Nous travaillons sous Excel durant cette séance.

REF 1 - Les éléments de calcul référencés sont accessibles dans ce support : http://chirouble.univ-lyon2.fr/~ricco/cours/cours/Regression_Lineaire_Multiple.pdf
REF 2 – Calcul des quantiles et des probabilités des lois statistiques d'usage courant - http://tutoriels-data-mining.blogspot.fr/2017/04/probabilites-et-quantiles-sous-excel-r.html
REF 3 – Sélection de variables - http://chirouble.univ-lyon2.fr/~ricco/cours/slides/Reg_Multiple_Colinearite_Selection_Variables.pdf
REF 4 – Etude des résidus - http://chirouble.univ-lyon2.fr/~ricco/cours/slides/Reg_Multiple_Etude_Des_Residus.pdf
REF 5 – Ouvrage « Econométrie – La régression linéaire simple et multiple » - http://chirouble.univ-lyon2.fr/~ricco/cours/cours/econometrie_regression.pdf

Régression multiple

Nous travaillons sur le fichier « **cars_acceleration_regression_multiple - sujet.xlsx** ». On cherche à expliquer la variable « accélération » (endogène) à partir des autres (p = 4) variables (cylinders, engine.displacement, horsepower, weight) (exogènes).

Calcul matriciel

1. De combien d'observations dispose-t-on ? (n = 230)
2. Copier les données initiales dans une nouvelle feuille. Rajouter la colonne de constante avec la valeur 1 pour chaque individu (REF 1, page 12)
3. Calculez les paramètres estimés â de la régression multiple en utilisant explicitement les formules matricielles (REF 1, pages 10 à 12). Attention : **[1]** TRANSPOSE permet de transposer une matrice, INVERSEMAT permet de l'inverser, PRODUITMAT permet de réaliser un produit matriciel ; **[2]** ce sont des fonctions matricielles, il faut valider la saisie avec CTRL + SHIFT + ENTREE (*aux fins de vérification : constante^ = 17.446..., a^_cylinders = -0.214..., a^_weight = 0.0027983*).
4. Calculez les valeurs estimées de l'endogène **Y^** à partir des coefficients estimés (REF 1, page 33) (PRODUITMAT). En déduire les résidus de la régression ($\hat{\varepsilon} = y - \hat{y}$).
5. Calculez à partir des informations disponibles la SCT, la SCR et la SCE (pour éviter d'avoir à créer des colonnes de valeurs supplémentaires, voir du côté des fonctions Excel :

SOMMES.CARRES.ECARTS et SOMME.CARRES). Elaborez le tableau d'analyse de variance (REF 1, page 22). Fournissez le R^2 (REF 1, page 23), le R^2 corrigé (REF 1, page 24) et l'écart-type estimé de l'erreur $\hat{\sigma}_\varepsilon$ (REF 1, page 16) (*respectivement : 0.6028… ; 0.5957… ; 1.6794*).

6. Le modèle est-il globalement significatif à 5% (REF 1, page 25). Vous pouvez décider en comparant le F calculé avec le quantile de la loi de Fisher (REF 2, page 19), ou en calculant la p-value et la comparer avec le risque (REF 2, page 18). (*F-calculé = 85.378… ; p-value ≈ 0*)
7. Testez la significativité globale du modèle au sens des critères AIC et BIC (REF 3, pages 8 et 9) (*AIC modèle = 243.432 ; AIC pire modèle = 447.81422 ; BIC modèle = 260.623…*). Remarque : Le pire modèle – qui sert de référence - est celui pour lequel nous utilisons seulement la constante pour prédire les valeurs de l'endogène. Dans ce cas, SCR = SCT.

Calcul avec DROITEREG

8. Copier les données initiales dans une autre feuille. Appliquez la fonction DROITEREG sur vos données. Vérifiez que les résultats concordent avec ceux obtenus durant les questions précédentes (â, R^2, SCE, SCR, F, etc.) (REF 1, page 23).

Toutes les variables (Régression 1)				
weight	horsepower	eng.disp.	cylinders	constante
0.002798339	-0.077640356	-0.00550159	-0.214369	17.446394
0.000358056	0.006847739	0.004654132	0.2116059	0.767001
0.602832874	1.679425459	#N/A	#N/A	#N/A
85.37803604	225	#N/A	#N/A	#N/A
963.2247136	634.6057212	#N/A	#N/A	#N/A

9. Testez la significativité à 5% de chaque coefficient. Quelles sont les variables exogènes qui ne semblent pas avoir d'impact sur l'accélération ? (REF 1, pages 19 et 20) (t_{weight} = 7.81… ; $t_{horsepower}$ = -11.338…, *etc.*) (*non significatives à 5% = eng.disp et cylinders*) (REF 2, pages 11 à 13 pour le calcul des quantiles et des probabilités de la loi de Student).
10. Calculez les intervalles de confiance au niveau 95% de chaque coefficient (REF 1, pages 19 et 20). Est-ce que les résultats sont cohérents avec ceux obtenus lors des tests de significativité individuelle ? (oui, sinon il y a un sérieux problème dans les calculs)
11. Construisez le graphique des résidus (résidus en ordonnée, endogène en abscisse). Avez-vous des commentaires particuliers à formuler ? (REF 4, pages 5 à 8) (Il y a un problème manifestement, les résidus ne sont pas dispersés au hasard, nous passons outre et enchaînons la suite du TD,

mais dans une étude réelle il faudrait s'en occuper, peut-être en appliquant une transformation sur l'endogène par exemple).

12. On souhaite opérer une sélection de variables. Adoptez la démarche BACKWARD avec un niveau de signification à 5% (REF 3, page 10). Quelles sont les variables retenues dans le modèle finalement ? Ce résultat n'est-il pas curieux par rapport à ce que nous avions observé à la question sur la significativité individuelle des coefficients ? (*variables sélectionnées : weight, horsepower, eng.disp*).

Prédiction ponctuelle et par intervalle

13. A partir du modèle simplifié (après sélection de variables), effectuez la prédiction ponctuelle pour l'individu situé dans la feuille « **à prédire** » (REF 1, page 33) (prédiction ponctuelle : 16.839).
14. Calculez l'intervalle de prédiction au niveau de confiance 95% (REF 1, pages 33 et 34). Vérifiez que votre fourchette contient bien la valeur d'accélération que vous pouvez lire dans la feuille « vraies valeurs » (fourchette de prédiction à 95% : [13.507 ; 20.173]).

Interprétation et quelques tests supplémentaires

15. Revenons sur le premier modèle contenant l'ensemble des variables. Triez les variables selon leur degré d'influence dans la prédiction de l'accélération (quelle est la variable la plus influente, la seconde, etc.). Quel critère avez-vous utilisé pour mesurer l'influence des variables ?
16. Calculez les coefficients standardisés des variables (REF 5, section 13.2). Constatez-vous la même conclusion concernant l'impact comparé des variables exogènes dans la régression ? Comment interpréteriez-vous ces coefficients standardisés ?
17. Testez l'hypothèse H_0 : $a_{engine.disp} = a_{cylinders} = 0$ (les deux coefficients sont simultanément nuls) vs. H_1 : un des deux est non nul (à 5%) (REF 1, pages 28 et 29) (F calculé = 3.6769)
18. Refaites le même test en utilisant la procédure décrite dans (REF 5, section 10.4) (F calculé = 3.6769 ; identique à la question précédente, heureusement).
19. Refaites le même test en utilisant la procédure des « q » contraintes linéaires sur les coefficients (REF 1, pages 30 et 31) (F calculé = 3.6769 ; encore une fois).