Trade-off sesgo-varianza

Para ponerlo en palabras:

- Cuando hablamos de sesgo tenemos en cuenta la cercanía al fenomeno real complejo que obtenemos al ajustar con un modelo lineal relativamente simple.

- Cuando hablamos de varianza de $\widehat{m}$ nos referimos a la variabilidad que tendrán nuestras estimaciones del modelo m cuando se basan en distintos conjuntos de datos.

Trade-off sesgo-varianza

Vayamos al ejemplo de R

Trade-off sesgo-varianza

- El ejemplo nos muestra que las predicciones basadas en el modelo lineal simple y en el cuadrático son estables, pero muy sesgadas.
- A la vez, las basadas en polinomios con grado más alto tienen mucho menor sesgo, pero una gran variabilidad.
- Esto pone de manifiesto que los polinomios de menor grado, al ser más rígidos, no pueden capturar la curvatura, pero tampoco el ruido de los datos. Los de mayor grado, al ser más flexibles captan la forma funcional, pero también el ruido.

Trade-off sesgo-varianza

Trade-off: los modelos más complejos tendrán menor sesgo, pero mayor variabilidad.

Trade-off sesgo-varianza

- El ejemplo nos muestra que las predicciones basadas en el modelo lineal simple y en el cuadrático son estables, pero muy sesgadas.
- A la vez, las basadas en polinomios con grado más alto tienen mucho menor sesgo, pero una gran variabilidad.
- Esto pone de manifiesto que los polinomios de menor grado, al ser más rígidos, no pueden capturar la curvatura, pero tampoco el ruido de los datos. Los de mayor grado, al ser más flexibles captan la forma funcional, pero también el ruido.

Trade-off sesgo-varianza

Trade-off: los modelos más complejos tendrán menor sesgo, pero mayor variabilidad.

¿Qué modelo elegimos?

Error de predicción

Para evaluar la bondad del ajuste de un modelo $m(\mathbf{X})$ podemos considerar el error de predicción.

Si $\widehat{m}(\mathbf{X})$ es el valor predicho que obtenemos a partir de postular y ajustar el modelo m podrímos computar

$$ECM = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{m}(\mathbf{X}_i))^2,$$

pero como ya hemos visto en otros contextos (cuales???) si calculamos el error cuadrático medio ECM sobre las mismas observaciones con las que ajustamos, podemos llegar a sobreestimar.....

Existen distintas estrategias para evitar esto, las más contemporáneas se basan en dividir los datos y usar una parte para estimar y otra para evaluar.

Splitting the data $\mathcal{M}$

- $\mathcal{T}$: Muestra de entrenamiento: es usada para ajustar el modelo
- $\mathcal{V}$: Muestra de validación: es usada para seleccionar el modelo

Cuando hay muchos datos

- $\mathcal{T}$: Muestra de entrenamiento: es usada para ajustar el modelo.(80%)
- $\mathcal{V}$: Muestra de validación: es usada para seleccionar el modelo. (20%)

$$\frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}(y_i-\widehat{m}_{\mathcal{T}}(\mathbf{X}_i))^2$$

Cuando hay muchos datos

- $\mathcal{T}$: Muestra de entrenamiento: es usada para ajustar el modelo.(80%)
- $\mathcal{V}$: Muestra de validación: es usada para seleccionar el modelo. (20%)

$$\frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}(y_i-\widehat{m}_{\mathcal{T}}(\mathbf{X}_i))^2$$

Entre los modelos m que compiten elegimos m de manera de

$$\min\frac{1}{|\mathcal{V}|}\sum_{i\in\mathcal{V}}(y_i-\widehat{m}_{\mathcal{T}}(\mathbf{X}_i))^2$$

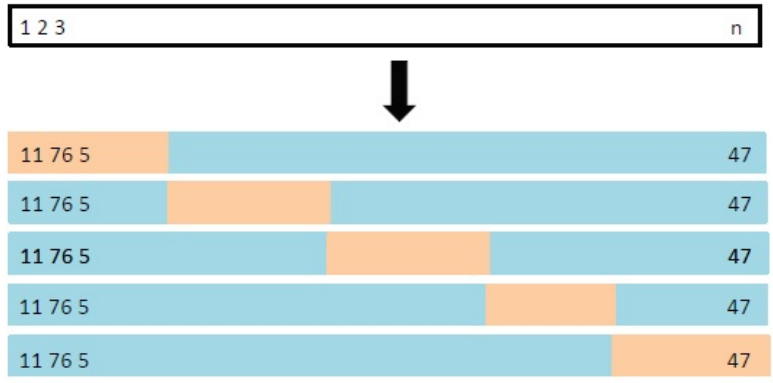
Menos datos

Convalidación Cruzada

Menos datos

Convalidación Cruzada

Cross Validation: K-fold - Representación esquemática (ISLR)



Cross Validation: K folders - Fórmulas

Función objetivo:

$$\mathsf{CV}(\widehat{m}) = \frac{1}{K}\sum_{k=1}^{K}\frac{1}{|\mathcal{T}_k^c|}\sum_{j\in\mathcal{T}_k^c}(y_j - \widehat{m}_{\mathcal{T}_k}(\mathbf{X}_j))^2$$

Elección de modelos

Los métodos de selección de modelos se pueden clasificar en

- **best all subsets:** el mejor de todos los modelos con $k = 1, 2, \ldots, p$ predictores, (sólo posible con p pequeño o muy moderado, orden 2^p modelos, con p=20, 1.048.576 modelos)
- **stepwise**: las variables se incorporan o eliminan paso a paso de acuerdo a cierto criterio.
  - **forward**: Comenzamos sin ningún predictor, probamos con todos los modelos de un predictor, elegimos (por ejemplo) el que minimiza el CV. Luego agregamos 1 covariable, sin quitar las anteriores y elegimos con el criterio de CV y así sucesivamente. La cantidad de modelos es del orden de $p(p+1)/2$.
  - **backward**: Comenzamos con el modelo completo y vamos sacando las variables de una en una.
- **penalización**: LASSO, Elastic Net permiten seleccionar variables automáticamente.

Pérdida de Precisión en la Prediccón

- Cuando el tamaño muestral n no es mucho más grande que p, puede haber sobreajuste y aumento de la variabilidad en el estimador de mínimos cuadrados.

- Si $p > n$ la estimación ya no es única y la varianza aumenta a infinito.

- Una solución es restringir a los estimadores, consiguiendo estimadores que, aunque sesgados, logran reducir la varianza.

- A la vez, cuando hay variables que no están relacionadas con la respuesta, es deseable no incluirlas (coeficiente nulo), de manera de bajar la complejidad del modelo y de facilitar la intepretabilidad.

Regularización o Penalización

Un camino para lograr esto es la regularización o penalización.

- En este enfoque se ajusta un modelo que contiene a todos las covariables.
- Sin embargo, el método de ajuste fuerza a encoger a los estimadores hacia 0 (*shrinkage*).
- El procedimiento se denomina *regularización o penalización* y tiene como propósito y efecto reducir la varianza estimada.
- Según la penalización que se elija, la estimación de algunos coeficientes puede ser exactamente 0, con lo que se logra otro objetivo: seleccionar variables.

Regresión de Ridge

Hoerl y Kennard (1970) propusieron obtener el estimador para el caso en que hay inestabilidad debida a colinealidad de la siguiente manera

$$\arg\min_{\mathbf{b}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[y_i - \mathbf{x}_i^{\mathsf{t}} \mathbf{b} \right]^2 \right\} \text{ sujeto a } \sum_{i=1}^{p} b_j^2 \leq c$$

siendo $\mathbf{x}_i^{\mathsf{t}} \mathbf{b} = b_0 + \sum_{i=1}^{p} x_{ij} b_j$.

Regresión de Ridge

Hoerl y Kennard (1970) propusieron obtener el estimador para el caso en que hay inestabilidad debida a colinealidad de la siguiente manera

$$\arg\min_{\mathbf{b}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[y_i - \mathbf{x}_i^{\mathsf{t}} \mathbf{b} \right]^2 \right\} \text{ sujeto a } \sum_{i=1}^{p} b_j^2 \leq c$$

siendo $\mathbf{x}_i^{\mathsf{t}} \mathbf{b} = b_0 + \sum_{i=1}^{p} x_{ij} b_j$.

Usualmente en la regresión Ridge se utilizan los datos estandarizados: la media y desvíos muestrales iguales a 0 y 1, respectivamente y así los coefidentes del modelo están en escalas semejantes y no de acuerdo a una escala relacionada a las unidades de medición de cada variable.

Típicamente, si el modelo incluye intercept, no se la incluye en el término de regularización. Si centramos los y_i de manera de que $\overline{Y} = 0$, podemos omitir a la intercept de la optimización y luego estimarla por $\overline{Y} - \sum_j \overline{X}_j \widehat{\theta}_j$ usando la solución hallada para los datos centrados.

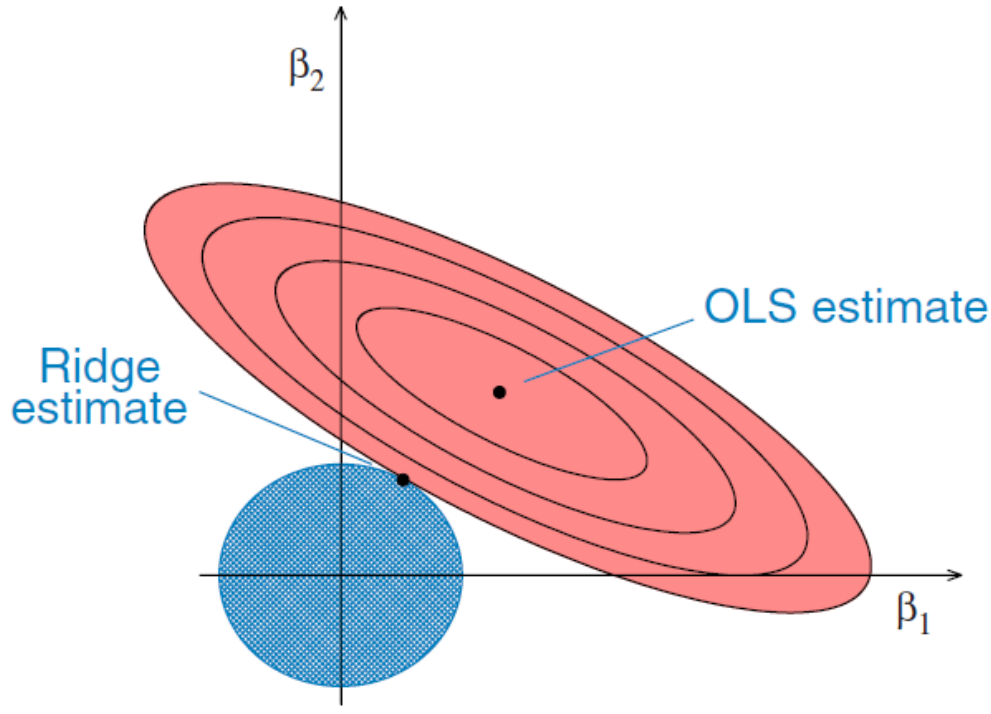


FIGURE 5.6. *The ridge regression estimator, $\widehat{\boldsymbol{\beta}}_{\mathrm{rr}}(k)$, as the solution of a penalized least-squares problem. The ellipses show the contours of the error sum-of-squares function, and the circle shows the boundary of the penalty function, $\beta_1^2+\beta_2^2 \leq c$, where c is the radius of the circle. The ridge estimator is the point at which the innermost elliptical contour touches the circular penalty.*

Regresión de Ridge

Alternativamente puede pensarse como minimizamos

$$\arg\min_{\mathbf{b}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[y_i - \mathbf{x}_i^{\mathbf{t}} \mathbf{b} \right]^2 + \lambda \sum_{j=1}^{p} b_j^2 \right\}$$

para algún $\lambda > 0$.

O bien, podríamos formularlo como

$$\arg\min_{\mathbf{b}} \left\{ \|\mathbf{Y} - \mathbf{X}^{\mathbf{t}} \mathbf{b}\|^2 + \lambda \|\mathbf{b}\|^2 \right\}$$

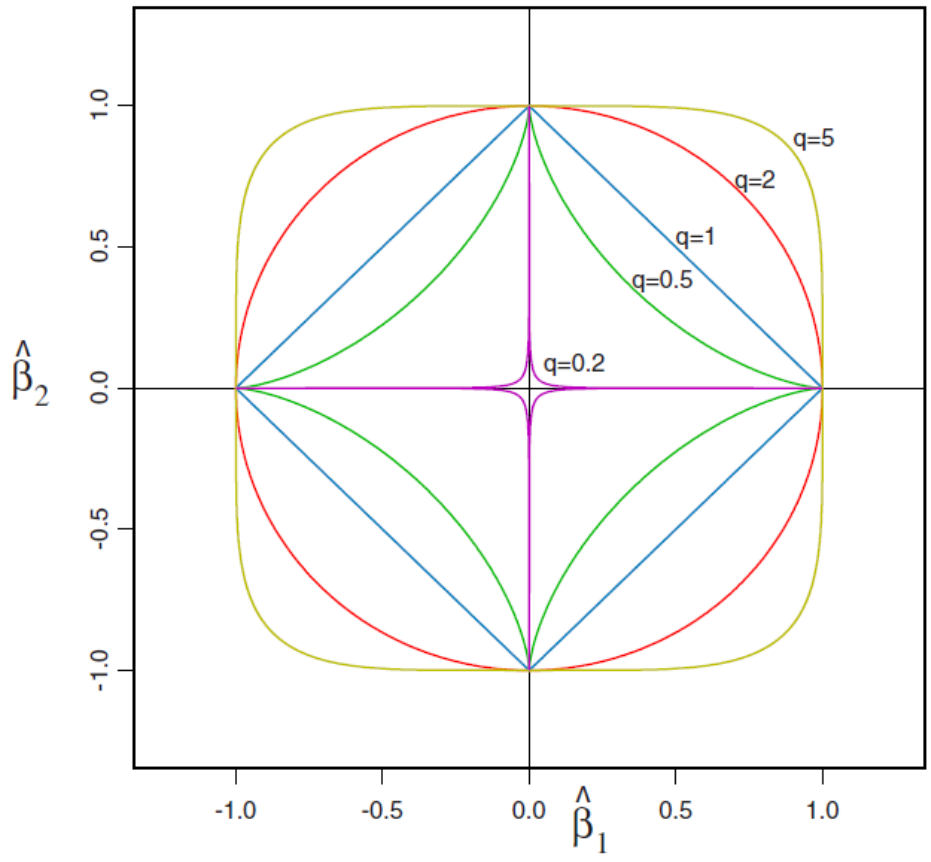


FIGURE 5.10. *Two-dimensional contours of the symmetric penalty function* $p_q(\boldsymbol{\beta}) = |\beta_1|^q + |\beta_2|^q = 1$ *for* $q = 0.2, 0.5, 1, 2, 5$. *The case* $q = 1$ *(blue diamond) yields the lasso and* $q = 2$ *(red circle) yields ridge regression.*

La forma de la región determina el punto donde se alcanza el óptimo.

LASSO:Least Absolute Selection and Shrinkage Operator

El método LASSO es un método en el que se penaliza a la estimación de los coeficientes cuando toman valores grandes.

$$\widehat{\boldsymbol{\beta}}^{\text{LASSO}} = \arg\min_{\mathbf{b}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left[y_i - \mathbf{x}_i^{\mathbf{t}} \mathbf{b} \right]^2 + \lambda \sum_{j=1}^{p} |b_j| \right\}$$

para algún $\lambda > 0$.

Usualmente en la regresión LASSO se utilizan los datos estandarizados: la media y desvíos muestrales iguales a 0 y 1, respectivamente y así los coedificentes del modelo están en escalas semejantes y no de acuerdo a una escala relacionada a las unidades de medición de cada variable.

Típicamente, si el modelo incluye intercept, no se la incluye en el término de regularización. Asumimos que $\overline{Y} = 0$.

LASSO

Función Objetivo

$$\mathcal{S}^{\text{LASSO}}(\mathbf{b}) = \underbrace{\frac{1}{2n}\sum_{i=1}^{n}\left[y_i - \mathbf{x}_i^{\mathsf{t}}\mathbf{b}\right]^2}_{\text{penaliza mal ajuste}} + \underbrace{\lambda\sum_{j=1}^{p}|b_j|}_{\text{penaliza coeficientes no nulos}}$$

- El primer término se encarga de controlar la bondad del ajuste.
- El segundo término se encarga de controlar la parsimonia del modelo.

- ¿Qué pasa si $\lambda = 0$?
- ¿Qué pasa si $\lambda = \infty$?
- ¿Cómo elegimos λ?

LASSO

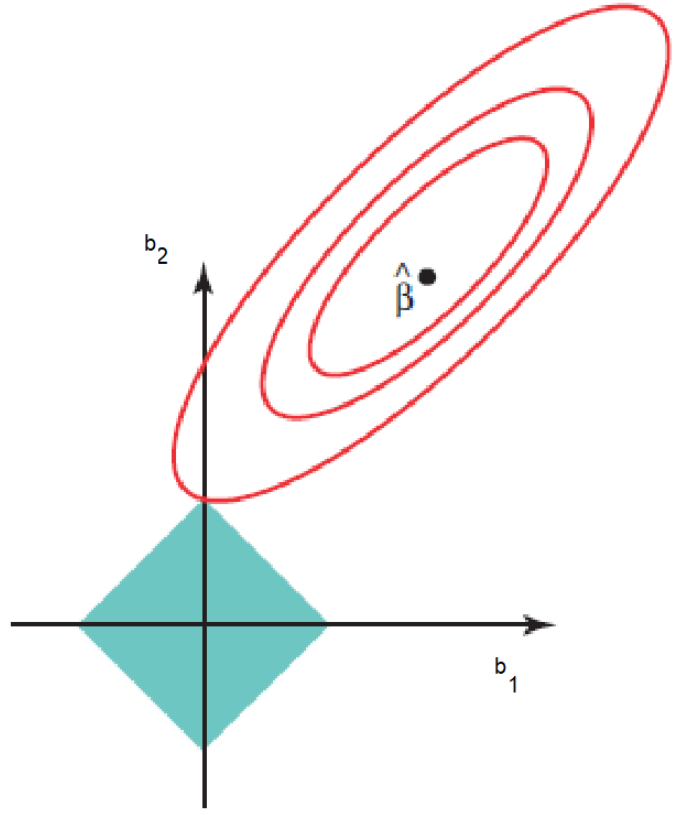
- Sorpresa: Bajo condiciones de regularidad LASSO elige automáticamente las variables que entran al modelo ($\widehat{\beta}_k \neq 0$) y cuales no ($\widehat{\beta}_k = 0$).
- ¿Por qué?

LASSO como optimización restringida

$$\begin{aligned} \widehat{\boldsymbol{\beta}}^{\text{LASSO}} &= \arg\min_{\mathbf{b}} \frac{1}{2n} \sum_{i=1}^{n} \left[y_i - \mathbf{x}_i^{\mathrm{t}} \mathbf{b} \right]^2 \\ &\quad \text{sujeto a } \sum_{j=1}^{p} |b_j| \leq s \end{aligned}$$

Los coeficientes anulados corresponden a soluciones de esquina, como muestra el siguiente gráfico tomado de ISL, James, Witten, Hastie & Tibshirani.

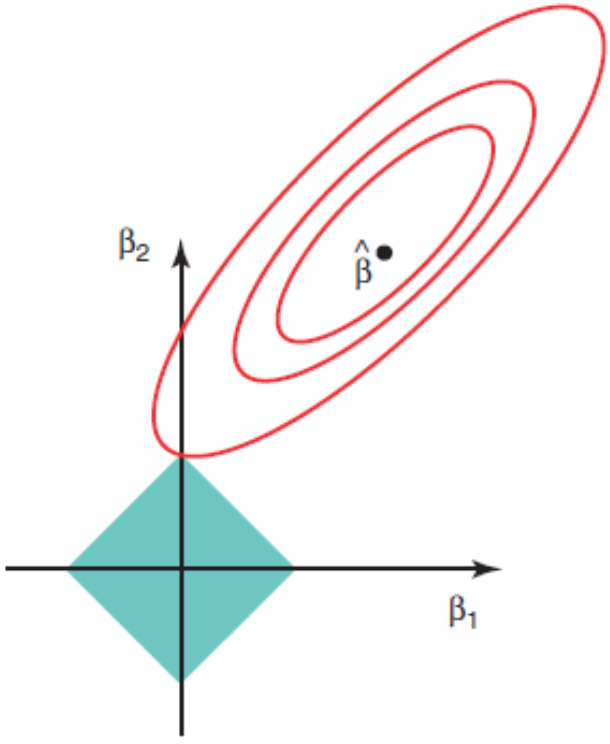
LASSO en $\mathbb{R}^2$



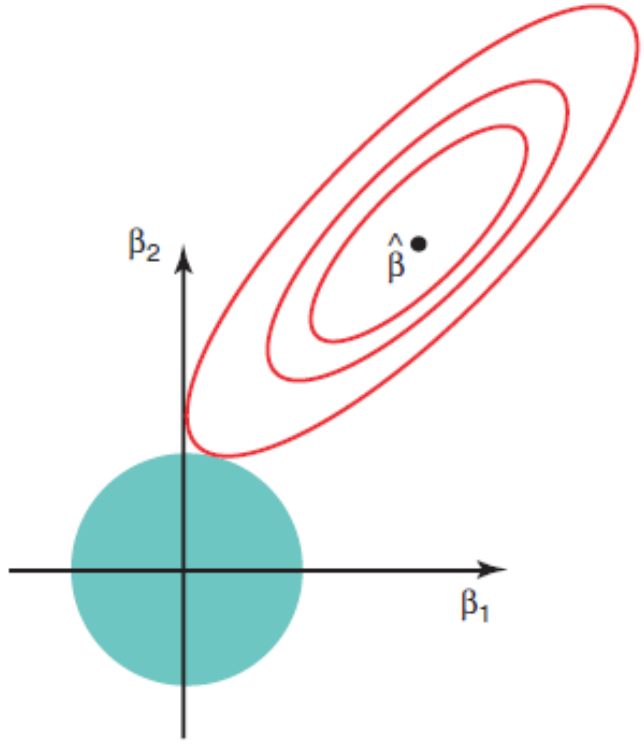
Las elipses representan las curvas de nivel de la función objetivo de mínimos cuadrados (LS), mientras que el $\widehat{\boldsymbol{\theta}}$ corresponde al estimador de LS. El rombo sólido celeste representa la restricción $|b_1| + |b_2| \leq c$. El método va a buscar al punto que cumple la restricción más cerca de la solución de LS, pero como el rombo tiene esquinas, si la solución ocurre en una de ellas, alguno de los coeficientes será 0!!!

LASSO y Ridge en $\mathbb{R}^2$

LASSO



Ridge



Elección de λ en LASSO

- La selección de λ se hace por convalidación cruzada.
- Se fija una grilla para λ.
- Dividimos aleatoriamente la muestra en $K-$folds (típicamente 5 o 10). Sea $k = 1, \ldots, K$:
  - Para el $k-$ésimo fold y cada λ, calculamos el estimador LASSO basado en la muestra de entrenamiento (o sea los otros $K-1$ folds.).
  - Predecimos las respuestas y_i del grupo de validación o sea del $k-$ésimo fold y calculamos el error de predicción: $ECM_k(\lambda)$.
- Una vez calculados los K $ECM_k(\lambda)$, para cada λ calculamos la pérdida de convalidación cruzada como:

$$CV(\lambda) = \frac{1}{K} \sum_{k=1}^{K} ECM_k(\lambda)$$

- Para cada λ además se calcula el desvío standard de CV: $SE(\lambda) = sd(ECM_1(\lambda), \ldots, ECM_K(\lambda))$

Elección de λ

- Elegiremos λ^{CV} como el valor de la grilla que minimiza a $CV(\lambda)$:

$$\lambda^{CV} = \arg\min CV(\lambda)$$

- Una vez computado λ^{CV} ajustamos el modelo usando todas las observaciones y el hiperparámetro λ^{CV}.

Elección de λ

- Elegiremos λ^{CV} como el valor de la grilla que minimiza a $CV(\lambda)$:

$$\lambda^{CV} = \arg\min CV(\lambda)$$

- Una vez computado λ^{CV} ajustamos el modelo usando todas las observaciones y el hiperparámetro λ^{CV}.

Elección de λ: regla de 1 desvío standard

Con la idea de obtener un modelo más parsimonioso, se suele usar la regla de 1 desvío standard para elegir el hiperparámetro λ.

La regla establece seleccionar el λ de acuerdo al siguiente criterio:

$$\lambda^{CV1sd} = \max\left\{\lambda : CV(\lambda) < CV\left(\lambda^{CV}\right) + SE\left(\lambda^{CV}\right)\right\}$$

Ejemplo

A data frame with 322 observations of major league players on the following 20 variables.

AtBat: Number of times at bat in 1986

Hits: Number of hits in 1986

HmRun: Number of home runs in 1986

Runs: Number of runs in 1986

RBI: Number of runs batted in in 1986

Walks: Number of walks in 1986

Years: Number of years in the major leagues

CAtBat: Number of times at bat during his career

CHits: Number of hits during his career

CHmRun: Number of home runs during his career

CRuns: Number of runs during his career

CRBI: Number of runs batted in during his career

CWalks: Number of walks during his career

League: A factor with levels A and N indicating player's league at the end of 1986

Division: A factor with levels E and W indicating player's division at the end of 1986

PutOuts: Number of put outs in 1986

Assists: Number of assists in 1986

Errors: Number of errors in 1986

Salary: 1987 annual salary on opening day in thousands of dollars

NewLeague: A factor with levels A and N indicating player's league at the beginning of 1987

References James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013) An Introduction to Statistical Learning with applications in R, www.StatLearning.com, Springer-Verlag, New York

Vamos a analizar los datos en R