Nous travaillons sous Excel durant cette séance.

Nous utiliserons les supports suivants :

REF 1 - http://chirouble.univ-lyon2.fr/~ricco/cours/cours/diapo_analyse_de_correlation.pdf

REF 2 - http://tutoriels-data-mining.blogspot.fr/2017/04/probabilites-et-quantiles-sous-excel-r.html

REF 3 - http://chirouble.univ-lyon2.fr/~ricco/cours/cours/Regression_Lineaire_Simple.pdf

Les résultats en gris sont fournis à titre indicatif pour que vous puissiez étalonner vos calculs.

Analyse des corrélations

Nous travaillons sur le fichier « **autos_correlation - sujet.xlsx** ».

1. La variable **city-mpg** correspond à la consommation en ville ; **mpg** indique le nombre de miles que l'on peut parcourir avec un gallon de carburant. Sachant que 1 miles = 1.609 km et 1 gallon = 3.785 litres. Créer la colonne **CONSO** qui indique la consommation des véhicules en ville en litres/100 km. [Indication : CONSO = (1 / CITYMPG) * (3.785/1.609) * 100]
2. Isolez les véhicules "fuel-type = gas ET aspiration = std" (Indication : *un tri à deux critères + copier/coller fait l'affaire, si vous connaissez les filtres Excel et que vous savez les utiliser c'est mieux* ; Réponse : 161 véhicules). Nous travaillons sur cet échantillon jusqu'à la fin de l'exercice.
3. Sur ce nouveau sous-échantillon. Calculer la corrélation entre CONSO et PRICE en construisant explicitement les éléments de la formule (REF 1, page 14) c.-à-d. construisez les colonnes XY, X^2, Y^2, XBarre, YBarre, etc. (Réponse : r = 0.8366)
4. Calculez la corrélation en utilisant la fonction COEFFICIENT.CORRELATION d'Excel. Obtenez-vous le même résultat ? (Oui a priori, il y a un sérieux problème dans le cas contraire).
5. La corrélation est-elle significativement non nulle à 5% ? (REF 1, pages 18 et 19) (t-calculé = 19.26) Remarque : vous avez deux pistes possibles pour la règle de décision, comparer la statistique de test avec le quantile de la loi de Student (LOI.STUDENT.INVERSE.N) (REF 2, page 12) ou calculer la p-value à partir de la statistique de test (LOI.STUDENT.BILATERALE) (REF 2, page 13).
6. Calculez l'intervalle de confiance au niveau 95% du coefficient de corrélation (REF 1, pages 21 et 22) (Indication : TANH correspond à la fonction inverse de la transformation de Fischer ;

Réponse : intervalle de confiance à 95% = [0.783, 0.877]). Pour les quantiles de la loi normale centrée et réduite sous Excel (LOI.NORMALE.STANDARD.INVERSE.N, voir REF 2, page 7).

7. D'autres variables peuvent interférer dans la relation CONSO et PRIX. Calculer la corrélation partielle r(CONSO,PRIX / ENGINE_SIZE) (Réponse : 0.465). Vérifier qu'elle est significative à 5% (Attention aux degrés de liberté, t-calculé = 6.60). Que faut-il en conclure ? (REF 1, pages 30 et 31).
8. Rajoutons une seconde variable de contrôle, calculer la corrélation partielle r(CONSO,PRIX / ENGINE_SIZE,HORSEPOWER) (Réponse : 0.4235) (REF 1, page 30 pour la formule de récurrence). Est-elle significative à 5% ? (Attention aux degrés de liberté, t-calculé = 5.858) Conclusion ?
9. Revenons à CONSO vs. PRICE. Construisez le graphique « nuage de points » avec CONSO en ordonnée et PRICE en abscisse. Que constate-t-on ? Quelle transformation de variables pourrait-on introduire pour améliorer le coefficient de corrélation ? Appliquez votre idée et calculez le coefficient de corrélation. Observez-vous une amélioration ? (Indication : la corrélation sur variables transformées devrait être plus élevée, ce qui légitimerait les transformations opérées).
10. Transformez CONSO et PRICE en rangs (RANG). Calculez et testez la significativité du coefficient ainsi calculé sur les rangs. Que constatez-vous ? (Indication : le passage au rang est une autre manière de dépasser le problème de non-linéarité, tant que la relation est monotone ; Réponse : r = 0.88079…)