Nous travaillons sous R (RStudio) durant cette séance.

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 sous R : http://tutoriels-data-mining.blogspot.fr/2009/05/diagnostic-de-la-regression-avec-r.html [**TUTO 2**]
- Calcul des quantiles et probabilités des lois statistiques d'usage courant : http://tutoriels-data-mining.blogspot.fr/2017/04/probabilites-et-quantiles-sous-excel-r.html [**TUTO 3**]
- Régression linéaire simple : http://eric.univ-lyon2.fr/~ricco/cours/cours/Regression_Lineaire_Simple.pdf [TUTO 4]

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

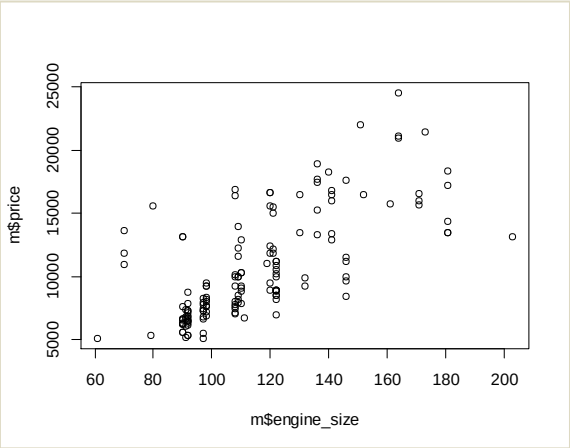
1. Chargement et manipulation des données

1. Nous traitons le fichier « autos.txt » qui décrit les caractéristiques d'un certain nombre de véhicules. Chargez le fichier dans un éditeur de texte quelconque pour observer l'organisation des données : que représente la première ligne (header = T) ? quel est le séparateur de colonnes (sep = ''\t'') ? quel est le point décimal (dec = ''.'') ?
2. Lancez R-Studio et créez un nouveau fichier script FILE / NEW FILE / R SCRIPT. Enregistrez votre script dans un fichier « exercices_td3.r ».
3. Modifiez le répertoire par défaut en indiquant le chemin de votre dossier (setwd) [TUTO 1, page 8 ; faites attention au séparateur dans l'écriture du chemin].
4. Chargez le fichier « autos.txt » dans une structure data.frame (read.table) que vous nommerez m.prim [TUTO 1, page 5].
5. Combien y a-t-il de variables dans le fichier ? Combien y a-t-il d'observations ? (nrow, ncol) (205, 19).
6. Affichez la liste des variables et leurs types (str).
7. Calculez les statistiques descriptives (summary) [TUTO 1, page 8].
8. Calculez la moyenne (mean), l'écart-type (sd) et les 1er et 3ème quartiles (quantile) de la variable price [TUTO 1, page 10]. Les résultats devraient concorder avec ce que l'on a observé avec summary().

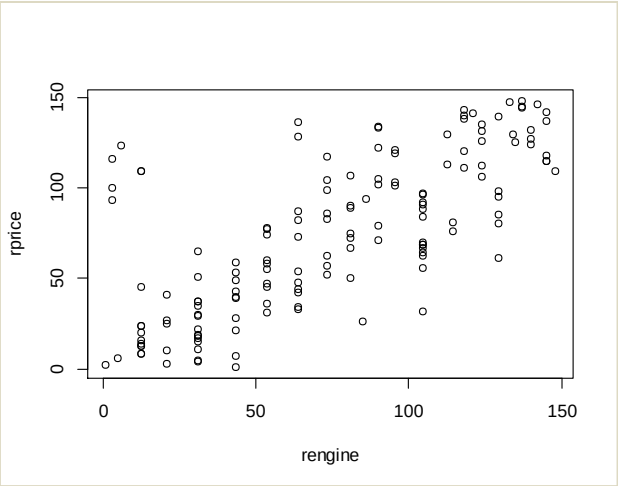
9. Calculez la fréquence des véhicules selon fuel_type (table) [TUTO 1, page 13] (idem, comparez avec ce que proposait summary)
10. Calculez la fréquence des véhicules selon aspiration (table) [TUTO 1, page 13]
11. Combien y a-t-il de véhicules « fuel_type = gas » et « aspiration = std » (table) [TUTO 1, pages 13 et 14] (161 observations)
12. Isolez dans un nouveau data frame, que vous nommerez m, les véhicules correspondant aux caractéristiques (« fuel_type = gas » et « aspiration = std » et « price < 30000 »). Conservez l'ensemble des variables. Combien d'observations correspondent à ces caractéristiques ? (nrow) (148 observations) [exemples TUTO 1, page 13]. A partir de maintenant, nous travaillerons sur ce sous-ensemble d'observations.

2. Corrélation

13. Créez le graphique « nuage de points » avec en abscisse « engine_size » et en ordonnée « price » (plot). Que constatez-vous ? [TUTO 1, page 15] (on peut supputer raisonnablement une forme de liaison positive)



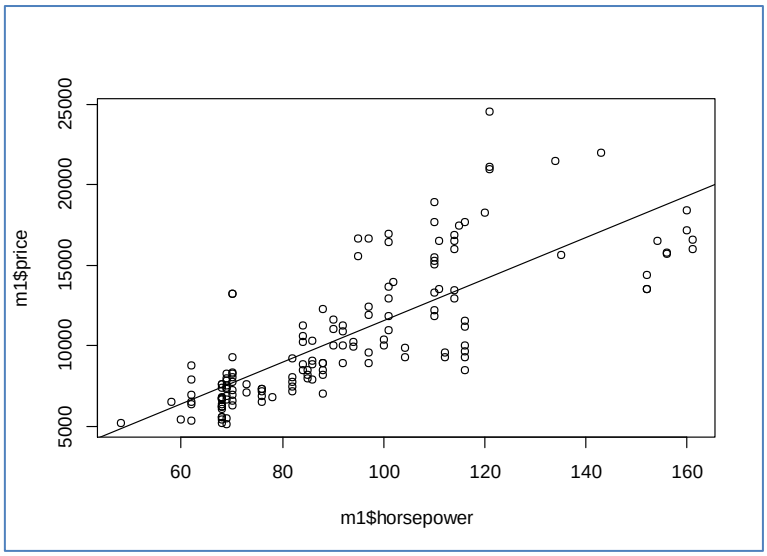
14. Calculez le coefficient de corrélation entre ces deux variables (cor) (0.699).
15. Le coefficient est-il significativement non nul au risque 5 % ? Quelle est la valeur du t-calculé ? (t-calculé = 11.81). Quelle est la valeur de la p-value (pt, pour la fonction de répartition de la loi de Student, attention il s'agit d'un test bilatéral ; TUTO 3, page 11) (p-value = 0.0000).
16. Calculez l'intervalle de confiance au niveau 95% du coefficient de corrélation (log = logarithme népérien ; qnorm fournit le quantile de la loi normale [TUTO 3, page 7] ; sqrt pour la racine carrée ; exp pour l'exponentielle ; tanh pour l'inverse de la fonction de transformation de Fisher) (intervalle de confiance = [0.6061 ; 0.7732]).
17. Transformez les 2 variables en rangs (rank). Refaites le graphique puis calculez le coefficient de corrélation (0.729).



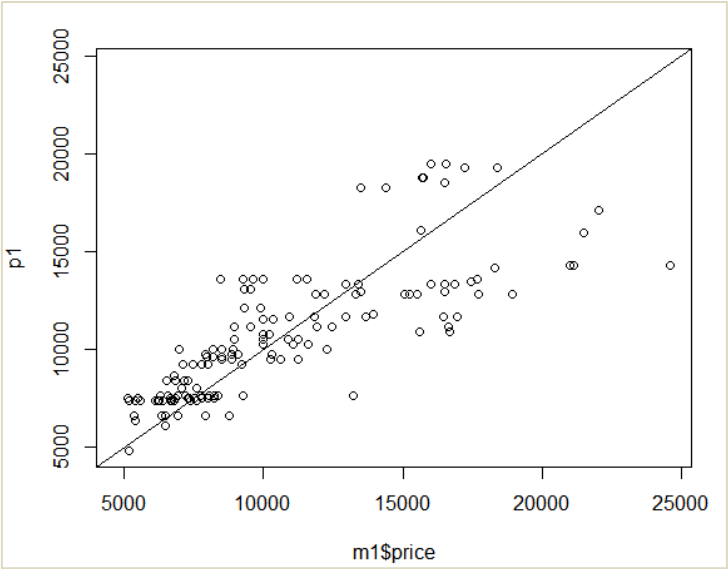
18. Remarquez-vous quelque chose d'étrange dans le graphique ? (*Vous devriez…*). Isolez les véhicules concernés et indiquez leurs marques (ce sont 6 voitures japonaises – et non pas 5 comme on pourrait le penser, 2 ''isuzu'' se superposent – après vérification).

3. Régression simple

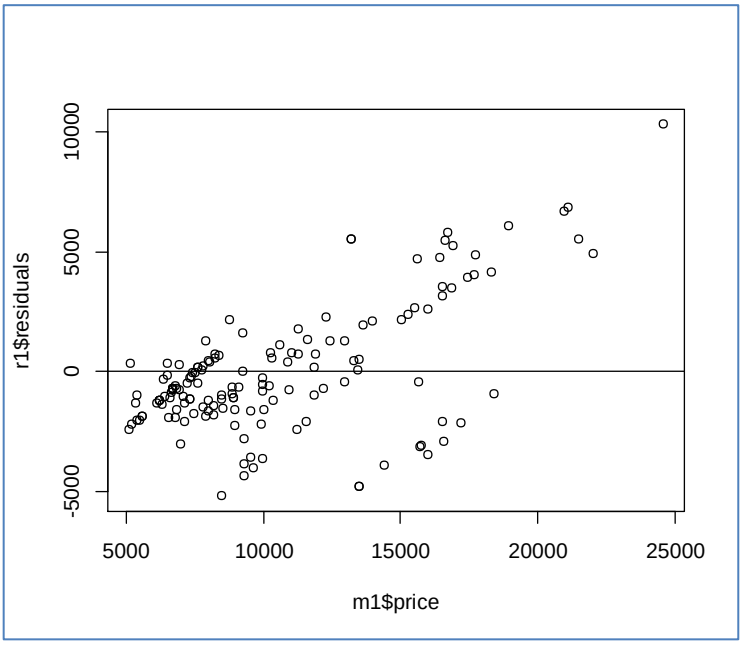
19. Revenons sur les variables non transformées en rangs. On souhaite expliquer le « price » (Y) à partir de la puissance « horsepower » (X). Créer le graphique « nuage de points » entre « horsepower » (abscisse) et « price » (ordonnée). Que remarquez-vous ? (il y a un point atypique énorme, une voiture très puissante !).
20. Quelle est la marque de ce point atypique ? (porsche). Nous décidons de la retirer de la base qui fait 147 observations maintenant. Votre nouveau data frame s'appelle m1.
21. Réalisez maintenant la régression linéaire simple de « price - Y » en fonction de « horsepower - X » (Y = a X + b) (lm) [TUTO 2, page 5]. Quelles valeurs des coefficients a et b obtenez-vous ? (a^ = 129.5951, b^ = -1415.8781).
22. Affichez les attributs de l'objet régression (attributes). Afficher explicitement les coefficients en accédant à la propriété « coefficients » de l'objet.
23. Créer le graphique « nuage de points » entre « horsepower » (abscisse) et « price » (ordonnée). Ajouter la droite de régression dans le graphique (abline)



24. Récupérez l'objet issu de la fonction summary() de la régression [TUTO 2, page 5]. Affichez l'objet lui-même (print).
25. En lisant la sortie de summary, quelles sont les valeurs du R^2 et R^2 ajusté ? (0.6176 et 0.6149) La régression est-elle globalement significative à 5% ? (F = 234.2, p-value = 2.2e-16).
26. Affichez la liste des attributs du résumé de la régression (attributes). Affichez explicitement le tableau des coefficients, quelles sont ses dimensions (dim) ? (2 lignes, 4 colonnes).
27. Calculez l'intervalle de confiance au niveau de confiance 90% de la pente de la régression (qt permet de calculer les quantiles de la loi de Student ; [TUTO 3, page 12]) (intervalle de confiance de a^ = [115.57, 143.61]).
28. Récupérez les valeurs de la variable cible prédites par le modèle ($fitted.values de l'objet régression). Afficher le graphique avec en abscisse les valeurs observées de « price » et en ordonnée les valeurs prédites par le modèle. Tracer une droite sur la diagonale principale pour situer la qualité de la prédiction. *Attention, mettez les mêmes limites en abscisse et ordonnée pour que le graphique soit carré (cf. les options de plot)*. Que constatez-vous ? (le modèle sous-estime les hautes valeurs de price)



29. Réalisez le graphique des résidus avec price en abscisse et les résidus en ordonnée. Constatez-vous la même chose que dans le graphique précédent ? (oui, forcément, ici le résidu est systématiquement positif lorsque que price dépasse un certain seuil)



30. Réalisez le graphique des résidus avec horsepower en abscisse et les résidus en ordonnée. Que constatez-vous maintenant ?

4. Prédiction

31. Pour horsepower = 100, quelle serait la valeur de price prédite par le modèle ?

    (prédiction ponctuelle = 11543.63)

32. Calculez son intervalle de prédiction au niveau de confiance 90% [TUTO 4, pages 30 à 34]

    (intervalle de prédiction = [7168.655, 15918.6])

33. Reconstruire le graphique (horsepower, price). Tracez la droite de régression. Placez le point à prédire avec la valeur prédite (en rouge), et matérialisez l'intervalle de prédiction (en bleu).