Nous travaillons sous Python (SPYDER) durant cet exercice

Régression linéaire multiple

Inspirez-vous des tutoriels suivants :

- Manipulation des données : http://tutoriels-data-mining.blogspot.fr/2017/02/python-manipulations-des-donnees-avec.html [**TUTO 1**]
- Réalisation de la régression : http://tutoriels-data-mining.blogspot.fr/2015/09/python-econometrie-avec-statsmodels.html [**TUTO 2**]
- Quantiles et probabilités des lois d'usage courant : http://tutoriels-data-mining.blogspot.fr/2017/04/probabilites-et-quantiles-sous-excel-r.html [**TUTO 3**]
- ANACONDA est téléchargeable ici : https://www.anaconda.com/distribution/ (nous utilisons la version Python 3.x).
- Installation et démarrage de l'EDI (éditeur de développement intégré) SPYDER de la distribution ANACONDA (http://tutoriels-data-mining.blogspot.com/2015/08/python-la-distribution-anaconda.html ; à partir de la page 9).

Objectif de l'étude : On souhaite expliquer la mortalité dans les métropoles américaines à partir des caractéristiques des villes (météo, pollution, etc.) et de la population (revenu, niveau d'éducation, etc.).

Voici la description de la base :

Description: Properties of some Standard Metropolitan Statistical Areas (a standard Census Bureau designation of the region around a city) in the United States, collected from a variety of sources. The data include information on the social and economic conditions in these areas, on their climate, and one indice of air pollution potentials.

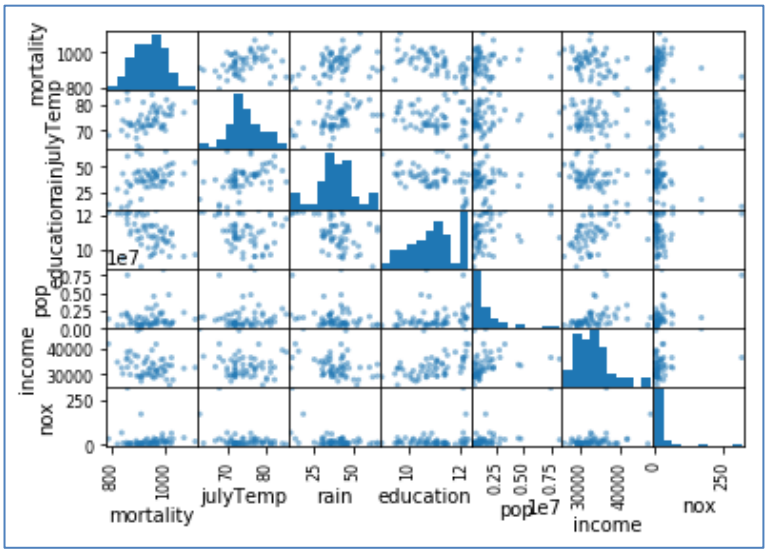
city: City name
Mortality: Age adjusted mortality
JulyTemp: Mean July temperature (degrees Farenheit)
Rain: Annual rainfall (inches)
Education: Median education
PopDensity: Population density
Pop: Population
income: Median income
NOx: Nitrous Oxide

« City » correspond au label des observations ; « **Mortality** » est la variable endogène ; les autres variables sont les exogènes potentielles.

A faire :

Importation et vérification des données

1. Ouvrez le fichier « **mortality.xlsx** » dans Excel pour appréhender sa structure. Dans quelle feuille Excel sont situées les données ? (la seconde feuille nommée ''data'').
2. Lancer l'environnement de développement intégré SPYDER. Créez un nouveau projet FICHIER / NOUVEAU FICHIER, enregistrez le sous le nom « exo_seance_5.py ».
3. Modifiez le répertoire par défaut (os.chdir) [**TUTO 2, page 3**]
4. Importez la librairie pandas et affichez sa version [**TUTO 1, page 2**]. Notez la version de pandas que vous utilisez, cette information sera importante lors de l'importation des données.
5. En utilisant la librairie « pandas », importez le fichier « **mortality.xlsx** » dans un objet que vous nommerez « **df** » (on peut lire directement un fichier Excel avec read_excel() : https://pandas.pydata.org/pandas-docs/stable/generated/pandas.read_excel.html ; attention aux différentes options, notamment header et index_col ; pour indiquer le nom de la feuille à lire, vous utiliserez sheetname [version 0.20 et précédentes] ou sheet_name [version 0.21 et suivantes])
6. Pour vérifier l'importation, affichez :
   a) Les premières lignes du DataFrame (head) [**TUTO 1, page 2**]
   b) Le nombre d'observations et de variables (shape) (58, 7)
   c) Le type des variables (info) (toutes doivent être numériques)
   d) Les statistiques descriptives (describe)
   e) La liste des labels (noms des villes) (index)
7. Réalisez le graphique croisant les variables prises deux à deux (scatter_matrix) [**TUTO 1, page 19**] [Remarque : à partir de Pandas 0.20, on peut faire directement `pandas.plotting.scattermatrix(…)`]. Quels commentaires ce graphique vous inspire-t-il ?



Régression et premières inspections des résultats

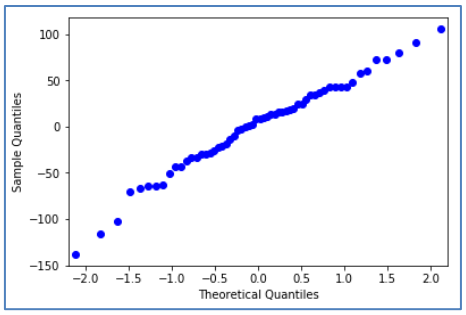
8. Lancez la régression avec l'ensemble des variables explicatives [**TUTO 2, pages 4 et 5**] (ols, fit). Affichez le détail des résultats [**TUTO 2, page 5**] (summary).

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              mortality   R-squared:                       0.371
Model:                            OLS   Adj. R-squared:                  0.297
Method:                 Least Squares   F-statistic:                     5.017
Date:                Fri, 11 Mar 2016   Prob (F-statistic):           0.000416
Time:                        14:33:10   Log-Likelihood:                -308.57
No. Observations:                  58   AIC:                             631.1
Df Residuals:                      51   BIC:                             645.6
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept   1115.5562    179.681      6.209      0.000       754.831  1476.282
julyTemp       1.4078      1.806      0.780      0.439        -2.218     5.033
rain           1.3946      0.791      1.763      0.084        -0.194     2.983
education    -30.3874     10.466     -2.903      0.005       -51.399    -9.376
pop         7.175e-06   5.74e-06      1.250      0.217     -4.34e-06  1.87e-05
income         -0.0004      0.002     -0.211      0.834        -0.004     0.003
nox             0.1081      0.203      0.534      0.596        -0.298     0.515
==============================================================================
Omnibus:                        2.416   Durbin-Watson:                   2.032
Prob(Omnibus):                  0.299   Jarque-Bera (JB):                1.755
Skew:                          -0.415   Prob(JB):                        0.416
Kurtosis:                       3.191   Cond. No.                     5.49e+07
==============================================================================
```

9. Quelles sont les valeurs du R^2 et R^2 ajusté ? Le modèle est-il globalement significatif à 10 % ? Quelles sont les variables pertinentes à 10 % (c.-à-d. dont les coefficients sont significatifs à 10%) ?

10. À partir des résultats fournis par **summary()**, pouvez-vous dire que la distribution des erreurs est compatible avec la loi normale au risque 10 % ?

11. Pour confirmer cette analyse, tracez le graphique quantile-quantile plot pour évaluer la compatibilité des résidus avec une distribution gaussienne [**TUTO 2, page 9**].



12. Construisez le graphique des résidus avec l'endogène. Utilisez la librairie matplotlib (http://www.python-simple.com/python-matplotlib/scatterplot.php). Que constatez-vous ?

13. Refaites le même graphique pour chaque exogène (en abscisse). Remarque : pour accéder aux valeurs de la colonne j d'un DataFrame Pandas, il faudrait écrire `df.iloc[:,j]`

14. On souhaite tester la nullité simultanée des coefficients de pop, income et nox :
    a) En utilisant la procédure des tests généralisés [**TUTO 2, page 7**] (F calculé = 1.195 ; p-value =0.3208…)
    b) En utilisant la procédure basée sur la comparaison des R^2 (https://eric.univ-lyon2.fr/~ricco/cours/cours/econometrie_regression.pdf ; section 10.4.3). Vous devriez obtenir une valeur identique de la statistique de test. Utilisez la librairie Scipy pour obtenir la p-value [**TUTO 3, page 18**] (même valeurs de stat de test et de p-value, sinon problème).

Détection des points atypiques

15. Nous souhaitons détecter les points atypiques et ou influents. Affichez le levier et le résidu studentisé des observations [**TUTO 2, page 9**].

16. Passez à la représentation graphique de ces informations. Quelle est l'observation qui semble la plus problématique dans les données ? [**TUTO 2, pages 11 et 12**]

17. Énumérez les villes à problèmes au sens du levier. [**TUTO 2, page 12**] (['Los Angeles, Long Beach, CA', 'San Francisco, CA', 'Bridgeport-Milford, CT', 'Worcester, MA', 'New York, NY']).

18. Énumérez les villes à problèmes au sens du résidu studentisé (au risque 10%) [**TUTO 2, page 13**] (['Los Angeles, Long Beach, CA', 'New Orleans, LA', 'Miami-Hialeah, FL', 'York, PA', 'Lancaster, PA', 'Chicago, IL']).

19. Quelles sont les villes problématiques au sens du levier **OU** du résidu studentisé ? (chacun fait comme il le souhaite, pour ma part j'ai réalisé un OU logique sur les deux vecteurs de booléens issus des tests sur le levier et le résidu studentisé -- https://docs.scipy.org/doc/numpy/reference/generated/numpy.logical_or.html) (10 observations).

20. Retirez ces villes de la base de données. Nommez **dfclean** le nouvel ensemble de données. De combien d'observations disposons-nous à présent ? (Il reste **48 observations** dans la base expurgée des observations à problèmes détectées précédemment) **Nous travaillons sur cette nouvelle base à partir de maintenant**.

Détection de la colinéarité

21. Relancez la régression et affichez les résultats détaillés.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:              mortality   R-squared:                       0.699
Model:                            OLS   Adj. R-squared:                  0.655
Method:                 Least Squares   F-statistic:                     15.85
Date:                Wed, 21 Mar 2018   Prob (F-statistic):           2.56e-09
Time:                        10:26:26   Log-Likelihood:                -233.75
No. Observations:                  48   AIC:                             481.5
Df Residuals:                      41   BIC:                             494.6
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   1192.0128    134.301      8.876      0.000     920.786    1463.239
julyTemp      -0.3944      1.307     -0.302      0.764      -3.034       2.245
rain           3.1994      0.635      5.035      0.000       1.916       4.483
education    -27.2583      8.520     -3.199      0.003     -44.465     -10.052
pop         1.413e-05    7.1e-06      1.988      0.053   -2.21e-07    2.85e-05
income        -0.0021      0.002     -1.066      0.293      -0.006       0.002
nox            0.7001      0.436      1.607      0.116      -0.180       1.580
==============================================================================
Omnibus:                        3.741   Durbin-Watson:                   2.064
Prob(Omnibus):                  0.154   Jarque-Bera (JB):                3.533
Skew:                           0.611   Prob(JB):                        0.171
Kurtosis:                       2.479   Cond. No.                     4.46e+07
==============================================================================
```

22. Quelles sont les variables associées à des coefficients significatifs à 10 % maintenant ? (rain, education, pop).
23. Détecter la colinéarité qui peut exister entre les variables explicatives du modèle en utilisant le critère VIF. Y a-t-il un problème de colinéarité dans cette régression ? [**TUTO 2, pages 15 et 16**] (non).
24. Testez à 10 % la nullité simultanée des coefficients associés aux variables « julyTemp » et « income ». Peut-on retirer ces variables du modèle ? (F = 0.6278, p-value = 0.5388).
25. **On décide de retirer ces variables**. **Réaliser la régression avec les exogènes : rain, education, pop, nox**. Interprétez les coefficients en analysant notamment leurs signes. Quelles sont les caractéristiques qui contribuent à la hausse (à la baisse) de la mortalité ?
26. Affichez le graphique qqplot pour les résidus de cette nouvelle régression. Que constate-t-on ?

Prédiction ponctuelle et par intervalle

27. Dans la feuille « prédiction » du fichier « **mortality.xls** » figurent les caractéristiques de la ville de Columbus. Calculer la prédiction ponctuelle du modèle pour cette observation (**[TUTO 2, à partir de la page 16]**) (prédiction ponctuelle : 895.5509…)
28. Calculez l'intervalle de prédiction au niveau de confiance 90 %. Est-ce qu'elle contient la vraie valeur de mortalité pour la ville de Colombus ? (835.8890 ; 955.2128).

Note finale : La lecture de ce support devrait vous être bénéfique :
http://statsmodels.sourceforge.net/devel/examples/notebooks/generated/ols.html