TP R Introduction à la statistique

Mohamed FOFANA | *L1 MIASHS - Université Grenoble Alpes*

Contents

1 Séance de TP3: Régressions linéaires **1**
1.1 **Dépendance non linéaire** . . . 1
1.2 **Saisie des séries x et y dans R** . . . 1
1.3 **Tableau de contingences et test du χ^2** . . . 2
1.4 **Régression linéaire simple** . . . 3
1.5 **Régression linéaire multiple** . . . 7

1 Séance de TP3: Régressions linéaires

1.1 Dépendance non linéaire

On considère les données suivantes :

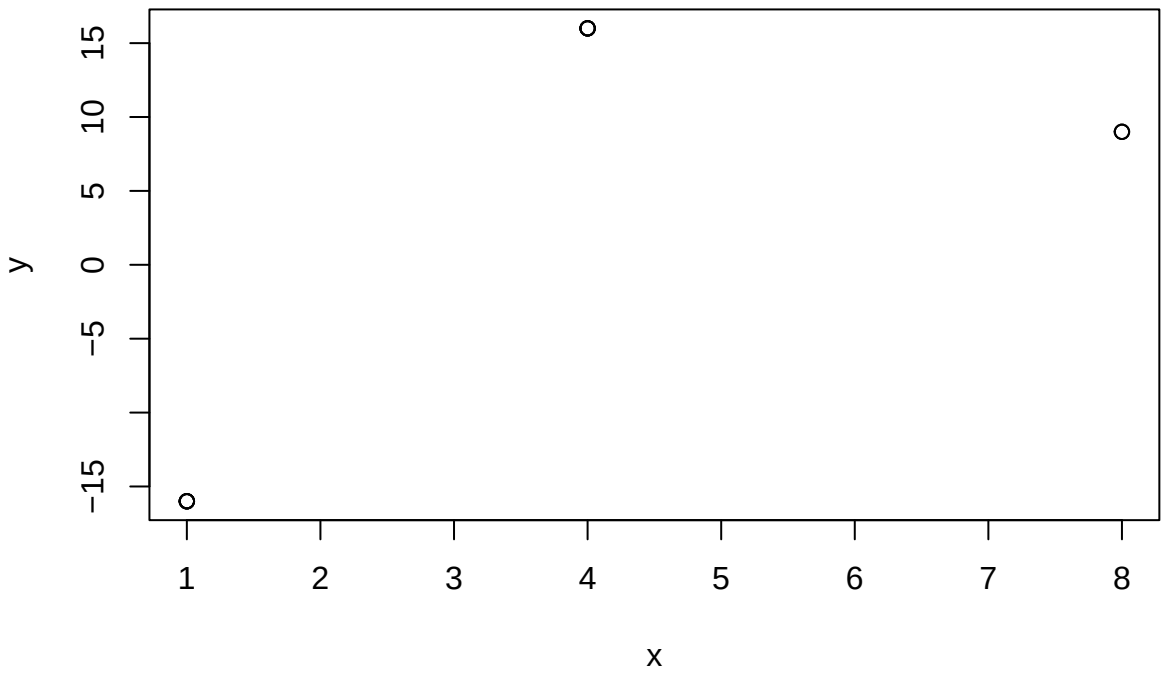
x_i	1	4	4	4	8	4	1	1	1	8
y_i	-16	16	16	16	9	16	-16	-16	-16	9

1.2 Saisie des séries x et y dans R

```
x=c(1,4,4,4,8,4,1,1,1,8)
y=c(-16,16,16,16,9,16,-16,-16,-16,9)
```

Tracer y en fonction de x,

```
plot(x,y)
```



Calcul du coefficient de corrélation

```
cor(x,y)
```

```
## [1] 0.7132086
```

1.3 Tableau de contingences et test du χ^2

```
contingence=table(x,y)
```

Calculer la statistique du Chi-2. Commenter.

On peut calculer la statistique du χ^2 avec R en utilisant `chisq.test(contingence)`*. Néanmoins, R nous informe qu'il ne peut pas calculer une p-valeur fiable à cause d'effectifs trop petits. On va chercher la valeur du quantile 0.95 du χ^2 :*

```
qchisq(0.95,df=4)
```

```
## [1] 9.487729
```

```
**Réponse** : On en déduit que les variables ne sont pas ind´ependantes
```

(au seuil 5%) car $D_{\chi^2} = 20$ est supérieure au quantile 0.95 du χ^2 qui vaut **9,49**, mais que leur dépendance n'est pas linéaire.

Remarque: On note que la p-value donnée par le test du χ^2 est très petite : *p-value = 0.0004994*. Inférieure à 5%, elle confirme la dépendance des variables.

1.4 Régression linéaire simple

On va utiliser le fichier GPA.csv qui contient des notes d'études secondaires et universitaires pour les diplômés en informatiques dans une école publique locale. Notre objectif est de déterminer la droite des moindres carrées permettant d'expliquer linéairement la note universitaire d'un étudiant par sa note du secondaire.

1.4.1 Acquisition des données

Quelle est la variable explicative (indépendante) et la variable à expliquer (dépendante) ?

`Réponse : Variable explicative : note secondaire, variable expliquée : note universitaire.`

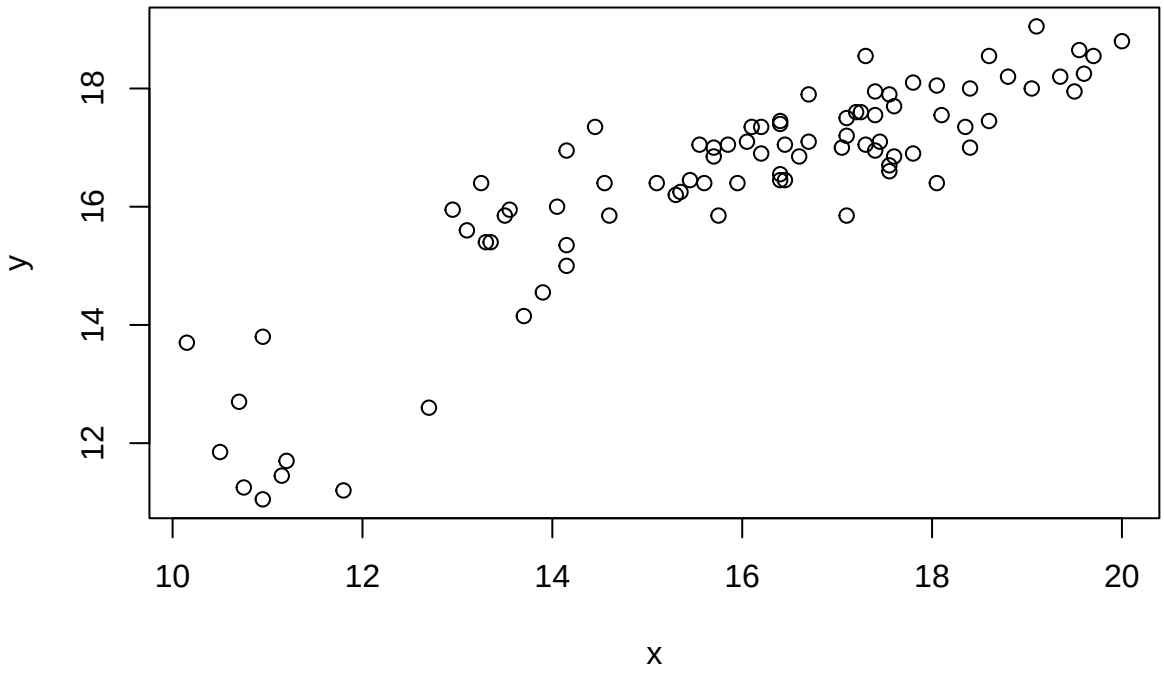
Stockons les variable. Nous pouvons voir que le format du fichier .csv suit les standards francophones (header, ";" pour séparer les colonnes, "," comme séparateur d´ ecimal). Pour éviter le détail de compléter tous les paramètres dans la fonction read.csv(), on peut utiliser la fonction `read.csv2()`.

```
data=read.csv2("C:/Users/DVE ICAMPUS/Desktop/MIASHS-UGA/STAT L1/Donnees/GPA.csv")
x=data$high_GPA
y=data$univ_GPA
```

1.4.2 Nuage de points

Tracer le nuage de points et commenter.

```
plot(x,y)
```



Réponse : Le nuage est relativement rectiligne, la régression linéaire a du sens, elle donnera une tendance moyennement précise de y en fonction de x.

1.4.3 Corrélation linéaire

la fonction suivante permet de calculer le coefficient de corrélation linéaire entre x et y.

```
correlation=function(u,v){
 cov=cov(u,v)      # Covariance entre u et v
 denom=sqrt(var(u) * var(v))  #Produit des écarts-types
 corr= cov/ denom    # Coefficient de corrélation
 corr
 }
```

Appliquer la fonction à x et y:

```
 correlation(x,y)
```

```
## [1] 0.8831408
```

Vérifier que l'on obtient la même valeur avec la fonction `cor` :

```
cor(x,y)
```

```
## [1] 0.8831408
```

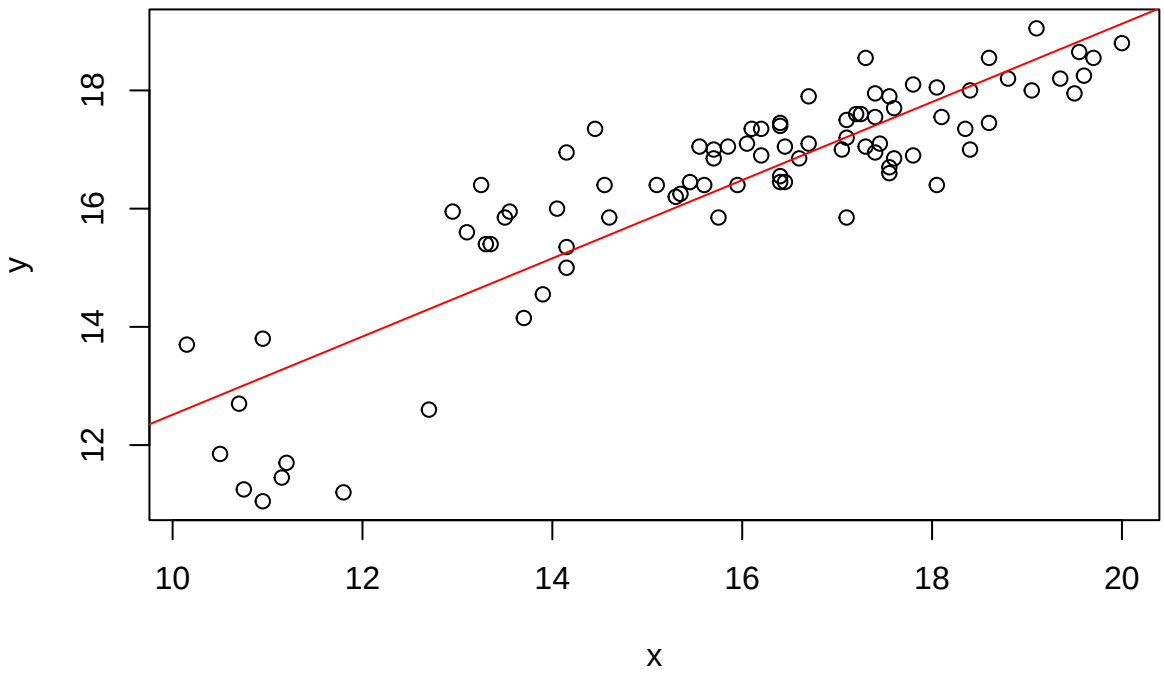
1.4.4 Régressions linéaire

Calculer les coefficients de la droite de régression $y = ax + b$ par la méthode des moindres carrés.

```
 a=cov(x,y)/var(x)
 b=mean(y)-a*mean(x)
```

On utilise la fonction `abline` pour tracer la droite de régression :

```
plot(x, y)
abline(b,a,col="red")
```



On peut calculer les valeurs ajust´ ees et les erreurs :

```
yajust=a*x+b
erreurs=y-yajust
```

La fonction `lm` permet déffectuer directement ce que nous avons fait :

```
lm(y~x)
```

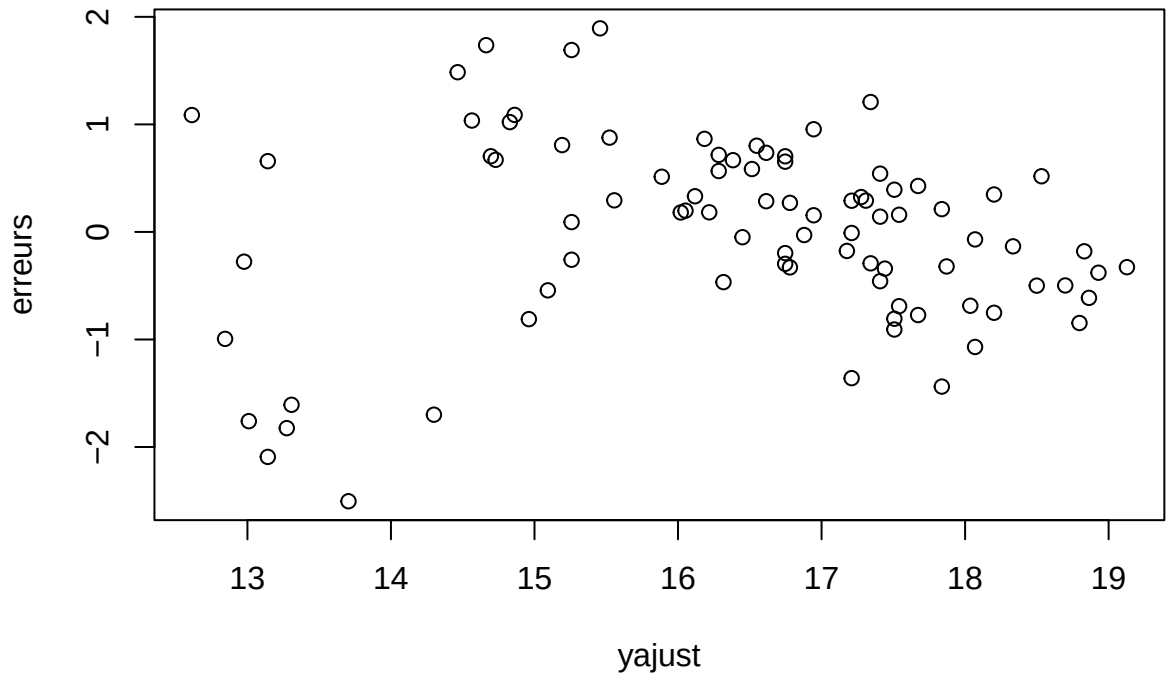
```
##
## Call:
## lm(formula = y ~ x)
##
## Coefficients:
## (Intercept)            x
##      5.8996       0.6614
```

```
model=lm(y~x)
b=model$coefficients[1]
a=model$coefficients[2]
yajustes=model$fitted.values
erreurs=model$residuals
```

1.4.5 Résidus

Représenter les résidus (erreurs) en fonction des valeurs ajustées.

```
plot(yajust,erreurs)
```



Voit-on une structure particulière des résidus ?

Réponse : Il ne semble pas y avoir de structure particulière des erreurs car dans une régression linéaire on a une hypothèse : les erreurs sont distribuées selon une loi normale N(0,σ) et sont indépendantes.

1.4.6 Prévision

Un diplomée a eu 2,5. Donner une prédiction de sa note universitaire.

```
x_new = 2.5
y_new = a * x_new + b
y_new
```

```
##        x
## 7.553131
```

Réponse : On prévoit 7,55.

1.4.7 Anova

Nous avons vu que la proportion de variance de y expliquée par la droite est assez faible. A l'aide de **l'ANOVA** décider si la variable x explique significativement la variable y.

```
anova(lm(y~x))
```

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## x          1 227.199 227.199  297.71 < 2.2e-16 ***
```

```
## Residuals 84  64.105   0.763
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Réponse : Puisque l'hypothèse nulle du test de l'ANOVA est "x n'explique pas y" et que la p−value est proche de 0, on peut conclure que x a un effet significatif sur y.

1.5 Régression linéaire multiple

Pour mesurer les performances auditives d'un individu, on le soumet à un signal sonnore d'une fréquence donnée dont l'intensié va croissante. On note alors le seuil à partir duquel le signal est perçu. Ce seuil, exprimée en décibel, mesure la performance d'audition à partir du seuil moyen. Ainsi, un seuil haut correspond à un trouble de l'audition pour la fréquence. Les données représentent les mesures obtenues pour quatre fréquences (500Hz, 1000Hz, 2000Hz et 4000Hz). On prendra le seuil maximal pour l'oreille gauche et l'oreille droite pour 1000 personnes de plus de 39 ans. On ajoutera la donnée globale qui est une auto-évaluation par le patient. Le tableau **audition2.csv** donne les valeurs des 1000 relevés auditifs. L'enjeu est de prédire globale par les variables *A5, A10, A20, A40* et de savoir si le patient peut prédire lui-même l'état de son audition.

1.5.1 Import des données

Importons les données et stockons les dans les variables du même nom.

```
audition=read.csv2("C:/Users/DVE ICAMPUS/Desktop/MIASHS-UGA/STAT L1/Donnees/audition2.csv")
A5=audition$A5
A10=audition$A10
A20=audition$A20
A40=audition$A40
globale=audition$globale
```

1.5.2 Les coefficients de la régression linéeaire

Donner l'équation de la régression lin´eaire multiple et discuter de l'effet des variables explicatives sur la variable expliquée.

```
lm(globale~A5+A10+A20+A40)
```

```
##
## Call:
## lm(formula = globale ~ A5 + A10 + A20 + A40)
##
## Coefficients:
## (Intercept)           A5          A10          A20          A40
##      4.1423       0.7082       0.8602       0.6355       0.5106
```

```
Réponse : Ainsi, on peut lier les variables par l'équation globale = 4,14 +0,7A5+0,86A10+0,64A20+0,51A4
On voit que le niveau d'audition globale évalué par le sujet augmente quand chaque seuil augmente avec
```

1.5.3 Tests statistiques

Faire les tests statistiques classiques de la régression linéaire et conclure.

```
summary(lm(globale~A5+A10+A20+A40))
```

```
##
## Call:
## lm(formula = globale ~ A5 + A10 + A20 + A40)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.63918 -0.19985  0.04163  0.27016  0.72287
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.14226    0.01274  325.10   <2e-16 ***
## A5           0.70820    0.01808   39.17   <2e-16 ***
## A10          0.86024    0.01948   44.17   <2e-16 ***
## A20          0.63553    0.01439   44.17   <2e-16 ***
## A40          0.51059    0.01333   38.30   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3613 on 995 degrees of freedom
## Multiple R-squared:  0.9653, Adjusted R-squared:  0.9651
## F-statistic:  6911 on 4 and 995 DF,  p-value: < 2.2e-16
```

Réponse :

- Le test de Fisher (ANOVA), donnée dans la dernière ligne du résumé, montre que, dans leur globalité

- Tous les tests de Student amènent à dire que les coefficients de la régression linéaire multiple sont
Chaque variable explique significativement la variable **globale**.

- De plus, le coefficient R^2 vaut 0,97. Ainsi, 97% de la variance de globale est expliquée par le mo