SY19
TP 4: Sélection de variables

Données `Residential_building`

Les données `Residential_building` se rapportent au coût et au prix de 372 appartements à Teheran, en fonction de caractéristiques des projets et de variables économiques. Les variables à expliquer sont le coût et le prix.

1. Tracez des histogrammes des variables `V9` (coût) et `V10` (prix). Que constatez-vous ? Appliquer une transformation logarithmique à ces variables et renommez-les.
2. Appliquez la régression linéaire à ces données en prenant le prix comme variable à expliquer. Que constatez-vous ? Quel est le rang de la matrice $\mathbf{X}$ contenant les valeurs des prédicteurs ?
3. Estimez la racine de l'erreur quadratique moyenne (*root mean squared error, RMSE*) par validation croisée, et l'erreur standard de l'estimation. Représentez graphiquement les valeurs prédites en fonction des valeurs observées de la variable à expliquer.
4. Même question pour la régression linéaire avec sélection de variables à l'aide des critères AIC et BIC. La sélection de variables réduit-elle significativement l'erreur de prédiction ?
5. Quelles sont les variables sélectionnées par le critère BIC ?
6. Reprendre l'analyse précédent en prenant cette fois la variable coût comme variable à expliquer.

Données `sonar`

Les données `sonar` ont été obtenues dans le cadre d'étude sur la classification de signaux sonar. La tâche consiste à discriminer entre les signaux sonar réfléchis par un cylindre métallique et ceux réfléchis par une roche à peu près cylindrique. La classe est codée dans la variable `V61`.

Analysez ces données en reprenant la méthodologie de l'exercice précédent, avec cette fois la régression logistique le taux d'erreur comme critère de performance.