TELECOM ParisTech **MDI720**
2016/2017 **Joseph Salmon**

Devoir maison n° 1 : Introduction à python et modèle linéaire

---

Pour ce travail vous devez déposer un **unique** fichier **anonymisé** (votre nom ne doit apparaître nulle part y compris dans son nom lui-même) sous format `ipynb` sur le site http://peergrade.enst.fr/. Vous devez charger votre fichier, avant le dimanche 9/10/2016 23h59. La correction sera disponible sur EOLE le lundi 10 et donc les personnes qui n'auront pas déposé leur travail avant la limite obtiendront **zéro**.

Entre le lundi 10 et le vendredi 14 octobre, 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 25).
- aspect global du code : indentation, Style PEP8, lisibilité du code, commentaires adaptés (Question 26)
- Point particulier : absence de bug sur votre machine (Question 27)

Des commentaires adaptés pourront être ajoutés question par question si vous en sentez le besoin ou l'utilité pour aider la personne notée à s'améliorer. Enfin, veillez à rester polis et courtois dans vos retours.

Les personnes qui n'auront pas rentré leurs notes avant la limite obtiendront également **zéro**.

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.

- Régression linéaire univariée -

1) Récupérer les données du fichier http://www.math.uah.edu/stat/data/Galton.csv et les charger avec `Pandas`. On utilisera `read_csv` pour cela et on transformera les tailles en cm [1], en arrondissant 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.

1. pour cela on pourra consulter la description des données proposées en http://www.math.uah.edu/stat/data/Galton.html

3) Afficher un estimateur de la densité de la population des pères en bleu, et de celles des mères en violet.

4) Afficher sur un même graphique 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 `'MeanParents'` qui contient la taille du "parent moyen", c'est-à-dire `('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 deux couleurs différentes les garçons et les 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 ? On ajoutera ensuite un estimateur par genre de la densité des résidus (en mettant un facteur proportionnel au nombre de personnes de chaque genre).

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

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

- Régression linéaire multiple -

On travaille ici avec la même base de données, mais cette fois on considère un modèle de régression avec les deux variables explicatives `'Father'` et `'Mother'`.

11) Calculer $\hat{\boldsymbol{\theta}}$, $\hat{\mathbf{y}}$ pour ce modèle, respectivement l'estimateur des moindres carrés et le vecteur de prédiction.

12) Afficher les points et les prédictions sur un même graphique 3D.

13) Calculer le carré de la norme du vecteur des résidus $\|\mathbf{r}\|^2$, avec $r_i = y_i - \hat{y}_i$.

14) Visualiser un estimateur de la densité des résidus. L'hypothèse de normalité est-elle crédible selon vous ? On ajoutera ensuite un estimateur par genre de la densité des résidus (en mettant un facteur proportionnelle au nombre de personnes de chaque genre).

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

15) Comparer l'influence des deux variables. Laquelle semble la plus explicative ? Tester avant et après centrage et réduction des données.

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

On travaille maintenant 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`.

16) 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. Quelle est le marqueur utilisé pour les données manquantes dans le fichier utilisé ? Enlever les lignes possédant des valeurs manquantes dans la base de données si besoin.

17) Calculer l'estimateur des moindres carrés $\hat{\boldsymbol{\theta}}$ 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` ?

18) 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 ?

19) Calculer $\|\mathbf{r}\|^2$ (le carré de la norme du vecteur des résidus), puis $\|\mathbf{r}\|^2/(n-p)$. Vérifier numériquement que :
$$\|\mathbf{y} - \bar{y}_n \mathbf{1}_n\|^2 = \|\mathbf{r}\|^2 + \|\hat{\mathbf{y}} - \bar{y}_n \mathbf{1}_n\|^2.$$

20) 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].

21) 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.

22) 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].

23) Procéder de même cette fois en fonction de la marque de la voiture. On pourra utiliser `str.split` et `str.replace` pour créer une nouvelle variable `'brand'`.

24) Reprendre la matrice X obtenue sans variables catégorielles. Obtenez 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]. La diagonale cette matrice H, forme le vecteur des "leviers", qu'on ajoutera comme nouvelle variable. Trier la base de données en fonction de cette nouvelle variable, et expliquer en quoi les voitures ayant les trois valeurs de "levier" maximales sont atypiques.

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