Statistiques : Robot

Nous travaillons sur la base de données `diabetes` de python. La base initiale comporte $n = 442$ patients et $p = 10$ covariables. La variable Y à expliquer est un score correspondant à l'évolution de la maladie. Pour s'amuser, un robot malicieux a contaminé le jeu de données en y ajoutant 200 variables explicatives inappropriées. Ensuite, non-content d'avoir déjà perverti notre jeu de données, il a volontairement mélangé les variables entre elle de façon aléatoire. Bien entendu le robot a ensuite pris soin d'effacer toute trace de son acte crapuleux si bien que nous ne connaissons pas les variables pertinentes. La nouvelle base de données comporte $n = 442$ patients et $p = 210$ covariables, notés X. Saurez-vous déjouer les plans de ce robot farceur et retrouver les variables pertinentes ?

1) Importer la base de données `data_dm3.csv`. La dernière colonne est la variable à expliquer. Les autres colonnes sont les variables explicatives. Préciser le nombre de variables explicatives et le nombre d'observations.
2) Les variables explicatives sont-elles centrées ? Normalisées ? (indications : des variables normalisées ont toutes le même écart-type) Qu'en est-il de la variable à expliquer ? Tracer un scatter plot de la base de données avec 4 covariables prises au hasard et la variable à expliquer (un scatterplot regroupe les graphes de chacune des variables en fonction de chacune des autres). Commenter les graphiques obtenus.
3) Echantillon d'apprentissage et de test. Créer 2 échantillons : un pour apprendre le modèle X_{train}, un pour tester le modèle X_{test}. On mettra 25% de la base dans l'échantillon 'test'. Donner les tailles de chacun des 2 échantillons. On notera que le nouvel échantillon de covariables X_{train} n'est pas normalizé. Dans la suite, on fera donc bien attention à inclure l'intercept dans nos régressions.
4) Donner la matrice de covariance calculée sur X_{train}. Tracer le graphe de la décroissance des valeurs propres de la matrice de covariance (ou de corrélation). Expliquer pourquoi il est légitime de ne garder que les premières variables de l'ACP. On gardera $k = 10$ variables dans la suite.
5) Suivant les observations de la question (4), appliquer la méthode de "PCA before OLS" qui consiste à appliquer OLS avec Y et $X_{\text{train}}V_{(1:k)}$, où $V_{(1:k)}$ contient les vecteurs propres (associés aux k plus grandes valeurs propres) de la matrice de covariance. Faire une régression linéaire (avec intercept), puis tracer les valeurs des coefficients (hors intercept). Sur un autre graphique, faire de même avec la méthode des moindres carrés classique.
6) Donner la valeur moyenne de la variable Y (sur le train set). Uniquement pour cette question, centrer et réduire les variables explicatives après ACP (de petite dimension). Faire une régression avec ces variables et vérifier que l'intercept est bien égal à la moyenne de Y sut le train. Commenter.
7) Pour les 2 méthodes (OLS et PCA before OLS) : Tracer les résidus de la prédiction sur l'échantillon test. Tracer leur densité (on pourra par exemple utiliser un histogramme). Calculer le coefficient de détermination sur l'échantillon test. Calculer le risque de prédiction sur l'échantillon test, i.e.,

$$n_{\text{test}}^{-1} \sum_{i \in \text{test}} (Y_i - X_i^T \hat{\theta}^{\text{train}})^2$$

ou $\hat{\theta}^{\text{train}}$ est le OLS calculé précedement et n_{test} est le nombre de points dans l'échantillon test.

(X1, ... X210)
pour i =1...210
Regression avec Xi avec intercept
=> O0,i , O1,i
=> calculer la p-value associés au test de nullité de O1,i

=> garder la plus petite p-value : pi°
y-O1i°Xi°° <— y & on écrit Xi° de (X1,...X210)
on répète avec les 219 covariance restantes
(voir photo)

8) Nous utilisons maintenant la statistique du test de nullité du coefficient (comme vu en cours).

9) Appliquer OLS sur les variables sélectionnées. Donner le risque de prédiction obtenu l'échantillon test et le comparer à ceux de OLS et PCA before OLS.

10) Afin de préparer la validation croisée, séparer l'échantillon train en 5 parties (appelées "folds") de façon aléatoire. On affichera les numéros d'échantillon sélectionnés dans chaque fold.

11) Appliquer la méthode de la régression ridge. Pour le choix du paramètre de régularisation, on fera une validation croisée sur les "folds" définies lors de la question précédente. A tour de rôle chacune des "folds" servira pour calculer le risque de prédiction alors que les autres seront utilisées pour estimer le modèle. On moyennera ensuite les 5 risques de prédictions. On donnera la courbe du risque de validation croisée en fonction du paramètre de régularisation (on veillera à bien choisir l'espace de définition pour le graphe). Donner le paramètre de régularisation optimal et la valeur du risque sur le test.

12) A l'aide de la fonction `lassoCV` de sklearn, choisir le paramètre de régularisation du LASSO. Donner le risque de prédiction associé.

13) Donner les variables selectionées par le Lasso. Combien y-en a t-il ? Appliquer la méthode OLS aux variables sélectionnées. Cette méthode est appelée Least-square LASSO. Donner le risque de prédiction associé

14) Le robot a avoué ! Les variables qui étaientt initialement présentes dans le jeux de donnée sont $[119, 121, 85, 28, 185, 149, 11, 88, 179, 155]$. Combien le Lasso en a-t-il retrouvé ? Combien Forward en a-t-il retrouvé ?