TP 5–Régressions linéaires

Exercice 5.1 *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

(i) Saisir les 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)
```

.

(ii) Calculer le coefficient de corrélation r_{xy} et commenter.

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

.

(iii) Faire le tableau de contingences.

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

.

(iv) 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)
```

.

Réponse :

On en déduit que les variables ne sont pas indépendantes (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 5.2 *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.*

Exercice 5.3 *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 carrés permettant d'expliquer linéairement la note universitaire d'un étudiant par sa note secondaire.*

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

Réponse :

Variable explicative : note secondaire, variable expliquée : note universitaire

(ii) Stocker les variable :

```
data=read.csv("GPA.csv",header =TRUE,sep=";",dec=",")
x=data$high_GPA
y=data$univ_GPA
```

.

(iii) 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.

(iv) Compléter la fonction suivante permettant de calculer le coefficient de corrélation linéaire entre x et y.

```
correlation= function(u,v){
cov= cov(x,y)
denom= sqrt(var(u) *......)
corr= ......./.......
return(......)
}
```

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

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

(vi) 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)
```

(vii) On utilise la fonction `abline` pour tracer la droite de rgression :

```
abline(b,a,col="red")
```

(viii) On peut calculer les valeurs ajustées et les erreurs :

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

(ix) La fonction `lm` permet d'effectuer directement ce que nous avons fait :

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

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

`plot(yajust,erreurs)` .

(xi) Voit-on une structure particulière des résidus ?

Réponse :

Il ne semble pas y avoir de structure particulière des erreurs
car dans une régression linéire on a une hypothèse : les erreurs sont distribuées
selon une loi normale $\mathcal{N}(0, \sigma)$ et sont indépendantes.

(xii) Un diplomé a eu 2,5. Donner une prédiction de sa note universitaire. On prévoit 2,8