Nous travaillons sous R (RStudio) durant cette séance. Vous créerez un projet MARKDOWN (voir https://rmarkdown.rstudio.com/articles_docx.html).

Plusieurs **tutoriels** devraient vous aider :

- Manipulation des données sous R : http://tutoriels-data-mining.blogspot.fr/2012/08/manipulation-des-donnees-avec-r.html [TUTO 1]
- Régression et diagnostic de la régression : http://tutoriels-data-mining.blogspot.fr/2009/05/diagnostic-de-la-regression-avec-r.html [TUTO 2]
- Econométrie – La régression linéaire simple et multiple : http://eric.univ-lyon2.fr/~ricco/cours/cours/econometrie_regression.pdf [TUTO 3]
- Etude des résidus : http://eric.univ-lyon2.fr/~ricco/cours/slides/Reg_Multiple_Etude_Des_Residus.pdf [TUTO 4]
- Quantiles et probabilités des lois usuelles : http://tutoriels-data-mining.blogspot.fr/2017/04/probabilites-et-quantiles-sous-excel-r.html [TUTO 5]
- Régression sur variables qualitatives : http://eric.univ-lyon2.fr/~ricco/cours/slides/Reg_Multiple_Exogenes_Qualitatives.pdf [TUTO 6]

Nos **supports de cours** sont en ligne : https://eric.univ-lyon2.fr/~ricco/cours/cours_econometrie.html

ANALYSE DES CIGARETTES

1. Inspection des données

On souhaite expliquer la nocivité des cigarettes CO à partir de ses caractéristiques (TAR, NICOTINE et WEIGHT).

1. Chargez le fichier « **cigarettes_pour_regression.txt** » dans un data frame [cf. read.table() ; attention aux options de la procédure, dans notre fichier la première colonne correspond aux identifiants des véhicules, ce n'est pas une variable ; attention également au point décimal ; enfin, le séparateur de colonnes est le caractère tabulation « \t »] (TUTO 2, page 3).
2. Affichez les observations, puis affichez le nombre de lignes et de colonnes du data frame (TUTO 2, page 3) (dim) (24 observations et 4 variables).
3. Affichez les noms des observations et des variables (rownames, colnames).
4. Calculez les statistiques descriptives pour chaque variable (summary).
5. Réalisez les graphiques nuages de points en croisant deux à deux les variables (pairs) (TUTO 2, page 4). Que constatez-vous ? Les variables sont-elles liées entres elles ? Y a-t-il des points atypiques ?

2. Régression linéaire multiple

6. Réalisez une régression linéaire multiple expliquant la variable CO à partir de toutes les autres (lm) (TUTO 2, page 5).

7. Récupérez l'objet summary() issu de lm(). Affichez-le. Quelle est la valeur du R2 de la régression (R^2 = 0.935). Le modèle est-il globalement significatif à 5% ? (F = 95.86) Quelles sont les coefficients significatifs à 5% ? Ces résultats corroborent-ils les constations issues des graphiques précédents ?

8. Affichez le champ $coefficients de l'objet issu de summary(). Quel est le type de cet objet ? (class) Quelles sont ses dimensions ? (dim)

9. Affichez les écarts-type des coefficients estimés.

10. Pour chaque coefficient, calculez sont intervalle de confiance au niveau 95% (TUTO 5, page 12 pour le calcul des quantiles de la loi de Student).

Bornes basses :

```
(Intercept)         TAR    NICOTINE      WEIGHT
 -6.7496811   0.4798127  -6.2657743  -4.5507177
```

Bornes hautes :

```
(Intercept)         TAR    NICOTINE      WEIGHT
   5.646286    1.295348    7.302713    8.709406
```

3. Analyse des résidus

11. Récupérez les résidus de la régression ($residuals). Calculez sa moyenne. Que constatez-vous ? (TUTO 2, page 6).

12. Construisez le graphique nuage de points en croisant en abscisse la variable cible (CO) et en ordonnée le résidu (plot). Y a-t-il des éléments saillants dans le graphique ?

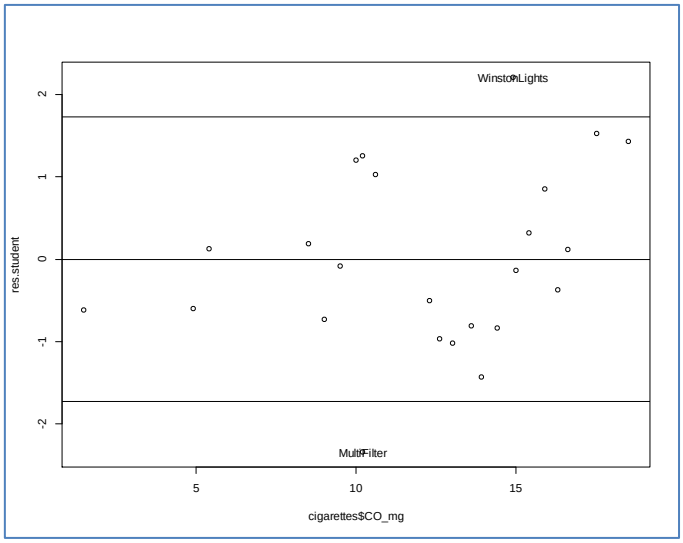
13. Réalisez la droite de Henry pour vérifier la compatibilité des résidus avec l'hypothèse de normalité (qqnorm). Que constatez-vous ? (TUTO 4, page 18)

14. Calculer le coefficient d'asymétrie g_1 (TUTO 4, page 19), puis le coefficient d'aplatissement g_2 (TUTO 4, page 20). Calculez alors la statistique de Jarque-Bera T (TUTO 4, page 21). Est-ce que les résidus sont

compatibles avec l'hypothèse de normalité ? (TUTO 5, page 15 ; pour le calcul de la p-value pour la loi du Khi-2). (g_1 = 0.1608, g_2 = -0.8123, T = 0.6361, p-value = 0.7275)

4. Détection des points atypiques et influents

15. Calculez le résidu studentisé de la régression (rstudent) (TUTO 2, page 9).
16. Calculez le seuil critique pour le résidu studentisé pour un risque de 10% (TUTO 2, page 9) (1.729133)
17. Quelles sont les marques de cigarette atypiques au sens de ce seuil ? (MultiFilter, WinstonLights).
18. Construisez le graphique nuage de points croisant en abscisse la variable cible CO et en ordonnée le résidu studentisé (plot). Insérez dans le graphique les lignes matérialisant les seuils critiques (text). Faites apparaître nommément les cigarettes atypiques [*Par rapport au tutoriel TUTO 2 page 9, il y a certainement mieux à faire que de passer par une boucle*].



19. Calculez le levier de chaque observation (TUTO 2, page 10) (influence.measures, voir la colonne « hat »).
20. Quels sont les points atypiques au sens du levier ? (Now).
21. Créez un nouveau data frame excluant les observations atypiques au sens du résidu studentisé **OU** du levier. De combien d'observations dispose-t-on maintenant ? (21 obs.). On travaille sur ce nouveau data frame à partir de maintenant.
22. Réalisez de nouveau la régression CO vs. les autres variables à partir de ce nouvel ensemble de données. Quelle est la valeur du R2 maintenant ? (0.9382)

5. Sélection de variables

23. Testez la significativité simultanée des coefficients de NICOTINE et WEIGHT en opposant les R^2 des régressions CO = f(TAR, NICOTINE, WEIGHT) et CO = f(TAR) (TUTO 3, section 10.4) (F = 0.04223…, p-value = 0.9587…).

24. Réalisez une sélection de variables « backward » optimisant le critère AIC (TUTO 2, page 14) (stepAIC). Quelles sont les variables finalement pertinentes pour l'explication du taux de CO des cigarettes ? Est-ce que ce résultat confirme la réponse obtenue dans la question précédente ?

6. Prédiction sur un nouveau fichier

25. Charger les données du fichier « **autres_cigarettes.txt** ». Combien y a-t-il de marques de cigarettes dans ce fichier ? (4)

26. Pour ces nouvelles observations, calculez les prédictions ponctuelles ainsi que leurs intervalles de confiance à 90% du modèle simplifié. Utilisez la commande predict(). Regardez du côté des paramètres pour produire automatiquement les intervalles de prédiction sans avoir à passer par les formules du cours (voir la doc. de R https://stat.ethz.ch/R-manual/R-devel/library/stats/html/predict.lm.html ; voir aussi http://www.sthda.com/english/articles/40-regression-analysis/166-predict-in-r-model-predictions-and-confidence-intervals/).

27. Sachant les vraies valeurs de l'endogène sont respectivement (Benz : 13.5, GoodLook : 21.3, RiverPlate : 8.25, Melia : 6.0). Pour quelles cigarettes l'intervalle de confiance de prédiction couvrent les bonnes valeurs de l'endogène ? (faites la vérification sous R !).

28. Accolez ces nouvelles variables (prédictions et bornes des intervalles de prédiction) au jeu de données « autres_cigarettes » (cbind).

29. Sauvegardez ce nouvel ensemble de données (data frame) dans le fichier « **output_regression.txt** ». Attention, vous devez respecter les conventions initiales (noms des variables sur la première ligne, étiquettes des observations sur la première colonne, séparateur tabulation, « . » comme point décimal) (write.table). Vérifiez dans un éditeur de texte que votre base a été exportée correctement.