Régression sur données réelles

1 Données

On va travailler sur les données `Boston`. Ces données sont directement accessibles sous R, via la librairie `MASS`. Elles contiennent des informations sur 506 différents arrondissements de Boston (une ligne de la base de données représente donc un arrondissement). Ces données sont issues du recensement effectué en 1970 aux États-Unis. Dans ce TP nous nous intéresserons uniquement aux deux variables `lstat` et `medv`. `lstat` représente le pourcentage d'adultes dans l'arrondissement qui n'ont pas atteints un certain niveau d'éducation au lycée et `medv` représente la valeur médiane (en milliers de dollars) des maisons de l'arrondissement. On cherche à expliquer `medv` (la variable Y) en fonction de `lstat` (la variable X).

2 Ajustements linéaire, polynomiaux, régressogramme et courbe de régression

1. Représentez graphiquement Y en fonction de X en utilisant la fonction `plot` de R.
2. Estimer les paramètres de la droite des moindres carrés de deux façons et tracez la droite :
   — calculez les coefficients avec les formules du cours.
   — utilisez la fonction `lm` de R. La commande `summary` permet alors d'afficher les coefficients des moindres carrés (ainsi que le R^2).
   — tracez l'ajustement linéaire sur le nuage de points en utilisant la commande `abline`.
3. Calculez le coefficient de corrélation ainsi que le R^2.
4. Effectuez le test de corrélation à partir de la fonction `cortest` du package `robusTest`. Pour cela tapez les commandes suivantes :

   ```
   install.packages("robusTest")
   library(robusTest)
   cortest(X,Y)
   ```

   où X et Y représentent les deux variables pour lesquelles on souhaite effectuer le test de corrélation.
5. Calculez les valeurs ajustées du modèle avec la commande `fitted`.
6. Calculez les résidus du modèle ainsi que leur écart-type. Représentez l'histogramme des résidus, leur boîte à moustache, puis représentez le nuage de points des résidus en fonction de X (avec les droites horizontales à +/- deux fois l'écart-type estimé). Commentez l'ajustement.
7. Pour une nouvelle valeur de X, donnez une prévision de Y à l'aide de la commande `predict`.
8. On cherche à présent à modéliser Y en fonction de X par régressogramme ou à l'aide de la droite de régression.
   — Pour cela, créez une nouvelle variable qualitative à partir de `lstat` qui contient 6 modalités : une modalité pour chacun des évènements $X \leq 5$, $5 < X \leq 10$, $10 < X \leq 15$, $15 < X \leq 20$, $20 < X \leq 25$, $25 < X$. On pourra pour cela s'aider de la fonction `cut` de R.

— Calculez la moyenne de Y pour chacune des modalité de cette nouvelle variable.
— Ajustez les valeurs calculées précédemment par une courbe constante par morceaux (ce qui donne le régressogramme) ou reliez les valeurs calculées précédemment par une droite (ce qui donne la courbe de régression).

9. Ajustez Y en fonction de X par un polynôme d'ordre 2. Donnez le R^2 à l'aide de la commande `summary`. Mêmes questions pour des polynômes d'ordres supérieurs. Tracez ces polynômes sur le nuage de points et commentez.
10. Calculez le critère AIC de chacun des différents modèles (le modèle linéaire ainsi que les différents polynômes qui ont été implémentés). On pourra utiliser la fonction `AIC` de R. Conclure sur le modèle qui vous semble le plus pertinent.