TP – Sur la Régression Ridge

On propose d'étudier la régression d'arête dans le cas simple de deux prédicteurs et à partir de données simulées.

Le modèle de simulation : on choisit le modèle linéaire suivant

$$\mathbf{Y = \beta_1 X^{(1)} + \beta_2 X^{(2)} + \varepsilon}$$

avec les variables centrées (pas d'intercept β_0 pour simplifier) et avec des prédicteurs de même variance simulés de la manière suivante :

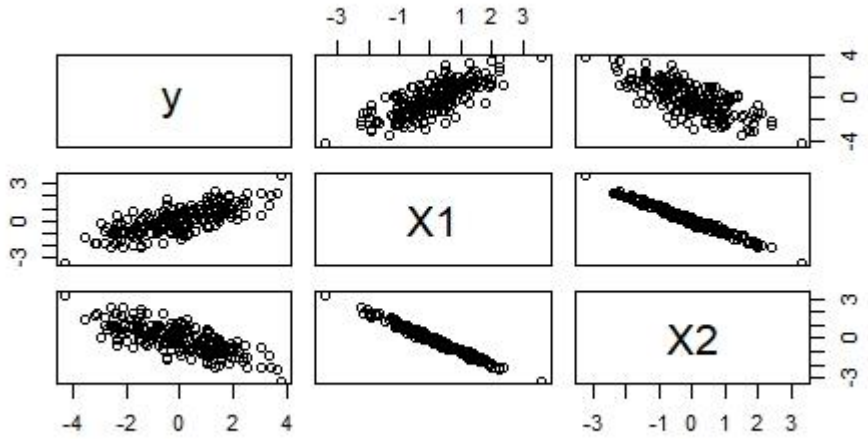
$\mathbf{X^{(1)} = \varepsilon^{(1)}}$; $\mathbf{X^{(2)} = \rho \times \varepsilon^{(1)} + \sqrt{1-\rho^2}\ \varepsilon^{(2)}}$ où $\varepsilon^{(1)}$, $\varepsilon^{(2)}$ sont **indépendantes** de même loi normale $\mathbf{N(0, \sigma_X^2)}$ et **indépendantes** du résidu $\boldsymbol{\varepsilon \sim N(0, \sigma^2)}$. Dans ce modèle gaussien, ρ est le **coefficient de corrélation linéaire entre $X^{(1)}$ et $X^{(2)}$** et mesure donc le **degré de colinéarité** entre les deux prédicteurs.

Pour débuter l'étude, on pourra fixer les valeurs suivantes des paramètres de simulation :

$$\mathbf{\rho = -0.99 \ ; \ \sigma_X = \sigma = 1 \ ; \ \beta_1 = 0.5 \text{ et } \beta_2 = -0.5}$$

Q1. Simuler un jeu de données de taille **n = 200** en créant une matrice ou un data frame de taille n×3 avec en première colonne une variable nommée « y » correspondant aux valeurs simulées de la réponse Y puis dans les deux colonnes suivantes les valeurs deux prédicteurs nommés « X1 » et « X2 ». Visualiser vos données simulées à l'aide la fonction **pairs()** pour vérifier la cohérence avec le modèle.

On donne ci-dessous un exemple de telle simulation :



L'estimation sous R du modèle de régression linéaire $\mathbf{y = \beta_1 x^{(1)} + \beta_2 x^{(2)} + \varepsilon}$ pour les données de la figure précédente donne les résultats suivants (on utilise la formule ***y ~ X1 + X2 – 1*** pour estimer le modèle **sans intercept**) :

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
X1	**0.5991**	**0.5044**	**1.188**	**0.236**
X2	**-0.2790**	**0.4966**	**-0.562**	**0.575**

Residual standard error: 0.9848 on 198 degrees of freedom
Multiple R-squared: 0.4332, Adjusted R-squared: 0.4274
F-statistic: 75.65 on 2 and 198 DF, p-value: < 2.2e-16

Comment expliquez-vous le fait qu'aucun des deux coefficients β_1 et β_2 ne soit significativement non nul en contradiction avec le modèle simulé et aussi le test global de Fisher ?

Q2. On rappelle que l'estimateur des moindres carrés ordinaires (MCO) du vecteur β est donné par la formule $\hat{\boldsymbol{\beta}} = \mathbf{(X'X)^{-1} (X'y)}$. Calculer $\hat{\boldsymbol{\beta}}$ à l'aide la fonction **solve()** puis comparer avec le résultat de la fonction **lm()**. Quelles sont les estimations obtenues pour les écarts-types des estimateurs $\hat{\beta}_1$ et $\hat{\beta}_2$ (erreurs standard) ?

Q3. On considère la régression Ridge consistant à minimiser le critère suivant

$$\mathbf{J(\beta) = \| y - X\beta \|_2^2 + \lambda \| \beta \|_2^2} \text{ avec } \boldsymbol{\lambda > 0} \text{ appelé paramètre de régularisation}$$

A partir de l'expression analytique de l'estimateur Ridge $\hat{\boldsymbol{\beta}}\mathbf{(\lambda) = (X'X + \lambda I_2)^{-1} (X'y)}$, tracer sur une même figure les valeurs obtenues pour les deux coefficients $\hat{\beta}_1(\lambda)$ **et** $\hat{\beta}_2(\lambda)$ en faisant varier λ de 0 (estimateur MCO) à une valeur suffisamment grande (à déterminer expérimentalement).

Q4. Simuler maintenant un jeu de données test de taille convenable (à partir du même modèle que celui utilisé pour apprendre, soit avec les mêmes valeurs des paramètres) et tracer l'erreur de prédiction Ridge en fonction de λ où l'erreur est mesurée de manière classique (RMSE ou Root Mean Square Error) selon

$$\mathbf{RMSE} = \sqrt{\mathbf{MSE}} \text{ avec } \mathbf{MSE} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \left(y_i - \hat{y}_i(\lambda) \right)^2 = \mathbf{MSE}(\lambda)$$

et $\hat{y}_i(\lambda)$ la prédiction Ridge $\hat{y}_i(\lambda) = \hat{\beta}_1(\lambda)\, x_i^{(1)} + \hat{\beta}_2(\lambda)\, x_i^{(2)}$ pour la donnée test n° i.

Analyser la courbe obtenue sachant que l'erreur initiale (associée à $\lambda = 0$, soit aucune régularisation) est celle de la régression linéaire multiple ordinaire.

Q5. Obtenir la valeur de λ minimisant le RMSE.

Q6. Pour la valeur de λ retenue, obtenir par simulation le biais ainsi que la variance de l'estimateur Ridge correspondant pour les deux coefficients β_1 et β_2 et comparer avec les MCO. On pourra comparer également le risque des deux estimateurs avec le risque des estimateurs MCO compte tenu de la relation bien connue

$$\textbf{Risque} = \textbf{Variance} + \textbf{Biais}^2$$

Q7. Illustrer le fameux **compromis « biais-variance »** cher à l'apprentissage statistique en traçant sur une même courbe le risque des deux estimateurs en fonction de λ (faire varier λ de 0 à une valeur qui permette de visualiser ce compromis).