

TP 1 économétrie - Licence 3 EM + Magistère

Thomas Chuffart

21 octobre 2013

1 Introduction

L'objectif de ce premier TP est de se familiariser avec les logiciels libres d'économétrie. Pour cela, il y a deux options envisageables :

- R project avec si vous le souhaitez une interface graphique Rcmdr
- GRETl

Le premier est intéressant car il permet d'agir en ligne de commande et via des scripts modifiables facilement. Si l'aspect programmation vous effraie, n'hésitez pas à installer Rcmdr, une interface graphique puissante qui prend en charge tout ce que l'on fera durant les séances machines. Rcmdr a aussi l'avantage de traduire les interactions de l'utilisateur en ligne de commandes permettant une double utilisation.

GRETl est quant à lui un user friendly freeware. Très utile pour celui qui ne souhaite pas s'investir dans un peu de programmation (pourtant indispensable sur le marché du travail). Il possède aussi un système de script mais qui moins puissant que celui de R. Passons aux choses sérieuses.

Pour illustrer ces logiciels, nous utiliserons une enquêtes américaine de 1985 `cps86.txt` composée de 534 observations. Nous étudierons, à l'aide d'un modèle de régression linéaire simple l'impact de l'éducation sur le salaire.

2 Importation et présentation d'un jeu de données

Hormis dans un cadre de simulation où l'économètre va simuler lui même ses propres données, la première étapes d'une étude sera toujours d'importer une base de données. Toutes les données dont on se servira en sécantes machine sont sur www.thomaschuffart.fr dans la rubrique teaching.

Via R cela dépend du jeu de données mais en général on utilise la commande :

```
Dataset <- read.table("nomdufichier.csv", header = TRUE, sep = ",", na.strings = "NA", dec = ".")
```

Vous devez avoir placé votre base de données dans le répertoire courant de R ou renseigner l'adresse complète sinon. Les fichiers csv (comma separated values) peuvent avoir des éléments séparés par des ',' des ';' ou juste un espace. L'option `sep=""` spécifie cette séparation. L'option `dec=""` renseigne le séparateur des décimales ici, un point. `Dataset` sera le nom de notre jeu de données. Pour avoir un résumé statistique de `Dataset`, il suffit de taper la commande :

```
summary(Dataset)
```

Vous pouvez aussi passer par l'interface graphique de Rcmdr dans l'onglet statistiques -> Résumés -> jeu de données actif. Il est aussi possible de sortir des graphiques telle que la densité estimée.

Question 1. Via l'interface graphique ou, en cherchant la commande adéquate, grapher la densité estimée des salaires horaires ainsi qu'un histogramme de la variable du niveau d'éducation. Testez les lignes de commande suivantes en modifiant les options (hint : `help(density)` `help(plot)` `help(hist)` .

```
d <- density( variable )
```

`plot(d)`

Ou un histogramme

`hist(Dataset$wage, breaks = "Sturges", col = "darkgray")`

Dans cette dernière commande vous remarquerez le signe \$ entre Dataset et la variable dont on souhaite grapher l'histogramme. Sous R pour appeler une variable on utilise cette méthode.

Question 2. Quel est le niveau moyen d'étude de l'échantillon ?

Pour créer une nouvelle variable à partir d'une variable existante, il suffit de rentrer la commande suivante :

`newvar <- Dataset$var[cond1]`

cond1 permet de mettre une condition sur la nouvelle variable par exemple, si je ne souhaite regarder que le salaire des femmes, j'effectuerai la commande suivante :

`wagef <- Dataset$wage[Dataset$FE == 1]`

Question 3 Calculez, pour chaque niveau d'éducation (ie de 6 à 18) le salaire moyen. Tracez la figure qui représente les moyennes des salaires horaires en fonction du niveau d'éducation. Pouvez vous dégager une tendance générale de ce graphique ? Combien de commandes avez vous lancé pour obtenir ces résultats ? Comment réduire ce nombre ? (hint : boucle ?)

3 Modélisation

La régression linéaire avec R, c'est possible, pour cela je vous invite à rentrer la commande

`help(lm)`

Question 4. Nous désirons analyser l'impact de l'éducation sur le salaire horaire. Quelle étude proposeriez vous ? Lancez la régression et commentez.

Question 4 bis. Retrouvez les résultats de l'estimation par calcul en utilisant la formule du TD précédent.

La commande lm retourne un objet. Dans un objet, il y a plein de choses. Récupérez les informations utiles à votre régression. Aidez vous des résultats obtenus via Rcmdr et l'interface graphique.

Question 5. Transformez votre variable dépendante en prenant le log de celle ci et relancez votre régression, quelles remarques pouvez vous faire ?

Vous pouvez récupérer les résidus de la régression. Je vous invite une nouvelle fois à refaire la commande `help(lm)`.

Question 6. Récupérez les résidus de la régression, analysez les. Que remarquez vous ?

Question 7. Effectuez un test de Student sur votre paramètre d'intérêt. Que concluez vous ? Les informations dont vous disposez sont-elles suffisantes pour conclure qu'une année d'étude supplémentaire, lorsqu'on a déjà fait 16 ans d'études, permet d'espérer un gain de salaire significatif ?

Question 7 bis. Calculez la statistique de test à l'aide des formules du TD précédent.

Question 8. Refaites ce TD sous GRETL!! (10 minutes top chrono)