ACT6100	**Analyse de données**
H2019	**Série 5**

Régularisation

1. On considère un base de données de taille $n = 8$ pour laquelle on a

$$\boldsymbol{X} = \begin{bmatrix} -2 \\ -1 \\ -1 \\ -1 \\ 0 \\ 1 \\ 2 \\ 2 \end{bmatrix} \qquad \boldsymbol{Y} = \begin{bmatrix} 35 \\ 40 \\ 36 \\ 38 \\ 40 \\ 43 \\ 45 \\ 43 \end{bmatrix}$$

et un modèle de régression linéaire classique $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, $i = 1, \ldots, 8$ et $\epsilon \sim$ Normale$(0, \sigma^2)$.

(a) En utilisant la régression Ridge avec $\lambda = 0$, calculer l'équation du modèle, c'est-à-dire les valeurs des paramètres β_0 et β_1.

(b) En utilisant la régression Ridge avec $\lambda = 4$, calculer l'erreur quadratique moyenne.

2. On définit les matrices

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \quad \boldsymbol{X} = \begin{bmatrix} 1 & X_{11} & \ldots & X_{1p} \\ \vdots & & \ddots & \vdots \\ 1 & X_{n1} & \ldots & X_{np} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_p \end{bmatrix}$$

$$\boldsymbol{Y}^* = \begin{bmatrix} Y_1 - \bar{Y} \\ \vdots \\ Y_n - \bar{Y} \end{bmatrix} \qquad \boldsymbol{X}^* = \begin{bmatrix} X_{11} - \bar{X}_1 & \ldots & X_{1p} - \bar{X}_p \\ \vdots & \ddots & \vdots \\ X_{n1} - \bar{X}_1 & \ldots & X_{np} - \bar{X}_p \end{bmatrix}.$$

Les estimateurs du modèle de régression Ridge sont la solution de

$$\widehat{\boldsymbol{\beta}}^{\boldsymbol{R}} = \arg\min_{\beta_0, \ldots, \beta_p} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2 + \lambda \sum_{j=1} \beta_j^2, \tag{1}$$

pour une valeur de $\lambda > 0$.

(a) Démontrer que $\widehat{\beta}_0^R = \bar{Y} - \sum_{j=1}^{p} \beta_j \bar{X}_j$.

(b) En intégrant le résultat trouvé à la sous-question précédente, vérifier que l'Équation (1) peut s'écrire comme étant

$$\arg\min_{\beta_1, \ldots, \beta_p} \sum_{i=1}^{n} \left(Y_i^* - \sum_{j=1}^{p} \beta_j X_{ij}^* \right)^2 + \lambda \sum_{j=1} \beta_j^2, \tag{2}$$

pour une valeur de $\lambda > 0$.

(c) Vérifier que l'Équation (2) peut s'écrire sous la forme matricielle

$$\arg\min_{\boldsymbol{\beta}^*} \left(\boldsymbol{Y}^* - \boldsymbol{X}^*\boldsymbol{\beta}^*\right)^T \left(\boldsymbol{Y}^* - \boldsymbol{X}^*\boldsymbol{\beta}^*\right) + \lambda \boldsymbol{\beta}^{*T}\boldsymbol{\beta}^*,$$

où $\boldsymbol{\beta}^* = \begin{bmatrix}\beta_1 & \ldots & \beta_p\end{bmatrix}^T$ et que la solution est donnée par

$$\widehat{\boldsymbol{\beta}}^* = \left(\boldsymbol{X}^{*T}\boldsymbol{X}^* + \lambda \boldsymbol{I}_p\right)^{-1} \boldsymbol{X}^{*T}\mathbf{Y}^*.$$

3. La base de données *FREMPL3* disponible dans la libraire *CASdatasets* contient le montant total des réclamations pour plusieurs contrats ainsi que des caractéristiques.
   (a) Construire une base de données *data* ne conservant que les dossiers ayant une réclamation strictement supérieure à 0$.
   (b) À partir de la base de données *data*, utiliser un modèle de régression linéaire généralisé de type Gamma avec fonction de lien logarithmique pour modéliser la variable **ClaimAmount** à partir des variables explicatives **LicAge**, **VehAge**, **Gender**, **MariStat**, **SocioCateg** et **DrivAge** (conserver toutes les variables et toutes les modalités). Utiliser correctement l'exposition (variable **Exposure**) dans le modèle. Répondre aux questions suivantes.
      i. Pour un assuré (genre masculin) célibataire dont le véhicule est dans la catégorie d'âge 3, dont la variable **LicAge** vaut 400, dans la catégorie socio-économique CSP50 et âgé de 45 ans, quel cout moyen des sinistres sera prédit par le modèle si l'exposition de cet assuré est unitaire ?
      ii. Est-ce que l'affirmation « l'âge du conducteur fait augmenter le cout moyen d'un sinistre » est vraie ou fausse ? Expliquer.
   (c) À partir de la base de données *data*, utiliser un modèle de régression polynomiale avec $K \leq 10$ de type Gamma avec fonction de lien logarithmique pour modéliser la variable **ClaimAmount** à partir de la variable explicative **DrivAge** et de l'exposition. L'équation du modèle est alors

$$E[Y] = (\text{Exposure}) \exp\left(\beta_0 + \sum_{j=1}^{K} \beta_j \left(\textbf{DrivAge}\right)^j\right), \quad K \leq 10.$$

   Déterminer la valeur de K en utilisant **(i)** la valeur du critère AIC **(ii)** la 2-validation croisée qui minimise l'erreur quadratique moyenne de validation.
   (d) À partir de la base de données *data*, utiliser un modèle de régression Ridge de type Poisson avec fonction de lien logarithmique pour modéliser la variable **ClaimAmount** à partir des variables explicatives **DrivAge** et **LicAge** et de l'exposition. Quelle valeur de λ est sélectionnée par 10-validation croisée ?
   (e) À partir de la base de données *data*, utiliser un modèle de régression Ridge de type Poisson avec fonction de lien logarithmique pour modéliser la variable **ClaimAmount** à partir de la variable explicative catégorielle **VehUsage** et de l'exposition. Quelles sont les valeurs des paramètres obtenues par 20-validation croisée ?
4. On sait que l'estimateur du modèle de régression linéaire $\boldsymbol{\beta} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y}$ est sans biais et que l'estimateur dans le modèle de Ridge est donné par $\boldsymbol{\beta}^R = \left(\mathbf{X}^T\mathbf{X} + \lambda \mathbf{I}_p\right)^{-1}\mathbf{X}^T\mathbf{Y}$. Vérifier que, si $\lambda \neq 0$, l'estimateur du modèle de Ridge est biaisé.
5. On considère [1] la base de données *mtcars* (une base de données non actuarielles mais classique en analyse de données) disponible en R pour laquelle on souhaite prédire la valeur de la variable **mpg** à l'aide des autres variables.

1. Exercice et code librement inspirés de *Data Analysis with R : Load, wrangle, and analyze your data using the world's most powerful statistical programming language* par Tony Fischetti.

(a) On considère un modèle *elastic net*, c'est-à-dire un modèle qui combine la régression Ridge et la régression Lasso en optimisant la fonction

$$\widehat{\boldsymbol{\beta}}^{\boldsymbol{R}} = \arg\,\min_{\beta_0,\ldots,\beta_p} \sum_{i=1}^{n} \left(Y_i - \beta_0 - \sum_{j=1}^{p} \beta_j X_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 + \alpha \sum_{j=1}^{p} |\beta_j| \,.$$

En utilisant la validation croisée *leave-one-out* (équivalente à une 32-validation croisée ici), calculer l'erreur quadratique moyenne pour des valeurs de $\alpha = 0.00, 0.01, 0.02, \ldots, 1.00$ et, à chaque fois, pour la valeur optimale de λ. Présenter les résultats sur un graphique. Quelle est la meilleure option ?

(b) Télécharger l'environnement de travail *QuickStartExample.RData* disponible sur le site *Moodle* du cours et refaire le même exercice (les matrice $\mathbf{x}$ et $\mathbf{y}$ contiennent déjà les variables explicatives et la variable réponse).