ECM 3A 2015/2016

OMIS-Statistiques
Feuille de travaux dirigés numéro 1
Régression linéaire

Exercice 1

La méthode Ridge consiste à chercher $\hat{\beta}$ qui minimise la fonction

$$\beta \mapsto \|Y - X\beta\|^2 + \lambda\|\beta\|^2.$$

Ce minimiseur est appelé l'estimateur Ridge.

1. Soit A une matrice carré. Montrer que pour $\lambda > 0$ suffisament petit, $\lambda Id + A$ est inversible.
2. Calculer l'estimateur Ridge $\hat{\beta}$.
3. Quelle est sa limite lorsque $\lambda \to \infty$?
   On suppose que X est de plein rang.
4. Quelle est la limite de $\hat{\beta}$ lorsque $\lambda \to 0$?
5. Que vaut $\hat{\beta}$ lorsque les colonnes de X sont orthonormales ?
6. Calculer le biais et la variance de cette estimateur.

Exercice 2

On examine l'évolution d'une variable Y en fonction de deux variables X_1 et X_2. On dispose de n observations de ces variables. On note $X = (\mathbf{1}, X_1, X_2)$ où $\mathbf{1}$ est le vecteur constant.

1. Nous obtenons le résultat suivant :

$$X^T X = \begin{pmatrix} 30 & 0 & 0 \\ ? & 10 & 7 \\ ? & ? & 15 \end{pmatrix}$$

   a) Donner les valeurs manquantes.
   b) Que vaut n ?
   c) Calculer le coefficient de corrélation empirique entre X_1 et X_2.

2. La régression linéaire de Y sur X donne

$$Y = -2\mathbf{1} + X_1 + 2X_2 + \hat{\varepsilon}, \text{ avec RSS} = \|\hat{\varepsilon}\|^2 = 12.$$

   a) Déterminer la moyenne arithmétique $\bar{Y}$.
   b) Calculer la somme des carrés expliquée (ESS), la somme des carrés totale (TSS) et le coefficient de corrélation R^2.

Exercice 3

Soit $Y = X\beta + \varepsilon$ un modèle de régression linéaire pour la concentration en ozone (Y) en fonction de la température X_1, du vent X_2 et de la nébulosité X_3. On veut tester la pertinence de l'influence du vent sur la concentration en ozone. Les données permettent d'obtenir

$$\begin{aligned} \hat{\sigma}^2 &= 9, \\ \hat{Y} &= (4, 4, 5, 3.2, 3.4, 7.3), \\ \hat{Y}^* &= (3, 3.2, 5, 3, 3.7, 7). \end{aligned}$$

Peut-on rejeter l'impertinence du vent ?

Exercice 4

Nous considérons le modèle de régression linéaire

$$Y = X\beta + \varepsilon,$$

où $Y \in \mathbb{R}^n$, X est une matrice de taille $n \times p$ composée de p vecteurs orthogonaux, $\beta \in \mathbb{R}^p$ et $\varepsilon \in \mathbb{R}^n$. Considérons U la matrice des q premières colonnes de X et V la matrice des $p-q$ dernières colonnes de X. On obtient par les MCO les estimations suivantes :

$$\begin{aligned}
\hat{Y}_X &= \hat{\beta}_1^X X_1 + ... + \hat{\beta}_p^X X_p \\
\hat{Y}_U &= \hat{\beta}_1^U X_1 + ... + \hat{\beta}_q^U X_q \\
\hat{Y}_V &= \hat{\beta}_{q+1}^V X_{q+1} + ... + \hat{\beta}_p^V X_p
\end{aligned}$$

On note ESS(A) la norme au carré de la projection de Y sur $Im(A)$.

1. Montrer que ESS(X) = ESS(U) + ESS(V).
2. Pour une variable X_i, montrer que l'estimation de β_i est identique quel que soit le modèle utilisé.

Exercice 5

Soit un modèle de régression $Y = X\beta + \varepsilon$. La dernière colonne (la p^{eme}) de X est le vecteur **1**.

1. Soient les variables $\{X_j\}_{j=1,\ldots,p}$ et Y et celles centrées notées $\{X'_j\}_{j=1,\ldots,p}$ et Y'. Montrer que la dernière colonne de X' regroupant les variables $\{X'_p\}$ vaut 0.
2. On note maintenant X' la matrice X centrée et privée de sa dernière colonne de 0 (elle est donc de dimension $n \times p - 1$.
   Soit le modèle suivant : $Y' = X'\beta' + \varepsilon$. En identifiant ce modèle avec le modèle de régression $Y = X\beta + \varepsilon$, trouver la valeur de β_p en fonction de $\beta'_1, ...\beta'_{p-1}$ et des moyennes empiriques de Y et de X. Ce coefficient est appelé coefficient constant (ou *intercept* en anglais).