TP2 : Analyse bivariée
Croisement Quantitatif-Quantitatif

Durée : 2h30

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

Exercice 1 = Permet de mettre en place qques les outils de la régression linéaire.

Exercice 1 — Un peu de théorie

Dans une population Ω de taille n, on observe deux caractéristiques 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.

La *covariance* entre les deux caractéristiques est définie par

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

C'est une forme bilinéaire symétrique, à valeurs réelles, telle que

$$c_{xx} = s_x^2$$

$$s_{x+y}^2 = s_x^2 + s_y^2 + 2c_{xy}$$

$$c_{xy}^2 \le s_x^2 s_y^2 \text{ (inégalité de Cauchy-Schwartz)}$$

On définit alors le *coefficient de corrélation linéaire* par

$$r_{xy} = \frac{c_{xy}}{s_x s_y}$$

1) Montrer que le coefficient de corrélation linéaire est symétrique, à valeurs dans [-1,1] et correspond à la covariance des observations réduites et centrées. A quoi correspondent les valeurs -1 et +1 ?

D'après l'inégalité de Cauchy-Schwartz, on a

$$0 \le c_{xy}^2 \le s_x^2 s_y^2 \Rightarrow 0 \le \frac{c_{xy}^2}{s_x^2 s_y^2} \le 1 \Rightarrow 0 \le r_{xy}^2 \le 1 \Rightarrow -1 \le r_{xy} \le 1$$

Les valeurs +1 et -1 correspondent à une relation linéaire parfaite entre x et y, c'est-à-dire à l'existence de deux coefficients a et b tels que : y=ax+b.

Notons $x'=\{\frac{x_k-\bar{x}}{s_X}\}_{k=1,\dots,n}$ et $y'=\{\frac{y_k-\bar{y}}{s_Y}\}_{k=1,\dots,n}$ les séries observées centrées et réduites.

$$c_{x'y'} = \frac{1}{n}\sum_{k=1}^{n}\left(\frac{x_k-\bar{x}}{s_X}\right)\left(\frac{y_k-\bar{y}}{s_Y}\right) = \frac{1}{s_X}\frac{1}{s_Y}\frac{1}{n}\sum_{k=1}^{n}(x_k-\bar{x})(y_k-\bar{y}) = \frac{1}{s_x}\frac{1}{s_y}c_{xy}$$

La droite de régression de y sur x est construite en minimisant,

$$S(a,b)=\frac{1}{n}\sum_{i=1}^{n}(y_i-(ax_i+b)^2\,,$$

C'est-à-dire en minimisant la moyenne des écarts au carré entre l'observation y_i et la valeur de la droite au point x_i. On obtient alors

$$\hat{a}=\frac{c_{xy}}{s_x^2}=r_{xy}\frac{s_y}{s_x} \quad \text{et} \quad \hat{b}=\bar{y}-\hat{a}\bar{x}\,.$$

La série des valeurs prédites par la droite de régression est donnée par

$$\hat{y}_i=\hat{a}x_i+\hat{b}\,,$$

et les *résidus* par

$$\hat{e}_i=y_i-(\hat{a}x_i+\hat{b})\,.$$

2) Montrer que la moyenne des valeurs prédites est égale à la moyenne de la série observée.

$$\bar{\hat{y}}=\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i=\frac{1}{n}\sum_{i=1}^{n}\hat{a}x_i+\hat{b}=\hat{a}\frac{1}{n}\sum_{i=1}^{n}x_i-\hat{b}=\hat{a}\bar{x}-\hat{b}=\bar{y}$$

3) Montrer que les résidus sont de moyenne nulle et sont non corrélés avec la série X. Qu'est-ce que cela signifie ?

Centrés :

$$\bar{e}=\frac{1}{n}\sum_{i=1}^{n}\hat{e}_i=\frac{1}{n}\sum_{i=1}^{n}[y_i-(\hat{a}x_i+\hat{b})]=\frac{1}{n}\sum_{i=1}^{n}y_i-\hat{a}\frac{1}{n}\sum_{i=1}^{n}x_i-\hat{b}=\bar{y}-\hat{a}\bar{x}-\hat{b}$$

$$= \bar{y} - \hat{a}\bar{x} - (\bar{y} - \hat{a}\bar{x}) = 0$$

Pour montrer qu'ils sont non corrélés, il suffit de montrer que la covariance est nulle,

$$c_{xe} = \frac{1}{n}\sum_{i=1}^{n} x_i \hat{e}_i - \bar{x}\,\bar{e} = \frac{1}{n}\sum_{i=1}^{n} x_i \hat{e}_i = \frac{1}{n}\sum_{i=1}^{n} x_i (y_i - \hat{a}x_i - \hat{b}) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \hat{a}\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \hat{b}\frac{1}{n}\sum_{i=1}^{n} x_i$$

$$= \left(c_{xy} + \bar{x}\,\bar{y}\right) - \hat{a}\left(s_x^2 + \bar{x}^2\right) - \hat{b}\bar{x} = \left(c_{xy} + \bar{x}\,\bar{y}\right) - \hat{a}\left(s_x^2 + \bar{x}^2\right) - (\bar{y} - \hat{a}\bar{x})\bar{x}$$

$$= c_{xy} - \hat{a}s_x^2 = c_{XY} - \frac{c_{xy}}{s_x^2}s_x^2 = 0$$

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

Exercice 2

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.

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

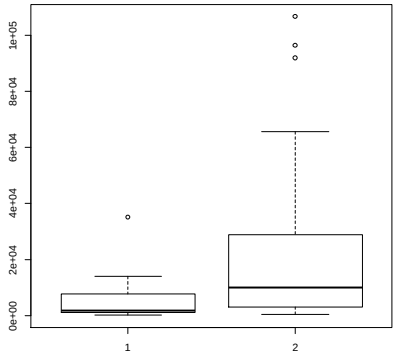
Exercice 3 — Données : DepensesEdu.xls

Le fichier DepencesEdu.xls 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)
```

```
   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
```

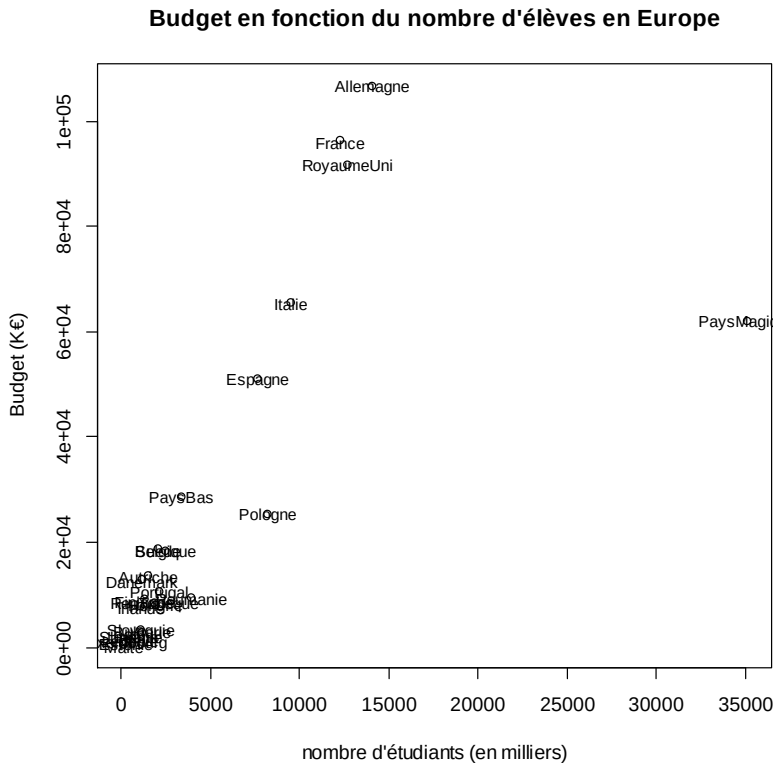


On constate qu'il y a des indidus atypiques dans les deux series qu'il faudra peut être supprimer

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

```
> 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)
           nbEleves  Depenses
nbEleves  1.0000000 0.7236921
Depenses  0.7236921 1.0000000
```

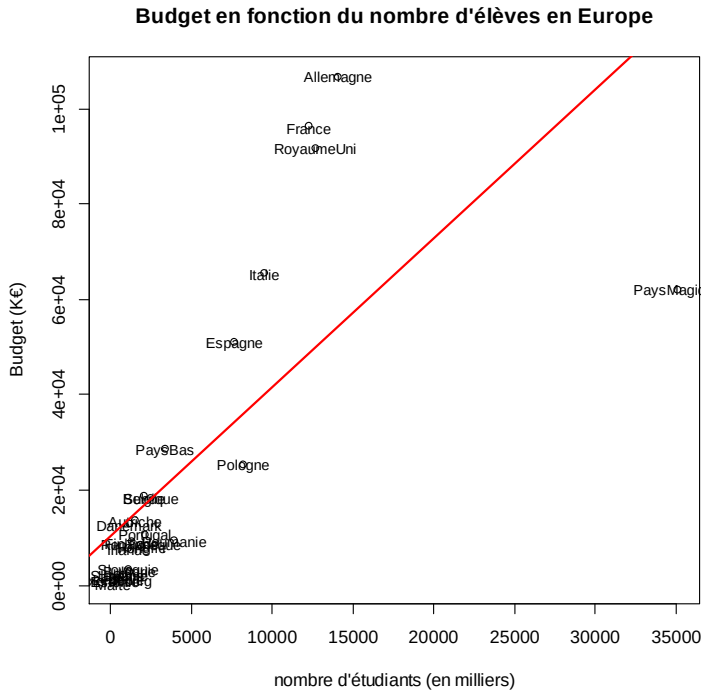
- $|r_{xy}|$~1 donc relation droite entre le nombre d'élèves et les dépenses
- r_{xy}>0 donc plus le nombre d'élèves augmente plus la dépense augmente

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

```
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 .
```

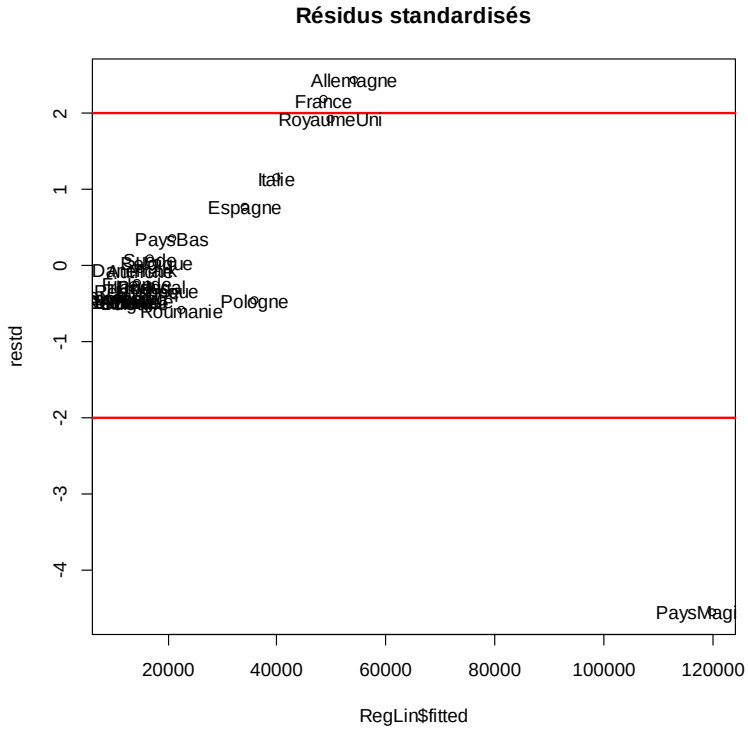
```
x           3.125e+00  6.083e-01   5.137 2.94e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

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 ?



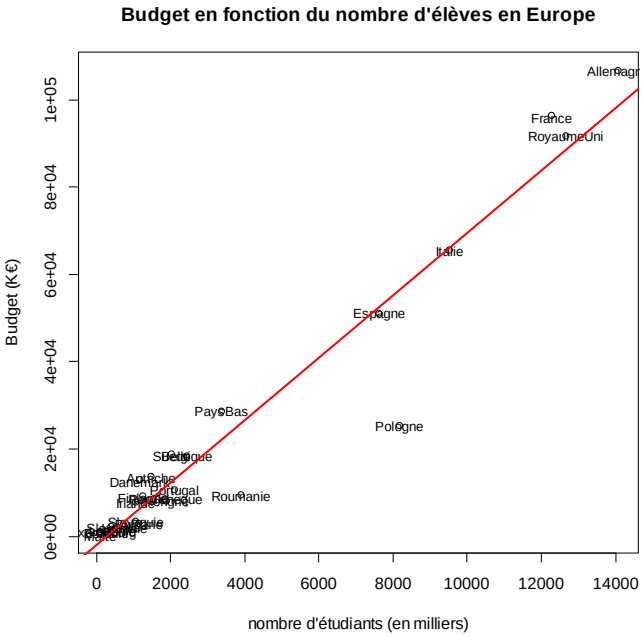
Là encore le Pays Magique est détecté comme atypique. On le supprime de l'étude.

5)

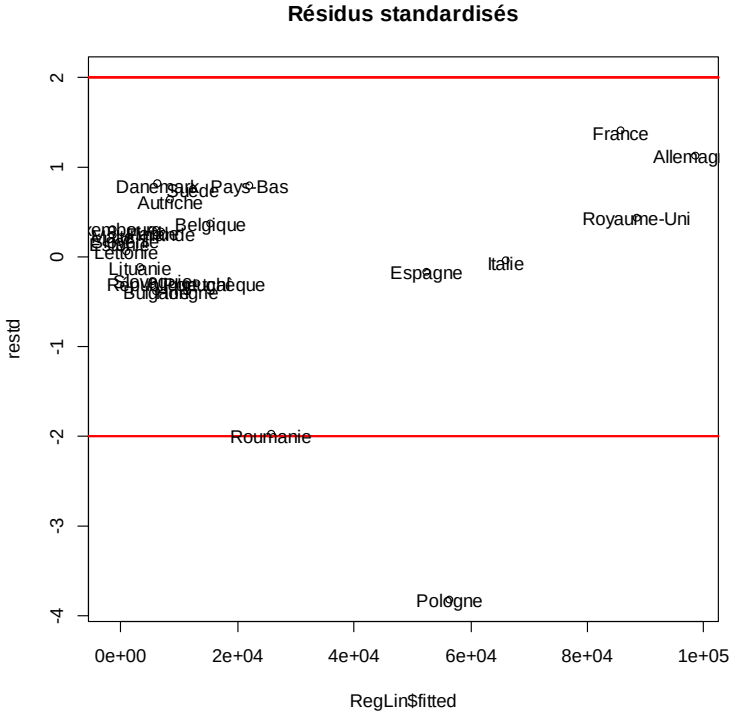
```
x=as.vector(tab$nbEleves)[1:25]
 y=as.vector(tab$Depenses)[1:25]
plot(x,y,main="Budget en fonction du nombre d'élèves en Europe", xlab="nombre d'étudiants
(en milliers)",ylab="Budget (K€)")
 text(x,y,row.names(tab),cex=0.8)          # cex=taille de la police
RegLin <- lm(y~x) #  construit le modèle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

```
> 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.

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

Exercice 4

(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.

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(tab$Vente~tab$Prix)
> 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
```

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

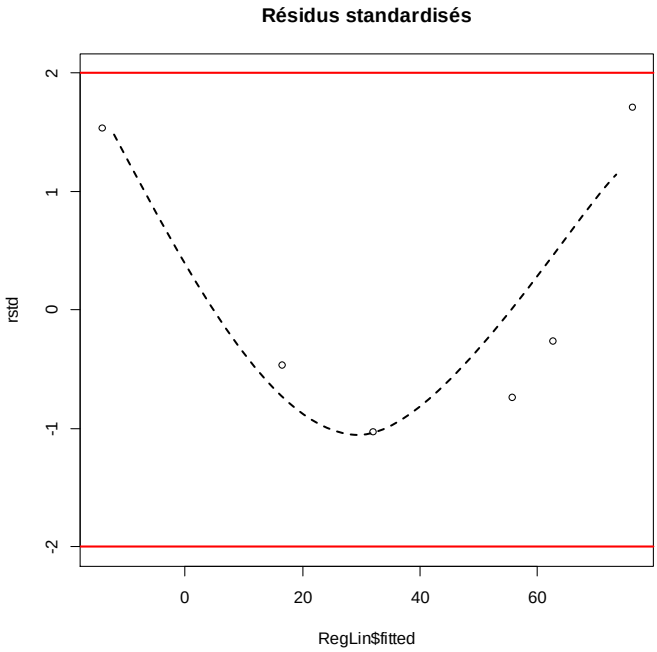
⇒ 75% de la variabilité des commandes s’explique par la droite de régression :

Commandes= 112.74-0.38×Prix

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)
```

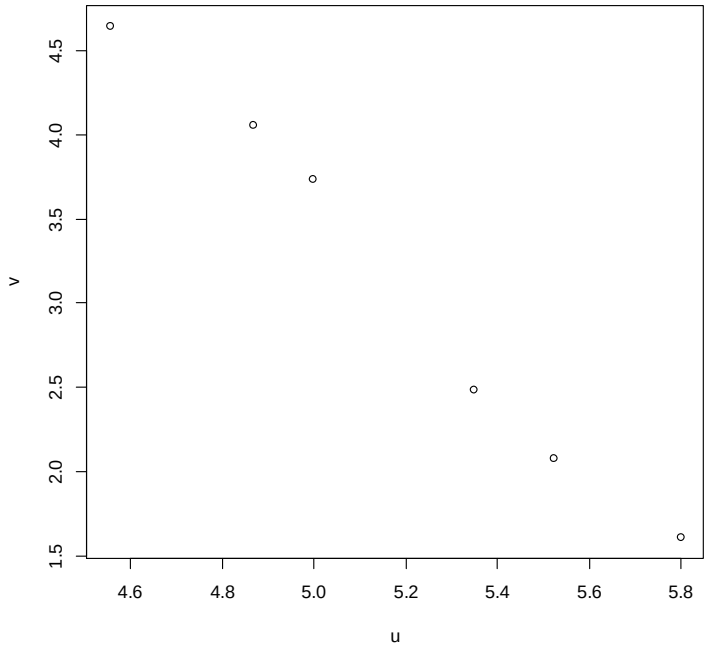
```
abline(h=-2,col="red",lwd=2)
```



⇒ 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.

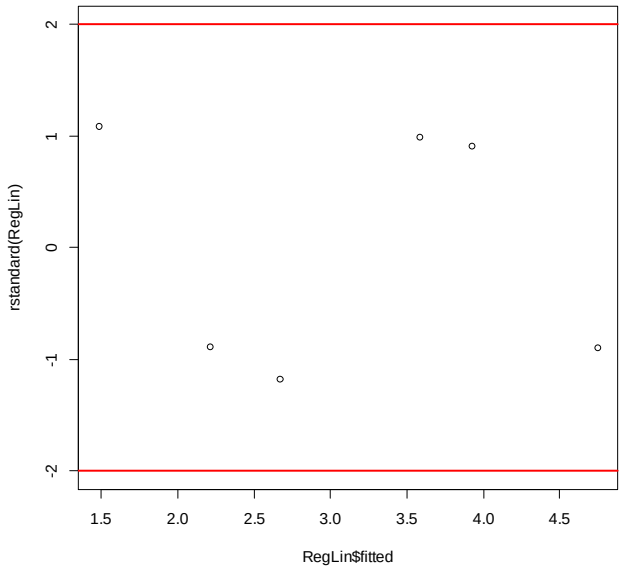
```
> RegLin=lm(v~u)
> summary(RegLin)

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 – facultatif)

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.*

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

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

$$= \frac{1}{n}\sum_{i=1}^{n} e_i(\hat{a}x_i + \hat{b}) = \hat{a}\frac{1}{n}\sum_{i=1}^{n} e_i x_i + \hat{b}\bar{e} = 0 \text{ d'après 3)}$$

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, $R^2 = r_{xy}^2$.