Polytech'Lille/GIS4

Fiche TP Modélisation Avancée

C. Preda / Q. Grimonprez

L'objectif de ce TP est de mettre en œuvre sous R et SAS les modèles de régression PCR, PLS et Ridge. Pour SAS on utilisera que PCR et Ridge. Le jeux de données considéré est décrit ci-dessous.

Description des jeux de données

donnees_cornell.xls

On trouve dans le livre de Michel Tenenhaus ([10] page 78) l'exemple suivant tiré de Cornell (1990). On cherche à connaître l'influence des proportions de sept composants sur l'indice d'octane moteur de douze différents mélanges d'essences. Les variables sont les suivantes :

- y : indice d'octane moteur
- x_1 : distillation directe (entre 0 et 0.21)
- x_2 : reformat (entre 0 et 0.62)
- x_3 : naphta de craquage thermique (entre 0 et 0.12)
- x_4 : naphta de craquage catalytique (entre 0 et 0.62)
- x_5 : polymère (entre 0 et 0.12)
- x_6 : alkylat (entre 0 et 0.74)
- x_7 : essence naturelle (entre 0 et 0.08)

TABLE 4.1 – *Données Cornell*

x_1	x_2	x_3	x_4	x_5	x_6	x_7	y
0	0,23	0	0	0	0,74	0,03	98,7
0	0,1	0	0	0,12	0,74	0,04	97,8
0	0	0	0,1	0,12	0,74	0, 04	96,6
0	0,49	0	0	0,12	0,37	0,02	92
0	0	0	0,62	0,12	0,18	0,08	86,6
0	0,62	0	0	0	0,37	0,01	91,2
0,17	0,27	0,1	0,38	0	0	0,08	81,9
0,17	0,19	0,1	0,38	0,02	0,06	0,08	83,1
0,17	0,21	0,1	0,38	0	0,06	0,08	82,4
0,17	0,15	0,1	0,38	0,02	0,1	0,08	83,2
0,21	0,36	0,12	0,25	0	0	0,06	81,4
0	0	0	0,55	0	0,37	0,08	88,1

En R :

- Réaliser les statistiques descriptives univariées et bivariées (y versus les autres variables)
- Réaliser le modèle de régression linéaire entre y et toutes les autres variables (fonction R : lm). Analyser la validité et les performances du modèle complet (R2, significativité coefficiants
- Réaliser un sélection des variables pas-à-pas (fonction R : stepAIC).
- On souhaite un modèle avec toutes les variables présentes ! En effet, il est difficile, dans ce contexte, de présenter un modèle partiel (un composante n'aurait pas d'influence sur l'indice d'octane!). Pour cela on va réaliser la régression PCR, PLS et Ridge.

Explorez sous R la fonction « pcr » et la fonction « pls ». Un exemple d'utilisation des ces techniques a été fait en cours et est disponible sur :
http://math.univ-lille1.fr/~preda/GIS4/ModAv/car.r

- Réaliser la régression Ridge à l'aide de la fonction « ridge.lm » du package MASS. Choisissez le paramètre de regularisation lambda qui minimise le PRESS (GCV). Explorer l'objet retourné par ridge.lm.

Comparer les performances de PCR, PLS et Ridge et présenter vos conclusions.

En SAS :

- introduire le jeux de données dans une étape data à l'aide de l'option « datalines» ou « cards » (http://www.ats.ucla.edu/stat/sas/library/SASRead_os.htm)

- réaliser les statistiques univariées et bivariées

- réaliser le modèle de régression complete (PROC REG) et celui obtenu par sélection selon le R2 ajusté.

Voir pour la sélection de modèle :
http://analytics.ncsu.edu/sesug/2005/SA01_05.PDF ou
https://support.sas.com/documentation/cdl/en/statug/63347/HTML/default/viewer.htm#statug_reg_sect030.htm

- Réaliser un modèle avec toutes les variables à l'aide de la régression ridge : Proc REG et option « ridge ». Voir aussi la note vue en cours de Patrick Breheny sur la régression ridge :
http://web.as.uky.edu/statistics/users/pbreheny/764-F11/notes/9-1.pdf

- Réaliser les régressions PLS et PCR grace à la procedure PROC PLS.

Consulter pour la procedure PLS le support.sas.com :

https://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_pls_sect013.htm

Comparer les sorties R et SAS.

Enjoy !