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

BE1- Régression linéaire - A RENDRE

C. Helbert

Exercice 1 On s'intéresse au prix de mise en vente des appartements à Grenoble. Pour ce faire, dans le fichier "immo.txt", on a récolté 27 appartements pour lesquels on connait la surface et le prix de mise en vente (en Keuros). On voudrait savoir s'il y a un lien entre surface et prix de mise en vente.

1. Proposer un premier modèle de régression.
   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 l'hypothèse $\mathcal{H}_0$, 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 prix moyen de mise en vente pour un appartement de 90 m^2. Donner un intervalle de confiance pour cette grandeur au niveau de confiance 95%.

3. Utiliser un intervalle de prédiction à 95% pour savoir si c'est statistiquement acceptable de mettre en vente un appartement de 90 m^2 à 280 Keuros.

4. Etudier les résidus et proposer un deuxième modèle de régression plus adapté à ces données.
   — Vérifier la qualité de la régression
   — L'incertitude de prédiction pour un appartement de 90 m^2 a-t-elle été réduite ?

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.