Ecole Centrale de Lyon
Statistique appliquée aux sciences de l'ingénieur

BE1- Régression linéaire - A RENDRE

C. Helbert

Exercice 1 On s'intéresse à l'acidification de l'océan suite aux activités humaines.
Le fichier "ph.txt" contient trois colonnes :
- *days* : le nombre de jours depuis le début de l'acquisition des mesures. Le jour 1 correspond au 2/10/1988, le jour 11028 au 11/12/2018.
- *temp* : la température de l'océan
- $pHcalc\_insitu$: le pH de l'océan calculé à partir de mesures in situ.

1. Proposer un modèle de régression de $pHcalc\_insitu$ en fonction de *days*
   a) Représenter la droite de régression et les données sur le même graphique.
   b) Quel est le pourcentage de variance expliquée par cette régression ?
   c) Analyser le test de student. On explicitera les hypothèses $\mathcal{H}_0$ et $\mathcal{H}_1$, la statistique du test, sa loi sous $\mathcal{H}_0$, la p_{value} et la conclusion du test.

2. Avec ce modèle que peut-on prévoir comme pH moyen en 2050 ($days = 19000$) ? Donner un intervalle de confiance pour cette grandeur au niveau de confiance 95%.

3. Utiliser un intervalle de prédiction à 95% pour savoir si le pH en 2050 pourra être observé en dessous de 8 ?

4. Etudier les résidus.

5. L'océan se réchauffe-t-il ? On répondra à cette question par un modèle statistique qu'on prendra soin de valider et dont on analysera les différents tests.

Exercice 2 Le jeu de données étudié ici concerne la valeur des logements des villes aux alentours de Boston. On cherche à identifier les variables dont dépend la valeur des logements .

Les variables utilisées sont les suivantes :
- CRIM taux de criminalité par habitant
- ZN proportion de terrains résidentiels
- INDUS proportion de terrains industriels
- CHAS 1 si ville en bordure de la rivière Charles 0 sinon
- NOX concentration en oxydes d'azote
- RM nombre moyen de pièces par logement
- AGE proportion de logements construit avant 1940
- DIS distance du centre de Boston
- RAD accessibilité aux autoroutes de contournement
- TAX taux de l'impôt foncier
- PTRATIO rapport élèves-enseignant par ville
- LSTAT % de la population à faibles revenus

— *class* valeur du logement en 1000$

On commence par mettre en place un modèle complet permettant de modéliser *class* en fonction d'une combinaison linéaire des autres variables.

1. Quelle est la part de variance expliquée par ce modèle ?
2. Le modèle de régression est-il significatif dans son ensemble (prendre un risque de première espèce $\alpha = 1\%$) ? Donner l'hypothèse H_0, la statistique du test, sa loi sous H_0 et la conclusion.
3. Quelles sont les variables significatives (prendre un risque de première espèce $\alpha = 1\%$) ? Est-on sûr qu'il n'y en a pas d'autres ?
4. Proposer une méthode pour simplifier le modèle. Expliquer la méthode. La mettre en oeuvre.
5. Le modèle obtenu est-il satisfaisant ?
6. Proposer un meilleur modèle.