Tomado de ISLR Hastie et al Cap 6.

Métodos de contracción de los coeficientes

Regularización

Se tiene un modelo de la forma $\underline{Y} = \underline{X}\underline{\beta} + \underline{\varepsilon}$. Se define el error residual como $\underline{r} = \underline{Y} - \underline{\hat{Y}} = \underline{Y} - \underline{X}\underline{\hat{\beta}}$ y se busca

$$\underline{\hat{\beta}} = \arg\min_{\underline{\beta}} \frac{\|\underline{r}\|^2}{2} = \arg\min_{\underline{\beta}} \frac{(\underline{Y} - \underline{X}\underline{\hat{\beta}})^T (\underline{Y} - \underline{X}\underline{\beta})}{2}$$

La función $l(\underline{\beta}) = \frac{(\underline{Y} - \underline{X}\underline{\hat{\beta}})^T (\underline{Y} - \underline{X}\underline{\beta})}{2}$ es la función de costo. $\underline{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^T \in \mathbb{R}^{p+1}$

La regularización consiste en modificar $l(\underline{\beta})$ sumándole un $p(\underline{\beta})$, donde $p(\cdot)$ es un factor que depende de la magnitud de $\underline{\beta}$ (o sea de su norma). La nueva función de costo es $l_p(\underline{\beta}) = l(\underline{\beta}) + \lambda p(\underline{\beta})$.

$\lambda \geq 0$ es la constante de regularización.

Ejemplos

i) $p(\underline{\beta}) = \|\underline{\beta}\|_1 = \sum_{i=1}^{p} |\beta_i|$ **Lasso**

ii) $p(\underline{\beta}) = \|\underline{\beta}\|_2^2 = \sum_{i=1}^{n} \beta_i^2$ **Ridge**

↙ Aquí hay un abuso de notación, porque el β_0 no se regulariza

Cuando $\lambda = 0$ no hay regularización.

Cuando $\lambda > 0$ hay regularización

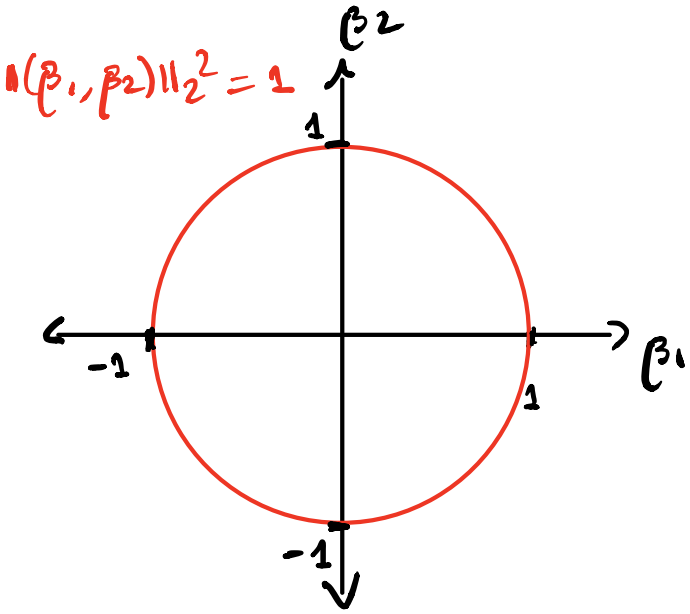
$p(\underline{\beta})$ penaliza las variables con coeficientes grandes. Si la variable no reduce el error entonces se castiga que tenga un coeficiente grande

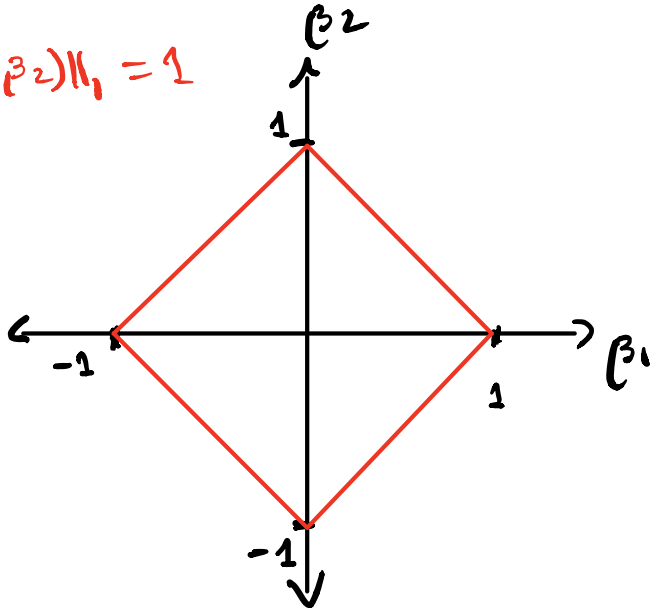
$$l_p(\underline{\beta}) = l(\underline{\beta}) + \lambda p(\underline{\beta})$$

↓ Ajuste

↓ Complejidad

Supongamos que $\underline{\beta} = (\beta_0, \beta_1, \beta_2)$

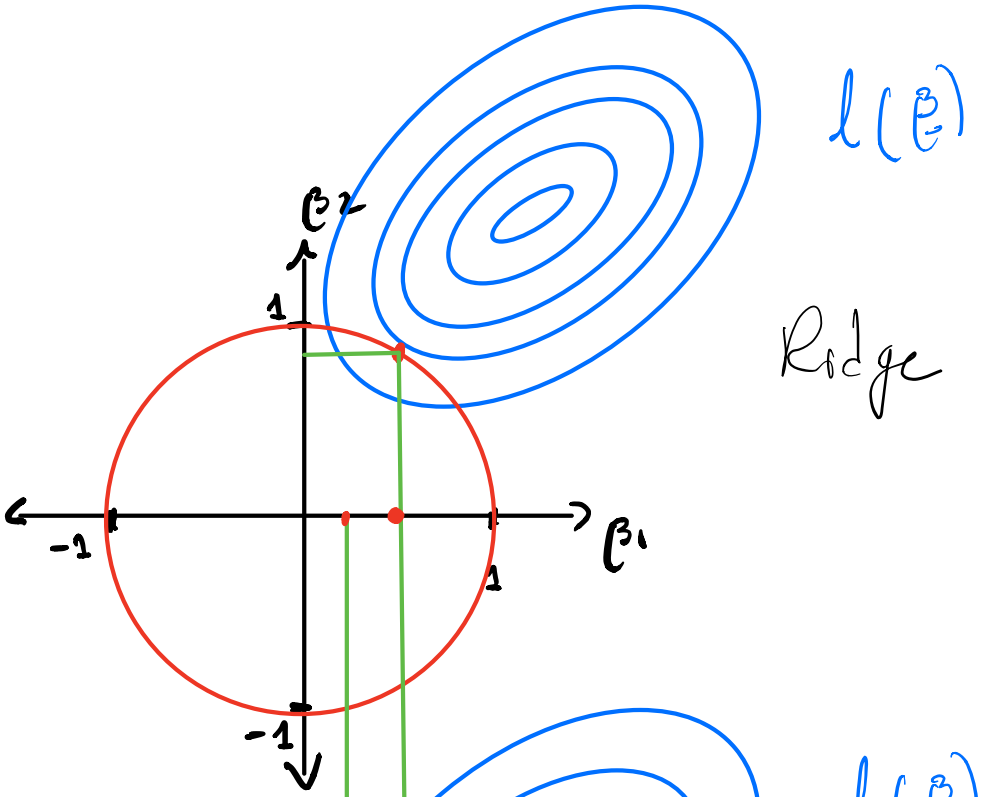




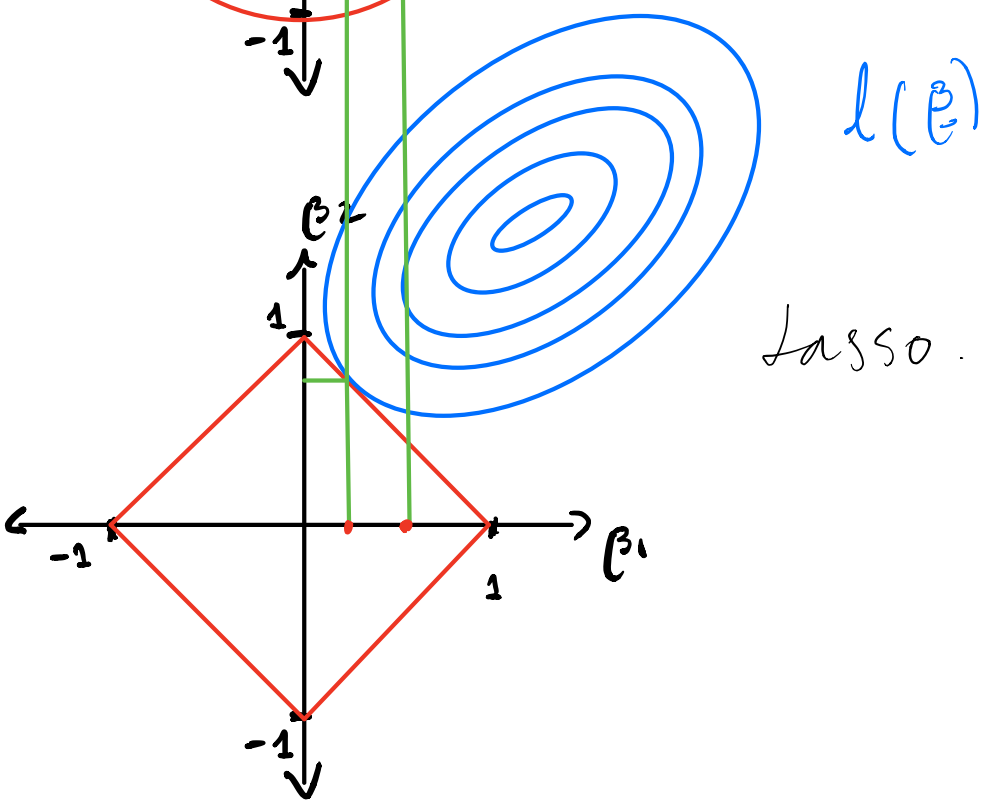
Con predictores centrados se tiene $\hat{\beta}_0 = \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$

Nota: tanto $\|\cdot\|_1$ como $\|\cdot\|_2$ son sensibles a la escala. Esta aproximación solo debe usarse con variables estandarizadas. (centradas y escaladas)

¿Por qué lasso es mejor para selección de variables? Porque encoje más rápido los coeficientes.



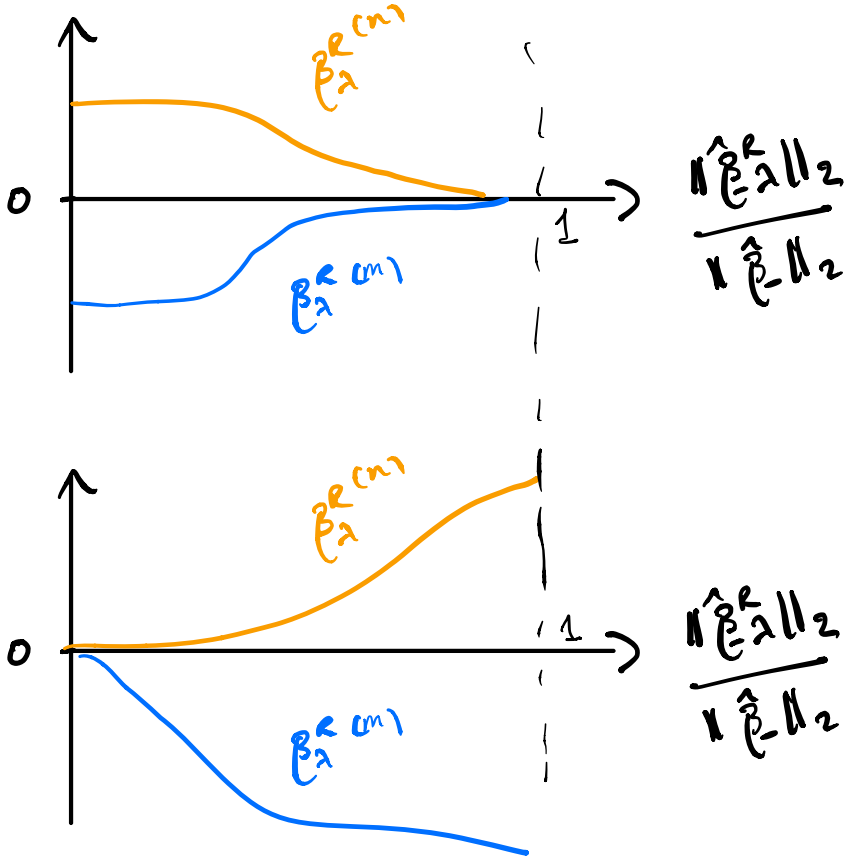
Ridge



Lasso.

Denotemos por $\hat{\beta}^R_\lambda$ y $\hat{\beta}^L_\lambda$ los estimadores Ridge y Lasso para un valor particular de λ y $\hat{\beta}$ el estimador para $\lambda = 0$ (mínimos cuadrados o máxima verosimilitud)

$\hat{\beta}^R_\lambda(i)$ es el estimador para el i-ésimo predictor. (que debe estar estandarizado)



$\hat{\beta}^R_1 = \hat{\beta}$

¿Cómo escoger λ? Por validación cruzada.