Regularización

agosto 2017

Los **métodos de selección de variables** implican usar el estimador de mínimos cuadrados para ajustar un modelo lineal con sólo un subconjunto de las variables predictoras.

Los métodos de **regularización** implican ajustar el modelo lineal incluyendo todas las variables disponibles usando una técnica que restringe o regulariza los estimadores de los coeficientes "encogiendolos" hacia el cero.

Los **métodos de selección de variables** implican usar el estimador de mínimos cuadrados para ajustar un modelo lineal con sólo un subconjunto de las variables predictoras.

Los métodos de **regularización** implican ajustar el modelo lineal incluyendo todas las variables disponibles usando una técnica que restringe o regulariza los estimadores de los coeficientes "encogiendolos" hacia el cero.

Esta técnica reduce la varianza de los estimadores significativamente.

Los **métodos de selección de variables** implican usar el estimador de mínimos cuadrados para ajustar un modelo lineal con sólo un subconjunto de las variables predictoras.

Los métodos de **regularización** implican ajustar el modelo lineal incluyendo todas las variables disponibles usando una técnica que restringe o regulariza los estimadores de los coeficientes "encogiendolos" hacia el cero.

Esta técnica reduce la varianza de los estimadores significativamente.

Los más conocidos: **Regresión ridge** y **Lasso**.

Los **métodos de selección de variables** implican usar el estimador de mínimos cuadrados para ajustar un modelo lineal con sólo un subconjunto de las variables predictoras.

Los métodos de **regularización** implican ajustar el modelo lineal incluyendo todas las variables disponibles usando una técnica que restringe o regulariza los estimadores de los coeficientes "encogiendolos" hacia el cero.

Esta técnica reduce la varianza de los estimadores significativamente.

Los más conocidos: **Regresión ridge** y **Lasso**.

Regresión ridge

Recordemos que el estimador de mínimos cuadrados se define como

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

Regresión ridge

Recordemos que el estimador de mínimos cuadrados se define como

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

El estimador ridge se define como

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

- Como el estimador de mínimos cuadrados, el estimador ridge busca estimadores que ajusten bien a los datos, buscando que la RSS sea pequeña. Sin embargo, el término de penalización $\lambda \sum_{j=1}^{p} \beta_j^2$ es chico cuando los coeficientes β_j están cerca de cero y tiene el efecto de encojer los coeficientes hacia el cero.

- Como el estimador de mínimos cuadrados, el estimador ridge busca estimadores que ajusten bien a los datos, buscando que la RSS sea pequeña. Sin embargo, el término de penalización $\lambda \sum_{j=1}^{p} \beta_j^2$ es chico cuando los coeficientes β_j están cerca de cero y tiene el efecto de encojer los coeficientes hacia el cero.
- El parámetro de calibración λ sirve para controlar el impacto relativo de estos dos términos en los estimadores

- Como el estimador de mínimos cuadrados, el estimador ridge busca estimadores que ajusten bien a los datos, buscando que la RSS sea pequeña. Sin embargo, el término de penalización $\lambda \sum_{j=1}^{p} \beta_j^2$ es chico cuando los coeficientes β_j están cerca de cero y tiene el efecto de encojer los coeficientes hacia el cero.
- El parámetro de calibración λ sirve para controlar el impacto relativo de estos dos términos en los estimadores
- Elegir un buen valor de λ es crítico. Para eso usamos validación cruzada.

Escala de las covariables

Los estimadores regularizados no son equivariantes por cambios de escala. Esto quiere decir que si se cambian las unidades de medición de las variables explicativas, las predicciones cambiarán.

Como se penaliza usando la norma de beta, los tamaõs de los coeficientes deben ser comparables.

Conviene estandarizar las variables explicativas para que estén todas en la misma escala reemplazando cada x_{ij} por

$$\tilde{x}_{ij} = \frac{x_{ij}}{\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \overline{x_j})^2}$$

Datos de cancer de próstata

Este conjunto de datos se recolectó con el objetivo de indentificar factores de riesgo para cancer de próstata Se quiere predecir el logaritmo of PSA (lpsa) a partir de las siguientes variables

- `lcavol` log del volumen del cancer
- `lweight` log del peso de la próstata
- `age` edad
- `lbph` log of benign prostatic hyperplasia amount lbph
- `svi` seminal vesicle invasion
- `lcp` log de penetración capsular
- `gleason` índice de Gleason
- `pgg45` percent of Gleason scores 4 or 5

Aplicación a los datos de cancer de próstata

```
>  prostate<-read.table("prostate.txt",header=TRUE)
>  prostatelm<-lm(lpsa~.,data=prostate)
>  summary(prostatelm)

Call:
lm(formula = lpsa ~ ., data = prostate)

Residuals:
     Min       1Q   Median       3Q      Max
-1.76644 -0.35510 -0.00328  0.38087  1.55770
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.181561   1.320568   0.137  0.89096
lcavol       0.564341   0.087833   6.425 6.55e-09
lweight      0.622020   0.200897   3.096  0.00263
age         -0.021248   0.011084  -1.917  0.05848
lbph         0.096713   0.057913   1.670  0.09848
svi          0.761673   0.241176   3.158  0.00218
lcp         -0.106051   0.089868  -1.180  0.24115
gleason      0.049228   0.155341   0.317  0.75207
pgg45        0.004458   0.004365   1.021  0.31000

---
Residual standard error: 0.6995 on 88 degrees of freedom
Multiple R-squared:  0.6634,    Adjusted R-squared:  0.6328
F-statistic: 21.68 on 8 and 88 DF,  p-value: < 2.2e-16
```

```
>   set.seed(12)
>   prostateglm1<-glm (lpsa~.,data=prostate)
>   cverr1<-cv.glm(prostate,prostateglm1,K=5)
>   cverr1$delta
[1] 0.6061653 0.5861579
>
```

Aplicación a los datos de cancer de próstata

```
> fitridge<-glmnet(x,y,alpha=0)
> names(fitridge)
 [1] "a0"         "beta"       "df"         "dim"        "lambda"
 [6] "dev.ratio" "nulldev"    "npasses"    "jerr"       "offset"
[11] "call"       "nobs"
```

Aplicación a los datos de cancer de próstata

```
> fitridge
Call:  glmnet(x = x, y = y, alpha = 0)

      Df      %Dev    Lambda
 [1,]  8 3.484e-36 843.40000
 [2,]  8 5.118e-03 768.50000
 [3,]  8 5.613e-03 700.20000
 [4,]  8 6.156e-03 638.00000
 [5,]  8 6.750e-03 581.30000
 [6,]  8 7.402e-03 529.70000
 [7,]  8 8.115e-03 482.60000
 [8,]  8 8.897e-03 439.80000
 [9,]  8 9.753e-03 400.70000
[10,]  8 1.069e-02 365.10000
```

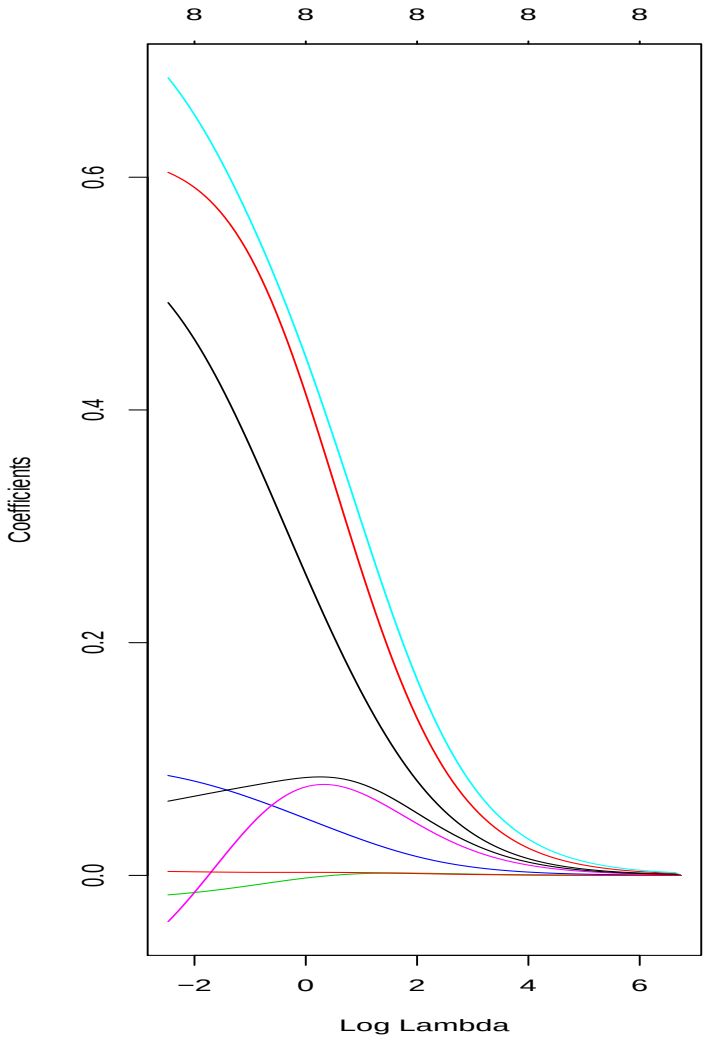
```
[11,]  8 1.171e-02 332.70000
[12,]  8 1.284e-02 303.10000
[13,]  8 1.406e-02 276.20000
[14,]  8 1.541e-02 251.60000
[15,]  8 1.687e-02 229.30000
[16,]  8 1.848e-02 208.90000
[17,]  8 2.023e-02 190.40000
[18,]  8 2.214e-02 173.50000
[19,]  8 2.423e-02 158.00000
[20,]  8 2.650e-02 144.00000
[21,]  8 2.898e-02 131.20000
[22,]  8 3.168e-02 119.60000
[23,]  8 3.463e-02 108.90000
[24,]  8 3.783e-02  99.26000
[25,]  8 4.130e-02  90.44000
[26,]  8 4.508e-02  82.40000
[27,]  8 4.917e-02  75.08000
```

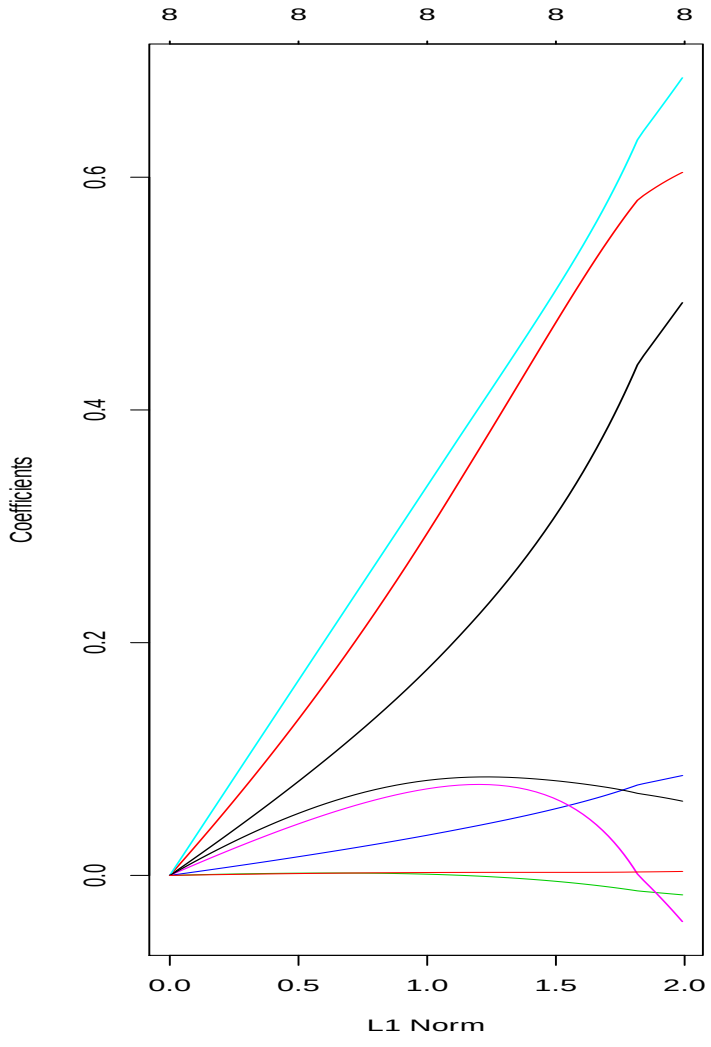
```
 [87,]   8 6.408e-01   0.28270
 [88,]   8 6.432e-01   0.25760
 [89,]   8 6.455e-01   0.23470
 [90,]   8 6.475e-01   0.21380
 [91,]   8 6.494e-01   0.19480
 [92,]   8 6.510e-01   0.17750
 [93,]   8 6.525e-01   0.16180
 [94,]   8 6.539e-01   0.14740
 [95,]   8 6.551e-01   0.13430
 [96,]   8 6.561e-01   0.12240
 [97,]   8 6.571e-01   0.11150
 [98,]   8 6.579e-01   0.10160
 [99,]   8 6.587e-01   0.09257
[100,]   8 6.593e-01   0.08434
```

Aplicación a los datos de cancer de próstata

```
> par(mfrow=c(1,2))
> plot(fitridge,xvar="lambda")
> plot(fitridge,xvar="norm")
```

Aplicación a los datos de cancer de próstata





Comentarios de la figura anterior

- En la figura de la izquierda, cada curva corresponde al estimador ridge de uno de los coeficientes, en función de λ. A medida que λ aumenta, los estimadores se acercan a cero de manera suave.
- En la figura de la izquierda, cada curva corresponde al estimador ridge de uno de los coeficientes, en función de la norma de $\hat{\beta}$.

¿Cómo elegimos λ?

Decidir qué valor de λ producirá mejores predicciones para nuevos pacientes.

El compromiso entre sesgo y varianza

En aprendizaje automático:
Sesgo: es el error que se produce al aproximar un problema de la vida real, que puede ser extremadamete complicado, por un modelo relativamente simple.

El compromiso entre sesgo y varianza

En aprendizaje automático:
Sesgo: es el error que se produce al aproximar un problema de la vida real, que puede ser extremadamete complicado, por un modelo relativamente simple.

Varianza: es una medida de cuánto variarían las pedicciones si se usara otra muestra de entrenamiento

El compromiso entre sesgo y varianza

En aprendizaje automático:
Sesgo: es el error que se produce al aproximar un problema de la vida real, que puede ser extremadamete complicado, por un modelo relativamente simple.

Varianza: es una medida de cuánto variarían las pedicciones si se usara otra muestra de entrenamiento

Mayor flexibilidad $\Rightarrow$ menor sesgo y mayor varianza.
Menor flexibilidad $\Rightarrow$ mayor sesgo y menor varianza.

El compromiso entre sesgo y varianza

En aprendizaje automático:
Sesgo: es el error que se produce al aproximar un problema de la vida real, que puede ser extremadamete complicado, por un modelo relativamente simple.

Varianza: es una medida de cuánto variarían las pedicciones si se usara otra muestra de entrenamiento

Mayor flexibilidad $\Rightarrow$ menor sesgo y mayor varianza.
Menor flexibilidad $\Rightarrow$ mayor sesgo y menor varianza.

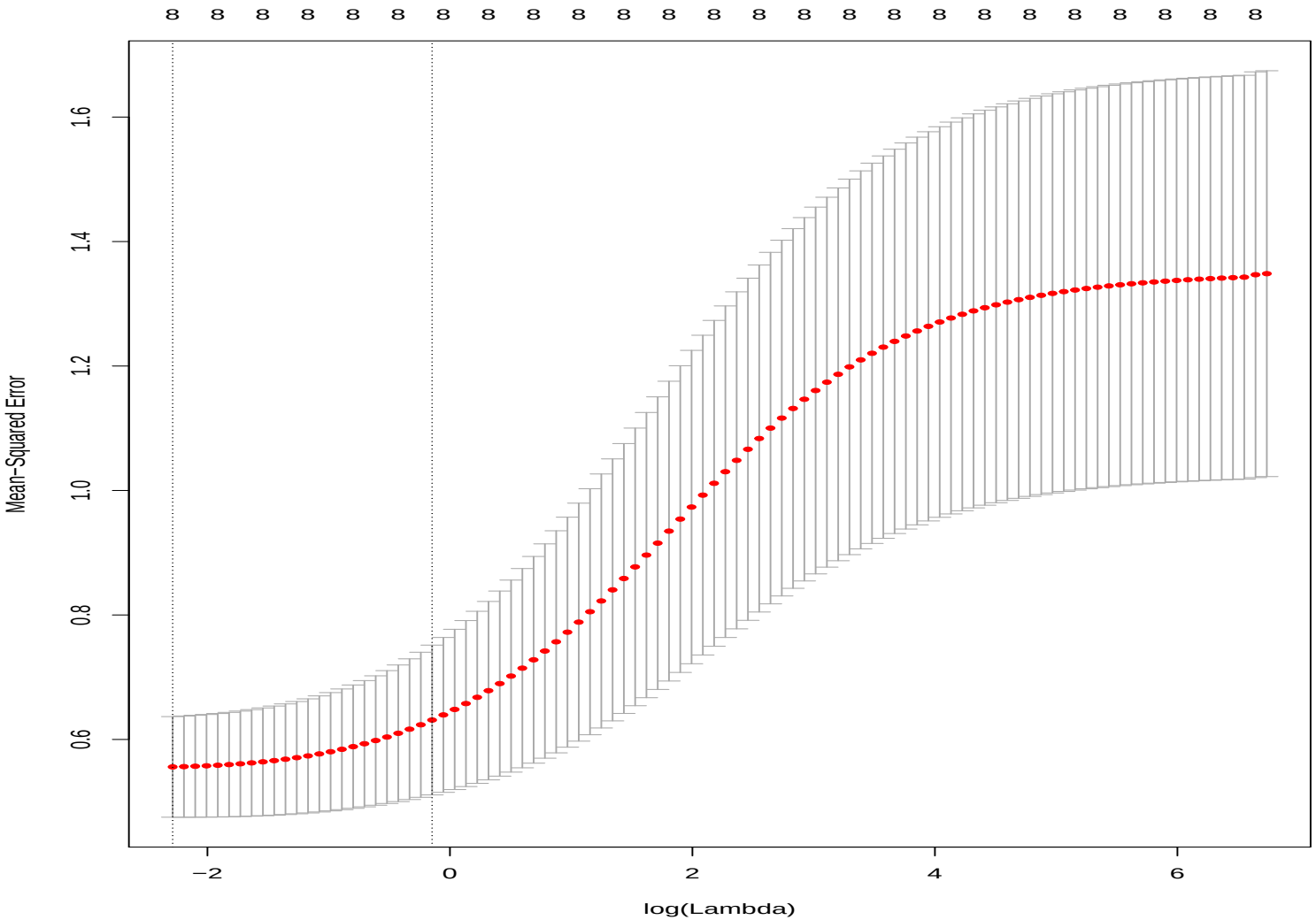
Ridge y lasso ajustan este compromiso automáticamente. Logran un equilibrio entre sesgo y varianza de manera de minimizar el error de pedicción. ¿Como? Eligiendo λ apropiadamente.

Elección del valor de λ por validación cruzada

1. Tomamos una grilla de valores de λ y calculamos el error de validación cruzada para cada λ.
2. Elegimos el valor de λ para el cual el error de validación cruzada es menor.
3. Finalmente se ajusta el modelo utilizando todas las covariables y el valor elegido de λ.

Aplicación a los datos de cancer de próstata

```
> set.seed(12)
> cv0=cv.glmnet(x,y,alpha=0,nfolds=5)
> cv0$lambda.min
[1] 0.09256606
> which(cv0$lambda==cv0$lambda.min)
[1] 99 ## este es el valor de lambda que minimiza el error
## de testeo estimado.
> cv0$cvm[99]
[1] 0.5068303 ## este es el estimador del error de testeo
## cuando lambda=98.
> plot(cv0)
```


8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8 8
Mean-Squared Error
0.6
0.8
1.0
1.2
1.4
1.6
−2
0
2
4
6
log(Lambda)

El Lasso

Recordemos que el estimador de mínimos cuadrados se define como

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

El Lasso

Recordemos que el estimador de mínimos cuadrados se define como

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

y el estimador ridge se define como

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

El Lasso

Recordemos que el estimador de mínimos cuadrados se define como

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

y el estimador ridge se define como

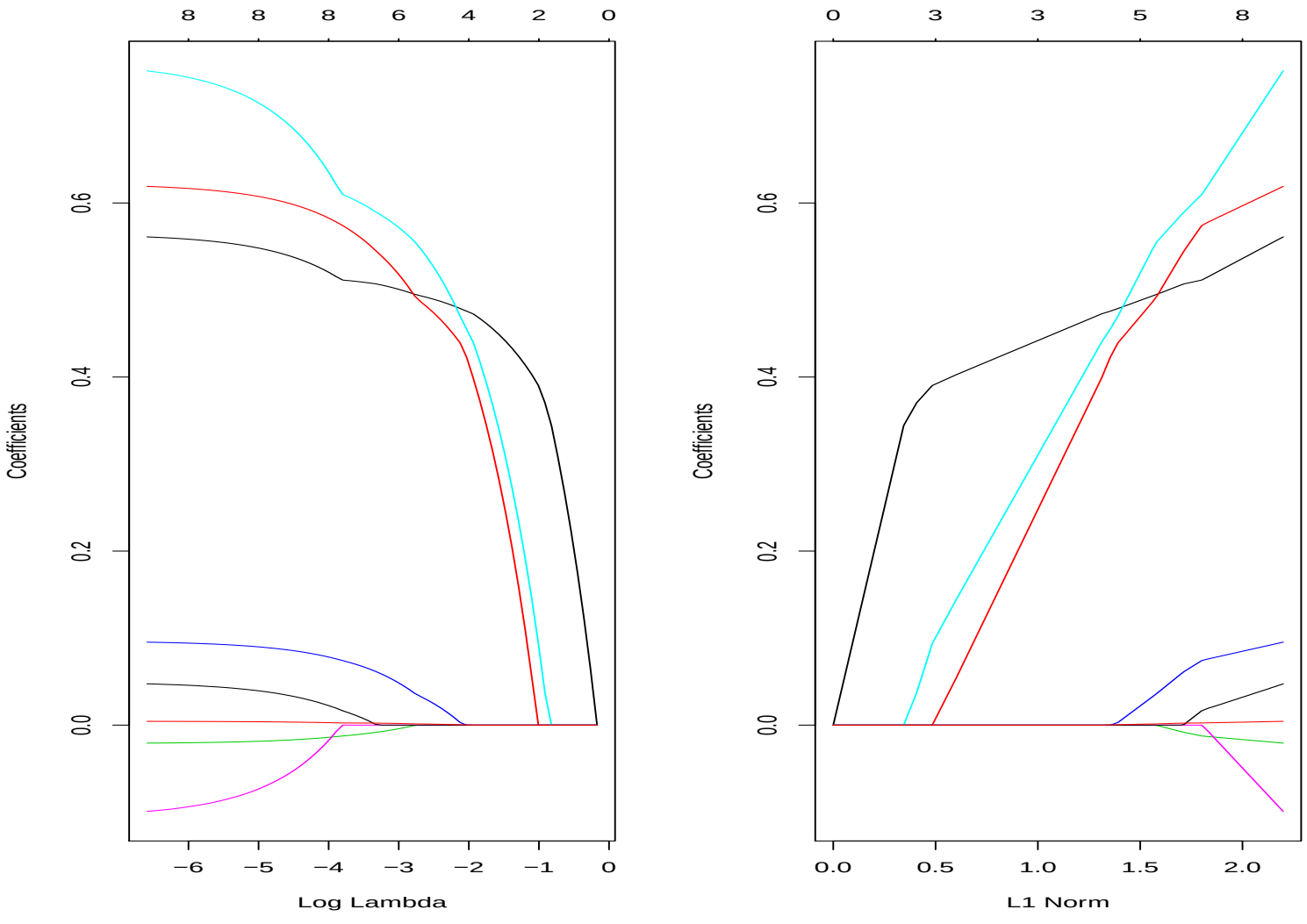
$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

el estimador lasso se define como

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Aplicación a los datos de cancer de próstata

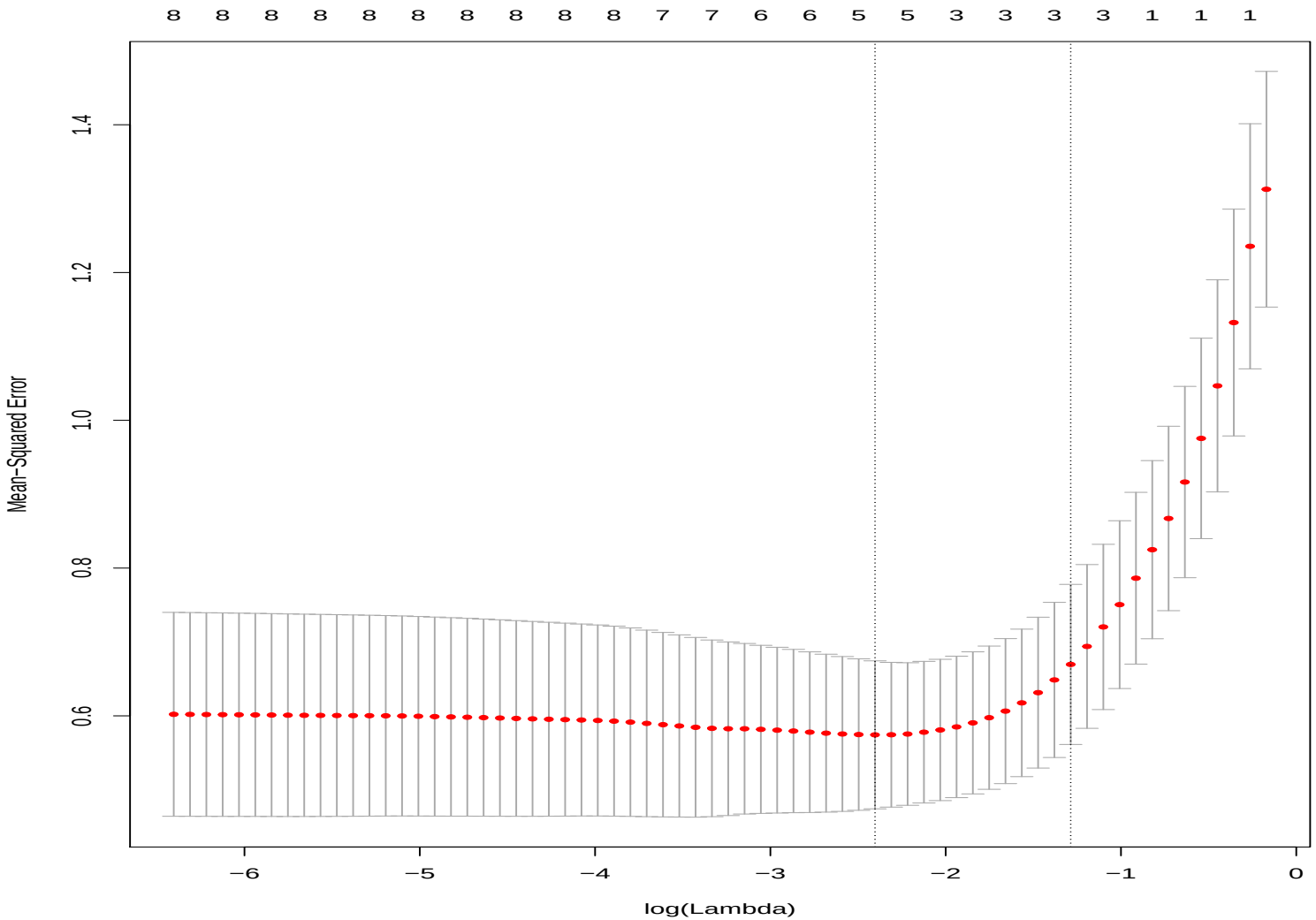
```
set.seed(11)
fitlasso<-glmnet(x,y,alpha=1)
par(mfrow=c(1,2))
plot(fitlasso,xvar="lambda")
plot(fitlasso,xvar="norm")
```


8 8 8 6 4 2 0
Coefficients
0.0
0.2
0.4
0.6
−6 −5 −4 −3 −2 −1 0
Log Lambda
0 3 3 5 8
Coefficients
0.0
0.2
0.4
0.6
0.0 0.5 1.0 1.5 2.0
L1 Norm

Elección de λ por validación cruzada

```
> cv1=cv.glmnet(x,y,alpha=1,nfolds=5)
> which(cv1$lambda==cv1$lambda.min)
[1] 26
>   cv1$cvm[25]
[1] 0.5784757

> plot(cv1)
```


8 8 8 8 8 8 8 8 8 8 7 7 6 6 5 5 3 3 3 3 1 1 1
Mean-Squared Error
0.6
0.8
1.0
1.2
1.4
−6
−5
−4
−3
−2
−1
0
log(Lambda)

Coeficientes estimados

```
>    coef(cv1, s ="lambda.min" , exact = FALSE)
9 x 1 sparse Matrix of class "dgCMatrix"
                             1
(Intercept)  0.085443850
lcavol       0.503903656
lweight      0.531274516
age         -0.006222137
lbph         0.054941054
svi          0.580620260
lcp          .
gleason      .
pgg45        0.002086482
```

Coeficientes estimados

```
>    coef(cv1, s ="lambda.1se" , exact = FALSE)

9 x 1 sparse Matrix of class "dgCMatrix"
                      1
(Intercept) 1.1009298
lcavol      0.4327040
lweight     0.2024125
age         .
lbph        .
svi         0.2714114
lcp         .
gleason     .
pgg45       .
```

Propiedad de selección de variables del lasso.

Decimos que el lasso da modelos ralos, es decir donde sólo algunos coeficientes son no nulos.

Propiedad de selección de variables del lasso.

Decimos que el lasso da modelos ralos, es decir donde sólo algunos coeficientes son no nulos.

Muchas veces se necesita identificar un subconjunto de covariables lo más pequeño posible con el cual poder predecir correctamente (por ejemplo, para usar en diagnósticos futuros).

Geometría de ridge y lasso

Puede probarse que

$$\hat{\boldsymbol{\beta}}^R = \arg\max_{\boldsymbol{\beta}} L(\beta_0, \beta_1, \ldots, \beta_p)$$

sujeto a

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

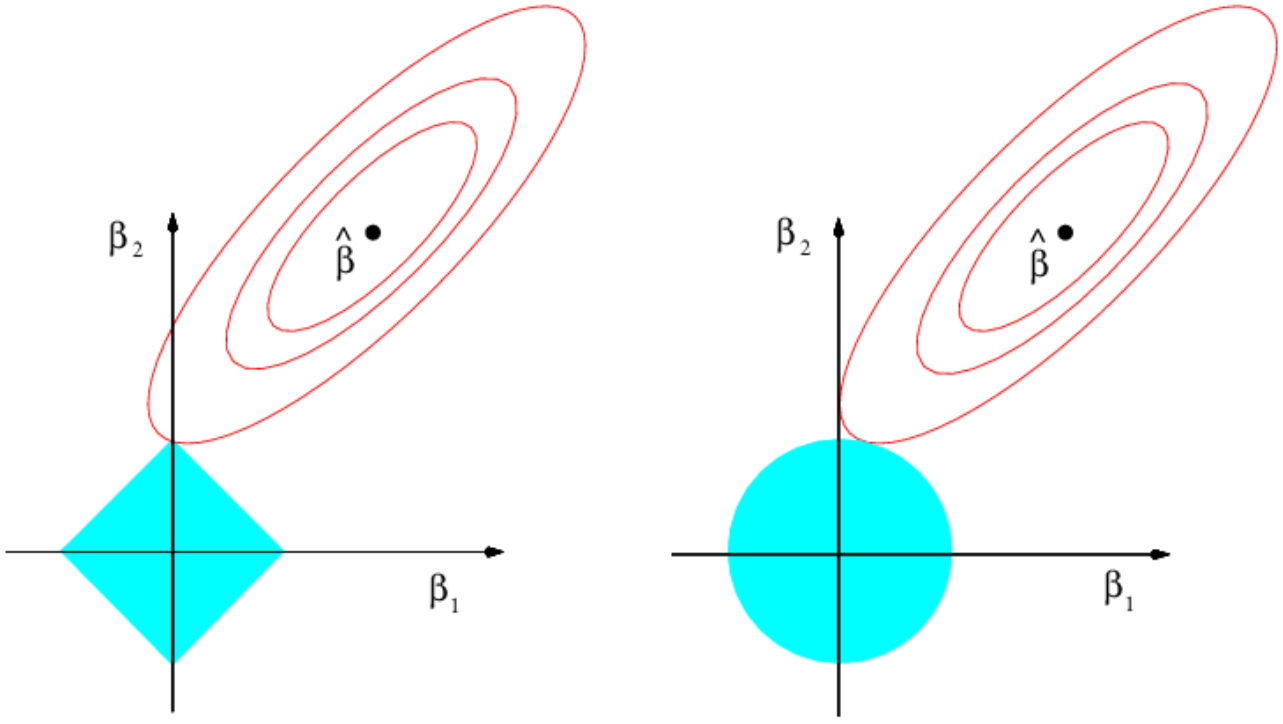
y

$$\hat{\boldsymbol{\beta}}^L = \arg\max_{\boldsymbol{\beta}} L(\beta_0, \beta_1, \ldots, \beta_p)$$

sujeto a

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

Geometría de ridge y lasso



Este gráfico fue tomado del libro Witten, Hastie, y Tibshirani (2013).

Comparación entre ridge y lasso

- Ninguno de los dos métodos supera al otro en todo contexto.
- En general si solo una pequeña proporción de las covariables se relaciona con la respuesta funcionará mejor lasso.
- Si una gran proporción de las covariables se relaciona con la respuesta funcionará mejor ridge.

Ejercicio de aplicación: Datos de bateadores

Considerar los datos `Hitters` del paquete ISLR.
1- Hacer una selección de variables utilizando los métodos tradicionales. Con las variables elegidas, ajustar un modelo lineal y utilizar los estimadores obtenidos para predecir el salario de los bateadores. Estimar el error de testeo usando validación cruzada.
2- Utilizando todas las variables disponibles calcular los estimadores ridge y utilizarlos para predecir el salario de los bateadores. Estimar el error de testeo usando validación cruzada.
3- Utilizando todas las variables disponibles calcular los estimadores lasso y utilizarlos para predecir el salario de los bateadores. Estimar el error de testeo usando validación cruzada. Comparar la selección de variables que realiza el lasso con la del item 1.

Bibliografía

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112). New York: springer.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, pp. 241-249). New York: Springer series in statistics.