haber mucha variabilidad en el ajuste de mínimos cuadrados, lo que da como resultado un ajuste excesivo y, en consecuencia, predicciones deficientes.

- Y si P> N, entonces ya no hay una estimación única del coeficiente de mínimos cuadrados: la varianza es infinita, por lo que el método no se puede utilizar en absoluto.

- **Al restringir o reducir los coeficientes estimados, a menudo podemos reducir sustancialmente la varianza a costa de un aumento insignificante del sesgo. Esto puede conducir a mejoras sustanciales en la precisión con la que podemos predecir la respuesta para las observaciones que no se utilizan en el entrenamiento del modelo**

Interpretabilidad del modelo:

- A menudo ocurre que algunas o muchas de las variables utilizadas en un modelo de regresión múltiple, de hecho, no están asociadas con la respuesta.

- La inclusión de tales variables irrelevantes conduce a una complejidad innecesaria en el modelo resultante. Al eliminar estas variables, es decir, al establecer las estimaciones de coeficientes correspondientes en cero, podemos obtener un modelo que se interpreta más fácilmente.

- Sin embargo, es muy poco probable que los mínimos cuadrados produzcan estimaciones de coeficientes que sean exactamente cero.

- **Veremos algunos enfoques para realizar automáticamente la selección de características o la selección de variables, es decir, para excluir variables irrelevantes de un modelo de regresión múltiple.**

- Hay muchas alternativas, tanto clásicas como modernas, para usar mínimos cuadrados para el ajuste de modelos. Se discutirán tres clases importantes de métodos:

Subset Selection	Este enfoque implica identificar un subconjunto de los predictores P que creemos que están relacionados con la respuesta. Luego ajustamos un modelo usando mínimos cuadrados en el conjunto reducido de variables
Shrinkage	Este enfoque implica ajustar un modelo que involucre a todos los predictores. Sin embargo, los coeficientes estimados se reducen a cero en relación con las estimaciones de mínimos cuadrados. Esta contracción (también conocida como regularización) tiene el efecto de reducir la varianza. Los métodos de contracción también pueden realizar una selección de variables.
Dimension Reduction	Este enfoque implica proyectar los predictores P en un subespacio de dimensión M, donde M<P. Esto se logra calculando M diferentes combinaciones lineales, o proyecciones de las variables. Luego, estas proyecciones M se utilizan como predictores para ajustar un modelo de regresión lineal por mínimos cuadrados

1. Subset Selection

- Consideramos algunos métodos para seleccionar subconjuntos de predictores. Estos incluyen los procedimientos de selección: i) Best Subset; y ii) Stepwise model

Best Subset Selection

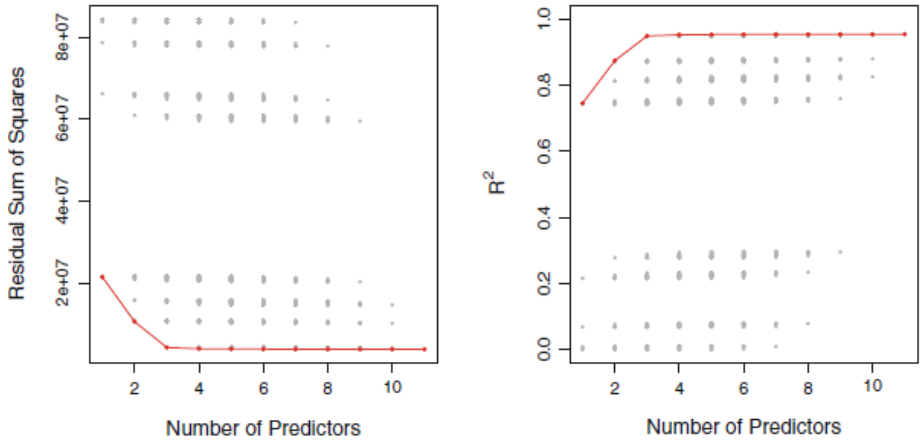
- Para realizar la mejor selección de subconjuntos, ajustamos un mejor subconjunto de regresión de mínimos cuadrados para cada combinación posible de los predictores P. Todos los $\binom{p}{2} = p(p-1)/2$ modelos que contienen exactamente dos predictores, y así sucesivamente.
- Luego, analizamos todos los modelos resultantes, con el objetivo de identificar el mejor. El problema de seleccionar el mejor modelo entre las posibilidades 2p consideradas por la selección del mejor subconjunto no es trivial. Esto generalmente se divide en dos etapas, como se describe en el Algoritmo:

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Del algoritmo:

- El paso 2 identifica el mejor modelo (en los datos de entrenamiento) para cada tamaño de subconjunto, con el fin de reducir el problema de uno de 2p modelos posibles a uno de p + 1 modelos posibles. Estos modelos forman la frontera inferior representada en rojo del gráfico.
- Ahora, para seleccionar un único mejor modelo, simplemente debemos elegir entre estas opciones p + 1. Esta tarea debe realizarse con cuidado, porque el RSS de estos modelos p + 1 disminuye monótonamente y el R2 aumenta monótonamente a medida que aumenta el número de características incluidas en los modelos.
- Por lo tanto, si usamos estas estadísticas para seleccionar el mejor modelo, siempre terminaremos con un modelo que involucra todas las variables. El problema es que un RSS bajo o un R2 alto indican un modelo con un error de entrenamiento bajo,

Para cada modelo posible que contiene un subconjunto de los diez predictores en el conjunto de datos de crédito, se muestran el RSS y R2. La frontera roja rastrea el mejor modelo para un número determinado de predictores, según RSS y R2. Aunque el conjunto de datos contiene solo diez predictores, el eje x varía de 1 a 11, ya que uno de las variables es categórica y toma tres valores, lo que lleva a la creación de dos variables ficticias.

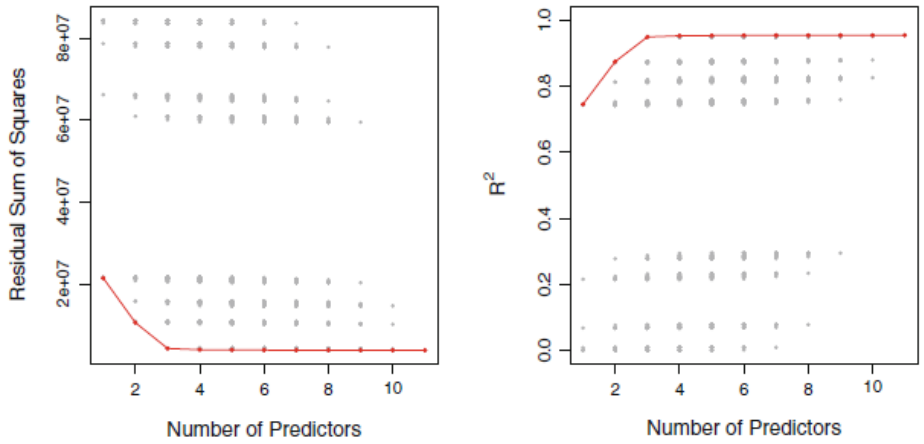


Interpretación del gráfico:

- Cada punto representado corresponde a un ajuste del modelo de regresión de mínimos cuadrados utilizando un subconjunto diferente de los 11 predictores en el conjunto de datos.

- Las curvas rojas conectan los mejores modelos para cada tamaño de modelo, según RSS o R2. La figura muestra que, como se esperaba, estas cantidades mejoran a medida que aumenta el número de variables; sin embargo, a partir del modelo de tres variables, hay poca mejora en RSS y R2 como resultado de incluir predictores adicionales

- Si bien la selección del mejor subconjunto es un enfoque simple y conceptualmente atractivo, adolece de limitaciones computacionales. El número de posibles modelos que deben considerarse crece rápidamente a medida que aumenta p. En general, hay modelos 2p que involucran subconjuntos de predictores p.



1.1. Stepwise Selection

- Por razones computacionales, la mejor selección de subconjuntos no se puede aplicar con P muy grandes. La mejor selección de subconjuntos también puede sufrir problemas estadísticos cuando P es grande.

- Cuanto mayor sea el espacio de búsqueda, mayor será la posibilidad de encontrar modelos que se vean bien en los datos de entrenamiento, aunque es posible que no tengan poder predictivo sobre datos futuros.

- Por lo tanto, un espacio de búsqueda enorme puede llevar a un ajuste excesivo y una alta varianza de las estimaciones de coeficientes. Por ambas razones, Stepwise methods (los métodos escalonados), que exploran un conjunto de modelos mucho más restringido, son alternativas atractivas para la mejor selección de subconjuntos.

Forward Stepwise Selection:

- La selección progresiva hacia adelante es una alternativa computacionalmente eficiente a la mejor selección de subconjuntos. Mientras que el mejor procedimiento de selección de subconjuntos (best subset selection) considera todos los modelos 2p posibles que contienen subconjuntos de los predictores P, el avance por pasos considera un conjunto de modelos mucho más pequeño.

- La selección por pasos hacia adelante comienza con un modelo que no contiene predictores y luego agrega predictores al modelo, uno a la vez, hasta que todos los predictores están en el modelo. En particular, en cada paso se agrega al modelo la variable que da la mayor mejora adicional al ajuste. Más formalmente, el procedimiento de selección por pasos hacia adelante se da con el siguiente algoritmo:

Algorithm 6.2 *Forward stepwise selection*

1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.
2. For $k = 0, \ldots, p-1$:
   (a) Consider all $p-k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   (b) Choose the *best* among these $p-k$ models, and call it $\mathcal{M}_{k+1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

- la selección por pasos hacia adelante implica el ajuste de un modelo nulo, junto con P - K modelos en la k-ésima iteración, para k = 0, ..., p 1. Esto equivale a un total de $\sum_{k=0}^{p-1}(p-k) = 1+p(p+1)/2$ modelos. Ésta es una diferencia sustancial: cuando P = 20, la mejor selección de subconjuntos requiere el ajuste de 1.048.576 modelos, mientras que la selección por pasos hacia adelante requiere el ajuste de solo 211.

- En el Paso 2 del Algoritmo, debemos identificar el mejor modelo de entre aquellos P - K que aumentan k con un predictor adicional. Podemos hacer esto simplemente eligiendo el modelo con el RSS más bajo o el R2 más alto. Sin embargo, en el Paso 3, debemos identificar el mejor modelo entre un conjunto de modelos con diferentes números de variables

- La selección progresiva hacia adelante se puede aplicar incluso en la configuración de alta dimensión donde n <p; sin embargo, en este caso, es posible construir submodelos M0, ..., Mn − 1 solamente, ya que cada submodelo se ajusta al uso de mínimos cuadrados, que no producirá una solución única si p ≥ n.

Backward Stepwise Selection:

- Al igual que la selección por pasos hacia adelante, la selección por pasos hacia atrás proporciona una alternativa eficaz a la mejor selección de subconjuntos. Sin embargo, a diferencia de la selección progresiva hacia adelante, comienza con el modelo de mínimos cuadrados completo que contiene todos los predictores p, y luego elimina iterativamente el predictor menos útil, uno a la vez. Los detalles se dan en el algoritmo.
- La selección hacia atrás requiere que el número de muestras n sea mayor que el número de variables p (para que se pueda ajustar el modelo completo). Por el contrario, se puede usar progresivamente hacia adelante incluso cuando n <p, por lo que es el único método de subconjunto viable cuando p es muy grande.

Algorithm 6.3 *Backward stepwise selection*

1. Let $\mathcal{M}_p$ denote the *full* model, which contains all p predictors.
2. For $k = p, p-1, \ldots, 1$:
   (a) Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   (b) Choose the *best* among these k models, and call it $\mathcal{M}_{k-1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Hybrid Approaches:

- Los mejores enfoques de selección de subconjuntos, hacia adelante y hacia atrás por pasos, generalmente dan modelos similares pero no idénticos. Como otra alternativa, se encuentran disponibles versiones híbridas de selección progresiva hacia adelante y hacia atrás, en las que las variables se agregan al modelo secuencialmente, en analogía a la selección hacia adelante.

- Sin embargo, después de agregar cada nueva variable, el método también puede eliminar cualquier variable que ya no mejore el ajuste del modelo. Este enfoque intenta imitar más de cerca la mejor selección de subconjuntos mientras conserva las ventajas computacionales de la selección por pasos hacia adelante y hacia atrás.

1.2. Escogiendo el modelo óptimo

- La mejor selección de subconjuntos, la selección hacia adelante y la selección hacia atrás dan como resultado la creación de un conjunto de modelos, cada uno de los cuales contiene un subconjunto de predictores p.

- Para implementar estos métodos, necesitamos una forma de determinar cuál de estos modelos es el mejor. El modelo que contiene todos los predictores siempre tendrá el RSS más pequeño y el R2 más grande, ya que estas cantidades están relacionadas con el error de entrenamiento.

- En cambio, deseamos elegir un modelo con un error de prueba bajo. El error de entrenamiento puede ser una estimación pobre del error de prueba. Por lo tanto, RSS y R2 no son adecuados para seleccionar el mejor modelo entre una colección de modelos con diferentes números de predictores.

- Para seleccionar el mejor modelo con respecto al error de prueba, necesitamos estimar este error de prueba. Hay dos enfoques comunes:
    1. Podemos estimar indirectamente el error de la prueba haciendo un ajuste al error de entrenamiento para tener en cuenta el sesgo debido al sobreajuste.
    2. Podemos estimar directamente el error de prueba, utilizando un enfoque de conjunto de validación o un enfoque de validación cruzada, como se discutió en el Capítulo

Cp, AIC, BIC, and Adjusted R^2:

- El conjunto de entrenamiento MSE es generalmente una subestimación del MSE de prueba. Esto se debe a que cuando ajustamos un modelo a los datos de entrenamiento usando mínimos cuadrados, estimamos específicamente los coeficientes de regresión de manera que el RSS de entrenamiento (pero no el RSS de prueba) sea tan pequeño como posible.

- En particular, el error de entrenamiento disminuirá a medida que se incluyan más variables en el modelo, pero es posible que el error de prueba no. Por lo tanto, el conjunto de entrenamiento RSS y el conjunto de entrenamiento R2 no se pueden usar para seleccionar entre un conjunto de modelos con diferentes números de variables.

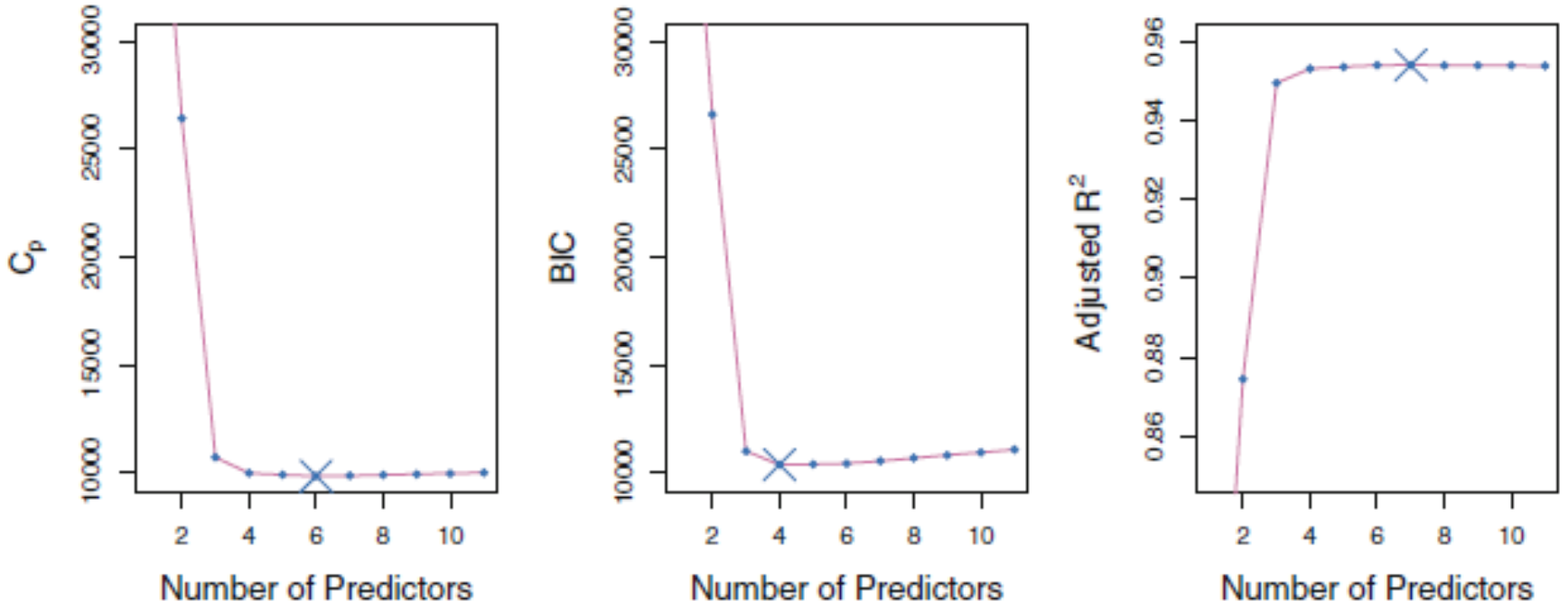
- Sin embargo, se encuentran disponibles varias técnicas para ajustar el error de entrenamiento para el tamaño del modelo. Estos enfoques se pueden utilizar para seleccionar entre un conjunto de modelos con diferentes números de variables. Ahora consideramos cuatro de estos enfoques: Cp, AIC, BIC, and Adjusted $\boldsymbol{R}$^2.

- El conjunto de entrenamiento MSE es generalmente una subestimación del MSE de prueba. Esto se debe a que cuando ajustamos un modelo a los datos de entrenamiento usando mínimos cuadrados, estimamos específicamente los coeficientes de regresión de manera que el RSS de entrenamiento (pero no el RSS de prueba) sea tan pequeño como posible.

- En particular, el error de entrenamiento disminuirá a medida que se incluyan más variables en el modelo, pero es posible que el error de prueba no. Por lo tanto, el conjunto de entrenamiento RSS y el conjunto de entrenamiento R2 no se pueden usar para seleccionar entre un conjunto de modelos con diferentes números de variables.

- Sin embargo, se encuentran disponibles varias técnicas para ajustar el error de entrenamiento para el tamaño del modelo. Estos enfoques se pueden utilizar para seleccionar entre un conjunto de modelos con diferentes números de variables. Ahora consideramos cuatro de estos enfoques: Cp, AIC, BIC, and Adjusted $\boldsymbol{R}$^$\mathbf{2}$.

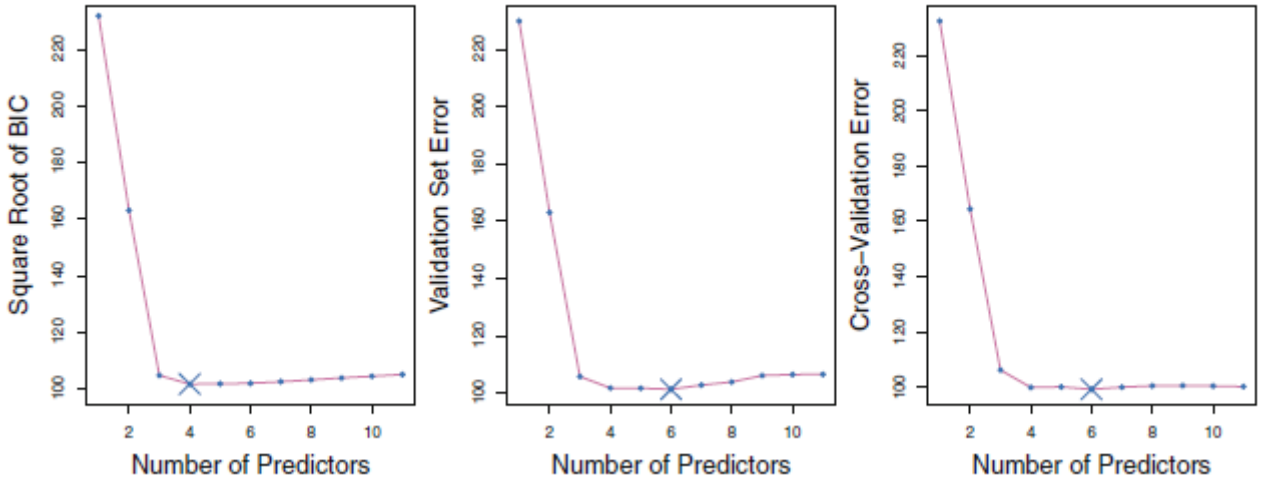
- Cp, BIC y R2 ajustado se muestran para los mejores modelos de cada tamaño para el conjunto de datos Credit (la frontera inferior en la Figura). Cp y BIC son estimaciones de la prueba MSE. En la gráfica del medio, vemos que la estimación BIC del error de prueba muestra un aumento después de seleccionar cuatro variables. Las otras dos parcelas son bastante planas después de incluir cuatro variables



Validation and Cross-Validation:

- Como alternativa a los enfoques que se acaban de discutir, podemos estimar directamente el error de prueba usando el conjunto de validación y los métodos de validación cruzada. Podemos calcular el error del conjunto de validación o el error de validación cruzada para cada modelo en consideración, y luego seleccionar el modelo para el cual el error de prueba estimado resultante es más pequeño. Este procedimiento tiene una ventaja en relación con AIC, BIC, Cp y R2 ajustado, en que proporciona una estimación directa del error de prueba y hace menos suposiciones sobre el verdadero modelo subyacente.

- Hoy en día con computadoras rápidas, los cálculos necesarios para realizar la validación cruzada casi nunca son un problema. Por lo tanto, la validación cruzada es un enfoque muy atractivo para seleccionar entre varios modelos en consideración.



- La figura muestra, en función de d, el BIC, los errores del conjunto de validación y los errores de validación cruzada en los datos de crédito, para el mejor modelo de variable d. Los errores de validación se calcularon seleccionando aleatoriamente tres cuartas partes de las observaciones como conjunto de entrenamiento y el resto como conjunto de validación. Los errores de validación cruzada se calcularon utilizando k = 10 pliegues.

- El fundamento aquí es que si un conjunto de modelos parece ser más o menos igualmente bueno, entonces también podríamos elegir el modelo más simple, es decir, el modelo con el menor número de predictores. En este caso, la aplicación de la regla de un error estándar al conjunto de validación o al enfoque de validación cruzada conduce a la selección del modelo de tres variables.

2. Shrinkage Methods

- Los métodos de selección de subconjuntos descritos implican el uso de mínimos cuadrados para ajustar un modelo lineal que contiene un subconjunto de predictores.

- Como alternativa, podemos ajustar un modelo que contenga todos los p predictores usando una técnica que restrinja o regularice las estimaciones de coeficientes, o de manera equivalente, que reduzca las estimaciones de coeficientes hacia cero.

- La reducción de las estimaciones de coeficientes puede reducir significativamente su varianza. **Las dos técnicas más conocidas para reducir los coeficientes de regresión hacia cero son la regresión de crestas (ridge regression) y de Lasso.**

Ridge Regression:

- El procedimiento de ajuste de mínimos cuadrados estima β0, β1, ..., βp utilizando los valores que minimizan.

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 .$$

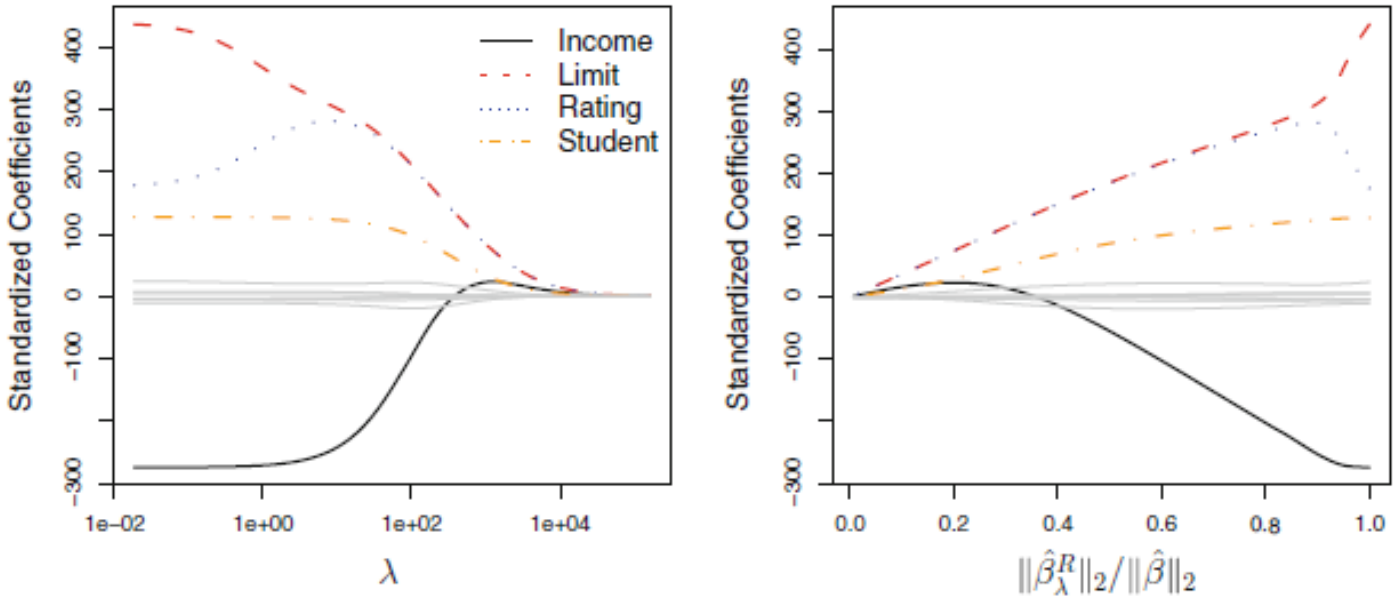
- La regresión de crestas es muy similar a los mínimos cuadrados, excepto que los coeficientes se estiman minimizando una cantidad ligeramente diferente. En particular, las estimaciones del coeficiente de regresión de la cresta βˆR son los valores que minimizan.

$$\sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 = \text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2, \qquad (6.5)$$

- Observe que en el siguiente gráfico, la penalización por contracción se aplica a β1, ..., βp, pero no al intercepto β0. Queremos reducir la asociación estimada de cada variable con la respuesta; sin embargo, no queremos reducir la intersección, que es simplemente una medida del valor medio de la respuesta cuando xi1 = xi2 = ... = xip = 0.

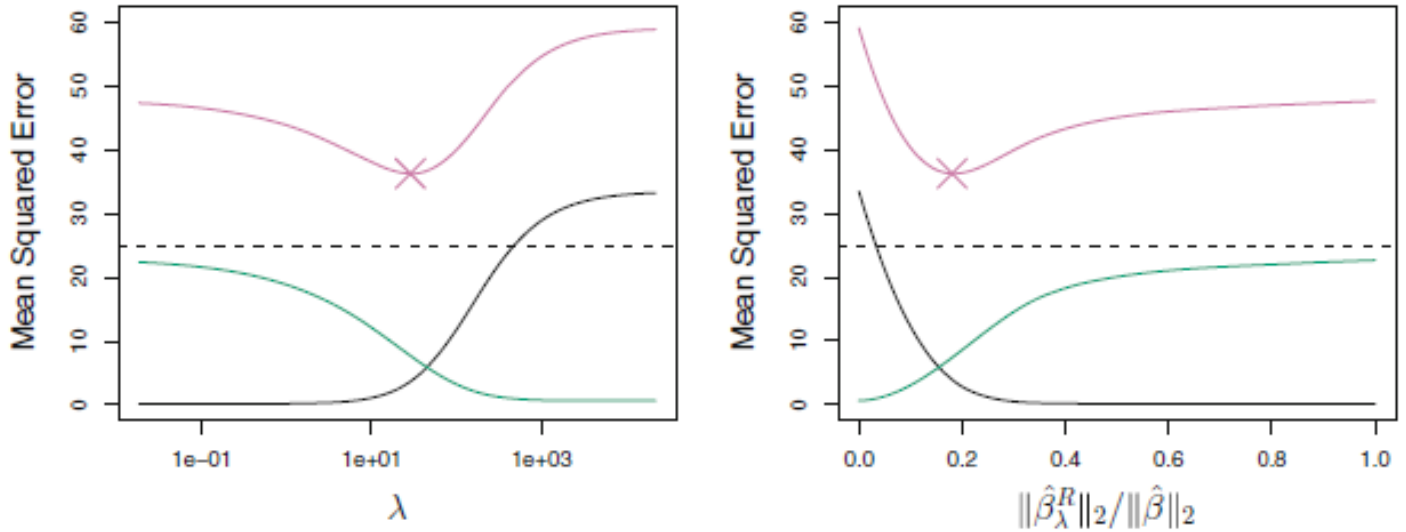
- Si asumimos que las variables, es decir, las columnas de la matriz de datos X: se han centrado para tener una media cero antes de realizar la regresión de la cresta, entonces la intersección estimada tomará la forma de $\hat{\beta}_0 = \bar{y} = \sum_{i=1}^{n} y_i/n.$

Los coeficientes de regresión de cresta estandarizados se muestran para el conjunto de datos Credit, como una función de λ y ‖βˆR‖2 / ‖βˆ‖2



¿Por qué es mejor la regresión de crestas (ridge regression) con respecto a los mínimos cuadrados?

- La ventaja de la regresión de crestas sobre los mínimos cuadrados se basa en el intercambio sesgo-varianza. A medida que aumenta λ, la flexibilidad del ajuste de regresión de la cresta disminuye, lo que conduce a una disminución de la varianza pero a un mayor sesgo.
- Esto se ilustra en el gráfico de la izquierda, utilizando un conjunto de datos simulados que contiene p = 45. La curva verde en el muestra la varianza de las predicciones de regresión de la cresta en función de λ. A medida que aumenta λ, la contracción de las estimaciones del coeficiente de cresta conduce a una reducción sustancial de la varianza de las predicciones, a expensas de un ligero aumento del sesgo. Recuerde que el error cuadrático medio de la prueba (MSE), trazado en púrpura, es una función de la varianza más el sesgo al cuadrado



Regresión Lasso:

- La regresión de crestas tiene una desventaja obvia. A diferencia de la selección del mejor subconjunto, paso hacia adelante y paso hacia atrás, que generalmente seleccionará modelos que involucran solo un subconjunto de las variables, la regresión de cresta incluirá todos los predictores p en el modelo final.

- La penalización λ β2 reducirá todos los coeficientes hacia cero, pero no establecerá ninguno de ellos exactamente en cero (a menos que λ =). Esto puede no ser un problema para la precisión de la predicción, pero puede crear un desafío en la interpretación del modelo en entornos en los que el número de variables p es bastante grande.

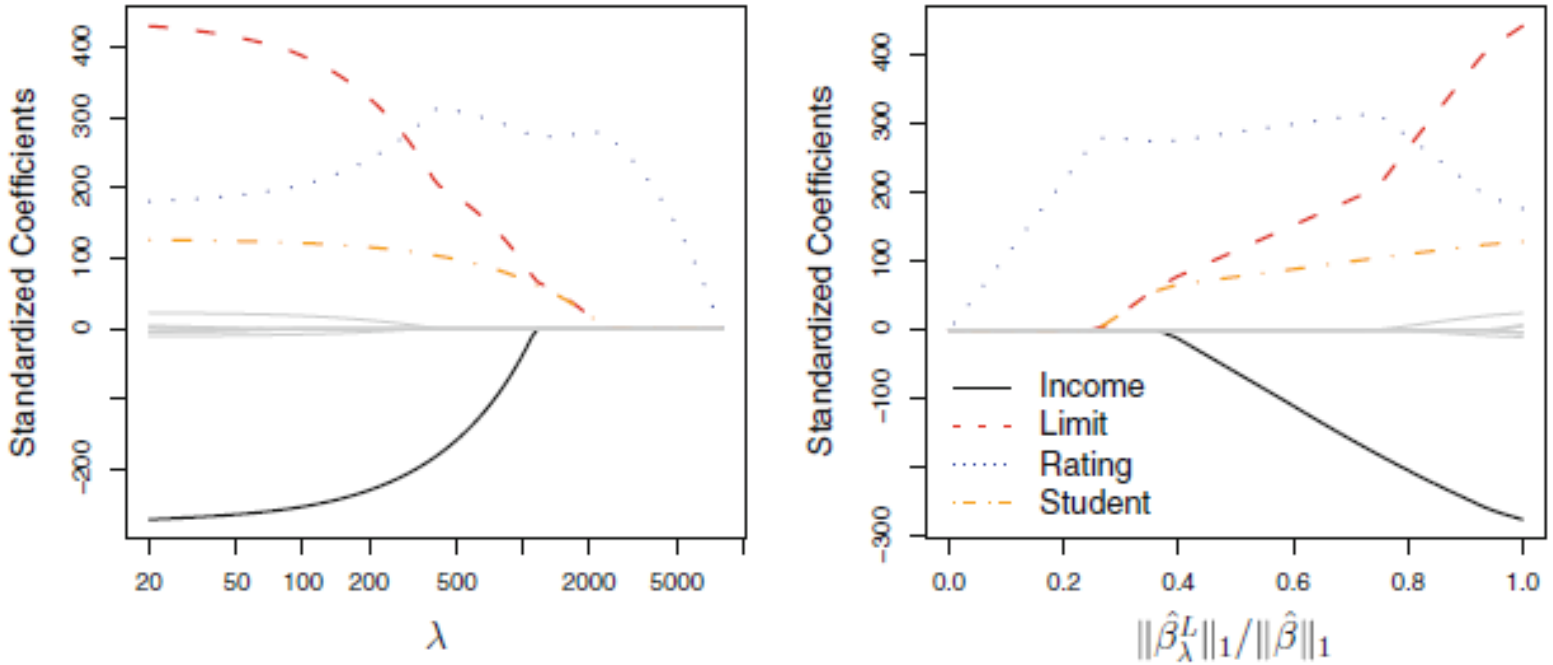
- Sin embargo, la regresión de crestas siempre generará un modelo que incluya los diez predictores. Incrementar el valor de λ tenderá a reducir las magnitudes de los coeficientes, pero no resultará en la exclusión de ninguna de las variables.

- El lazo es una alternativa relativamente reciente a la regresión de la cresta que supera esta desventaja. Los coeficientes de lazo, βˆL, minimizan la cantidad.

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{p}\beta_j x_{ij}\right)^2+\lambda\sum_{j=1}^{p}|\beta_j| = \mathrm{RSS}+\lambda\sum_{j=1}^{p}|\beta_j|. \qquad (6.7)$$

- La regresión de Lasso y cresta tienen formulaciones similares. La única diferencia es que el término β2 en la penalización por regresión de la cresta ha sido reemplazado por | βj | en la pena de lazo. En el lenguaje estadístico, el lazo usa una penalización A1 (pronunciada "ell 1") en lugar de una penalización A2. La norma A1 de un vector de coeficientes β está dada por β 1 = βj.

- En el caso del lazo, la penalización A1 tiene el efecto de forzar algunas de las estimaciones de coeficientes a ser exactamente iguales a cero cuando el parámetro de ajuste λ es suficientemente grande. Por lo tanto, al igual que la selección del mejor subconjunto, el Lasso realiza la selección de variables. Como resultado, los modelos generados a partir del Lasso son generalmente mucho más fáciles de interpretar que los producidos por regresión de crestas.



- Como ejemplo, considere las gráficas de coeficientes en la Figura 6.6, que se generan al aplicar el Lasso al conjunto de datos de Crédito. Cuando λ = 0, entonces el Lasso simplemente da el ajuste de mínimos cuadrados, y cuando λ se vuelve lo suficientemente grande, el Lasso o da el modelo nulo en el que todas las estimaciones de coeficientes son iguales a cero. Sin embargo, entre estos dos extremos, los modelos de regresión de cresta y lazo son bastante diferentes entre sí. Moviéndose de izquierda a derecha.
- Por lo tanto, dependiendo del valor de λ, el lazo puede producir un modelo que involucre cualquier número de variables. Por el contrario, la regresión de crestas siempre incluirá todas las variables del modelo, aunque la magnitud de las estimaciones de los coeficientes dependerá de λ

Otra formulación para la regresión de crestas y el lazo:

- Se puede demostrar que las estimaciones de coeficientes de regresión de Lasso y cresta resuelven los problemas. para cada valor de λ, hay unos s tales que darán las mismas estimaciones de coeficientes de Lasso. De manera similar, para cada valor de λ hay un s correspondiente tal que las ecuaciones darán las mismas estimaciones de coeficientes de regresión de cresta:

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} |\beta_j| \leq s \tag{6.8}$$

and

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} \beta_j^2 \leq s, \tag{6.9}$$

- Aquí I (βj = 0) es una variable indicadora: toma un valor de 1 si βj = 0, y es igual a cero en caso contrario. Entonces equivale a encontrar un conjunto de estimaciones de coeficientes tales que RSS sea lo más pequeño posible, sujeto a la restricción de que no más de s coeficientes pueden ser distintos de cero. El problema es equivalente a la mejor selección de subconjuntos.

- Desafortunadamente, resolver (6.10) es computacionalmente inviable cuando p es grande, ya que requiere considerar todos los modelos p que contienen predictores s.

- **Por lo tanto, podemos interpretar la regresión de crestas y el Lasso como alternativas computacionalmente factibles para la mejor selección de subconjuntos que reemplazan la forma intratable del presupuesto en (6.10) con formas que son mucho más fáciles de resolver. Por supuesto, el Lasso está mucho más relacionado con la mejor selección de subconjuntos, ya que solo el Lasso realiza la selección de características para un tamaño suficientemente pequeño.**

$$\underset{\beta}{\text{minimize}} \left\{ \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} I(\beta_j \neq 0) \leq s. \tag{6.10}$$

La propiedad de selección de variable del Lasso:

- ¿Por qué el Lasso, a diferencia de la regresión de crestas, da como resultado estimaciones de coeficientes que son exactamente iguales a cero? La solución de mínimos cuadrados se marca como $\hat{\beta}$, mientras que el diamante azul y el círculo representa las restricciones de regresión de Lasso y cresta. Si s es suficientemente grande, entonces las regiones de restricción contendrán $\hat{\beta}$, por lo que la regresión de la cresta y las estimaciones de lazo serán las mismas que las estimaciones de mínimos cuadrados.

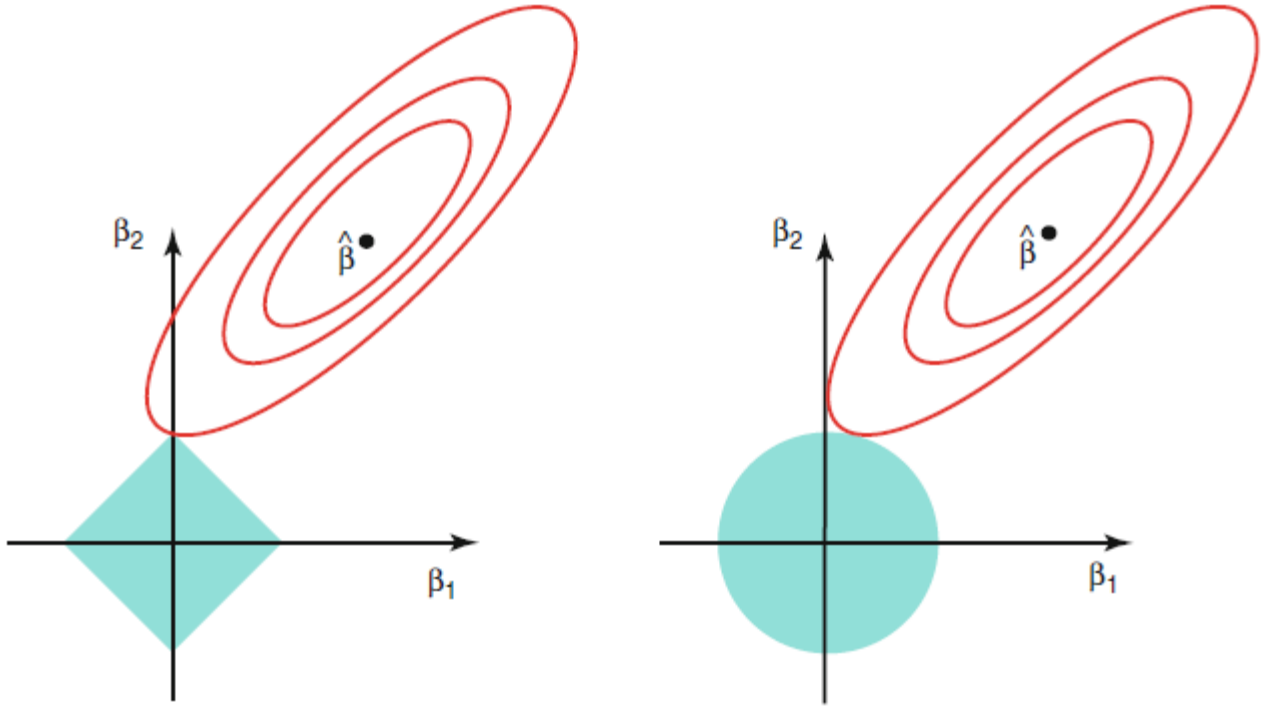
- Sin embargo, en la Figura 6.7 las estimaciones de mínimos cuadrados se encuentran fuera del diamante y del círculo, por lo que las estimaciones de mínimos cuadrados no son las mismas como las estimaciones de regresión de Lasso y cresta.

- Las elipses que se centran alrededor de $\hat{\beta}$ representan regiones de constante RSS. En otras palabras, todos los puntos de una elipse determinada comparten un valor común del RSS. A medida que las elipses se expanden alejándose de las estimaciones de coeficientes de mínimos cuadrados, la RSS aumenta.

- Dado que la regresión de la cresta tiene una restricción circular sin puntos agudos, esta intersección generalmente no ocurrirá en un eje, por lo que las estimaciones del coeficiente de regresión de la cresta serán exclusivamente distintas de cero.

- Sin embargo, la restricción de lazo tiene esquinas en cada uno de los ejes, por lo que la elipse a menudo intersecará la región de restricción en un eje. Cuando esto ocurre, uno de los coeficientes será igual a cero.

Contornos de las funciones de error y restricción para Lasso (izquierda) y regresión de cresta (derecha). Las áreas azules sólidas son las regiones de restricción,| β1 | + | β2 | ≤ sy β21 + β22 ≤ s, mientras que las elipses rojas son los contornos de la RSS.



Comparación de la regresión de Lasso y Ridge

- Está claro que el Lasso tiene una gran ventaja sobre la regresión de crestas, ya que produce modelos más simples e interpretables que involucran solo a un subconjunto de predictores. Sin embargo, ¿qué método conduce a una mejor precisión de predicción?

- La regresión de lazo y cresta da como resultado sesgos casi idénticos. Sin embargo, la varianza de la regresión de la cresta es ligeramente menor que la varianza del Lasso. En consecuencia, el MSE mínimo de la regresión de la cresta es ligeramente más pequeño que el del Lasso.

- Sin embargo, los datos de la Figura 6.8 se generaron de tal manera que los 45 predictores se relacionaron con la respuesta, es decir, ninguno de los coeficientes verdaderos β1, ..., β45 fue igual a cero. El Lasso asume implícitamente que un número de coeficientes realmente es igual a cero. En consecuencia, no es sorprendente que la regresión de la cresta supere al Lasso en términos de error de predicción en este entorno.

- Ni la regresión de la cresta ni el Lasso dominarán universalmente al otro. En general, uno podría esperar que el Lasso funcione mejor en un entorno donde un número relativamente pequeño de predictores tiene coeficientes sustanciales, y los predictores restantes tienen coeficientes que son muy pequeños o iguales a cero. La regresión de crestas funcionará mejor cuando la respuesta sea una función de muchos predictores, todos con coeficientes de aproximadamente el mismo tamaño.

- Sin embargo, el número de predictores relacionados con la respuesta nunca se conoce a priori para conjuntos de datos reales. Se puede utilizar una técnica como la validación cruzada para determinar qué enfoque es mejor en un conjunto de datos en particular.

- Al igual que con la regresión de crestas, cuando las estimaciones de mínimos cuadrados tienen una varianza excesivamente alta, la solución de lazo puede producir una reducción de la varianza a expensas de un pequeño aumento del sesgo y, en consecuencia, puede generar predicciones más precisas. A diferencia de la regresión de crestas, el Lasso realiza una selección de variables y, por lo tanto, da como resultado modelos que son más fáciles de interpretar.

- Hay algoritmos muy eficientes para ajustar modelos tanto de cresta como de lazo; en ambos casos, todas las trayectorias de los coeficientes se pueden calcular con aproximadamente la misma cantidad de trabajo que un único ajuste de mínimos cuadrados

Un caso especial simple para la regresión de crestas y el lazo

- Para obtener una mejor intuición sobre el comportamiento de la regresión de la cresta y el lazo, considere un caso especial simple con n = p, y X una matriz diagonal con unos en la diagonal y ceros en todos los elementos off-diagonals.

- Para simplificar aún más el problema, suponga también que estamos realizando una regresión sin una intersección. Con estos supuestos, el problema habitual de mínimos cuadrados se simplifica a encontrar β1, ..., βp que minimizan

$$\sum_{j=1}^{p}(y_j - \beta_j)^2.$$

- En este caso, la solución de mínimos cuadrados viene dada por:

$$\hat{\beta}_j = y_j.$$

- Y en este escenario, la regresión de la cresta equivale a encontrar β1,. . . , βp tal que

$$\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

- Se minimiza, y el lazo equivale a encontrar los coeficientes tales que:

$$\sum_{j=1}^{p}(y_j - \beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

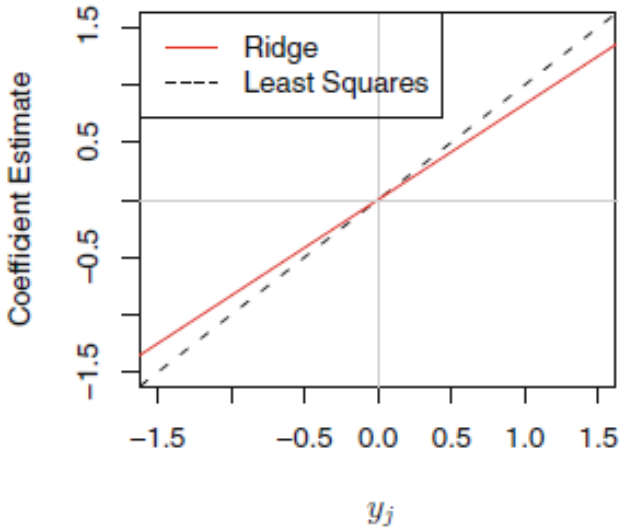
- Se minimiza. Se puede demostrar que en este escenario, las estimaciones de regresión de la cresta toman la forma

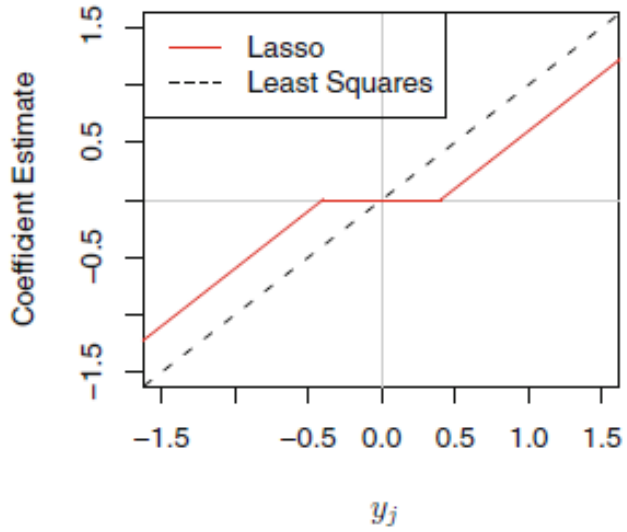
$$\hat{\beta}_j^R = y_j/(1 + \lambda),$$

- Y las estimaciones de Lasso toman la forma:

$$\hat{\beta}_j^L = \begin{cases} y_j - \lambda/2 & \text{if } y_j > \lambda/2; \\ y_j + \lambda/2 & \text{if } y_j < -\lambda/2; \\ 0 & \text{if } |y_j| \leq \lambda/2. \end{cases}$$

- La figura 6.10 muestra la situación. Podemos ver que la regresión de la cresta y el Lasso realizan dos tipos de contracción muy diferentes. En la regresión de crestas, cada estimación de coeficiente de mínimos cuadrados se reduce en la misma proporción. En contraste, el Lasso encoge cada coeficiente de mínimos cuadrados hacia cero en una cantidad constante, λ / 2; los coeficientes de mínimos cuadrados que son menores que λ / 2 en valor absoluto se reducen completamente a cero. El tipo de encogimiento
- La contracción por el Lasso en esta configuración simple (6.15) se conoce como umbral suave. El hecho de que algunos coeficientes de lazo se reduzcan por completo a cero explica por qué el lazo realiza la selección de características.
- En el caso de una matriz de datos X más general, la regresión de crestas reduce más o menos cada dimensión de los datos en la misma proporción, mientras que el Lasso reduce más o menos todos los coeficientes hacia cero en una cantidad similar, y los coeficientes suficientemente pequeños se reducen hasta cero





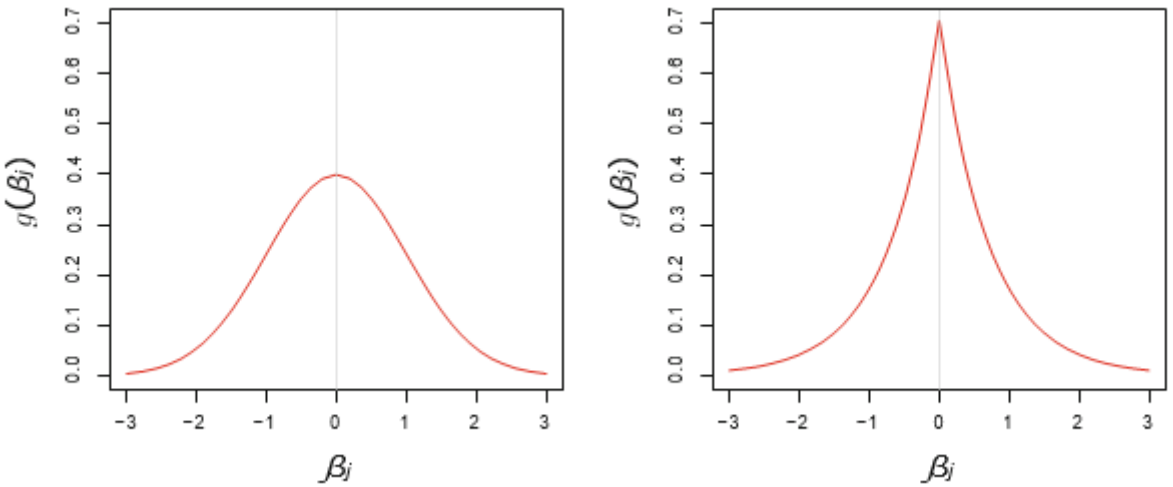
Interpretación bayesiana para la regresión de crestas y el Lasso

- Ahora mostramos que se puede ver la regresión de la cresta y el Lasso a través de una lente bayesiana. Un punto de vista bayesiano para la regresión supone que el vector de coeficientes β tiene alguna distribución previa, digamos p (β), donde β = (β0, β1, ..., βp) T. La probabilidad de los datos se puede escribir como f (Y | X, β), donde X = (X1, ..., Xp). Multiplicando la distribución anterior por la probabilidad campana nos da (hasta una constante de proporcionalidad) la distribución posterior, que toma la forma

$$p(\beta|X,Y) \propto f(Y|X,\beta)p(\beta|X) = f(Y|X,\beta)p(\beta),$$

- Donde la proporcionalidad anterior se sigue del teorema de Bayes, y la igualdad anterior se deriva del supuesto de que X es fijo. Asumimos el modelo lineal habitual

$$Y = \beta_0 + X_1\beta_1 + \ldots + X_p\beta_p + \epsilon,$$



Selección del parámetro de afinación

- Implementar la regresión de crestas y el Lasso requiere un método para seleccionar un valor para el parámetro de ajuste λ, o de manera equivalente, el valor de la restricción s.
- La validación cruzada proporciona una forma sencilla de abordar este problema. Elegimos una cuadrícula de valores de λ y calculamos el error de validación cruzada para cada valor de λ. Luego seleccionamos el valor del parámetro de ajuste para el cual el error de validación cruzada es más pequeño. Finalmente, el modelo se reajusta utilizando todas las observaciones disponibles y el valor seleccionado del parámetro de ajuste.
- La Figura 6.12 muestra la elección de λ que resulta de realizar una validación cruzada de dejar uno fuera en los ajustes de regresión de la cresta del conjunto de datos Credit. Las líneas verticales punteadas indican el valor seleccionado de λ. En este caso, el valor es relativamente pequeño, lo que indica que el ajuste óptimo solo implica un

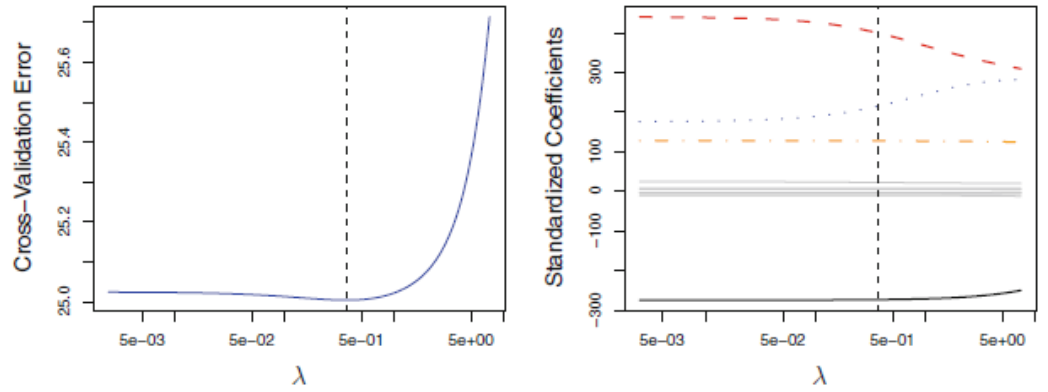


FIGURE 6.12. Left: *Cross-validation errors that result from applying ridge regression to the* `Credit` *data set with various value of* λ. Right: *The coefficient estimates as a function of* λ. *The vertical dashed lines indicate the value of* λ *selected by cross-validation.*

¡GRACIAS!

INFOX
educación sin fronteras