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

TP-4 : Corrélation et régréssion (suite)

Le quatrième TP est dédié à la corrélation et à la régression linéaire et non 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 Condition d'application de la régression

Un point est à garder en mémoire quand on effectue une régression linéaire : pour qu'une droite soit un "bon" modèle pour la relation entre y et x, il faut que les points sur le graphe soient raisonnablement alignés. Si on calcule le coeffcient de corrélation entre x et $y = e^x$, il sera positif, significativement différent de 0, mais une droite ne sera pas un bon modèle pour la relation observée ! Un des exemples les plus connus de présentation de la relation entre un nuage de points et un coefficient de corrélation concerne les données de `Anscombe`. Le data.frame est pré-existant dans `R`, contient 8 colonnes, à l'abscisse $x1$ correspond l'ordonnée $y1$ et ainsi de suite :

```
> data(anscombe)
> names(anscombe)
```

1. Tracer les quatres nuages de points.
2. Calculer les quatre coefficients de corrélation linéaire.
3. Comparez la variabilité de ces représentations graphiques en relation avec la proximité de leurs coefficients de corrélation linéaire. Que pouvez-vous conclure ? Identifiez ce qui, dans chaque graphique, pourrait violer les conditions d'application de l'analyse par régression linéaire.

2 Quelques exemples et exercices

2.1 À propos de la sécurité routière

Prenons la relation entre la vitesse des voitures (en miles par heure) et la distance de freinage avant l'arrêt du véhicule (en pieds). Les données ont été collectées en 1920 mais restent d'actualité.

1. À l'aide de la fonction `data()`, chargez le jeu de données `cars`, puis convertissez les vitesses en km/h ($1\,miles/h = 1.609344\,km/h$) et les distances en mètres ($1\,p = 0.3048\,m$).
2. Tracer le nuage de points représentant la distance de freinage en fonction de la vitesse.
3. Semble-t-il y avoir une corrélation linéaire entre ces deux variables ?

4. Effectuer la régression linéaire. Quelle est l'équation de la droite de régression ?

5. Ajouter la droite de régression sur le nuage de points.

6. On souhaite maintenant tester un modèle non linéaire selon lequel le distance de freinage dépendrait du carré de la vitesse. On pose alors $t = cars\$vitesse * cars\$vitesse$. Calculer le coefficient de corrélation linéaire entre ces deux variable. Effectuer la régression linéaire de la distance de freinage sur cette nouvelle variable. Quelle est l'équation de la courbe obtenue. Ajouter cette courbe sur le graphique précédent en utilisant la commande `lines`. Conclure.

2.2 Tension artérielle et fumeurs

Dans une population, on a tiré au sort 34 sujets (17 fumeurs et 17 non fumeurs) à qui on a mesuré la tension artérielle (en mmHg) et demandé l'âge (en années). Les résultats sont dans le tableau disponible sur `Elearn` et nommé `Donnees_tension_fumeurs.xls`.

1. Faites ce qu'il faut pour charger ces données dans une variable du nom de `epidemio`.

2. Construire le nuage de points en posant en abscisse l'âge et en ordonnée la tension artérielle. Superposer le modèle linéaire correspondant.

3. Calculer le coefficient de corrélation linéaire liant ces deux variables. Conclure.

4. L'information "fumeur ou non fumeur" n'a pas été introduite. Coloriez les points du nuage par cette information en utilisant les instructions suivantes :

```
> plot(x=epidemio$Age[epidemio$Fumeur==0],
     y=epidemio$Tension[epidemio$Fumeur==0],
     xlim=range(epidemio$Age),
     ylim=range(epidemio$Tension),
     xlab="Age",
     ylab="Tension",
     pch=1)
> points(x=epidemio$Age[epidemio$Fumeur==1],
     y=epidemio$Tension[epidemio$Fumeur==1],
     pch=16)
> legend("topleft",
       legend=c("Non fumeur", "Fumeur"),
       pch=c(1,16))
```

   Les points noirs représentent les fumeurs, les blancs les non fumeurs. Sur la base de ce graphique, que peut-on conclure ?

5. Reprendre le graphique précédent et y ajouter les droites de régression séparément pour les fumeurs et non fumeurs.

6. Comment interpréter les pentes et les ordonnées à l'origine de ces deux droites de régression ici ? Biologiquement, que pouvez-vous conclure à partir du graphe précédent ?

2.3 Piraterie et réchauffement climatique

Un des dogmes de la religion (parodique) pastafariste est que le réchauffement climatique est une conséquence directe de la diminution du nombre de pirates. Cette assertion est prouvée par les données disponibles sur `Elearn` dans une fichier nommé `Pirates_temperature.xls`.

1. Charger ces données dans `R` dans une variable du nom de `pirates`.
2. Tracer le nuage de points.
3. Calculer le coefficient de corrélation linéaire entre le nombre de pirates et la température mondiale moyenne.
4. Ajouter le modèle linéaire sur le graphique. Quelle est l'équation de la droite de régression ?
5. Discutez de cette affirmation d'un point de vue scientifique.

2.4 Conclusion

L'existence d'une corrélation élevée entre deux variables x et y ne conduit pas à l'existence d'une relation cause - effet. On utilise la connaissance de x pour prédire des valeurs de y. Cela n'implique pas qu'un changement de x cause un changement de y. Considérons un exemple classique du genre. Dans "Une logique de la communication", Paul Watzlawick `http://www.evoweb.net/stat.htm` raconte que la plus forte corrélation trouvée dans les années 1950 a été celle entre la consommation de bière sur la côte ouest des USA, et la mortalité infantile au Japon. Cet exemple a été fréquemment repris pour montrer les limites des statistiques et démontrer "qu'on peut leur faire dire n'importe quoi". Et en effet beaucoup feront remarquer qu'on ne peut accuser les Américains assoiffés de tuer les Japonais (on remarquera d'ailleurs que personne n'accuse les enfants Japonais d'assoiffer les Américains). Ici les causes de la variation commune de ces deux variables sont à chercher dans le contexte historique de l'après-guerre...