Université de Pau et des Pays de l'Adour **Année 2019-2020**
Département de Mathématiques
M1-Génie Pétrolier - Analyse de données

TP-3 : Corrélation et régréssion linéaire

Le troisième TP est dédié à la corrélation et à la régression linéaire.

Dans les travaux pratiques précédents, nous avons vu comment mettre un titre et des noms aux axes lorsque vous effectuez un graphique. À partir de cette séance, il sera impératif de mettre un titre et des noms aux axes pour chaque graphique, même si cela n'est pas mentionné dans la question.

1 La corrélation

1.1 Rappel

On rappelle que le coefficient de corrélation linéaire de Pearson est donné par

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Ce coefficient varie entre -1 et $+1$. Il est nul quand les variables sont indépendantes, négatif quand les variables sont corrélées négativement et positif quand les variables sont corrélées positivement.

1.2 Un premier exemple

1. Nous avons 5 nuages de points :

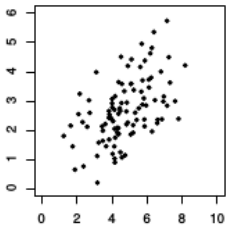

Figure 1

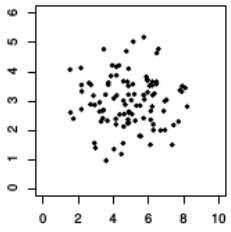

Figure 2

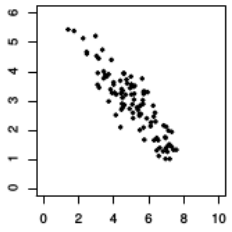

Figure 3

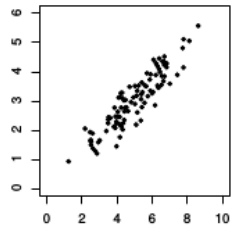

Figure 4

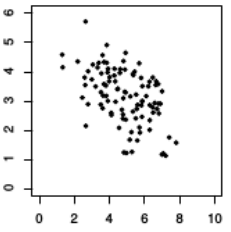

Figure 5

et les 5 coefficients de corrélation linéaire suivants :

$$a) -0.90 \qquad b) -0.49 \qquad c) 0.00 \qquad d) 0.52 \qquad e) 0.91.$$

Associer à chaque figure son coefficient de corrélation linéaire.

1.3 Taille des parents et des enfants

Examinons les données de Francis Galton (1822-1911), un des pionniers de l'analyse des corrélations, sur la relation entre la taille (en pouces) de 928 enfants et la taille de leurs parents (en pouces). Dans le jeu de données, la première colonne contient la taille moyenne des parents, dit mid-parent, la seconde colonne celle des enfants.

Vous trouverez ces données sur `Elearn` dans le cours *Analyse de données - M1 Génie Pétrolier*.

Ces données sont dans un fichier Excel, que `R` ne sait pas lire directement. Il faut les convertir à un format lisible. Pour cela :

- Sauvegardez le fichier Excel dans votre répertoire, et ouvrez-le avec OpenOffice.
- Dans "Enregistrer sous", choisissez dans la rubrique "Filtre" l'option "Texte CSV"
- OpenOffice vous demande comment représenter les changements de colonne, avec l'option "Séparateur de champ". Choisissez "Tabulation".
- Sauvegardez. Vous avez un fichier avec l'extension .csv qui est un fichier au format texte, donc lisible par `R` ainsi que par n'importe quel autre ordinateur ou logiciel d'analyse.

Pour importer ce fichier dans `R`, une petite précision : vous avez vu que les nombres comportent des virgules, et pas des points. Il faut le préeciser à `R` lors de la lecture avec l'option `dec`, pour qu'il reconnaisse des nombres ; la commande de lecture donne donc :

```
> taille = read.table("taimdc.csv",
       header=TRUE,
       dec=",")
```

Selon la façon dont vous avez enregistré vos données, vous devrez peut-être utilisiler l'option `sep` pour préciser le séparateur.

Si vous êtes sur votre ordinateur personnel, vous devrez certainement préciser dans quel dossier se trouve votre fichier.

Notez comme `R` a converti les nombres à virgule en nombres avec un point, à l'anglaise.

1. Autre question nationale, les tailles sont en pouces, ce n'est pas très lisible pour nous français, aussi convertissez-les en centimètres (1 pouce vaut 2.54 cm).
2. Tracer le nuage de points. Quel problème remarque-t-on ?
3. Pour palier à ce problème, on va utiliser une carte de densité. Utiliser la fonction `kde2d` pour obtenir une carte de densité et la fonction `image` pour la représenter. Notons qu'il faut inclure la librairie `MASS` pour pouvoir utiliser la fonction `kde2d`.
4. Avec la formule donnée en cours, calculer le coefficient de corrélation linéaire (de Pearson) entre la taille du mid-parent et celle des enfants.
5. À l'aide de la fonction `cor` calculer le coefficient de corrélation linéaire (de Pearson) entre la taille du mid-parent et celle des enfants.
6. D'après vous, existe-t-il une corrélation linéaire entre ces deux variables ?
7. Le coefficient de corrélation de Pearson nous renseigne sur l'existence d'une corrélation linéaire. Quel coefficient peut-on calculer afin d'obtenir une information sur la corrélation (linéaire ou non linéaire) ?

2 Régression linéaire

Nous allons maintenant chercher à expliquer la taille des enfants par celles des parents via un modèle linéaire. Nous allons donc effectuer une régression linéaire.

1. À quoi correspond la droite de régression linéaire de la taille des enfants sur la taille des parents ?
2. Avec les formules du cours, calculer les coefficients de cette droite. Quelle est l'équation de cette droite ?
3. Retrouver ce résultat en utilisant la fonction `lm` (pour *linear model*) pour effectuer la régression linéaire de y sur x.
4. En utilisant la fonction `abline`, ajouter la droite de régression linéaire sur le graphique précédent.
5. Ajouter sur le graphique la droite d'équation $y = x$ (les enfants font la même taille que leurs parents). Conclure.