Université Mostefa Ben Boulaïd - Batna 2
Faculté de Mathématiques et d'Informatique
Département de Mathématiques
Master 1 Mathématiques Appliquées

Solution TD 01

Exercice 1 *On veut prédire la hauteur H d'un arbre en fonction de son diamètre D. Pour faire une régression linéaire, on effectue un changement de variable en posant $Y = lnH$ et $X = lnD$. Voici les mesures faites sur 5 arbres.*

X	-1.61	-1.20	-0.97	-0.51	-0.42
Y	2.22	2.27	2.38	2.60	2.65

1. *Donner le coefficient de corrélation linéaire empirique entre X et Y.*
2. *Donner l'équation de la droite de régression empirique de Y par rapport à X.*
3. *Tester la signification de cette régression au seuil 5%.*
4. *Donner la hauteur prévue d'un arbre de diamètre 0.7.*
5. *Donner un intervalle de confiance de niveau 95% pour la prédiction d'un arbre de diamètre 0.7.*

Solution de l'exercice : 1 *1. D'après la définition 1.5 page 5, le coefficient de corrélation est donné par :*

$$r = Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_x \sigma_y}$$

On a : $n = 5$

$Cov(X,Y) = \overline{XY} - \overline{X}.\overline{Y}$	$\overline{X} = \frac{1}{n}\sum_{i=1}^{n=5} x_i$	$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n=5} y_i$	$\overline{XY} = \frac{1}{n}\sum_{i=1}^{n=5} x_i y_i$
$\sigma_x = \sqrt{Var(X)}$	$Var(X) = \overline{X^2} - \overline{X}^2$	$\overline{X^2} = \frac{1}{n}\sum_{i=1}^{n=5} x_i^2$	
$\sigma_Y = \sqrt{Var(Y)}$	$Var(Y) = \overline{Y^2} - \overline{Y}^2$	$\overline{Y^2} = \frac{1}{n}\sum_{i=1}^{n=5} y_i^2$	

— $\overline{X} = \frac{1}{n}\sum_{i=1}^{n=5} x_i = \frac{1}{5}[(-1.61) + (-1.20) + (-0.97) + (-0.51) + (-0.42)] = -0.942$

— $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n=5} y_i = \frac{1}{5}[2.22 + 2.27 + 2.38 + 2.60 + 2.65] = 2.424$

—

$$\begin{aligned}\overline{XY} &= \frac{1}{n}\sum_{i=1}^{n=5} x_i y_i \\ &= \frac{1}{5}[(-1.61 \times 2.22) + (-1.20 \times 2.27) + (-0.97 \times 2.38) + (-0.51 \times 2.60) + (-0.42 \times 2.65)] \\ &= -2.2091\end{aligned}$$

— $\overline{X^2} = \frac{1}{n}\sum_{i=1}^{n=5} x_i^2 = \frac{1}{5}[(-1.61)^2 + (-1.20)^2 + (-0.97)^2 + (-0.51)^2 + (-0.42)^2] = 1.0819$

— $\overline{Y^2} = \frac{1}{n}\sum_{i=1}^{n=5} y_i^2 = \frac{1}{5}[2.22^2 + 2.27^2 + 2.38^2 + 2.60^2 + 2.65^2] = 5.9056$

— $Var(X) = \overline{X^2} - \overline{X}^2 = 1.0819 - 0.942^2 = 0.1945$

— $\sigma_x = \sqrt{Var(X)} = \sqrt{0.1945} = 0.441$

— $Var(Y) = \overline{Y^2} - \overline{Y}^2 = 5.9056 - 2.424^2 = 0.0298$

— $\sigma_Y = \sqrt{Var(Y)} = \sqrt{0.0298} = 0.1726$

— $Cov(X,Y) = \overline{XY} - \overline{X}.\overline{Y} = -2.2091 - (-0.942)(2.424) = 0.0743$

Alors, le coefficient de corrélation est :

$$r = \frac{Cov(X,Y)}{\sigma_x \sigma_y} = \frac{0.0743}{0.441 \times 0.1726} = 0.9761$$

Donc la corrélation entre X et Y est très forte.

2. *D'après la définition 1.2, page 2. On a :*

$$\begin{cases} \hat{b}_1 = \overline{y} - \hat{b}_2 \overline{x} \\ \hat{b}_2 = \dfrac{\sum_{i=1}^n (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2} = \dfrac{Cov(x,y)}{Var(x)} \end{cases}$$

est donc :

$$\begin{cases} \hat{b}_1 = \overline{y} - \hat{b}_2 \overline{x} = 2.424 - (0.382)(-0.942) = 2.7838 \\ \hat{b}_2 = \dfrac{Cov(x,y)}{Var(x)} = \dfrac{0.0743}{0.1945} = 0.382 \end{cases}$$

Donc la droite de régression de Y en fonction de X est donnée par :

$$Y = 2.7838 + 0.382X$$

3. *D'après le test du caractère significatif du modèle, citer dans la page 10, On utilise dans ce exercice la première méthode, en utilisant une variable de Student.*
En utilisant l'hypothèse suivante :

$$(\mathcal{H}_0) : b_2 = 0 \quad contre \quad (\mathcal{H}_1) : b_2 \neq 0$$

Alors, on sait que :

$$Y_i = b_1 + b_2 x_i + \epsilon_i$$

et

$$T = \frac{\beta_2 - b_2}{\dfrac{\Sigma}{\sqrt{nVar(x)}}} \sim \mathcal{S}t(n-2).$$

Sous $(\mathcal{H}_0)$ cette hypothèse devient

$$Y_i = b_1 + \epsilon_i$$

et

$$T = \frac{\beta_2}{\dfrac{\Sigma}{\sqrt{nVar(x)}}} \sim \mathcal{S}t(n-2).$$

Par définition de la loi de Student on sait que

$$\mathbb{P}\left[\mid T \mid \leq t_{1-\alpha/2}^{n-2} \right] = 1 - \alpha.$$

Puisque on veut tester la signification de cette régression au seuil de 5%, alors $\alpha = 5\%$.

— *la première étape, on calcule $t_{1-\alpha/2}^{n-2}$ pour $\alpha = 5\%$ et degrè de liberté $= ddl = n - 2 = 5 - 2 = 3$, en utilisant la table de la loi de Student, d'après le fichier pdf des tableaux statistique page 2, On a :*

ddl/α	.....	5%	.
.	.	.	.
3	.	3.1824	.

alors, pour $\alpha = 5\%$ *et* $n-2=3$ *on a,* $t^{n-2}_{1-\alpha/2} = 3.1824$

— *la deuxième étape, on calcule la valeur* $t = \dfrac{\hat{b}_2}{\dfrac{\hat{\sigma}}{\sqrt{nVar(x)}}}$ *de la variable aléatoire* T *sur les données* $(x_i, y_i)_{1\leq i\leq n}$, *on a :*

$$t = \frac{\hat{b}_2}{\dfrac{\hat{\sigma}}{\sqrt{nVar(x)}}}$$

D'abord, il faut calculer la valeur $\hat{\sigma}^2$ *(la valeur de la variable aléatoire* Σ^2 *:=estimateur de la variable aléatoire* σ^2*). D'après la proposition 1.8 page 7, on a :*

$$\Sigma^2 := \frac{1}{n-2}\sum_{i=1}^{n} E_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

d'après l'exercice 6 de la série de TD01, on a :

$$\Sigma^2 := \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \frac{n}{n-2}Var(Y)(1-r^2)$$

Alors,

$$\hat{\sigma}^2 = \frac{n}{n-2}Var(Y)(1-r^2) = \frac{5}{3}(0.0298)(1-0.9761^2) = 0.0023$$

Alors,

$$t = \frac{\hat{b}_2}{\dfrac{\hat{\sigma}}{\sqrt{nVar(x)}}} = \frac{0.382\times\sqrt{5\times 0.1945}}{0.0479} = 7.8645$$

On adopte alors la règle de décision suivante

— *Si* $\mid t \mid \leq t^{n-2}_{1-\alpha/2}$ *alors on accepte* $(\mathcal{H}_0)$ *au risque* α *(il n'y a pas de lien linéaire entre les deux variables* X *et* Y*, avec un risque de* α*)*

— *si* $\mid t \mid > t^{n-2}_{1-\alpha/2}$ *alors on rejette l'hypothèse* $(\mathcal{H}_0)$ *et on accepte* $(\mathcal{H}_1)$ *(il y a un lien linéaire entre les deux variables)*

Puisque, $|t| = 7.8645 > 3.1824 = t^{n-2}_{1-\alpha/2}$, *on rejette l'hypothèse* $(\mathcal{H}_0) : b_2 = 0$ *et on accepte l'hypothèse* $(\mathcal{H}_1) : b_2 \neq 0$ *au seuil de* 5%, *et donc il y a un lien linéaire entre* X *et* Y *au seuil de* 5%.

4. *D'après l'énoncé de l'exercice, le diamètre est la variable* D *avec* $X = lnD$ *et* $Y = lnH$ *avec* H *est la hauteur de l'arbre. alors on pose :*

$$D_6 = 0.7, \quad X_6 = lnD_6 = -0.3567$$

D'après la définition 1.3, page 3. on a :

$$\hat{y}_i := \hat{b}_1 + \hat{b}_2 x_i$$

Alors,

$$\hat{y}_6 := \hat{b}_1 + \hat{b}_2 x_6 = 2.7838 + 0.382\times(-0.3567) = 2.6475$$

et puisque $Y = lnH$, *alors* $H = e^Y$. *Donc*

$$H_6 = e^{y_6} = e^{2.6475} = 14.1186$$

5. *D'après la section 1.8.2 Prévision d'une valeur, page 13, l'intervalle de confiance de niveau de confiance $1-\alpha$ est donnée par :*

$$IC_{1-\alpha}(y_{n+1}) = \left[\hat{y}_{n+1} - t^{n-2}_{1-\alpha/2}\hat{\sigma}_{e_{n+1}}, \hat{y}_{n+1} + t^{n-2}_{1-\alpha/2}\hat{\sigma}_{e_{n+1}}\right]$$

avec

$$\hat{\sigma}^2_{e_{n+1}} = \hat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(x_{n+1}-\overline{x})^2}{\sum_{i=1}^n (x_i-\overline{x})^2}\right)$$

comme :

$$\sum_{i=1}^{n}(x_i-\overline{x})^2 = nVar(X)$$

Donc :

$$\hat{\sigma}^2_{e_6} = \hat{\sigma}\left(1 + \frac{1}{5} + \frac{(x_6-\overline{x})^2}{5Var(X)}\right) = 0.0023\left(1 + \frac{1}{5} + \frac{(-0.3567-(-0.942))^2}{5\times 0.1945}\right) = 0.00357$$

et, pour $\alpha = 5\%$ et $t^{n-2}_{1-\alpha/2} = 3.1824$

$$\begin{aligned} IC_{95\%}(y_6) &= \left[\hat{y}_6 - t^{n-2}_{1-\alpha/2}\hat{\sigma}_{e_6}, \hat{y}_6 + t^{n-2}_{1-\alpha/2}\hat{\sigma}_{e_6}\right] \\ &= \left[2.6475 - 3.1824\times\sqrt{0.00357}, 2.6475 + 3.1824\times\sqrt{0.00357}\right] \\ &= [2.4573, 2.8376] \end{aligned}$$

Finalement

$$IC_{95\%}(H_6) = \left[e^{2.4573}, e^{2.8376}\right] = [11.6732, 17.0747]$$

Exercice 2 *On a mesuré pour 16 malades atteints de leucémie :*

— x_i *le logarithme décimal du nombre de globules blancs le jour du diagnostic de la maladie,*

— y_i *le nombre de semaines de survie après le diagnostic.*

On suppose que chaque y_i est une observation d'une loi $\mathcal{N}(b_2x_i + b_1, \sigma^2)$ et que les 16 mesures sont indépendantes. On donne les résultats suivants :

$$\sum_{i=1}^{16} x_i = 64.63, \quad \sum_{i=1}^{16} y_i = 1061, \quad \sum_{i=1}^{16} x_i^2 = 266.457, \quad \sum_{i=1}^{16} y_i^2 = 113611, \quad \sum_{i=1}^{16} x_iy_i = 3972.26$$

1. *Estimer les paramètres b_1, b_2 et σ^2*
2. *Calculer un intervalle de confiance pour le paramètre σ, au seuil de 5%.*
3. *Calculer un intervalle de confiance pour le paramètre b_2, au seuil de 5%.*
4. *Déduire si la régression est significative ou bien non, au seuil de 5%.*
5. *Tester la signification de cette régression au seuil 5%, en utilisant une variable de Fisher.*

Solution de l'exercice : 2 *On a :*

$\overline{X} = \frac{1}{16}\sum_{i=1}^{16} x_i = \frac{64.63}{16} = 4.0393$	$\overline{Y} = \frac{1}{16}\sum_{i=1}^{16} y_i = \frac{1061}{16} = 66.3125$
$\overline{XY} = \frac{1}{16}\sum_{i=1}^{16} x_iy_i = \frac{3972.26}{16} = 248.2662$	$\overline{X^2} = \frac{1}{16}\sum_{i=1}^{16} x_i^2 = \frac{266.457}{16} = 16.6535$
$Var(X) = \overline{X^2} - \overline{X}^2 = 0.3375$	$\sigma_x = \sqrt{Var(X)} = 0.5809$
$\overline{Y^2} = \frac{1}{16}\sum_{i=1}^{16} y_i^2 = \frac{113611}{16} = 7100.6875$	$Var(Y) = \overline{Y^2} - \overline{Y}^2 = 2703.3398$
$\sigma_Y = \sqrt{Var(Y)} = 51.9936$	$Cov(X,Y) = \overline{XY} - \overline{X}.\overline{Y} = -19.5898$

1. *D'après la définition 1.2, page 2. On a :*

$$\begin{cases} \hat{b}_1 = \overline{y} - \hat{b}_2\overline{x} \\ \hat{b}_2 = \dfrac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2} = \dfrac{Cov(x,y)}{Var(x)} \end{cases}$$

est donc :

$$\begin{cases} \hat{b}_1 = \overline{y} - \hat{b}_2\overline{x} = 66.3125 - (-58.0438) \times 4.0393 = 300.7688 \\ \hat{b}_2 = \dfrac{Cov(x,y)}{Var(x)} = \dfrac{-19.5898}{0.3375} = -58.0438 \end{cases}$$

Donc la droite de régression de Y en fonction de X est donnée par :

$$Y = 300.7688 - 58.0438X$$

D'après la proposition 1.8 page 7, on a :

$$\Sigma^2 := \frac{1}{n-2}\sum_{i=1}^{n} E_i^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

d'après l'exercice 6 de la série de TD 01, on a :

$$\Sigma^2 := \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \frac{n}{n-2}Var(Y)(1-r^2)$$

Alors,

$$\hat{\sigma}^2 = \frac{n}{n-2}Var(Y)(1-r^2)$$

D'après la définition 1.5 page 5, le coefficient de corrélation est donné par :

$$r = Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_x\sigma_y}$$

Alors

$$r = \frac{-19.5898}{0.5809 \times 51.9936} = -0.6486$$

par conséquent

$$\hat{\sigma}^2 = \frac{1}{16} \times 2703.3398 \times (1 - (-0.6486)^2) = 1789.8211$$

Donc la corrélation entre X et Y est presque forte.

2. *D'après la proposition 1.11, page 8. On a :*

$$Z = (n-2)\frac{\Sigma^2}{\sigma^2} \sim \mathcal{X}^2(n-2)$$

D'après la densité de la loi de $\mathcal{X}^2(n-2)$, voir figure 1 on a :

$$\mathbb{P}\left[\mathcal{X}_1^2 \leq Z = (n-2)\frac{\Sigma^2}{\sigma^2} \leq \mathcal{X}_2^2\right] = 1 - \alpha.$$

avec $\mathcal{X}_1^2$ et $\mathcal{X}_2^2$ sont les quantiles d'ordre $\alpha/2$ et $1-\alpha/2$ respectivement de la loi de $\mathcal{X}^2(n-2)$, Ici le degré de liberté(ddl) est $(n-2)$. Alors

$$\mathbb{P}\left[(n-2)\frac{\Sigma^2}{\mathcal{X}_2^2} \leq \sigma^2 \leq (n-2)\frac{\Sigma^2}{\mathcal{X}_1^2}\right] = 1 - \alpha.$$

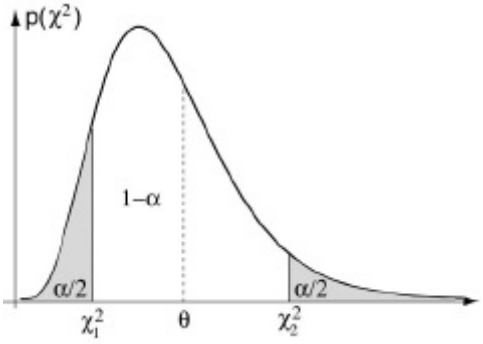

FIGURE 1 – Densité de la loi de $\mathcal{X}^2$

Par conséquent,

$$IC_{1-\alpha}(\sigma^2) = \left[(n-2)\frac{\hat{\sigma}^2}{\mathcal{X}_2^2}, (n-2)\frac{\hat{\sigma}^2}{\mathcal{X}_1^2}\right]$$

Pour calculer $\mathcal{X}_1^2$ et $\mathcal{X}_2^2$ un utilisant la table de la loi de $\mathcal{X}^2$, (voir le fichier pdf des tableaux statistique, page 3). Alors pour $\alpha = 5\%$ et $ddl = n-2 = 16-2 = 14$ on a :

ddl/α	.....	$\alpha/2 = 2.5\%$	.....	$1-\alpha/2 = 97.5\%$	..
.	.....	.	.....	.	..
14	.....	$\mathcal{X}_1^2 = 5.629$	.....	$\mathcal{X}_2^2 = 26.119$	..

Finalement,

$$\begin{aligned} IC_{95\%}(\sigma^2) &= \left[(n-2)\frac{\hat{\sigma}^2}{\mathcal{X}_2^2}, (n-2)\frac{\hat{\sigma}^2}{\mathcal{X}_1^2}\right] \\ &= \left[14\frac{1789.8211}{26.119}, 14\frac{1789.8211}{5.629}\right] \\ &= [959.3589, 4451.5003] \end{aligned}$$

3. *D'après la définition 1.8, page 10 on a : On obtient*

$$IC_{1-\alpha}(b_2) = \left[\hat{b}_2 - \frac{\hat{\sigma}}{\sqrt{nVar(x)}}t_{1-\alpha/2}^{n-2}, \hat{b}_2 + \frac{\hat{\sigma}}{\sqrt{nVar(x)}}t_{1-\alpha/2}^{n-2}\right]$$

On calcule $t_{1-\alpha/2}^{n-2}$ pour $\alpha = 5\%$ et degrè de liberté $= ddl = n-2 = 16-2 = 14$, en utilisant la table de la loi de Student, (voir le fichier pdf des tableaux statistique page 2), On a :

ddl/α	.....	5%	.
.	.	.	.
14	.	$t_{1-\alpha/2}^{n-2} = 2.1448$	.

alors,

$$\begin{aligned} IC_{95\%}(b_2) &= \left[-58.0438 - \frac{42.3062}{\sqrt{16\times 0.3375}} \times 2.1448, -58.0438 + \frac{42.3062}{\sqrt{16\times 0.3375}} \times 2.1448\right] \\ &= [-97.0913, -18.9962] \end{aligned}$$

4. *Pour tester si la régression est significative, en utilisant les hypothèses suivants :*

$$(\mathcal{H}_0) : b_2 = 0 \quad \textit{contre} \quad (\mathcal{H}_1) : b_2 \neq 0$$

On remarque que

$$b_2 = 0 \notin IC_{95\%}(b_2) = [-97.0913, -18.9962]$$

alors on rejette $(\mathcal{H}_0) : b_2 = 0$ *et on accepte l'hypothèse* $(\mathcal{H}_1) : b_2 \neq 0$ *au seuil de* 5%*. Donc il y a un lien linéaire entre* X *et* Y *au seuil de* 5%*.*

5. *Il s'agit toujours de test l'hypothèse suivante*

$$(\mathcal{H}_0) : b_2 = 0 \quad \textit{contre} \quad (\mathcal{H}_1) : b_2 \neq 0$$

en utilise cette fois une variable de Fisher. D'après le test du caractère significatif du modèle, la deuxième méthode, citer dans la page 11. On a :

$$T^2 = \left(\frac{\beta_2 - b_2}{\frac{\sigma}{\sqrt{nVar(x)}}} \right)^2 \left(\frac{(n-2)}{(n-2)\frac{\Sigma^2}{\sigma^2}} \right) \sim \mathcal{F}(1, n-2)$$

Alors, sous l'hypothèse $\mathcal{H}_0$

$$T^2 = \frac{nVar(x)\beta_2^2}{\Sigma^2} \sim \mathcal{F}(1, n-2)$$

Par définition de la loi de Fisher

$$\mathbb{P}(T^2 \leq f_{1-\alpha}^{1,n-2}) = 1 - \alpha$$

Où $f_{1-\alpha}^{1,n-2}$ *est la fractile d'ordre* $1-\alpha$ *de la loi de Fisher.*

(a) *la première étape : premièrement on remarque que la variable qui suit une loi de Fisher ayant deux degrés de liberté, Ici*

$$T^2 \sim \mathcal{F}(1, n-2)$$

Alors, $ddl_1 = 1$ *et* $ddl_2 = n - 2 = 14$*.*
On calcule la valeur $f_{1-\alpha}^{1,n-2}$*, Pour* $\alpha = 5\%$*,* $ddl_1 = 1$ *et* $ddl_2 = 14$*. En utilisant la table de la loi de Fisher (voir le fichier pdf des tableaux statistiques, page 4). on trouve :*

ddl_2/ddl_1	1	.....
.	.	.....
14	$f_{1-\alpha}^{1,n-2} = 4.600$	.....

(b) *la deuxième étape : on calcule la valeur*

$$t^2 = \frac{nVar(x)\hat{b}_2^2}{\hat{\sigma}^2}$$

de la variable aléatoire T^2 *sur les données* $(x_i, y_i)_{1 \leq i \leq n}$*. d'après l'exercice 06 de la série de Td 01, on a :*

$$t^2 = \frac{nVar(x)\beta_2^2}{\Sigma^2} = (n-2)\frac{R^2}{1-R^2},$$

avec R^2 *est le coefficient de détermination et* $R^2 = r^2$ *Donc :*

$$t^2 = (n-2)\frac{r^2}{1-r^2} = 14\frac{(-0.6486)^2}{1-(-0.6486)^2} = 10.16663$$

On adopte alors la règle de décision suivante (page 11) :

- *Si* $0 \leq t^2 \leq f_{1-\alpha}^{1,n-2}$ *alors on accepte* $(\mathcal{H}_0)$ *au risque* α *(il n'y a pas de lien linéaire entre les deux variables* X *et* Y*, avec un risque de* α*)*
- *Sinon on rejette l'hypothèse* $(\mathcal{H}_0)$ *et on accepte* $(\mathcal{H}_1)$ *(il y a un lien linéaire entre les deux variables)*

Puisque $t^2 = 10.1663 > f_{1-\alpha}^{1,n-2} = 4.600$*, alors on rejette l'hypothèse* $(\mathcal{H}_0)$ *et on accepte* $(\mathcal{H}_1)$ *au seuil de* 5% *(il y a un lien linéaire entre les deux variables* 5%*) (qui affirme la réponse de la question 4.)*

Exercice 3 *On dit souvent que le pouls* Y *d'une personne est relié à l'âge* X *par* $Y = 220 - X$*. Supposons que l'on veuille le prouver empiriquement et que pour cela on dispose des observations suivantes :*

x_i	18	23	25	35	65	54	34	56	72	19	23	42	18	39	37
y_i	202	186	187	180	156	169	174	172	153	199	193	174	198	183	178

Les données du pouls et de l'âge confirment-elles la règle indiquée ci-dessus, au seuil de 5%*.*

Solution de l'exercice : 3 *Il s'agit de tester le modèle linéaire spécifié* $Y = 220 - X$*. D'après le test d'un modèle linéaire spécifié, page 11 et 12. On testera l'hypothèse suivante :*

$$(\mathcal{H}_0) : b_1 = b_1^* \quad et \quad b_2 = b_2^* \quad contre \quad (\mathcal{H}_1) : b_1 \neq b_1^* \quad ou \quad b_2 \neq b_2^*$$

avec $b_1^* = 220$ *et* $b_2^* = -1$ *(d'après l'énoncé de l'exercice) Sous* $(\mathcal{H}_0)$ *on a :*

$$Y_i = 220 - x_i + \epsilon_i$$

Sous l'hypothèse $(\mathcal{H}_0)$*, nous avons*

$$Z = \frac{n-2}{2} \frac{\sum_{i=1}^n \left[(\beta_1 - b_1^*) + (\beta_2 - b_2^*)x_i\right]^2}{\sum_{i=1}^n \left[Y_i - \beta_1 - \beta_2 x_i\right]^2} \sim \mathcal{F}(2, n-2)$$

1. *la première étape, on calcule la valeur* $f_{1-\alpha}^{2,n-2}$ *pour* $\alpha = 5\%$*,* $ddl_1 = 2$ *et* ddl_2 *(puisque on veut tester cette hypothèse au seuil de* 5% *alors* $\alpha = 5\%$*, de plus on remarque que* $Z \sim \mathcal{F}(2, n-2)$ *donc* $ddl_1 = 2$ *et* $ddl_2 = n - 2 = 13$*). En utilisant la table de la loi de Fisher (voir le fichier pdf des tableaux statistiques, page 4). on trouve :*

ddl_2/ddl_1	.	2	.....
.	.	.	.....
13	.	$f_{1-\alpha}^{2,13} = 3.806$	.....

2. *la deuxième étape : On calcule la valeur de la variable* Z

$$z = \frac{n-2}{2} \frac{\sum_{i=1}^n \left[(\hat{b}_1 - b_1^*) + (\hat{b}_2 - b_2^*)x_i\right]^2}{\sum_{i=1}^n \left[y_i - \hat{b}_1 - \hat{b}_2 x_i\right]^2}$$

On calcule $\hat{b}_1$ *et* $\hat{b}_2$ *D'après la définition 1.2, page 2. On a :*

$$\begin{cases} \hat{b}_1 = \overline{y} - \hat{b}_2 \overline{x} \\ \hat{b}_2 = \dfrac{\sum_{i=1}^n (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2} = \dfrac{Cov(x,y)}{Var(x)} \end{cases}$$

et d'après les données $(x_i, y_i)_{1 \leq i \leq 15)}$ *de cet exercice on a :*

$\overline{X} = \frac{1}{15}\sum_{i=1}^{15} x_i = 37.3333$	$\overline{Y} = \frac{1}{15}\sum_{i=1}^{15} y_i = 180.2666$
$\overline{XY} = \frac{1}{15}\sum_{i=1}^{15} x_i y_i = 6502.2666$	$\overline{X^2} = \frac{1}{15}\sum_{i=1}^{15} x_i^2 = 1679.2$
$Var(X) = \overline{X^2} - \overline{X}^2 = 285.4247$	$\sigma_x = \sqrt{Var(X)} = 16.8945$
$\overline{Y^2} = \frac{1}{15}\sum_{i=1}^{15} y_i^2 = 32695.8666$	$Var(Y) = \overline{Y^2} - \overline{Y}^2 = 199.8195$
$\sigma_Y = \sqrt{Var(Y)} = 14.1357$	$Cov(X,Y) = \overline{XY} - \overline{X}.\overline{Y} = -227.6804$

par suite

$$\begin{cases} \hat{b}_1 = \overline{y} - \hat{b}_2\overline{x} = 180.2666 - (-0.7976) \times 37.3333 = 210.0436 \\ \hat{b}_2 = \dfrac{Cov(x,y)}{Var(x)} = \dfrac{-227.6804}{285.4247} = -0.7976 \end{cases}$$

de plus on a :

$$\begin{aligned} z &= \frac{n-2}{2}\frac{\sum_{i=1}^{n}\left[(\hat{b}_1 - b_1^*) + (\hat{b}_2 - b_2^*)x_i\right]^2}{\sum_{i=1}^{n}\left[y_i - \hat{b}_1 - \hat{b}_2 x_i\right]^2} \\ &= \frac{n-2}{2}\frac{\sum_{i=1}^{n}\left[(\hat{b}_1 - b_1^*) + (\hat{b}_2 - b_2^*)x_i\right]^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \\ &= \frac{1}{2}\frac{\sum_{i=1}^{n}\left[(\hat{b}_1 - b_1^*) + (\hat{b}_2 - b_2^*)x_i\right]^2}{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \\ &= \frac{1}{2}\frac{\sum_{i=1}^{n}(y_i^*)^2}{\hat{\sigma}^2} \end{aligned}$$

avec

$$y_i^* = (\hat{b}_1 - b_1^*) + (\hat{b}_2 - b_2^*)x_i, \quad et \quad \hat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

— *Calculon $\hat{\sigma}^2$. On a :*

$$\begin{aligned} \hat{\sigma}^2 &= \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \\ &= \frac{n}{n-2}Var(Y)(1-r^2) \\ &= \frac{n}{n-2}Var(Y)\left(1 - \frac{Cov(x,y)^2}{Var(x)Var(Y)}\right) \\ &= 21.0013 \end{aligned}$$

— *Calculons $y_i^* = (\hat{b}_1 - b_1^*) + (\hat{b}_2 - b_2^*)x_i$, $\forall 1 \leq i \leq 15$ on a $\forall 1 \leq i \leq 15$:*

$$\begin{aligned} y_i^* &= (\hat{b}_1 - b_1^*) + (\hat{b}_2 - b_2^*)x_i \\ &= (210.0436 - 220) + (-0.7976 - (-1))x_i \\ &= -9.9564 + 0.2024x_i \end{aligned}$$

Par suite

x_i	*18*	*23*	*25*	*35*	*65*	*54*	*34*	*56*
y_i^*	*-6.3132*	*-5.3012*	*-4.8964*	*-2.8724*	*3.1996*	*0.9732*	*-3.0748*	*1.378*
$(y_i^*)^2$	*37.3131*	*28.1027*	*23.9747*	*8.2506*	*10.2374*	*0.9471*	*9.4543*	*1.8988*

x_i	72	19	23	42	18	39	37
y_i^*	4.6164	-6.1108	-5.3012	-1.4556	-6.3132	-2.0628	-2.4676
$(y_i^*)^2$	21.3111	37.3418	28.1027	2.1187	39.8564	4.2551	6.089

donc :

$$\sum_{i=1}^{n}(y_i^*)^2 = 259.2535$$

par suite

$$z = \frac{1}{2}\frac{\sum_{i=1}^{n}(y_i^*)^2}{\hat{\sigma}^2} = \frac{1}{2}\frac{259.2535}{21.0013} = 6.1732$$

3. *la troisième étape : On adopte la règle de décision suivante :*
   - *Si* $0 \leq z \leq f_{1-\alpha}^{2,n-2}$ *alors on accepte* $(\mathcal{H}_0)$ *au risque* α *(on accepte le modèle linéaire* $Y_i = b_1^* + b_2^* x_i + \epsilon_i$*)*
   - *Sinon on rejette l'hypothèse* $(\mathcal{H}_0)$ *et on accepte* $(\mathcal{H}_1)$ *(on n'accepte pas le modèle linéaire* $Y_i = b_1^* + b_2^* x_i + \epsilon_i$*)*

   On a :

$$z = 6.1732 > f_{1-\alpha}^{2,13} = 3.806$$

alors on rejette l'hypothèse $(\mathcal{H}_0)$ *et on accepte* $(\mathcal{H}_1)$ *au risque* $\alpha = 5\%$ *,*

on n'accepte pas le modèle linéaire $Y_i = 220 - x_i + \epsilon_i$

Exercice 4 *On envisage de prévoir la taille des œufs de coucou susceptibles d'être pondus dans un nid, à partir de ses dimensions. La variable* Y *désigne la variable "longueur d'un œufs" et la variable* X *le "diamètre d'un nid" en mm. Pour chaque œuf d'un échantillon de* $n = 16$ *œufs, on a relevé d'une part la réalisation de* Y *et d'autre part celle de* X *pour le nid où il a été trouvé. Les valeurs observées* $(x_i, y_i)_{1\leq i\leq 16}$ *du couple* (X, Y) *sont les suivantes :*

x_i	100	113	110	106	112	105	107	108
y_i	19.8	22.1	21.5	20.9	22	20.8	21.2	21

x_i	122	126	121	122	110	116	118	120
y_i	23.8	24.9	24	23.8	21.7	22.8	23.1	23.5

1. *Préciser le modèle de régression linéaire approprié pour aborder le problème de prévision posé et les hypothèses de travail nécessaires pour appliquer l'analyse de ce modèle.*
2. *Donner le coefficient de corrélation linéaire empirique entre* X *et* Y*. Interpréter.*
3. *Calculer les estimations des paramètres* (b_1, b_2, σ^2) *du modèle correspondant aux données.*
4. *Tester la signification de cette régression au seuil* 1%*.*
5. *On s'intéresse aux nids de 128 mm de diamètre et on cherche à prévoir la taille des œufs de coucou qu'on peut trouver.*
   (a) *Quelle est la valeur prédite de la longueur de tels œufs.*
   (b) *Déterminer l'intervalle de confiance pour cette moyenne au niveau de confiance* 95%*.*

Solution de l'exercice : 4 -

1. *On a un couple de variables mesuré sur un échantillon de taille $n = 16$. On cherche à connaitre la taille des œufs Y à partir du diamètre des nids X. Le but est donc d'établir un modèle reliant X et Y avec, Y la variable expliquée et X la variable explicative. On considère le modèle linéaire*

$$Y_i = b_2 xi + b_1 + \epsilon_i, \quad i = 1, ..., n,$$

où les ϵ_i, sont des variables aléatoires indépendantes de loi $\mathcal{N}(0, \sigma^2)$.

2. *On a :*

$\overline{X} = \frac{1}{16}\sum_{i=1}^{16} x_i = 113.5$	$\overline{Y} = \frac{1}{16}\sum_{i=1}^{16} y_i = 22.3062$
$\overline{XY} = \frac{1}{16}\sum_{i=1}^{16} x_i y_i = 2541.7687$	$\overline{X^2} = \frac{1}{16}\sum_{i=1}^{16} x_i^2 = 12934.5$
$Var(X) = \overline{X^2} - \overline{X}^2 = 52.25$	$\sigma_x = \sqrt{Var(X)} = 7.2284$
$\overline{Y^2} = \frac{1}{16}\sum_{i=1}^{16} y_i^2 = 499.5043$	$Var(Y) = \overline{Y^2} - \overline{Y}^2 = 1.9377$
$\sigma_Y = \sqrt{Var(Y)} = 1.392$	$Cov(X, Y) = \overline{XY} - \overline{X}.\overline{Y} = 10.015$

$$r = Corr(X, Y) = \frac{Cov(X, Y)}{\sigma_x \sigma_y} = \frac{10.015}{7.2284 \times 1.392} = 0.9953$$

la corrélation entre X et Y est très forte.

3. $$\begin{cases} \hat{b}_1 = \overline{y} - \hat{b}_2 \overline{x} = 22.3062 - 0.1916 \times 113.5 = 0.5596 \\ \hat{b}_2 = \dfrac{Cov(x, y)}{Var(x)} = \dfrac{10.015}{52.25} = 0.1916 \end{cases}$$

Donc la droite de régression de Y en fonction de X est donnée par :

$$Y = 0.5596 + 0.1916X$$

Pour $\hat{\sigma}^2$, on a :

$$\hat{\sigma}^2 = \frac{n}{n-2} Var(Y)(1 - r^2)$$

par conséquent

$$\hat{\sigma}^2 = \frac{16}{14} \times 1.9377 \times (1 - 0.9953^2) = 0.0207$$

4. *Il s'agit de tester l'hypothèse suivante :*

$$(\mathcal{H}_0) : b_2 = 0 \quad contre \quad (\mathcal{H}_1) : b_2 \neq 0$$

Alors, on sait que :

$$Y_i = b_1 + b_2 x_i + \epsilon_i$$

et

$$T = \frac{\beta_2 - b_2}{\dfrac{\Sigma}{\sqrt{nVar(x)}}} \sim \mathcal{S}t(n-2).$$

Sous $(\mathcal{H}_0)$ cette hypothèse devient

$$Y_i = b_1 + \epsilon_i$$

et

$$T = \frac{\beta_2}{\dfrac{\Sigma}{\sqrt{nVar(x)}}} \sim \mathcal{S}t(n-2).$$

Par définition de la loi de Student on sait que

$$\mathbb{P}\left[\mid T \mid \leq t_{1-\alpha/2}^{n-2}\right] = 1-\alpha.$$

Puisque on veut tester la signification de cette régression au seuil de 1%, alors $\alpha = 1\%$.

— *la première étape, on calcule* $t_{1-\alpha/2}^{n-2}$ *pour* $\alpha = 1\%$ *et degrè de liberté* $= ddl = n-2 = 14$, *en utilisant la table de la loi de Student, (voir le fichier pdf des tableaux statistique page 2), On a :*

ddl/α	.....	1%
.	.	.
14	.	$t_{1-\alpha/2}^{n-2} = 2,9768$

alors, pour $\alpha = 1\%$ *et* $n-2 = 14$ *on a,* $t_{1-\alpha/2}^{n-2} = 2,9768$

— *la deuxième étape, on calcule la valeur* $t = \dfrac{\hat{b}_2}{\dfrac{\hat{\sigma}}{\sqrt{nVar(x)}}}$ *de la variable aléatoire* T *sur les données* $(x_i, y_i)_{1\leq i\leq n}$, *on a :*

$$t = \frac{\hat{b}_2}{\dfrac{\hat{\sigma}}{\sqrt{nVar(x)}}} = \frac{0.1916 \times \sqrt{16 \times 52.25}}{0.1438} = 38.5247$$

— *la troisième étape : on adopte alors la règle de décision suivante*

— *Si* $\mid t \mid \leq t_{1-\alpha/2}^{n-2}$ *alors on accepte* $(\mathcal{H}_0)$ *au risque* α *(il n'y a pas de lien linéaire entre les deux variables* X *et* Y, *avec un risque de* α*)*

— *si* $\mid t \mid > t_{1-\alpha/2}^{n-2}$ *alors on rejette l'hypothèse* $(\mathcal{H}_0)$ *et on accepte* $(\mathcal{H}_1)$ *(il y a un lien linéaire entre les deux variables)*

Puisque, $|t| = 38.5247 > 2,9768 = t_{1-\alpha/2}^{n-2}$, *on rejette l'hypothèse* $(\mathcal{H}_0) : b_2 = 0$ *et on accepte l'hypothèse* $(\mathcal{H}_1) : b_2 \neq 0$ *au seuil de 1%, et donc il y a un lien linéaire entre* X *et* Y *au seuil de 1%.*

5. (a) *On a :* $x_{17} = 128$, *et*

$$\hat{y}_{17} = 0.5596 + 0.1916 \times x_{17},$$

par suite,

$$\hat{y}_{17} := 0.5596 + 0.1916 \times 128 = 25.0844$$

(b) *L'intervalle de confiance de niveau de confiance* $1-\alpha = 95\%$ *de* y_{17} *est donnée par :*

$$IC_{1-\alpha}(y_{17}) = \left[\hat{y}_{17} - t_{1-\alpha/2}^{n-2}\hat{\sigma}_{e_{17}}, \hat{y}_{17} + t_{1-\alpha/2}^{n-2}\hat{\sigma}_{e_{17}}\right]$$

avec

$$\hat{\sigma}_{e_{17}}^2 = \hat{\sigma}^2\left(1 + \frac{1}{16} + \frac{(x_{17}-\overline{x})^2}{\sum_{i=1}^{1} 6(x_i-\overline{x})^2}\right)$$

comme :

$$\sum_{i=1}^{n}(x_i-\overline{x})^2 = nVar(X)$$

Donc :

$$\hat{\sigma}_{e_{17}}^2 = \hat{\sigma}^2\left(1 + \frac{1}{16} + \frac{(x_{17}-\overline{x})^2}{16Var(X)}\right) = 0.0207\left(1 + \frac{1}{16} + \frac{(128-113.5)^2}{16 \times 52.25}\right) = 0.0271$$

et, pour $\alpha = 5\%$ *et* $ddl = n-2 = 14$ *on a* $t^{n-2}_{1-\alpha/2} = 2.1448$

$$\begin{aligned} IC_{95\%}(y_{17}) &= \left[\hat{y}_{17} - t^{n-2}_{1-\alpha/2}\hat{\sigma}_{e_{17}}, \hat{y}_{17} + t^{n-2}_{1-\alpha/2}\hat{\sigma}_{e_{17}}\right] \\ &= \left[25.0844 - 2.1448 \times \sqrt{0.0271}, 25.0844 + 2.1448 \times \sqrt{0.0271}\right] \\ &= [2.4573, 2.8376] \end{aligned}$$

Exercice 5 *Soient* $(x_i, y_i)_{1\leq i\leq n}$, *n mesures pour les variables* Y *et* X, *on suppose qu'il existe un lien linéaire entre* X *et* Y.

$$Y_i = b_1 + b_2 x_1 + \epsilon_i, \quad 1 \leq i \leq n \quad \epsilon_i \sim \mathcal{N}(0,1)$$

On considère x_{n+1} *une nouvelle valeur de la variable* X *d'une prévision*

$$\hat{y}_{n+1} = \hat{b}_1 + \hat{b}_2 x_{n+1}$$

Alors, la précision de cette prédiction est

$$E_{n+1} = Y_{n+1} - \hat{Y}_{n+1} = (b_1 - \beta_1) + (b_2 - \beta_2)x_{n+1} + \epsilon_{n+1} \quad \sim \quad \mathcal{N}(0, \sigma^2_{e_{n+1}}),$$

avec Y_{n+1} *et* $\hat{Y}_{n+1}$ *sont des variables aléatoires indépendantes.*

1. *Montrer que* $\sigma^2_{e_{n+1}} = \sigma^2\left(1 + \frac{1}{n} + \frac{(x_{n+1}-\overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)$.
2. *Montrer que* $\Sigma^2_{e_{n+1}} = \Sigma^2\left(1 + \frac{1}{n} + \frac{(x_{n+1}-\overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)$ *est un estimateur sans biais de* $\sigma^2_{e_{n+1}}$.
3. *Déterminer la loi de la variable* $\frac{Y_{n+1} - \hat{Y}_{n+1}}{\Sigma_{e_{n+1}}}$.
4. *Déduire que :*

$$IC_{1-\alpha}(y_{n+1}) = \left[\hat{y}_{n+1} - t^{n-2}_{1-\alpha/2}\hat{\sigma}_{e_{n+1}}, \hat{y}_{n+1} + t^{n-2}_{1-\alpha/2}\hat{\sigma}_{e_{n+1}}\right]$$

Solution de l'exercice : 5 -

1. *puisque*

$$E_{n+1} = Y_{n+1} - \hat{Y}_{n+1} \quad \sim \quad \mathcal{N}(0, \sigma^2_{e_{n+1}}),$$

alors, $Var(E_{n+1}) = \sigma^2_{e_{n+1}}$. *Donc il suffit de montrer que :*

$$Var(E_{n+1}) = \sigma^2\left(1 + \frac{1}{n} + \frac{(x_{n+1}-\overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)$$

On a :

$$\begin{aligned} Var(E_{n+1}) &= Var(Y_{n+1} - \hat{Y}_{n+1}) \\ &= Var(Y_{n+1}) + Var(\hat{Y}_{n+1}) - 2Cov(Y_{n+1}, \hat{Y}_{n+1}) \end{aligned}$$

Puisque, si X *et* Y *deux variable aléatoires et si* α *et* β *deux réels on a :*

$$Var(\alpha X + \beta Y) = \alpha^2 Var(X) + \beta^2 Var(Y) + 2\alpha\beta Cov(X,Y). \tag{1}$$

Par suite, puisque Y_{n+1} *et* $\hat{Y}_{n+1}$ *sont des variables aléatoires indépendantes, alors*

$$Cov(Y_{n+1}, \hat{Y}_{n+1}) = 0$$

Donc

$$\begin{aligned} Var(E_{n+1}) &= Var(Y_{n+1}) + Var(\hat{Y}_{n+1}) \\ &= Var(b_1 + b_2 x_{n+1} + \epsilon_{n+1}) + Var(\beta_1 + \beta_2 x_{n+1}) \end{aligned}$$

de même, d'après l'équation (1)*, on a :*

$$Var(E_{n+1}) = Var(\epsilon_{n+1}) + Var(\beta_1) + x_{n+1}^2 Var(\beta_2) + 2x_{n+1} Cov(\beta_1, \beta_2)$$

D'après la proposition 1.6, page 6 (Voir Chapitre 01) et d'après l'hypothèse de normalité, page 2. on a :

$$\begin{aligned} Var(E_{n+1}) &= \sigma^2 + \frac{\sigma^2}{n}\frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \overline{x})^2} + x_{n+1}^2 \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2} - 2x_{n+1}\frac{\sigma^2 \overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2} \\ &= \sigma^2 \left(1 + \frac{1}{n}\frac{\sum_{i=1}^n x_i^2}{\sum_{i=1}^n (x_i - \overline{x})^2} + \frac{x_{n+1}^2}{\sum_{i=1}^n (x_i - \overline{x})^2} - 2\frac{x_{n+1}\overline{x}}{\sum_{i=1}^n (x_i - \overline{x})^2}\right) \\ &= \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right) \end{aligned}$$

2. *D'après la définition 1.4, page 4 on montre que*

$$\mathbb{E}(\Sigma_{e_{n+1}}^2) = \sigma_{e_{n+1}}^2$$

On a :

$$\begin{aligned} \mathbb{E}(\Sigma_{e_{n+1}}^2) &= \mathbb{E}\left(\Sigma^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)\right) \\ &= \mathbb{E}(\Sigma^2)\left(1 + \frac{1}{n} + \frac{(x_{n+1} - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right) \end{aligned}$$

D'après la proposition 1.8, page 7 on a : Σ^2 est un estimateur sans biais de σ^2, alors

$$\mathbb{E}(\Sigma^2) = \sigma^2$$

par suite :

$$\begin{aligned} \mathbb{E}(\Sigma_{e_{n+1}}^2) &= \sigma^2 \left(1 + \frac{1}{n} + \frac{(x_{n+1} - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right) \\ &= \sigma_{e_{n+1}}^2 \end{aligned}$$

3. *on a :*

$$Y_{n+1} - \hat{Y}_{n+1} \sim \mathcal{N}(0, \sigma_{e_{n+1}}^2),$$

alors

$$\frac{Y_{n+1} - \hat{Y}_{n+1}}{\sigma_{e_{n+1}}} \sim \mathcal{N}(0, 1) \tag{2}$$

et d'après la proposition 1.11 page 8, (voir chapitre 1 cours)

$$(n-2)\frac{\Sigma^2}{\sigma^2} \sim \mathcal{X}^2(n-2)$$

par suite

$$(n-2)\frac{\Sigma^2\left(1 + \frac{1}{n} + \frac{(x_{n+1} - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)}{\sigma^2\left(1 + \frac{1}{n} + \frac{(x_{n+1} - \overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\right)} \sim \mathcal{X}^2(n-2)$$

donc

$$(n-2)\frac{\Sigma^2_{e_{n+1}}}{\sigma^2_{e_{n+1}}} \sim \mathcal{X}^2(n-2) \tag{3}$$

Finalement, d'après la définition 1.8, page 9 et d'après l'équation (2) et l'équation (3) on a :

$$\frac{Y_{n+1}-\hat{Y}_{n+1}}{\Sigma_{e_{n+1}}} \sim \mathcal{S}t(n-2)$$

4. *D'après la réponse de la question 3. On a :*

$$\frac{Y_{n+1}-\hat{Y}_{n+1}}{\Sigma_{e_{n+1}}} \sim \mathcal{S}t(n-2),$$

et d'après la densité de la loi de $\mathcal{S}t(n-2)$, voir figure 1 on a :

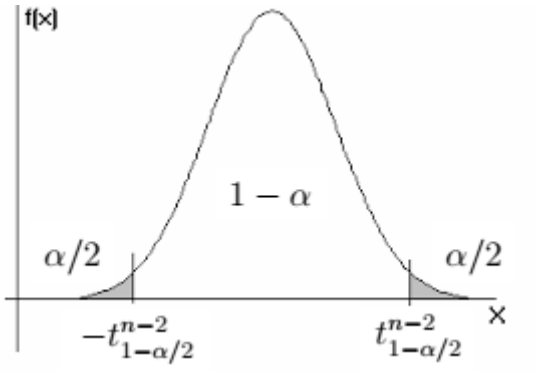

FIGURE 2 – Densité de la loi de $\mathcal{S}t(n-2)$

$$\mathbb{P}\left(-t^{n-2}_{1-\alpha/2} \leq \frac{Y_{n+1}-\hat{Y}_{n+1}}{\Sigma_{e_{n+1}}} \leq t^{n-2}_{1-\alpha/2}\right) = 1-\alpha$$

alors,

$$\mathbb{P}\left(\hat{Y}_{n+1} - \Sigma_{e_{n+1}} t^{n-2}_{1-\alpha/2} \leq Y_{n+1} \leq \hat{Y}_{n+1} + \Sigma_{e_{n+1}} t^{n-2}_{1-\alpha/2}\right) = 1-\alpha$$

Par conséquent

$$IC_{1-\alpha}(y_{n+1}) = \left[\hat{y}_{n+1} - t^{n-2}_{1-\alpha/2}\hat{\sigma}_{e_{n+1}}, \hat{y}_{n+1} + t^{n-2}_{1-\alpha/2}\hat{\sigma}_{e_{n+1}}\right]$$

Exercice 6 *Montrer que :*

1. $$\Sigma^2 := \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \frac{n}{n-2}Var(Y)(1-r^2),$$

avec r est le coefficient de corrélation.

2. $$T^2 = \frac{nVar(x)\beta_2^2}{\Sigma^2} = (n-2)\frac{R^2}{1-R^2},$$

avec R^2 est le coefficient de détermination

3.

$$R^2 = r^2$$

Solution de l'exercice : 6 *1. On a :*

$$\begin{aligned}
\Sigma^2 &= \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \\
&= \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \\
&= \frac{n}{n-2} \times \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \times \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \\
&= \frac{n}{n-2} Var(Y)(1 - r^2)
\end{aligned}$$

on a :

—

$$var(Y) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2.$$

— *D'après l'équation de la décomposition de la dispersion de Y (sous section 1.8.1 page 12 chapitre 1), on a :*

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 - \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2$$

— *d'après la définition 1.10 page 13 chapitre 1 de la coefficient de détermination, on a :*

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}$$

— *On accepte que*

$$R^2 = r^2$$

d'après cet exercice.

finalement,

$$\begin{aligned}
\Sigma^2 &= \frac{n}{n-2} \times \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \times \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \\
&= \frac{n}{n-2} \times \frac{1}{n}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \times \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2 - \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2}{\sum_{i=1}^{n}(Y_i - \overline{Y})^2} \\
&= \frac{n}{n-2} Var(Y)(1 - R^2) \\
&= \frac{n}{n-2} Var(Y)(1 - r^2)
\end{aligned}$$

2.

$$\begin{aligned}
T^2 &= \frac{nVar(x)\beta_2^2}{\Sigma^2} \\
&= \frac{nVar(x)\left(\frac{Cov(x,Y)}{Var(x)}\right)^2}{\frac{n}{n-2}Var(Y)(1-R^2)} \\
&= (n-2)\frac{Cov(x,Y)^2}{Var(x)Var(Y)}\frac{1}{1-R^2} \\
&= (n-2)\left(\frac{Cov(x,Y)}{\sigma_x\sigma_Y}\right)^2\frac{1}{1-R^2} \\
&= (n-2)\times r^2 \times \frac{1}{1-R^2} \\
&= (n-2)\frac{R^2}{1-R^2}
\end{aligned}$$

3.

$$\begin{aligned}
R^2 &= \frac{\sum_{i=1}^n(\hat{Y}_i-\overline{Y})^2}{\sum_{i=1}^n(Y_i-\overline{Y})^2} \\
&= \frac{\sum_{i=1}^n(\hat{b}_1+\hat{b}_2x_i-(\hat{b}_1+\hat{b}_2\overline{x}))^2}{nVar(Y)} \\
&= \frac{\sum_{i=1}^n\hat{b}_2^2(x_i-\overline{x})^2}{nVar(Y)} \\
&= \hat{b}_2^2\frac{\sum_{i=1}^n(x_i-\overline{x})^2}{nVar(Y)} \\
&= \hat{b}_2^2\frac{nVar(x)}{nVar(Y)} \\
&= \left(\frac{Cov(x,Y)}{Var(x)}\right)^2\frac{Var(x)}{Var(Y)} \\
&= \left(\frac{Cov(x,Y)}{\sigma_x\sigma_Y}\right)^2 \\
&= r^2
\end{aligned}$$