

---

TP2

---

On utilisera IPython Notebook pour réaliser ce TP. On déposera son compte-rendu sur le site pédagogique avant le 22 novembre, 23h59.

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` disponible depuis le lien [https://bitbucket.org/portierf/shared\\_files/downloads/data\\_dm3.csv](https://bitbucket.org/portierf/shared_files/downloads/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 ? 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  $X_{\text{train}}$ , un pour tester le modèle  $X_{\text{test}}$ . On mettra 20% de la base dans l'échantillon 'test'. Donner les tailles de chacun des 2 échantillons. On notera que le nouvel échantillon de covariables  $X_{\text{train}}$  n'est pas normalisé. Dans la suite, on fera donc bien attention à inclure l'intercept dans nos régression.
4. Donner la matrice de covariance calculée sur  $X_{\text{train}}$ . 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 60 variables dans la suite.
5. Suivant les observations de la question (Q4), appliquer la méthode de "PCA before OLS" qui consiste à appliquer OLS avec  $Y$  et  $X_{\text{train}} V_{(1:60)}$ , où  $V_{(1:60)}$  contient les vecteurs propres (associés aux 60 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 les valeurs des intercepts pour les 2 régressions précédentes. Donner la valeur moyenne de la variable  $Y$  (sur le train set). Les intercepts des 2 questions sont-ils égaux ? Commenter. Uniquement pour cette question, centrer et réduire les variables après ACP (de petite dimension). Faire une régression avec ces variables et vérifier que l'intercept est bien égal à la moyenne de  $Y$  sur le train.
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.

8. Coder la méthode de forward variable sélection. On pourra utiliser la statistique du test de nullité du coefficient (comme vu en cours). Pour l'instant, on ne met pas de critère d'arrêt sur la méthode. C'est à dire que l'on ajoute une variable à chaque étape jusqu'à retrouver la totalité des variables. Afficher l'ordre de sélection des variables.
9. Critère d'arrêt : On décide d'arrêter lorsque la p-valeur dépasse 0.1. Illustrer la méthode en donnant (i) les 3 graphes des statistiques obtenues lors de la sélection de la 1er, 2eme et 3eme variables (en abscisse : l'index des variables, en ordonné : la valeur des stats) , (ii) le graphe des 50 premières p-valeurs (dont chacune est associée à la variable sélectionnée). Sur ce même graphe, on tracera la ligne horizontale d'ordonnée 0.1. Enfin on donnera la liste des variables sélectionnées.
10. 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.
11. Afin de préparer la validation croisée, séparer l'échantillon train en 4 parties (appelées "folds") de façon aléatoire. On affichera les numéros d'échantillon sélectionnés dans chaque fold.
12. 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 4 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.
13. A l'aide de la fonction `lassoCV` de sklearn, choisir le paramètre de régularisation du LASSO. Donner le risque de prédiction associé.
14. Donner les variables sélectionnées par le lasso. Combien y-en a t-il ? Appliquer la méthode OLS aux variables sélectionnées. Cette méthode est appelé Least-square LASSO.
15. Cette dernière question est un question d'ouverture vers une approche non-linéaire. En utilisant les variables sélectionnées par le LASSO (Q13) ou par la méthode forward (Q9), mettre au point une méthode de regression non-linéaire. On apprendra les différents paramètres par validation croisée et l'on donnera la valeur du risque de prédiction calculé sur l'échantillon test. Des performances moindres par rapport à OLS peuvent se produire. Commenter.