Econometrie Appliquée (M1-S2)

TD1 - Introduction à RStudio et Régression Linéaire Simple et Multiple

Jean-Baptiste Guiffard

2025-02-04

Objectifs :

- Introduction à R ;
- Estimation d'un modèle de régression linéaire simple ;
- Interprétation des coefficients.

Avant tout...

Installation de R sur CRAN : https://cran.r-project.org/

Installation de RStudio : https://posit.co/download/rstudio-desktop/

Exercice 1

Voici une dataframe qui associe deux variables: - La taille du père - La taille de son fils aîné

```
data1 <- data.frame(height_father = c(175, 173, 177, 174, 178, 172, 180, 176, 178, 177, 179, 181),
                    height_son = c(178, 176, 178, 175, 179, 176, 178, 175, 181, 177, 178, 180))
```

1. Calculer avec R, et en utilisant les formules vues en cours :

- La droite des moindres carrés de la taille des fils en fonction de la taille des pères.
- Représenter les données et la droite (commande plot, et abline).

2. Calculer avec R, et en utilisant les formules vues en cours :

- la droite des moindres carrés de la taille des pères en fonction de la taile des fils
- Représenter les données et la droite (commande plot, et abline).

3. Montrer par le calcul que le produit des pentes des deux droites est égal au carré du coefficient de corrélation empirique entre les deux variables (coefficient de détermination). Le vérifier avec R sur les données.

$$r = \frac{Cov(X,Y)}{\sigma(X) * \sigma(Y)}$$

4. Vérifier vos calculs "à la main avec R" avec la commande lm

Exercice 2 :

En utilisant la librarie haven, chargé la base de données "school_perf.dta" dans l'environnement RStudio, pour rappel pour télécharger une librairie on utilise la commande install.packages(' ').

Pour charger la base de données :

```r
data_school <- read_dta('school_perf.dta')
```

La base de données est un échantillon de N = 10 854 élèves entrés en classe de 6e en France métropolitaine en septembre 1989 et à qui ont été administrées, à ce moment, des épreuves standardisées. On dispose ainsi de leur score de performance en français, de leur score de performance en mathématiques et de leur score moyen. On a observé par ailleurs le nombre total d'enfants dans la famille de l'élève. En plus, la base inclut le diplôme le plus élevé de chacun des parents (variables A52M et A52P). Pour ces deux dernières variables, les modalités précises sont : aucun diplôme (1), certificat d'études primaires (2), BEPC (3), CAP (4), BEP (5), Baccalauréat (6), DEUG, BTS ou DUT (7), second ou troisième cycle, grande école (8).

On s'interroge sur la liaison statistique entre le score de performance en français et différentes variables explicatives.

1. Décrire la base de données (commande ncol, nrow, colnames, str. . .)

2. Créez les variables quantitatives MOTYEAR et FATYEAR en associant au diplôme le plus élevé de chacun des parents un nombre théorique d'années d'études ; dans l'ordre des diplômes ci-dessus, il s'agit de : 5 ans, 6 ans, 9 ans, 10 ans, 11 ans, 12 ans, 14 ans et 16 ans.

Utiliser la commande as.factor puis la commande levels pour recoder la variable créée. N'oubliez aussi de re-transformer en format numeric la variable obtenue.

2. Analyse du biais

a. Déterminez la liaison statistique existant entre :

- le nombre d'enfants dans la famille et le score de performance en français
- le nombre d'années d'études du père ou de la mère et le score de performance en français
- le nombre d'enfants dans la famille et le nombre d'années d'études du père ou de la mère

b. Au vu de ces différentes liaisons, pensez-vous que la régression linéaire simple de MOYF sur NBENF fournira une estimation correcte de l'effet de la taille de la famille sur la performance scolaire ? Justifiez votre réponse.

3. Estimation d'un modèle de régression linéaire simple

a. Ecrivez et estimez le modèle de régression linéaire simple de MOYF sur NBENF.

b. Commentez la sortie R obtenue.

Exercice 3 : Regression lineaire Multiple et effets d'interaction

On s'interroge sur la liaison statistique entre le score de performance en français et différentes variables explicatives, et l'interaction entre certaines de ces variables explicatives.

1 - Ecrivez et estimez le modèle de régression linéaire multiple de MOYF sur NBENF, MOTYEAR, FATYEAR et le sexe de l'élève. Commentez la sortie R obtenue.

2- Testez l'hypothèse d'un éventuel effet « de saturation » de l'éducation de la mère étant donné le niveau d'éducation du père (l'impact d'une année d'études supplémentaire de la mère sur le score de performance en français diffère selon l'éducation du père).

3- Testez l'hypothèse d'un éventuel effet « sexué » de l'éducation de la mère (l'impact d'une année d'études supplémentaire de la mère sur le score de performance en français diffère selon le sexe de l'enfant). variable sexe == 1 (si garçon), ==2 (si fille).