G. Hunault
Licence MEF

Angers, mai 2011

Examen de Logiciels Statistiques

On s'intéresse ici au dossier EAEF01 qui contient un extrait des données du recensement américain. On trouvera ces données et leur descriptif à l'endroit habituel (`k:\stat_ad\`) mais aussi sur le Web à l'adresse

`http://forge.info.univ-angers.fr/~gh/Datasets/eaef01.dar`

1. Analyse de variables QT

Effectuez l'étude statistique *univariée* des variables quantitatives AGE et EARN (salaire horaire). Vous fournirez un tableau résumé **court** de cette analyse et vous rédigerez ce qu'on peut en conclure. On ne demande aucun graphique.

2. Analyse de variables QL

Etudiez maintenant séparément puis conjointement les variables ETH (ethnie) et MARI (statut marital) issues de ces mêmes données. On ne demande, là encore, aucun graphique. Que peut-on en conclure ?

Vous fournirez, parmi tous les tableaux résumés possibles, celui qui est à la fois le plus concis et le plus explicite possible.

Vous effectuerez le test adéquat pour tester l'indépendance entre les modalités des deux variables et vous détaillerez, si besoin est, les modalités les plus liées. Vous n'oublierez pas de commenter "scientifiquement" les résultats.

3. Une analyse de variance

On voudrait maintenant savoir s'il y a une différence au niveau des salaires horaires entre les hommes et les femmes. Réalisez le ou les tests correspondants. Vous fournirez l'interprétation statistique traditionnelle (p-value, hypothèse nulle...) et l'interprétation métier ("différence [*non*] significative..."). On ne demande bien sûr aucun graphique.

4. Une régression linéaire

Enfin, on veut savoir s'il y a une relation linéaire entre le nombre d'années d'études et le salaire horaire. Vous réaliserez la régression linéaire correspondante en prenant bien soin du sens de la relation. Vous discuterez d'une causalité éventuelle et de la présence d'une relation d'un type autre que linéaire. On ne demande aucun graphique.

5. Discussion

Essayer de défendre en au moins une dizaine de lignes le point suivant :

> *L'une des grandes qualités du logiciel R est sa capacité à produire des graphiques de grande qualité, soignés et très complets.*

Pour faire « *évolué(e)* », on utilisera au moins 3 mots de 4 syllabes ou plus. Vous fournirez plusieurs exemples de codes R concrets pour justifier le point de vue proposé.

ANNEXE 1 : DESCRIPTIF DES DONNEES *EAEF01*

Les données initiales sont issues du site

`http://econ.lse.ac.uk/courses/ec220/G/ec212.html`

Le dossier EAEF01 correspond à un échantillon de 540 personnes, prises au hasard, avec 9 colonnes de données dont 8 variables statistiques.

Variable	Description
IDEN	Cette colonne de données est un identificateur de la personne interrogée.
SEX	Sexe de la personne interrogée : 2=Homme ; 1=Femme.
AGE	Age de la personne interrogée (en années).
ETHNIE	Ethnie de la personne interrogée : 3=blanc ; 2=hispanique ; 1=noir.
SCHOOL	Nombre d'années de scolarité de la personne interrogée.
MSCHOOL	Nombre d'années de scolarité de la mère.
FSCHOOL	Nombre d'années de scolarité du père.
EARN	Salaire de la personne interrogée ; Il est exprimé en dollars par heure.
MARI	Situation maritale de la personne interrogée : 1=célibataire, 2=marié, 3=divorcé.

ANNEXE 2 : EXTRAIT DES DONNEES *EAEF01*

IDEN	SEX	AGE	ETH	SCHOOL	MSCHOOL	FSCHOOL	EARN	MARI
I000006	2	41	3	16	12	12	27.11	2
I000018	2	44	3	13	12	16	113.96	3
I000021	1	40	3	16	12	18	38.46	2
I000039	2	43	3	12	12	12	28.84	2
I000049	1	44	3	14	12	12	19.23	2
I000073	2	38	3	18	16	16	57.69	1
I000085	2	43	1	11	11	11	8.75	1
I000087	2	39	1	12	11	12	8.50	1
...								
I012096	1	38	3	15	8	11	38.46	3
I012099	1	37	3	18	14	18	11.00	2
I012100	2	40	3	12	8	16	16.62	2
I012104	1	45	3	15	13	16	16.32	2
I012106	2	38	3	18	13	16	31.25	1
I012122	2	45	3	8	6	10	24.73	2

ELEMENTS DE SOLUTION

Après la lecture des données et le transfert dans le *dataframe* nommé ici `ea` par la commande

```
ea <- lit.dar("k:/Stat_ad/eaef01.dar")
```

ou la commande

```
ea <- lit.dar("http://forge.info.univ-angers.fr/~gh/Datasets/eaef01.dar")
```

les commandes suivantes fournissent les résultats numériques demandés :

```
# Question 1

decritQT("AGE",ea[,"age"],"ans",TRUE)
decritQT("SALAIRE HORAIRE",ea[,"earn"],"$",TRUE)

# Question 2 (attention à l'ordre des modalités)

decritQL("ETHNIE",ea[,"eth"],"noir hispanique blanc",TRUE)
decritQL("STATUT MARITAL",ea[,"mari"],"célibataire marié divorcé",TRUE)

triCroise("ETHNIE",ea[,"eth"],"noir hispanique blanc",
          "STATUT MARITAL",ea[,"mari"],"célibataire marié divorcé",TRUE)

# Question 3

decritQTparFacteur("SALAIRE HORAIRE",ea[,"earn"],"$",
                   "SEXE",ea[,"sex"],"femme homme")

# Question 4

anaLin("SALAIRE HORAIRE",ea[,"earn"],"$",
       "ANNEES D'ETUDE",ea[,"school"],"ans")
```