Voir en particulier ce document pour cet exercice << Rakotomalala R., « Econométrie – La régression linéaire simple et multiple », 2016 ; http://eric.univ-lyon2.fr/~ricco/cours/cours/econometrie_regression.pdf >>. Les numéros de pages ou de sections indiqués dans le sujet y feront référence.

Le fichier « **crime.xlsx** » décrit les informations concernant n = 47 états américains en 1960 (Source : http://lib.stat.cmu.edu/DASL/Datafiles/USCrime.html). La variable « CrimeRate » joue un rôle particulier, elle indique le taux de criminalité dans les états.

Réalisez l'exercice exclusivement sous R (RStudio). Utilisez intensivement Google pour identifier les commandes adéquates. Des sites généralistes telles que Quick-R (http://www.statmethods.net/) devrait vous aider à avancer rapidement. A rendre, un fichier « .r » avec les commandes commentées.

1. Importez le fichier « **crime.xlsx** ». Combien d'observations et de variables comporte-t-il ?
2. Calculer la corrélation entre CrimeRate et W. Est-elle significativement différente de 0 au risque de 5 % (H0 : r = 0 vs. H1 : r ≠ 0) (cf. http://eric.univ-lyon2.fr/~ricco/cours/cours/Analyse_de_Correlation.pdf ; section 2.4).
3. Quelles sont les 5 variables les plus liées positivement à CrimeRate ? Pour répondre à cette question, vous devez calculer la corrélation de CrimeRate avec chacune des autres variables et trier les résultats par ordre décroissant.
4. On constatera que les dépenses de police (Ex0 et Ex1) sont très fortement liées au taux de criminalité, cela ne veut pas dire que plus on mise sur la sécurité, plus la criminalité augmente. Ce serait plutôt l'inverse. Elles ne peuvent pas expliquer la criminalité, nous allons les exclure de l'analyse. Créez un nouveau data frame sans Ex0 et Ex1.
5. Effectuez la régression expliquant CrimeRate à l'aide de W (**lm**). Observez attentivement les attributs de l'objet fourni par la régression (attributes), faites de même pour celui issu de la commande summary(). Quelles sont les valeurs des paramètres estimés de la régression ? Quelle est la valeur du R^2 (coefficient de détermination) de la régression. Que signifie cet indicateur ?
6. La régression est-elle globalement significative à 5% (section 3.1.1). La pente de la droite de régression est-elle significative (section 3.3.1) ? Rapprochez ces différents résultats avec celui du test de significativité de la corrélation (question n°2). Remarquez-vous quelque chose ?
7. Comment interpréter le coefficient lié à la variable W dans la régression ? Calculer son intervalle de confiance à 95 %.
8. Créer un graphique avec en abscisse W et en ordonnée CrimeRate. Faites-y figurer la droite de régression. Que penser de la qualité de la régression ?
9. Prédiction ponctuelle : quelle serait le taux de criminalité d'un état avec W = 507 (section 4.1) ?
10. Calculez pour la même observation supplémentaire l'intervalle de prédiction au niveau de confiance 90% (section 4.2.3).

11. On souhaite maintenant réaliser la régression en mettant à contribution toutes les variables (à l'exception de Ex0 et Ex1 que nous avions exclues). Quelle est la valeur du R^2 ? La régression est-elle globalement significative à 5% (section 10.2) ?
12. Quelles sont les variables significatives à 5% (section 10.3) ? (Remarque : exploitez les sorties de R, aucun calcul à faire explicitement).
13. On veut tester à 5% la nullité simultanée des coefficients liées aux variables (S, LF, N, NW) (section 10.4). Que peut-on conclure ?
14. Finalement, on préfère effectuer une sélection de variables Backward avec le critère Akaike AIC (http://eric.univ-lyon2.fr/~ricco/cours/slides/Reg_Multiple_Colinearite_Selection_Variables.pdf ; pages 8 et 9) (section 15.3.3, stepAIC). Quelles sont les variables retenues à l'issue du processus ?
15. A l'aide de modèle, effectuez la prédiction ponctuelle pour l'état américain avec les caractéristiques suivantes

Age	S	Ed	LF	M	N	NW	U1	U2	W	X
128	0	113	624	972	28	10	77	25	507	206

16. Calculez l'intervalle de prédiction à 90 %. Comparez l'intervalle obtenu avec celui de la question n°10. Que constatez-vous ? (section 12.2 ; **Note :** *n'insistez pas si vous rencontrez des difficultés, l'opération est assez ardue sous R*).