Linear Model Seleccion and Regularization

Matias Bajac

2024-09-24

Por que considerar alternativas al modelo lineal

- Mejorar la precisio: en especial cuando **p> n** para controlar la varianza.
- Interpretabilidad del modeo: removiendo predictores irrelevantes, facilita la interpretabilidad.

Veremos algunas aproximaciones para ver como seleccionar auotmaticamente el modelo.

Seleccion de modelos

Regresion lineal en altas dimensiones

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

con p muy grande.

dos grandes estrategias

1) Hallar el mejor subconjunto de predictoras que creemos se relacionan con la respuesta y ajustamos un modelo por minimos cuadrados con ella.
2) Esstimacion regularizada o **shrinkage**, ajustamos un modelo con p predictores pero los coneficientes estimados llevadoa a cero relativo a las estimaciones de minimios cuadrados. Objetivo reducir la variazna y tambiens e puede usar seleccion de variables

Datos sobre Daiabetes

- n = 442 pacientes con diabetes
- y avance de la enfermedad en 1 anio
- $x = x_1, ..., x_{10}$ caracteristicas de los pacientes
- Reservamos 100 observaciones de las 442 para test.

Queremos modelar la relacion y ~ x para realizar predicciones sobre como puede evolucionar la enfermedad.

Altas dimensiones

$$Y = f(X_1, ..., X_p) + \epsilon$$

- El valor de p puede ser muy alto
- Inicialmente en problemas de Biologia (pero luego se extiende a finanza, medicina,...)
- Los metodos estadisticos *tradicionales* estan pensados para contexto donde n > p, hay suficientes datos para cada parametro a estimar.
- Llamamos alta dimensiones a situaciones con n ~ p o incluso n < P.

Cuando p es alto

-Aproximaciones globales 'simples' tienen mucho error.

- Aproximaciones locales no son adecuadas

Regresion lineal

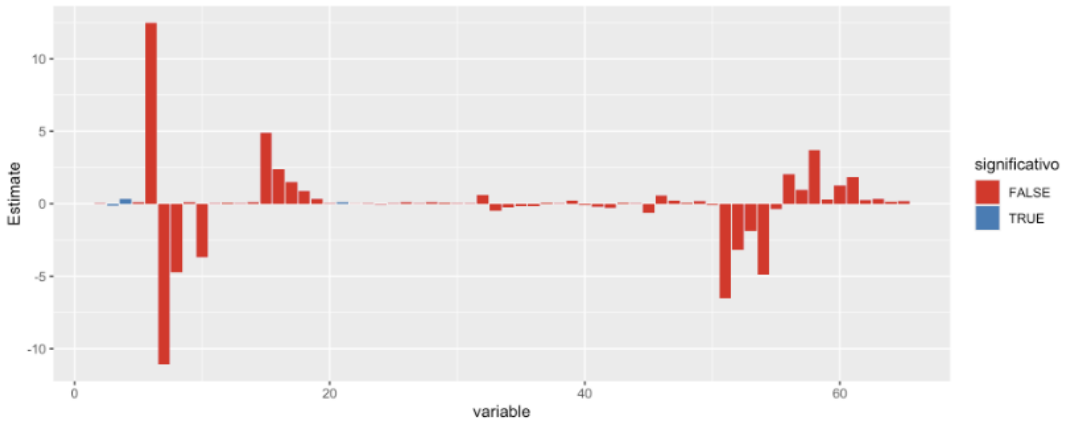
- La estimacion por minimos cuadrados es muy inestable si p ~ n

-Aumenta el error por varianza.

- Ejemplo trivial: con n=2, regresion simple tiene ajuste perfecto

Regresión lineal en los datos Diabetes

```
1  mm.reg <- lm(y~. , data=train) # modelo completo
```



- Muy pocas variables significativas
- Bajo poder predictivo

##Distancias a pares ##

Geometria en espacios de altas dimensiones es poco intuituva

Pequena simulacion

-$X\ Unif([0,1]^p)$

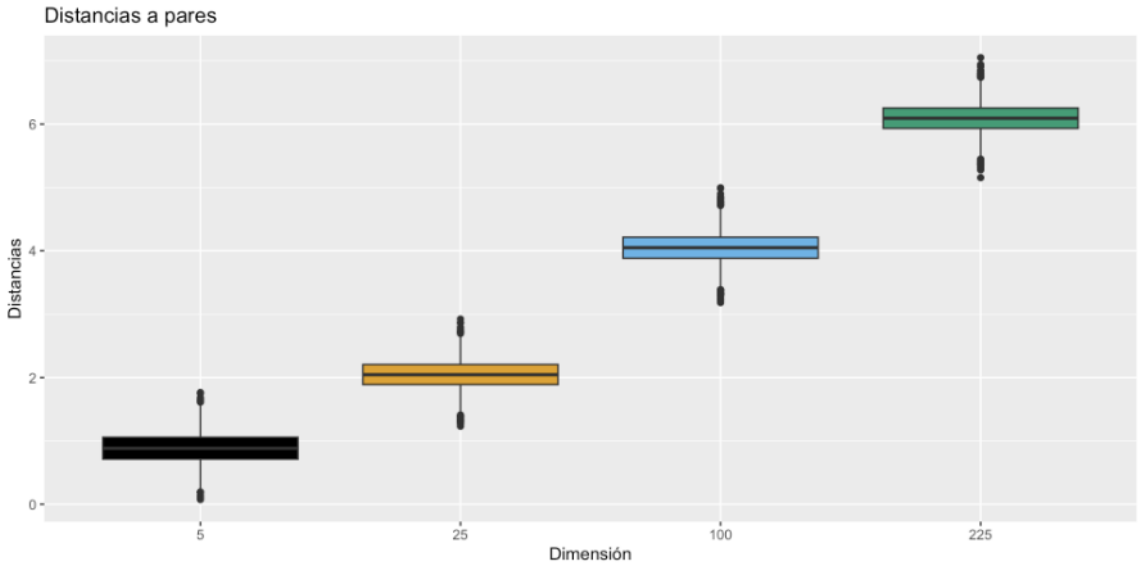
-Simular n $= 100$ replicas, para $p \in (5, 25, 100, 225)$

- obtener $d_2(i,j) = \sqrt{\sum (x_i - x_j)^2}$

```
# Fijamos parámetros
n = 100
p <- 25

# Simulamos los datos
xs <- matrix(runif(n*p), ncol=p )

# Calculamos distancias
#xs |> dist() |> as.numeric()
```

Distancias a pares

Cuando p aumenta, no hay vecinos *cercanos* !!



Podemos mejorar

-Modelo lineal no tiene buenos resultados, pero es un buen punto de partida.

-Podemos mejorar su performance, sustituyendo MCO por otras tecnicas de ajuste.

-Controlar el error por varianza cuando p es grande

Seleccion de modelos

Regresion lineal en altas dimensiones

$$Y = \beta_0 + \beta_1 X_1 \ + ... + \beta_P X_P \ + \epsilon$$

con p muy grande

1) Hallar el mejor subconjunto de predictores: recorrer automaticamente el espacio de modelos posibles.

2)Esstimacion regularizada o **shrinkage**: penalizar el valor de los coeficientes y llevarlos hacia 0 (los mata)

Las estimaciones de los parametros de los niveles mas bajos son tiradas.

El **shrinkage** ocurre porque los parametros de los niveles bajos (β_p) son influenciados por:

1) El conjunto de datos que dependen directamente de ese parametro
2) Los parametros de niveles mas altos de los cuales dependen los parametros de niveles mas bajos(y son afectados por todos los datos)

Modelos posibles

Si tenemos p_0 posibles variables explicativas La relacion puede ser no - aditiva y no lineal, ingenieria de explicativas

-Incluir **todas** las variables predictoras

- Incluir **todas** las interacciones a pares
- Incluir todas las variables numericas al cuadrado

Quedan $p = p_0 + \frac{p_0(p_0-1)}{2} + p_0$ posibles variables para incluir en el modelo.

En el ejemplo reuslta en 64 varoiables predictoras

En total hay 2^p posibles modelos para construir

Seleccionar subconjunto

-Elegir un subconjunto de los p predictores para incluir en el modelo

- Analogo a seleccionar un modelo entre los 2ˆp modelos posibles.

Estrategias mas comunes

-Seleccionar el mejor subconjunto

-Seleccionar hacia adelante (forward)

- Seleccionar hacia atras (backward)

Seleccionar el mejor subconjunto (meh)

1) M_0 es el modelo nulo que no tiene predictores. Predice la media muestral para cada observacion.
2) para k = 1,2,. . . .,p:

a) ajusto todos (combinaciones de p modelos con k variables) modelos que contienen k predicotres

b) Seleccionamos el mejor de las combinaciones de parametros y modelo y lo llamamos M_k, Mejor en base a menor MSE o mayor R cuadrado.

3) Seleccionamos el mejor modelo entre $M_0, ..., M_p$ usando cross validation para el error de prediccion, C_p, AIC, BIC, o R^2 ajustado.

Este metodo **no puede ser aplicado** para problemas con p grande porque es computacionalmente costoso.

Seleccion hacia adelante (forward)

-Ventajas computacionales claras respecto a seleccionar el mejor subconjunto

- No garantiza encontrar el mejor modelo posible dentro de todos los 2^p modelos contenidos subconjuntos de p predictores

Supongamos que tenemos 5 predicores

X_1, X_2, X_3, X_4, X_5 empieza siendo la mejor X_3 luego agrego otra variable X_3 X_4 , o X_3 X_5 etc y asi sucesivamente hasta encontrar el mejor model

1) $M_0 es el modelo nulo, el cual no contine predictores.

2) para k = 0,0,. . . p-1

a) Considerar todos los p - k modelos que aumenten los predictores en M_k con un predictor adicional.

b) Elijo el mejor entre los p - k modelos y lo llamamos $M_{k+1}. Mejor en base a menor MSE o mayor R^2

3) Seleccionamos el mejor modelo entre $M_0, ..., , M_p$ usando cross validation para el error de prediccion, AIC, BIC o R^2 ajustado.

Seleccion hacia atras

- Comienza con predictores y va quitando de a uno, hasta que se llega al modelo nulo.
- Es una busqueda guiada en el espacio de modelos posibles.

-Mucho menos computo que encontrar el mejor subconjunto.

Seleccion hacia atras: algoritmo

1)M_p reperesenta el modelo con todos los predictores

2) para $\forall\ k\ \in\ (p, p-1, ..., 1)$

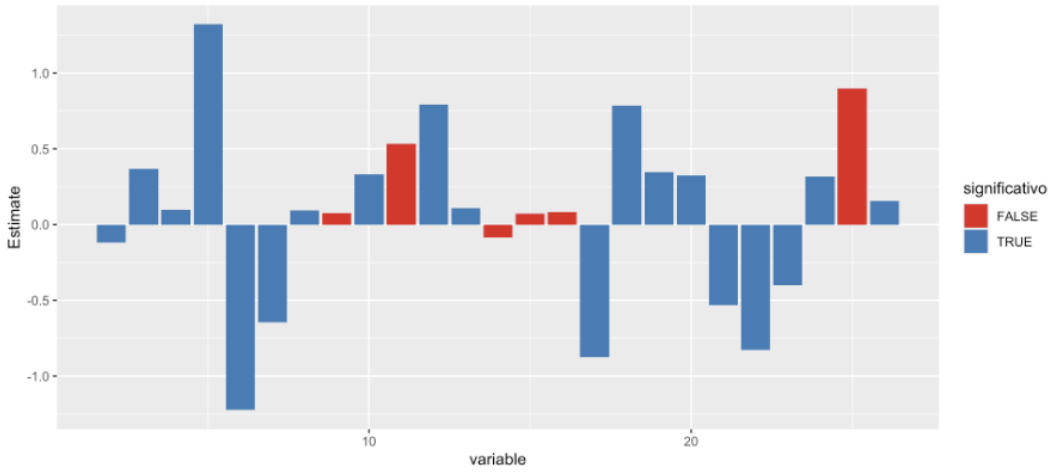
a) Considerar los (k) modelos con 1 variable menos que M_k, para el total de k -1 predictores.

b) Seleccionar el de menor MSE y llamarlo M_{k-1}

c) Seleccionamos el mejor modelo entre $M_0, ..., M_p$ usando cross validation para el error de prediccion, AIC,BIC, o R^2 ajustado.

En los datos

```
1  st <- step(mm.reg, trace = 0)
```



- 25 variables, la mayoría significativas

Regularizacion

##Estimacion penalizada ##

-Los metodos de seleccion de subconjuntos de modelos usan MCO (minimos cuadrados) para el ajuste de un modelo lineal con un subconjunto de predictores

- Una alternativa es ajustar un modelo que tenga **todos los p predictores** usando una tecnica para regularizar o restringir los coeficientes estimados. **Que restrinja la estimacion de los coeficientes a 0**
- Esta restriccion en los coeficientes estimados reduce la varianza

El metodo clasico de MCO partimos de en con modelo lineal.

$$Y = \beta_0 + \beta_1 X_1 \ + ... + \beta_P X_P \ + \epsilon$$

y obtenemos $\hat{\beta}$ minimizando

$$\hat{\beta} = argmin\beta = \{\sum_i (y_i - \beta X_i)^2\}$$

LA IDEA BASICA ES PENALIZAR EL VALOR DE LOS COEFICIENTES, PARA ACERCARLOS A 0

$$\hat{\beta} = argmin\beta = \{\sum_i (y_i - \beta X_i)^2 + \lambda c(\beta)\}$$

-Varios metodos, usnado distintos $c(\beta)$ factor de penalizacion

- λ controla el peso relativo de la penalidad
- **Penalidades** mas usadas: **Ridge** y **Lasso**

Regresion Ridge

Propone penalidad cuadratica, $c(\beta) = \sum_j \beta_j^2 = ||\beta||_2$

$$\hat{\beta} = argmin\beta = \{\sum_i (y_i - \beta X_i)^2 + \lambda \sum_j \beta_j^2\}$$

con $\lambda > 0$

- $\lambda = 0$ Nos queda la solucion de MCO
- $\lambda \to \infty$ nos queda $\beta_j \to 0$ penaliza muho
- λ es un parametro de **tuneo** que tiene que ser seleccionado separadamente
- El factor de penalizacion $\lambda \sum_{j=1}^{p} \beta_j^2$ es chico cuando los coeficientes β estan proximos a 0
- λ controla la importancia que lal funcion de perdida le asigna a la penalizacion
- Para cada valor de λ se obtienen diferentes estimaciones de los coeficientes, por lo que resulta fundamental seleccionar el valor de λ optimo (CV)
- El parametro λ de tuneo sirve para controlar el impacto en las esitmaciones de los coeeficientes

```
library(glmnet)
```

```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.1-8
```

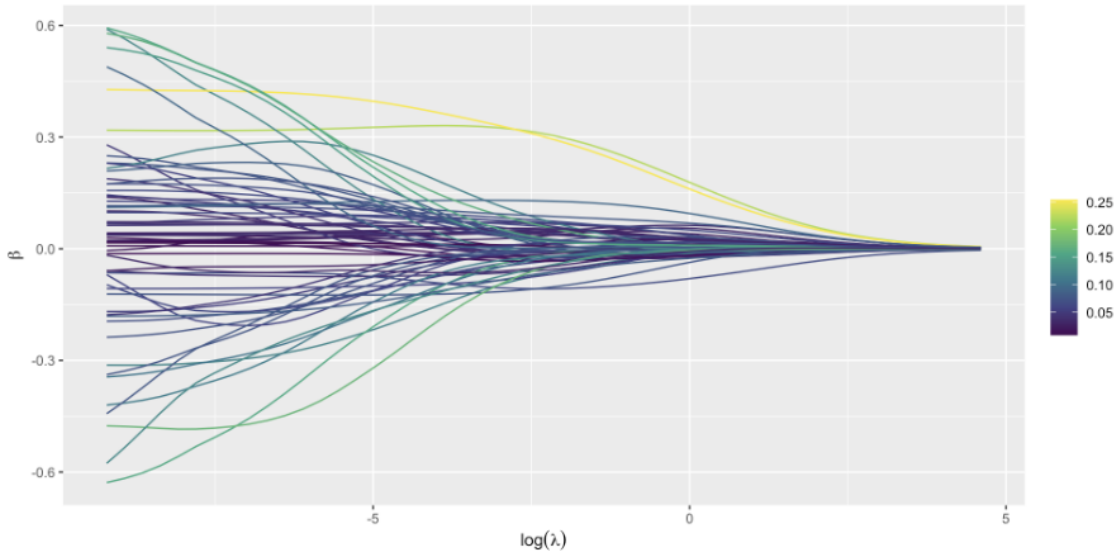
```
#xs <- model.matrix(y ~ ., data = train) # matriz de predictores
#gr <-  10^seq(2,-4, length = 50)      # valores de lambda

#rd <- glmnet(x = xs, y = train$y, alpha = 0, lambda = gr)
```

El resultado es una secuencia de coeficientes para los distintos valores de lambda

Podemos ver como cambian los coeficientes cuando cambia el peso de la penalidad.

Las que mas rapido se van son lasm menos importantes

Resultados Ridge



Escala de los predictores

-En regresion la escala de X_j no es muy importante

- Para estimar β_λ^{Ridge}, la escala de las predictoras tiene mucho impacto, debido a la penalidad.

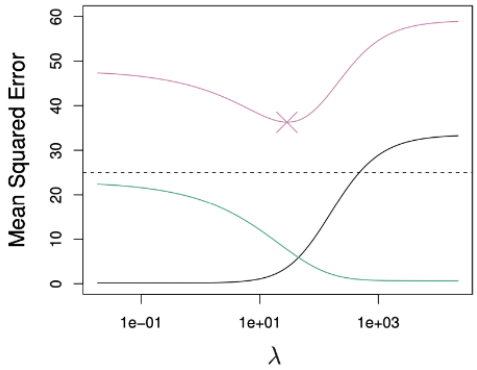
-Para evitar este efecto, es usual **estandarizar** los predictores

Error de prediccion en Ridge

Generalmente Ridge tiene **mejor** performance predictiva que **MCO**

Error de predicción en Ridge

Generalmente Ridge tiene mejor performance predictiva que MCO.



Datos simulados:

- n = 50, p = 45
- Negro: SESGO
- Verde: VARIANZA
- Rosa: Error Generalización

El optimo se da donde las dos se cruzan

Desventajas Ridge

- Inclue a todos los **predictores** en el modelo final, es decir no selecciona variables.
- La penalizacion **reduce** el valor de los coeficientes pero no los hace nulos, lo cual dificulta la interpretacion si d es grande

Regresion Lasso

$$\hat{\beta}^{Lasso} = argmin\beta = \{\sum_i (y_i - \beta X_i)^2 + \lambda \sum_j |\beta_j|\}$$

Un problema de Ridge, es que mantiene todos los B_j distinto a 0, no es lo mismo que **Seleccionar un modelo**

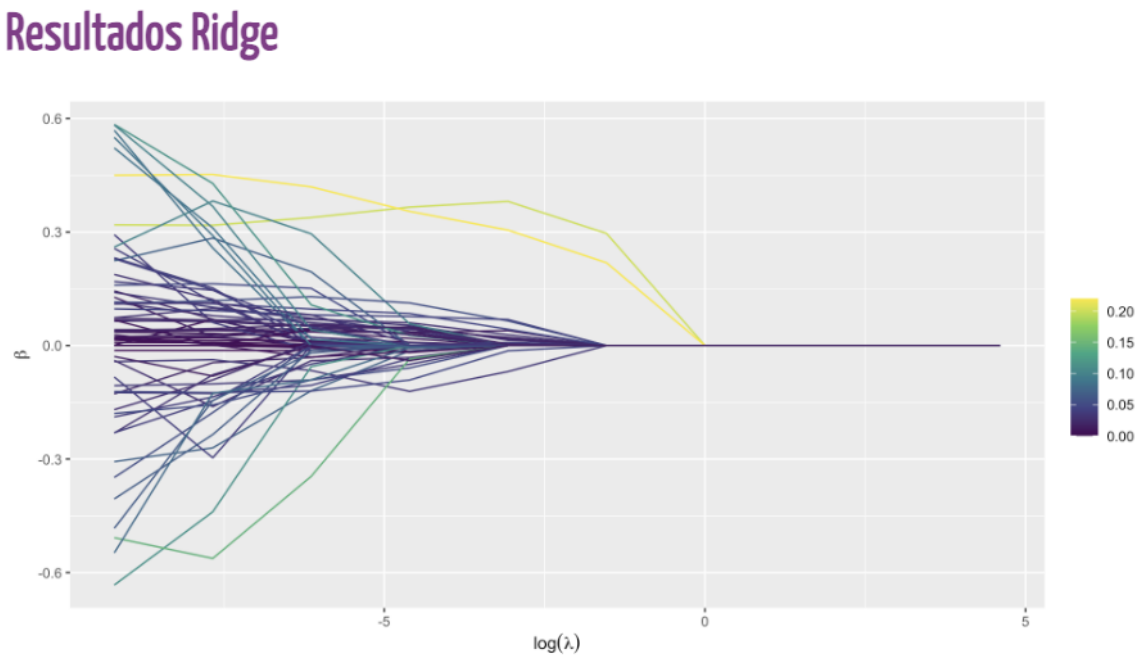
Cambiando la **penalidad** podemos atacar el problema .

- Reducce los coeficientes hacia 0 (los mata)
- Debido a la penalidad, cuando λ es muy grande, hay coficientes (β_j) exactamente iguales a 0.
- Se puede considerar como un procedimiento de **seleccion de modelos**
- Al igual que ridge, hay que elegir un valor para λ (me defino una grilla) y hago cross validation.

```
#xs <- model.matrix(y ~ ., data=train) # matriz de predictores
#gr <-  10^seq(2,-4, length = 10)      # valores de lambda

#laso <- glmnet(x=xs, y=train$y, alpha=1, lambda = gr)
```

El resultado es una secuencia de coeficientes para los diferentes valores de λ

Podemos ver como cambian los coeficientes cuando cambia el peso de la penalidad



Comparar penalidades

Comparar penalidades

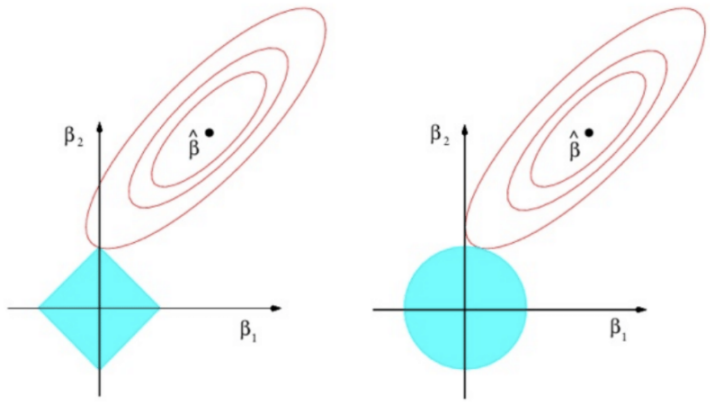


FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions,* $|\beta_1|+|\beta_2| \leq s$ *and* $\beta_1^2+\beta_2^2 \leq s$, *while the red ellipses are the contours of the RSS.*

Como en la regresion ridge, lasso contrae los coeficientes a 0, pero en este caso si se les asigna el valor 0 a lambda si lambda es lo suficientemente grande. por lo que **selecciona variables**

otra vez, es importante la eleccion de λ (CV)

Ultimo grafico:

Obs: $\hat{beta}$ indica la solucion de minimos cuadrados, las elipses las curvas de nivel de RSS las regiones corresponden a $|\beta_1| + |\beta_2| \leq s$ y $\beta_1^2 + \beta_2^2 \leq s$ respectivamente

Si s es suficientemente grande , entonces las regiones sombreadas contienen a $\hat{\beta}$ y los estimadores ridge y lasso coinciden con los minimos cuadrados.

Los estimadores de ridge y lasso se obtienen en la primera curva de nivel "toca" la region sobmreada (es decir cuando se minimiza RSS sujeto a als restricciones $|\beta_1| + |\beta_2| \leq s$ y $\beta_1^2 + \beta_2^2 \leq s$ respectivamente)

En la figura para laso, esto se da cando $\beta_1 = 0$

Para ridge, debido a la forma circular, la interseccion entre la curva de nivel y la frontera de la region generalmente no ocurre en un eje y por lo tanto los estimadores no seran nulos y no se seleccionan varables.