TP2 : Analyse bivariée : Croisement Quantitatif-Quantitatif

Durée : 3h

L'objectif de ce TP est d'étudier un lien éventuel entre deux variables quantitatives et de construire un modèle décrivant ce lien le cas échéant.

Le TP est assez long. Il faut absolument faire les exo 1, 2 et 3. L'exo 4 sera peut-etre fait en autonomie.

Exercice 1 — Un peu de géométrie

Dans une population Ω de taille n, on observe deux variables quantitatives continues, $x=\{x_k\}_{k=1,\dots,n}$, et $y=\{y_k\}_{k=1,\dots,n}$, de moyennes $\bar{x}$ et $\bar{y}$ et de variances s_x^2 et s_y^2.

On définit le produit scalaire,

$$<x,y>=\frac{1}{n}\sum_{k=1}^{n}x_k y_k .$$

1) Montrez que la covariance est le produit scalaire entre les vecteurs centrés $x-\bar{x}$ et $y-\bar{y}$. Puis exprimez le produit scalaire en fonction de C_{xy} et les moyennes.

Notons les vecteurs $\tilde{x}$ et $\tilde{y}$ tels que $\tilde{x}_k=x_k-\bar{x}$ et $\tilde{y}_k=y_k-\bar{y}$, pour k=1,…,n

$$C_{xy}=\frac{1}{n}\sum_{k=1}^{n}(x_k-\bar{x})(y_k-\bar{y})=\frac{1}{n}\sum_{k=1}^{n}\tilde{x}_k\tilde{y}_k=<\tilde{x},\tilde{y}>=<x-\bar{x},y-\bar{y}>$$

On a

$$C_{xy}=\frac{1}{n}\sum_{k=1}^{n}(x_k-\bar{x})(y_k-\bar{y})=\frac{1}{n}\sum_{k=1}^{n}x_k y_k-\bar{x}\underbrace{\frac{1}{n}\sum_{k=1}^{n}y_k}_{\bar{y}}-\bar{y}\underbrace{\frac{1}{n}\sum_{k=1}^{n}x_k}_{\bar{x}}-\bar{x}\,\bar{y}$$

$$=\frac{1}{n}\sum_{k=1}^{n}x_k y_k-\bar{x}\,\bar{y}=<x,y>-\bar{x}\,\bar{y}$$

$$\Leftrightarrow \quad <x,y>=C_{xy}+\bar{x}\,\bar{y}$$

2) Déterminez la norme du vecteur centré $x-\bar{x}$ puis la norme de x en fonction de sa variance et sa moyenne.

$$||x-\bar{x}||^2=<x-\bar{x},x-\bar{x}>=\frac{1}{n}\sum_{k=1}^{n}(x_k-\bar{x})^2=s_x^2$$

$$s_x^2=\frac{1}{n}\sum_{k=1}^{n}x_k^2-\bar{x}=<x,x>-\bar{x} \Leftrightarrow \|x\|^2=s_x^2+\bar{x}$$

3) Comment peut-on écrire la moyenne $\overline{x}$ à l'aide du produit scalaire ?

$$\overline{x} = \frac{1}{n}\sum_{k=1}^{n} x_k = < x,1 >$$

4) D'un point de vue géométrique à quoi correspond le coefficient de corrélation linéaire ?

$$r_{xy} = \frac{C_{xy}}{s_x s_y} = \frac{< x-\overline{x}, y-\overline{y} >}{\|x-\overline{x}\|\,\|y-\overline{y}\|} = \cos(x-\overline{x}, y-\overline{y})$$

5) On dit que les deux variables x et y sont non corrélées si $r_{xy}=0$ et entièrement corrélées si $r_{xy}=\pm1$. Qu'est-ce que cela signifie géométriquement ?

D'après ce qui précède,
si $r_{xy}=0$ alors il y a un angle droit entre les variables donc elles sont orthogonales
si $r_{xy}=\pm1$ alors il y a un angle plat ou nul entre les variables donc elles sont liées par une équation de droite.

6) A l'aide du produit scalaire, montrez que :

$$\overline{\hat{y}} = < \hat{y},1 > = < \hat{a}x+\hat{b},1 > = \hat{a}< x,1 > + \hat{b} = \hat{a}\overline{x}+\hat{b} = \hat{a}\overline{x}+(\overline{y}-\hat{a}\overline{x}) = \overline{y}$$

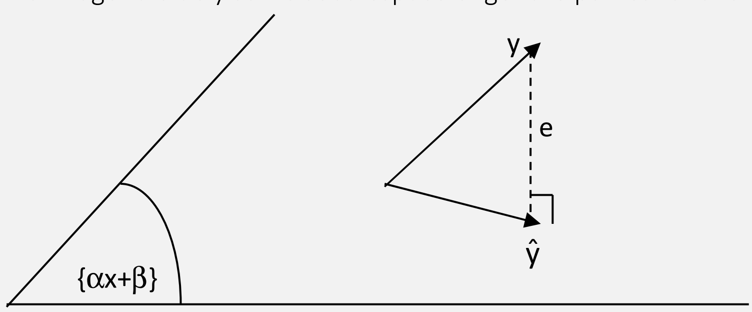
a) les résidus sont de moyenne nulle,

Immédiat d'après question a)

b) les résidus sont non corrélés avec la série observée x.

$$<e,x>=<y-\hat{y},x>=<y-\hat{a}x-\hat{b},x>=<y,x>-\hat{a}<x,x>-\hat{b}<1,x>$$
$$=\left(c_{xy}+\overline{x}\,\overline{y}\right)-\hat{a}\left(s_x^2+\overline{x}^2\right)-\hat{b}\overline{x}=\left(c_{xy}+\overline{x}\,\overline{y}\right)-\hat{a}\left(s_x^2+\overline{x}^2\right)-\left(\overline{y}-\hat{a}\overline{x}\right)\overline{x}$$
$$=c_{xy}-\hat{a}s_x^2=c_{xy}-\frac{c_{xy}}{s_x^2}s_x^2=0$$

En fait $\hat{y}$ est la projection orthogonale de y sur le sous-espace engendré par les fonctions affines de x.



Cela signifie qu'il ne reste plus « d'information » pour expliquer y par x dans les résidus.

Exercice 2 Chômage en 1982

Exercice pour mettre en place les formules. Pas de tableur. Très rapide à faire

On donne pour les six premiers mois de l'année 1982 les nombres d'offres d'emploi (concernant des emplois durables à temps plein) et de demandes d'emploi (déposées par des personnes sans emploi, immédiatement disponibles, à la recherche d'un emploi durable à plein temps). Les nombres sont exprimés en milliers.

Offres (x_i)	61	66,7	75,8	78,6	82,8	87,2
Demandes (y_i)	2034	2003,8	1964,5	1928,2	1885,3	1867,1

On a les résultats suivants

$$\bar{x} = 75{,}35 \qquad \bar{y} = 1947{,}15 \qquad s_x^2 = 97{,}49 \qquad s_y^2 = 4329{,}14 \qquad c_{xy} = -639{,}90$$

1) Calculer le coefficient de corrélation linéaire. Conclusion.

$$r_{xy} = \frac{c_{xy}}{s_x s_y} = \frac{-639{,}90}{\sqrt{97{,}49 \times 4329{,}14}} = -0{,}98$$

→ $|r_{xy}| \cong 1$ il y a donc une relation sous forme de droite entre l'offre et la demande.
→ $r_{xy} < 0$ donc plus l'offre augmente et plus la demande diminue.

2) Déterminer la droite de régression.

$$\hat{a} = \frac{c_{xy}}{s_x^2} = \frac{-639{,}90}{97{,}49} = -6{,}56 \qquad \hat{b} = \bar{y} - \hat{a}\bar{x} = 1947{,}15 - (-6{,}56) \times 75{,}35 = 2441{,}73$$

$$y = -6{,}56\,x + 2441{,}73$$

3) Calculer la prévision de la demande d'emploi s'il y a 61 milliers d'offres. Comparer avec la demande réelle.

$$y = -6{,}56 \times 61 + 2441{,}73 = 2041{,}34$$

4) Vérifier la formule de la décomposition de la variance. En déduire le coefficient de détermination.

$$s_{\hat{y}}^2 = r_{xy}^2 s_y^2 = (-0.98)^2 * 4329.14 = 4200.14$$

$$s_e^2 = s_y^2(1 - r_{xy}^2) = (1 - 0.98^2) * 4329.14 = 129$$

On a bien

$s_{\hat{y}}^2 + s_e^2 = 4200.14 + 129 = 4329.14 = s_y^2$

Le coefficient de détermination est

$R^2 = s_{\hat{y}}^2 / s_y^2 = 4200.14 / 4329.14 = 0.97 \quad (= r_{xy}^2)$

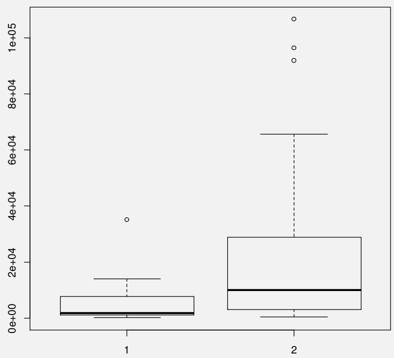
97% de la variabilité des demandes observées est expliquée par la droite de régression.

Exercice 3 Données : DepensesEduData.xls

Le fichier DepensesEduData.csv recense les dépenses publiques de certains états pour l'éducation ainsi que le nombre d'élèves (donnée Eurostat 2008).

```
tab <- read.table("DepensesEduData.csv",header=T,sep=";",dec=",")
summary(tab)
boxplot(tab$nbEleves,tab$Depenses)
```

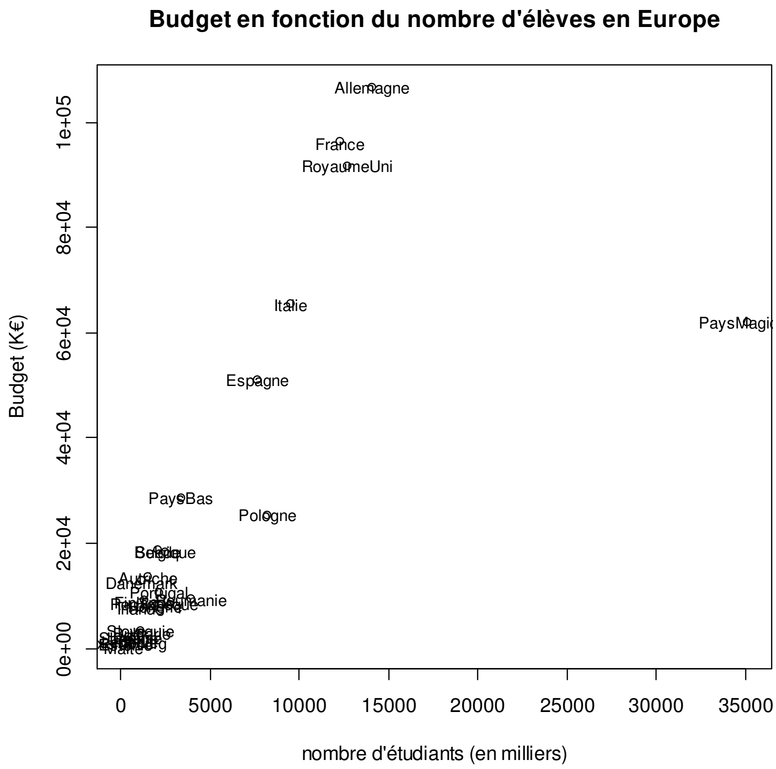
```
    nbEleves          Depenses
 Min.   :   74    Min.   :   492.9
 1st Qu.: 1055    1st Qu.:  3048.1
 Median : 1864    Median : 10030.8
 Mean   : 4847    Mean   : 25399.6
 3rd Qu.: 6687    3rd Qu.: 27863.8
 Max.   :35062    Max.   :106626.4
```



On constate qu'il y a des indidus atypiques dans les deux series qu'il faudra peut être supprimer

1) Tracer le nuage de points des dépenses en fonction du nombre d'élèves.

```
### nuage de points
plot(tab$nbEleves,tab$Depenses,main="Budget en fonction du nombre
d'élèves en Europe", xlab="nombre d'étudiants (en
milliers)",ylab="Budget (K€)")
text(tab$nbEleves,tab$Depenses,row.names(tab),cex=0.8)
# cex=taille de la police
```



Commenter le graphique : beaucoup de pays en amas proche de l'origine. Forme allongée avec pays ayant de fortes dépenses (France, Allemagne,..). Un pays qui est très éloigné des autres, ...

Prendre l'habitude de commenter et pas uniquement taper une ligne de code pour obtenir un graphique ou un chiffre.

2) Calculer le coefficient de corrélation linéaire. Conclusion

```
cor(tab)          # calcule la corrélation entre les variables
```

```
          nbEleves  Depenses
nbEleves 1.0000000 0.7236921
Depenses 0.7236921 1.0000000
```

- $|r_{xy}|$~1 donc relation droite entre le nombre d'élèves et les dépenses
- r_{xy}>0 donc plus le nombre d'élèves augmente plus la dépense augmente

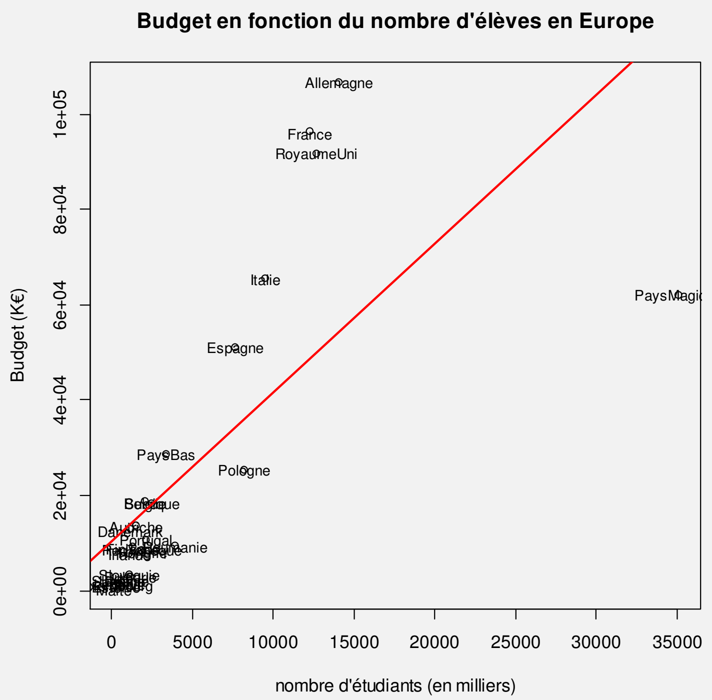
3) Déterminer la droite de régression. Tracer la droite sur le graphique.

```
#  construit le modèle de régression linéaire / lm  = linear model
modele <- lm(Depenses ~ nbEleves, data=tab)
summary(modele)    # résume toutes les caractéristiques du modèle
attributes(modele) # donne tous les attributs de l'objet « lm »
```

```
modele$coef         # donne les coefficients de la droite

# trace la droite sur le nuage de points
abline(modele$coef[1],modele$coef[2],col="red",lwd=2)
```

```
Call:
lm(formula = y ~ x)

Residuals:
   Min     1Q Median     3Q    Max 
-57769 -10012  -6835   1753  52420 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 1.025e+04  5.350e+03   1.917   0.0673 .  
x           3.125e+00  6.083e-01   5.137 2.94e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22760 on 24 degrees of freedom
Multiple R-squared:  0.5237,    Adjusted R-squared:  0.5039 
F-statistic: 26.39 on 1 and 24 DF,  p-value: 2.937e-05
```

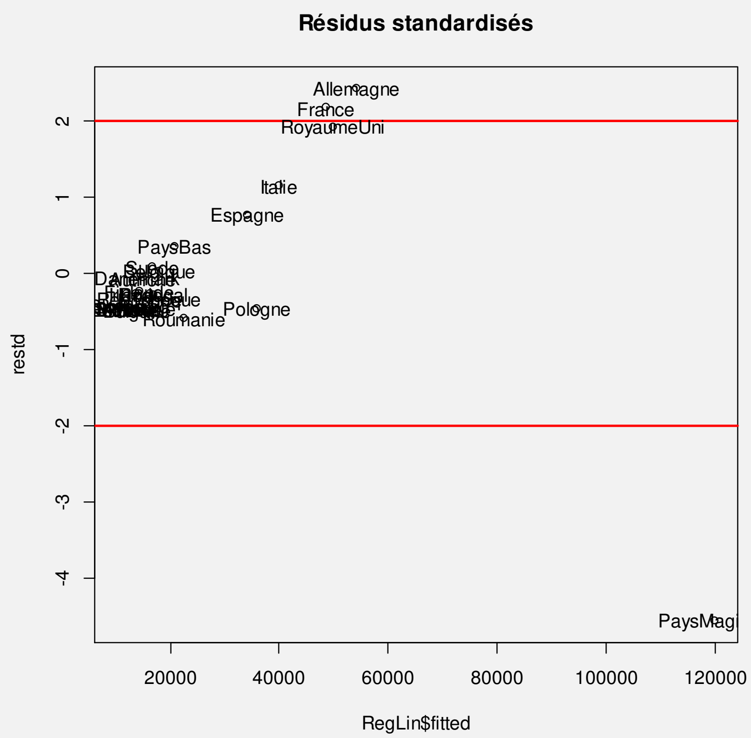
⇒ dépenses=3,125*Nb Elèves+1025



La droite ne passe pas « au milieu » du nuage de points car elle est « attirée » par le Pays Magique.

4) Vérifier les hypothèses sur les résidus. Quel pays semble atypique par rapport au modèle ?

```
modele$fitted          # affiche les prévisions données par le modèle
                       # aux points du tableau
modele$residuals       # affiche les résidus

restd <- rstandard(modele)      # affiche les résidus standardisés

X11()      # ouvre une nouvelle fenêtre graphique
plot(modele$fitted, restd ,ylim=range(-2,2,restd),
                                        main="Résidus standardisés")
# range donne le min et le max d'une série de nombres

abline(h=2,col="red",lwd=2)     # ajoute les lignes pour détecter les
                                # observations atypiques
abline(h=-2,col="red",lwd=2)
text(modele$fitted, restd ,row.names(tab))  # ajoute le nom des pays
```



Là encore le Pays Magique est détecté comme atypique. On le supprime de l'étude.

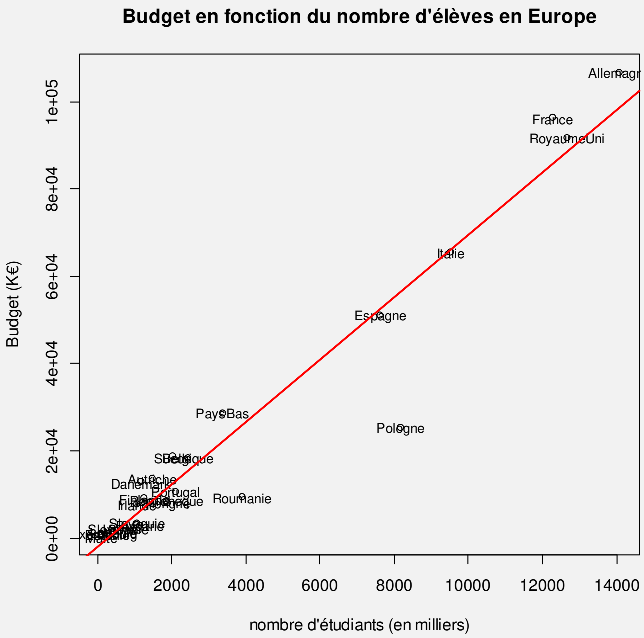
5) Supprimer le pays atypique et refaire la même chose.

```
x=as.vector(tab$nbEleves)[1:25]
y=as.vector(tab$Depenses)[1:25]
plot(x,y,main="Budget en fonction du nombre d'élèves en Europe",
xlab="nombre d'étudiants (en milliers)",ylab="Budget (K€)")
 text(x,y,row.names(tab),cex=0.8)           # cex=taille de la
police
```

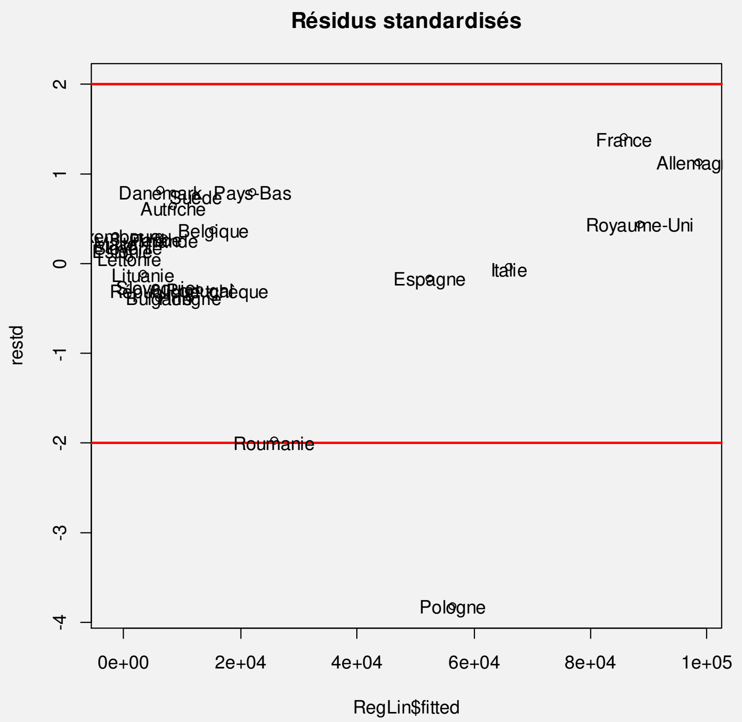
```
RegLin <- lm(y~x) #  construit le modèle de régression linéaire / lm
= linear model

RegLin$coef                          # donne les coefficients de la
droite
 (Intercept)           x
-2093.810192     7.153621
```

⇒ dépenses=7,15*Nb Elèves-2093,81

```
plot(x,y,main="Budget en fonction du nombre d'élèves en Europe",
xlab="nombre d'étudiants (en milliers)",ylab="Budget (K€)")
 text(x,y,row.names(tab)[1:25],cex=0.8)           # cex=taille de la
police
abline(RegLin$coef[1],RegLin$coef[2],col="red",lwd=2)         # trace
la droite
```



Le modèle semble plus représentatif des données excepté pour la Pologne.



Les résidus sont de moyenne nulle, le graphique des résidus ne présente pas de forme particulière. Cela signifie qu'il ne reste pas « d'information » dans les résidus.
La Pologne est détectée atypique par rapport au modèle. On pourrait la supprimer pour voir si le modèle change. Cela n’est pas le cas ici.

6) Quel pourcentage de variabilité des dépenses est expliqué par la droite de régression ? Est-ce que vous validez le modèle ?

```
summary(RegLin)       # résume toutes les caractéristiques du modèle

Call:
lm(formula = y ~ x)

Residuals:
   Min     1Q Median     3Q    Max 
-30943  -2196   1633   3284  10673 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -2093.8102  2230.1957  -0.939    0.358    
x              7.1536     0.3989  17.935 5.13e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8467 on 23 degrees of freedom
Multiple R-squared:  0.9333,    Adjusted R-squared:  0.9304 
F-statistic: 321.6 on 1 and 23 DF,  p-value: 5.131e-15
```

La droite de régression explique 93% (coef. de détermination) de la variabilité du budget.

7) Calculer les budgets prédits par le modèle pour 1000, 6000 et 9500 milliers d'étudiants. Placer les sur le graphique.

```
newx <- data.frame(c(1000,6500,9000)) # nouveaux points
names(newx)= "nbEleves"
prev <- predict(modele,newdata=newx)  # calcul les prévisions aux
                                      # nouveaux points
plot(tab$nbEleves,tab$Depenses,main="Budget en fonction du nombre
                + d'élèves en Europe", xlab="nombre d'étudiants (en
                + milliers)",ylab="Budget (K€)")
points(t(newx),prev,col="green",lwd=2) # t pour transposer le
                                       # vecteur newx
```

```
prev
        1         2         3
 5059.811 44404.724 62288.776
```

Exercice 4 Ventes

(PY Bernard, exercices corrigés de statistique descriptive, ed. economica)

Une étude a été menée auprès d'entreprises afin d'établir le lien entre les quantités commandés d'un bien, Y, et son prix, X et on obtient les observations suivantes (Commandes.csv).

Prix de vente (€)	Quantités commandées
95	104
130	58
148	42
210	12
250	8
330	5

1) Tracer le nuage de points.

```
tab <- read.table("Commandes.csv", header=T,sep=";")
plot(tab$Prix,tab$Vente, main="Nombre de commandes en fonction du
prix")
```

2) Calculer le coefficient de corrélation linéaire entre X et Y. Conclusion

```
cor(tab$Prix,tab$Vente)
[1] -0.8685478
```

⇒

- $|r_{xy}|$~1 donc relation droite
- r_{xy}<0 donc plus le prix augmente et moins il y a de commandes

3) Déterminer la droite de régression de Y en fonction de X.

```
RegLin=lm(Vente~Prix,data=tab)
summary(RegLin)

Call:
lm(formula = tab$Vente ~ tab$Prix)

Residuals:
      1       2       3       4       5       6 
 27.809  -4.726 -13.800 -19.947  -8.557  19.222 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept) 112.7414    22.9958   4.903  0.00803 **
tab$Prix     -0.3847     0.1098  -3.505  0.02478 * 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 21.37 on 4 degrees of freedom
Multiple R-squared:  0.7544,    Adjusted R-squared:  0.693 
F-statistic: 12.29 on 1 and 4 DF,  p-value: 0.02478
```

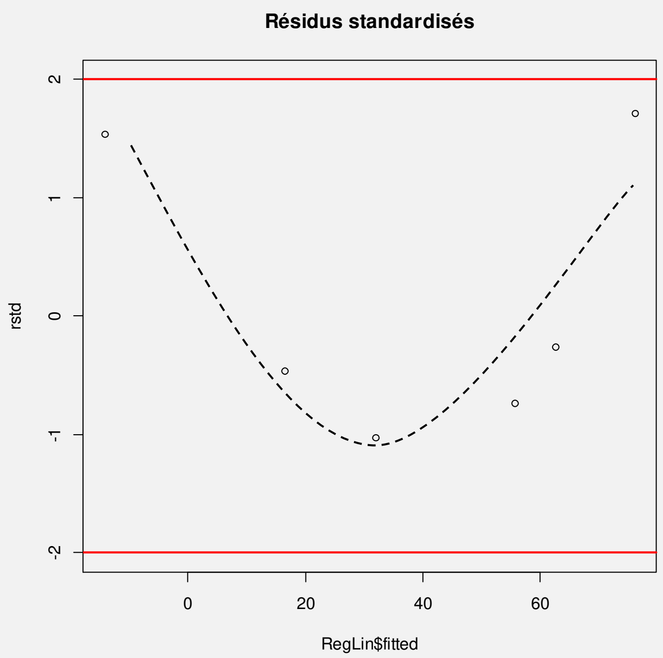
⇒ Vente = 112.74-0.38*Prix

4) Quel est le pourcentage de variation des quantités de commande expliquée par la droite de régression ?

75,44% de la variabilité des ventes est expliquée par ce modèle.

5) Calculer les résidus et vérifier les hypothèses sur les résidus. Conclusion.

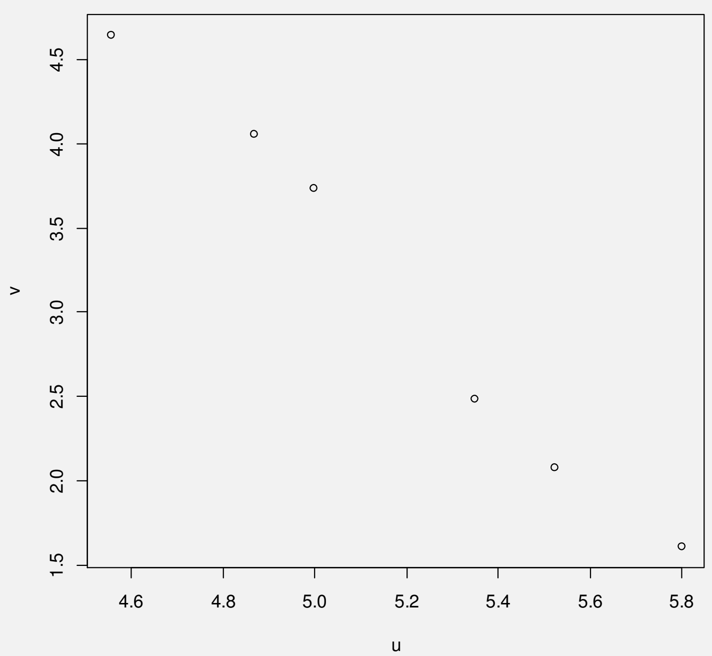
```
rstd=rstandard(RegLin)
plot(RegLin$fitted,rstd,ylim=range(-2,2,rstd),main="Résidus 
standardisés")
abline(h=2,col="red",lwd=2)
abline(h=-2,col="red",lwd=2)
```



⇒ Forme quadratique. Il reste de l'information dans les résidus. Malgré un pourcentage d'explication important, on ne peut pas valider le modèle

6) On pose u=log(x) et v=log(y). Quelle est la relation entre u et v ?

```
u=log(tab$Prix)
v=log(tab$Vente)
plot(u,v)
```



⇒ Relation linéaire entre u et v

7) Calculer le coefficient de corrélation linéaire entre u et v.

```
cor(u,v)
[1] -0.9918848
```

⇒

- $|r_{xy}|$~1 donc relation droite confirmée
- $r_{xy}<0$ donc plus u augmente et v diminue

8) Trouver la droite de régression de v sur u.

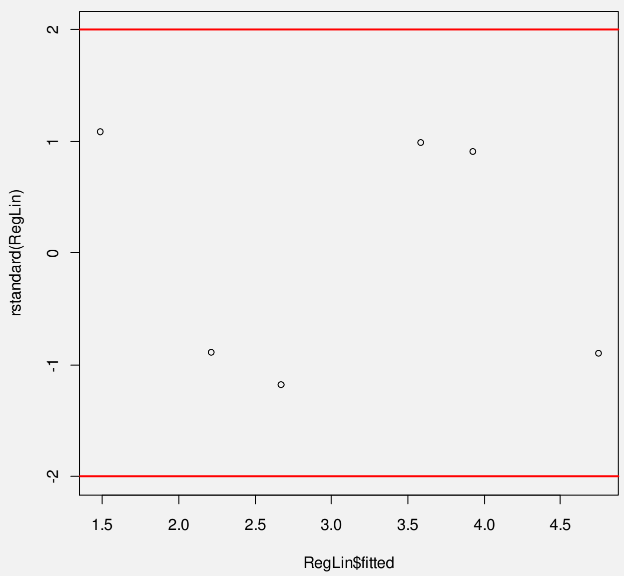
```
RegLin=lm(v~u)
summary(RegLin)

Call:
lm(formula = v ~ u)

Residuals:
      1       2       3       4       5       6 
-0.1042  0.1350  0.1525 -0.1820 -0.1299  0.1286 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  16.6990     0.8742    19.1 4.43e-05 ***
u            -2.6242     0.1682   -15.6 9.85e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1724 on 4 degrees of freedom
Multiple R-squared:  0.9838,    Adjusted R-squared:  0.9798 
F-statistic: 243.5 on 1 and 4 DF,  p-value: 9.852e-05
```

⇒ L'équation est v=16.7-2.6×u

9) Quel est le pourcentage de variation des quantités de commande expliquée par la droite de régression ?

Cette droite explique 98% de la variabilité de v

10) Valider le modèle.



Les résidus sont centrés, sans forme particulière et compris entre -2 et 2.

11) En déduire la quantité qui serait commandée si le prix était fixé à 75€.

```
newu <- data.frame(u = c(log(75)))
 newv <- predict(RegLin,newdata=newu)
 exp(newv)
       1
214.6269
```

Pour un prix de 75€, la quantité commandée serait 214

Exercice 1 (suite facultative)

7) Montrer que les résidus sont non corrélés avec la série X. Qu'est-ce que cela signifie ?
8) Montrer la formule de décomposition de la variance

$$s_y^2 = s_E^2 + s_R^2$$

où s_E^2 est la *variance expliquée* par la droite de régression, et s_R^2 est la *variance résiduelle.*

On peut alors montrer que le *coefficient de détermination*

$$R^2 = \frac{s_E^2}{s_y^2},$$

qui donne le taux de variance expliquée par la droite de régression, est égale au coefficient de corrélation linaire au carré, $R^2 = r_{xy}^2$.

$$s_y^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \overline{y})^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + (\hat{y}_i - \overline{y})^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \overline{y})$$

- $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}e_i^2 = s_e^2$ car $\overline{e} = 0$
- $\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 = s_{\hat{y}}^2$ car $\overline{\hat{y}} = \overline{y}$
- $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) = \frac{1}{n}\sum_{i=1}^{n}e_i(\hat{y}_i - \overline{y}) = \frac{1}{n}\sum_{i=1}^{n}e_i\hat{y}_i - \overline{y}\frac{1}{n}\sum_{i=1}^{n}e_i = \frac{1}{n}\sum_{i=1}^{n}e_i\hat{y}_i$ car $\overline{e} = 0$

$$= \frac{1}{n}\sum_{i=1}^{n}e_i(\hat{a}x_i + \hat{b}) = \hat{a}\frac{1}{n}\sum_{i=1}^{n}e_i x_i + \hat{b}\overline{e}$$

$$= \hat{a}\frac{1}{n}\sum_{i=1}^{n}e_i x_i \text{ car } \overline{e} = 0$$

=0 d'après 7)