TELECOM ParisTech **MDI720**
2017/2018 **Joseph Salmon**

DEVOIR MAISON : Introduction à python et modèle linéaire

Pour ce TP de test de la plate-forme "Classgrade" vous devez déposer un **unique** fichier **anonymisé** (votre nom ne doit apparaître nulle part y compris dans le nom du fichier lui-même) sous format `ipynb` sur le site http://peergrade.enst.fr/.

Vous devez charger votre fichier, avant le dimanche 15/10/2017 23h59. Entre le lundi 16/10/2017 et le dimanche 22/10/2017, 23h59, vous devrez noter trois copies qui vous seront assignées anonymement, en tenant compte du barème suivant pour chaque question :

- 0 (manquant/ non compris/ non fait/ insuffisant)
- 1 (passable/partiellement satisfaisant)
- 2 (bien)

Ensuite, il faudra également remplir de la même manière les points de notation suivants :

- aspect global de présentation : qualité de rédaction, d'orthographe, d'aspect de présentation, graphes, titres, etc. (Question 21).
- aspect global du code : indentation, Style PEP8, lisibilité du code, commentaires adaptés (Question 22)
- Point particulier : absence de bug sur votre machine (Question 23)

Des commentaires pourront être ajoutés question par question si vous en sentez le besoin ou l'utilité pour aider la personne notée à s'améliorer, et de manière obligatoire si vous ne mettez pas 2/2 à une question. Enfin, veillez à rester polis et courtois dans vos retours.

Rappel : aucun travail par mail accepté !

EXERCICE 1. (Expérience de Galton)

Le terme *régression* a été introduit par Sir Francis Galton (cousin de C. Darwin) alors qu'il étudiait la taille des individus au sein d'une descendance. Il tentait de comprendre pourquoi les grands individus d'une population semblaient avoir des enfants d'une taille plus petite, plus proche de la taille moyenne de la population ; d'où l'introduction du terme "régression". Dans la suite on va s'intéresser aux données récoltées par Galton.

1) Récupérer les données du fichier http://josephsalmon.eu/enseignement/TELECOM/MDI720/datasets/Galton.txt (voir aussi leur description ici[1] : http://josephsalmon.eu/enseignement/TELECOM/MDI720/datasets/Galton_description.txt)et charger les avec `Pandas`. On utilisera `read_csv` pour cela, et en arrondira les tailles sans chiffre après la virgule.
2) Combien de données manquantes y-t-il dans cette base de données ? Enlever si besoin les lignes ayant des données manquantes.
3) Afficher sur un même graphe un estimateur de la densité de la population des pères en bleu, et de celles des mères en orange.

1. http://www.randomservices.org/random/data/Galton.html

4) Afficher la taille du père en fonction de la taille de la mère pour les n observations figurant dans les données. Ajouter la droite de prédiction obtenue par la méthode des moindres carrés (avec constante et sans centrage/normalisation).

5) Afficher un histogramme du nombre d'enfants par famille.

6) Créer une colonne supplémentaire appelée `'MidParents'` qui contient la taille du « parent moyen », et valant `('Father'+ 1.08 * 'Mother')/2`.

Pour la i^{e} observation, on note x_i la taille du parent moyen et y_i la taille de l'enfant. On se base sur le modèle linéaire suivant : $y_i = \theta_0 + \theta_1 x_i + \varepsilon_i$ et on suppose que les variables ε_i sont centrées, indépendantes et de même variance σ^2 inconnue.

7) Estimer θ_0, θ_1, par $\hat{\theta}_0$, $\hat{\theta}_1$ en utilisant la fonction `LinearRegression` de `sklearn`, puis vérifier numériquement[2] les formules vues en cours pour le cas unidimensionnel

$$\hat{\theta}_0 = \overline{y}_n - \hat{\theta}_1 \overline{x}_n, \qquad \hat{\theta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x}_n)(y_i - \overline{y}_n)}{\sum_{i=1}^n (x_i - \overline{x}_n)^2}.$$

On fera attention aux normalisations utilisées pour la variance qui peuvent changer selon les packages.

8) Calculer et visualiser les valeurs prédites $\hat{y}_i = \hat{\theta}_0 + \hat{\theta}_1 x_i$ et les y_i sur un même graphique. On affichera de couleurs différentes les données correspondant aux garçons et celles correspondant aux filles.

9) Visualiser un estimateur de la densité des résidus $r_i = y_i - \hat{y}_i$. L'hypothèse de normalité est-elle crédible selon vous ? Calculer ensuite α_g (resp. α_f) les proportions de garçons (resp. de filles) dans la population. On ajoutera ensuite sur le graphique précédent, les fonctions $\alpha_g p_g$ et $\alpha_f p_f$, avec p_g (resp. p_f) les densités des résidus pour les garçons (resp. pour les filles).

10) Régresser cette fois les x_i sur les y_i (et non plus les y_i sur les x_i). On veut comparer numériquement les coefficients $\hat{\alpha}_0$ et $\hat{\alpha}_1$ ainsi obtenus par rapport aux $\hat{\theta}_0$ et $\hat{\theta}_1$ du modèle original. Vérifier numériquement que :

$$\begin{cases} \hat{\alpha}_0 = & \bar{x}_n + \dfrac{\bar{y}_n}{\bar{x}_n} \dfrac{\mathrm{var}_n(\mathbf{x})}{\mathrm{var}_n(\mathbf{y})} (\hat{\theta}_0 - \bar{y}_n), \\ \hat{\alpha}_1 = & \dfrac{\mathrm{var}_n(\mathbf{x})}{\mathrm{var}_n(\mathbf{y})} \hat{\theta}_1. \end{cases} \tag{1}$$

Exercice 2. (Analyse du jeu de données `auto-mpg`)

On travaille dans cette partie sur le fichier `auto-mpg.data`. On cherche à régresser linéairement la consommation des voitures sur leurs caractéristiques : nombre de cylindres, cylindrées (*engine displacement* en anglais), puissance, poids, accélération, année, pays d'origine et le nom de la voiture. Le vecteur contenant la consommation des voitures (plus précisément la distance parcourue, en miles, pour un gallon, ou mpg) est noté $\mathbf{y}$; les colonnes de X sont les régresseurs quantitatifs, donc pour le moment on laisse de côté les variables `origin` et `car name`.

11) Importer avec `Pandas` la base de données disponible ici https://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-mpg.data-original. On ajoutera le nom des colonnes en consultant l'adresse : https://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-mpg.names avec l'attribut `'name'` de `import_csv`. On pourra regarder l'intérêt de l'option `sep=r"\s\+"` si besoin. Y a-t-il un marqueur utilisé pour les données manquantes dans le fichier utilisé ? Si besoin, enlever les lignes possédant des valeurs manquantes dans la base de données.

2. On pourra utiliser par exemple `np.isclose`

12) Calculer l'estimateur des moindres carrés $\hat{\boldsymbol{\theta}}$ (avec ordonnée à l'origine) et sa prédiction $\hat{\mathbf{y}}$ sur une sous partie de la base obtenue en gardant les 9 premières lignes. Que constatez-vous pour les variables `cylinders` et `model year` ?

13) Calculer $\hat{\boldsymbol{\theta}}$ et $\hat{\mathbf{y}}$ cette fois sur l'intégralité des données, après les avoir centrées et réduites. Quelles sont les deux variables qui expliquent le plus la consommation d'un véhicule ?

14) Calculer $\|\mathbf{r}\|^2$ (le carré de la norme du vecteur des résidus), puis $\|\mathbf{r}\|^2/(n-p)$. Vérifier numériquement que :

$$\|\mathbf{y} - \bar{y}_n \mathbf{1}_n\|^2 = \|\mathbf{r}\|^2 + \|\hat{\mathbf{y}} - \bar{y}_n \mathbf{1}_n\|^2.$$

15) Supposons que l'on vous fournisse les caractéristiques suivantes d'un nouveau véhicule :

cylinders	displacement	horsepower	weight	acceleration	year
6	225	100	3233	15.4	76

Prédire sa consommation [3].

16) Utiliser la transformation `PolynomialFeatures` de `sklearn` sur les données brutes, pour ajuster un modèle d'ordre deux (avec les termes d'interactions : `interaction_only=False`). On normalisera et recentrera après avoir créé les nouvelles variables explicatives. Quelle est alors la variable la plus explicative de la consommation ?

17) On revient ici au modèle sans interactions. Proposer une manière de gérer la variable `origin`, par exemple avec `pd.get_dummies`. On ajustera un modèle linéaire sans constante dans ce cas. Déterminer laquelle des trois origines est la plus efficace en terme de consommation [4].

18) Procéder comme pour la question précédente, mais cette fois pour mesure l'influence de la marque de la voiture. On ne considère ici que les variables `'cylinders'`, `'displacement'`,`'horsepower'`, `'weight'`, `'acceleration'`, `'model year'` en plus de la marque. On pourra utiliser `str.split`, `str.replace` et `get_dummies`.

19) Reprendre la matrice X obtenue (sans variables catégorielles) question 13. Obtenir numériquement la SVD (partielle) de $X = USV^\top$ (par exemple en considérant l'option `full_matrices=False`) ; vérifier numériquement que $H = UU^\top$ est une projection orthogonale [5].

20) La diagonale de la matrice H, forme le vecteur des "leviers", qu'on ajoutera comme nouvelle variable. Trier la base de données en fonction de cette variable, et expliquer en quoi les voitures ayant les trois valeurs de "levier" maximales semblent atypiques.

RAPPEL :

21) aspect global de présentation : qualité de rédaction, d'orthographe, d'aspect de présentation, graphes, titres, etc.

22) aspect global du code : indentation, Style PEP8, lisibilité du code, commentaires adaptés.

23) Point particulier : absence de bug sur votre machine.

3. A titre d'information, la consommation effectivement mesurée sur cet exemple était de 22 mpg.
4. Pour info, 1 = usa ; 2 = europe ; 3 = japan
5. on admettra si besoin que c'est la matrice chapeau vue en cours