Université Iba Der Thiam **Année Académique 2020-2021**
U.F.R SET **Master 1 STAT ECO**

Cours de Modèles de Régression

Travail à rendre : Modèles linéaires

Exercice 1 Parmi les différentes substances chimiques qui polluent l'air, on compte l'ozone qui a un impact préoccupant sur la santé. Afin de mieux comprendre les conditions météorologiques qui favorisent une teneur élevée d'ozone dans l'air, nous allons analyser un jeu de donnes qui contient 13 variables au total dont la concentration en ozone pendant la journée et des informations sur la température, la nébulosité, le vent, la pluie. Nous disposons de 112 données relevées durant l'été 2001 à Rennes.

Dans ce TP nous nous limitons à analyser la concentration maximale en ozone par jour en fonction de la température prévue à midi. Ce sont les variables de nom maxO3 et T12 dans le fichier ozone.txt. Ainsi, nous considérons le modèle :

$$maxO3 = \beta_0 + \beta_1 * T12 + \varepsilon$$

1. Importer les données. Familiarisez-vous avec les données.
2. Tracer le nuage des points et superposer la droite de régression. Quelle est la conclusion du test $H_0 : \beta_k = 0$ contre $H_1 : \beta_k \neq 0$ pour chaque $k \in \{0,1\}$? Ajouter au graphique la droite de régression du modèle sans intercept à savoir $maxO3 = \beta * T12 + \varepsilon$ et comparer.
3. Vérifier par un calcul explicite que les valeurs des estimateurs $\hat{\beta}_0$, $\hat{\beta}_1$ renvoyées par la fonction **lm** sont exactes.
4. Est-ce que le jeu de données contient des valeurs aberrantes? Retracer le nuage des points en marquant les observations aberrantes.
5. Vérifier avec un QQ-plot approprié l'hypothèse gaussienne des résidus ε_i.
6. Comparer par un graphique les résidus estimés $\hat{\varepsilon}_i$ aux résidus standardisés t_i (accessible par la fonction **rstandard**) et aux résidus studentisés t_i^*.
7. Est-ce que le jeu de données contient des points leviers? Marquer les points leviers dans le nuage des points (**T12**,**max03**) (en plus des valeurs aberrantes).
8. Analyser la distance de Cook des observations.
9. Ajouter à la figure du nuage des points et de la droite de régression les intervalles de prédiction et de confiance en tout point xi observé. Comparer les deux intervalles. Interpréter la forme des bandes de ces intervalles. Interpréter la relation entre intervalle de confiance/prédiction avec les observations atypiques.

Exercice 2 On considère les données de l'exercice précédent. Ici on veut la concentration maximale en ozone par jour en fonction de tous les autres variables. On utilisera un modèle de régression multiple.

— En utilisant le critère du R^2, effectuez une sélection des variables.
— Effectuez l'analyse de résidus du modèle choisi.
— Étudier la colinéarité des variables explicatives du modèle choisi.

Exercice 3 Considérons le modèle linéaire $y_i = \beta + \epsilon_i, i = 1, ..., n$ où les ϵ_i sont i.i.d. de lois $\mathcal{N}(0, \sigma^2)$. Proposer deux estimateurs différents du paramètre β.