Modèles Linéaires Appliqués

Arthur Charpentier

Automne 2Q20

OLS #16 (multicolinéarité, Ridge)

Régression Linéaire sans $\mathcal{H}_1$

- $\mathbf{X}$ est mal conditionnée, i.e. certaines valeurs propres de $\mathbf{X}^\top\mathbf{X}$ sont proches de 0 $\Rightarrow \det(\mathbf{X}^\top\mathbf{X}) \simeq 0$, phénomène de colinéarité.
- (ou/et) $p > n$ le nombre de covariables est nettement supérieur au nombre d'observations.

L'hypothèse initiale $\mathcal{H}_1$ est donc largement remise en cause!

Régression Ridge

- Idée: perturber la matrice $\mathbf{X}$ pour éloigner ses valeurs propres de 0 et ainsi stabiliser l'inversion de $\mathbf{X}^\top\mathbf{X}$.
- Notation: $0 \leq \mu_1 \leq \ldots \mu_p$ valeurs propres ordonnées de $\mathbf{X}^\top\mathbf{X}$ et soit $\mathbf{P}$ la matrice orthogonale telle que $\mathbf{X}^\top\mathbf{X} = \mathbf{P}\mathbf{D}\mathbf{P}^\top$ avec $\mathbf{D} = \mathrm{diag}(\mu_1, \ldots, \mu_p)$.
- Soit $\lambda \geq 0$, $\mathbf{X}^\top\mathbf{X} + \lambda\mathbb{I}_p$ a les mêmes vecteurs propres que $\mathbf{X}^\top\mathbf{X}$ et pour valeurs propres $\mu_j + \lambda$ pour $j = 1, \ldots, p$.
- Hoerl and Kennard (1970): remplacer $\mathbf{X}^\top\mathbf{X}$ par $\mathbf{X}^\top\mathbf{X} + \lambda\mathbb{I}_p$ dans la définition des MCO:

$$\hat{\boldsymbol{\beta}}_{\mathrm{ridge}}(\lambda) = (\mathbf{X}^\top\mathbf{X} + \lambda\mathbb{I}_p)^{-1}\mathbf{X}^\top\mathbf{Y}$$

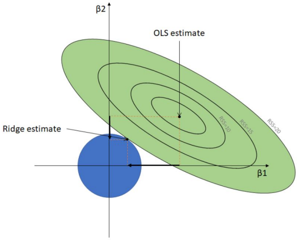
- Si $\lambda = 0$, alors $\hat{\boldsymbol{\beta}}_{\mathrm{ridge}} = \hat{\boldsymbol{\beta}}^{\mathrm{mco}}$; si $\lambda \to \infty$, alors $\hat{\boldsymbol{\beta}}_{\mathrm{ridge}} \to \mathbf{0}$.
- Fixer λ est important. Quelles sont les conséquences sur le biais et variance des estimateurs?

Régression Ridge

- Soit $\lambda \geq 0$, on peut montrer que $\hat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda)$ minimise

$$\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|^2$$

$$= \sum_{i=1}^{n}(Y_i - \mathbf{x}_i\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^{p}\beta_j^2$$



https://rstatisticsblog.com/

- Soit $\tilde{\boldsymbol{\beta}}$ l'estimateur minimisant

$$\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 \text{ sous la contrainte que } \|\boldsymbol{\beta}\|^2 \leq \delta.$$

$\forall\lambda > 0$, $\exists\delta$ tel que les deux solutions coïncident.

- $\hat{\boldsymbol{\beta}}_{\text{ridge}} = \left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbb{I}_p\right)^{-1}\mathbf{X}^\top\mathbf{X}\hat{\boldsymbol{\beta}}$.
- $\mathbb{E}\left(\hat{\boldsymbol{\beta}}_{\text{ridge}}\right) = \boldsymbol{\beta} - \lambda\left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbb{I}_p\right)^{-1}\boldsymbol{\beta}$.
- $\text{Var}(\hat{\boldsymbol{\beta}}_{\text{ridge}}) = \sigma^2\left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbb{I}_p\right)^{-1}\mathbf{X}^\top\mathbf{X}\left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbb{I}_p\right)^{-1}$.
- et l'erreur quadratique vaut

$$\text{EQM}(\hat{\boldsymbol{\beta}}_{\text{ridge}})\left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbb{I}_p\right)^{-1}\left(\sigma^2(\mathbf{X}^\top\mathbf{X}) + \lambda^2\boldsymbol{\beta}\boldsymbol{\beta}^\top\right)\left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbb{I}_p\right)^{-1}$$

alors que, pour rappel, $\text{EQM}(\hat{\boldsymbol{\beta}}) = \text{Var}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$.

Si μ_j valeur propre de $\mathbf{X}^\top\mathbf{X}$

- Sous $\mathcal{H}_1$ (nécessairement $r = p$) $\text{tr}(\text{EQM}(\hat{\boldsymbol{\beta}})) = \displaystyle\sum_{j=1}^{p} \frac{\sigma^2}{\mu_j}$;
- (potentiellement $r \leq p$):

$$\text{tr}(\text{EQM}(\hat{\boldsymbol{\beta}}_{\text{ridge}})) = \sum_{j=1}^{r} \frac{\sigma^2\mu_j + \lambda^2\left(\mathbf{P}^\top\boldsymbol{\beta}\right)_j^2}{(\mu_j+\lambda)^2};$$

- $\text{tr}(\text{EQM}(\hat{\boldsymbol{\beta}}_{\text{ridge}})) \leq \text{tr}(\text{EQM}(\hat{\boldsymbol{\beta}})) \Leftrightarrow \lambda \leq \dfrac{2\sigma^2}{\boldsymbol{\beta}^\top\boldsymbol{\beta}}$.

Rétrécissement / *shrinkage*

- Soit $\hat{\theta}$ un estimateur de θ sans biais et de variance σ^2, donc $\text{EQM}(\hat{\theta}) = \sigma^2$.
- Soit $\lambda > 0$ et $\tilde{\lambda} = \dfrac{\hat{\theta}}{1+\lambda}$ alors

$$\mathbb{E}(\tilde{\theta}) = \frac{\theta}{1+\lambda},\ \ \text{Var}(\tilde{\theta}) = \frac{\sigma^2}{(1+\lambda)^2} \text{ et } \text{EQM}(\tilde{\theta}) = \frac{\lambda^2\theta^2+\sigma^2}{(1+\lambda)^2}.$$

  On peut ainsi trouver $\lambda^\star$ telle que $\text{EQM}(\tilde{\theta})$ soit minimale
- $\tilde{\theta}$ est donc biaisé mais de variance plus faible que $\hat{\theta}$.

On peut trouver $\lambda^\star$ telle que $\lambda \mapsto \text{tr}(\text{EQM}(\hat{\boldsymbol{\beta}}_{\text{ridge}}(\lambda)))$ soit minimale

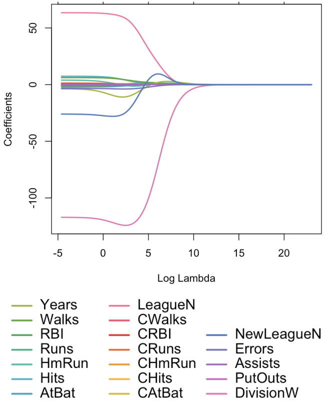
Régression Ridge

On peut utiliser

```
> library(MASS)
> ?lm.ridge
```

ou

```
> library(ISLR)
> library(glmnet)
> Hitters = na.omit(Hitters)
> x = model.matrix(Salary~.,
    Hitters)[,-1]
> y = Hitters$Salary
> ridge_mod = glmnet(x, y, alpha =
    0)
> plot(ridge_mod, var="lambda")
```

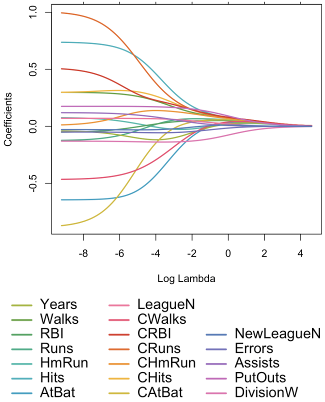


Régression Ridge

au lieu de $\|\beta\|_2$ (norme Euclidienne), on peut être tenté d'utiliser la norme de Mahalanobis (ou plutôt la norme euclidienne standardisée)
i.e. on centre et on réduit les variables explicatives

$$\mathbf{x}_j \mapsto \frac{\mathbf{x}_j - \overline{x}_j}{s_{x_j}}$$

```
> ys = (y-mean(y))/sd(y)
> xs = x
> for(i in 1:ncol(x)) xs[,i] = (x[,
    i]-mean(x[,i]))/sd(x[,i])
> ridge_mod_s = glmnet(xs, ys,
    alpha = 0)
> plot(ridge_mod_s, var="lambda")
```



Régression LASSO

- Ici, on ne contraint pas la norme euclidienne $\|\boldsymbol{\beta}\| = \|\boldsymbol{\beta}\|_2$ (i.e. la norme ℓ^2) mais la norme ℓ^1 des coefficients.
- La méthode Lasso consiste à minimiser

$$\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 \text{ sous la contrainte } \|\boldsymbol{\beta}\|_1 \leq \delta.$$

- Il n'existe pas de solution exacte (plusieurs algorithmes ont été proposés - LARS, coordinate descent algorithm).
- On peut montrer que le problème est équivalent à minimiser le problème

$$\hat{\boldsymbol{\beta}}_{\text{lasso}}(\lambda) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1$$

($\forall\lambda > 0$, $\exists\delta > 0$ tel que les solutions de ces deux problèmes coïncident)...