Département d'Informatique **Module : Analyse des Données**

TD N 01 (Régression Linéaire Simple)

Exercice 1. Une enquète auprès de 10 familles révèle la relation suivante entre le revenu monsuelle X est la consommation Y :

$X\left(\times 10^4 DA\right)$:	2	3	3.5	4	5	6	7	9	11.5	14
$Y\left(\times 10^4 DA\right)$:	1	2.5	2.7	2.9	3.6	3.5	6.8	6.7	8.2	9.7

1) Présenter le nuage des point (X, Y). En déduire la forme du modèle.
2) Pour estimer les paramètre du modèle nous utilisant les résultats du tableau suivant :

	x	y	x^2	y^2	xy	$\widehat{y}$	$\widehat{\varepsilon}$	$\widehat{\varepsilon}^2$
	2	1						
	:	:						
$\sum =$								

i) Compléter le tableau.
ii) Proposez des estimateurs pour les paramètres du modèle et la variance résiduelle σ^2.
iii) Etudier la signification et donner des intervalles de confiance des paramètres du modèle.
iv) Calculer la valeur du coefficient de détermination puis testez la validité du modèle.
v) Schant qu'une famille à un revenu qui vaut $6 \times 10^4 DA$, quelle est la valeur prédite de sa consommation monsuelle? Même question pour un revenu de $1.8 \times 10^4 DA$?

Exercice 2. La distribution suivante montre une relation linéaire simple entre l'évolution du taux de croissance économique Y(en %) et le prix de pétrole X(en \$) pendant 10 ans :

$$y_j = ax_j + b + \varepsilon_j, \quad j = 1, ..., 10 \text{ et } \varepsilon \sim N\left(0, \sigma_\epsilon^2\right).$$

	x	y	$(x-\overline{x})$	$(y-\overline{y})$	$(x-\overline{x})(y-\overline{y})$	$(x-\overline{x})^2$	$(y-\overline{y})^2$	$\hat{y}$	$\hat{\varepsilon}$	$\hat{\varepsilon}^2$
	111	1.9								
	120	4.4								
	133	5.6								
	105	4.1								
	101	3.2								
	95	-3								
	82	-2								
	70	0								
	66	-2.9								
	54	-3								
$\sum =$										

1) Donner la représentation graphique des données.
2) Completer le tableau et estimer les paramètres du modèle.
3) Calculer la valeur de $\hat{\sigma}_\epsilon^2$.
4) Calculer les quantités : $\hat{\sigma}_1^2$ *et* $\hat{\sigma}_0^2$ (i.e., estimateurs des variances de a et b resp.).
5) Testez la signification des paramètres du modèle au niveau 95%.
6) Déterminer l'intervalle de confiance à 95% des paramètres a et b.
7) Testez la validité du modèle.
8) Quelle quantité mesure la bonne explication de la variable à expliquer par la variable explicative? Que vaut-elle ici ? Est-ce une valeur satisfaisante ?
9) Donnez le taux de croissance lorsque le prix du pétrole est 130\$?
10) Considérons le cas du prix du pétrole 100\$, estimer la valeur correspondante du taux de croissance.

TD N 02 - 2018 : Régression Linéaire Multiple

Exercice 1. Dans une étude de régression multiple comportant quatre variables explicatives, nous avons obtenu le tableau d'AV pour 19 observations, comme suit:

Variation	*ddl*	*SC*	*MC*	*F*
Factorielle				
Résiduelle		2053		
Total		7654		

1) Compléter le tableau et tester la signification globale du modèle 95%.

2) Calculer le coefficient de détermination et $\hat{\sigma}_{\epsilon}^2$ du modèle.

Exercice 2. L'équation estimée ci-dessous a été obtenue par les MCO en utilisant des données trimestrielle sur la période de 2014 à 2017 (bornes inclues):

$$\widehat{Y}_t = \underbrace{2,23}_{(1.3)} + \underbrace{1,14}_{(0.05)} X_{t1} - \underbrace{3,48}_{(2.2)} X_{t2} + \underbrace{0,34}_{(0.15)} X_{t3}.$$

Les chifres entre parenthèses sont les écarts-types estimés. La SCE et la SCR étaient respectivement $109,16$ et $18,48$.
Calculez le coefficient de détermination et testez la significativité de chaque coefficient.

Exercice 3. Considérons la régression multiple avec constante et deux variables explicatives U et V :

$$Y_j = a + bU_j + cV_j + \varepsilon_j, \quad j = 1, ..., n \qquad \varepsilon \sim N\left(0, \sigma_\epsilon^2\right)$$

et soient les résultats empiriques :

$$X^tY = \begin{pmatrix} 11,5 \\ 24,8 \\ 17.3 \end{pmatrix}, \quad X^tX = \begin{pmatrix} 30 & 0 & 0 \\ 0 & 20 & 5 \\ 0 & 5 & 15 \end{pmatrix} \quad et \quad \sum_{i=1}^{n} \widehat{\varepsilon}_i^2 = 15,7.$$

En vous donne la matrice inverse $\left(X^tX\right)^{-1} = \begin{pmatrix} 0.033 & 0 & 0 \\ 0 & 0.545 & -0.018 \\ 0 & -0.018 & 0.0727 \end{pmatrix}$.

1) Ecrire le modèle sous forme matricielle, en déduire la taille de l'échantillon n ainsi que les moyenne empiriques : $\overline{U}$, $\overline{V}$.

2) Estimer les paramètres du modèle : $A = (a, b, c)^t$.

3) Calculer $\hat{\sigma}_{\epsilon}^2$.

4) Calculer les variances estimes de chacun des coefficients : $\hat{\sigma}_{bj}^2$, $\quad j = 0, 1, 2, 3$.

5) Donner les valeurs des covariances :

$$cov\left(\widehat{b}, \widehat{a}\right), \quad cov\left(\widehat{a}, \widehat{c}\right) \quad et \quad cov\left(\widehat{b}, \widehat{c}\right).$$

6) La variable explicative U est-elle à 90% significativement contributive dans l'explication de Y ?

7) Le coefficient c est-il à 90% significativement différent de 1 ?

8) Sachant que : $U_h = 13.3$, $V_h = 16.6$. Calculer la prévision $\widehat{Y}_h$ et son intervalle de confiance à 95%.

<u>Exercice 4.</u> On s'intéresse à l'effet du nombre d'années d'études des parents (M : *mère* et P : *père*) sur le nombre d'années d'études de leurs enfants noté Y. On dispose du nombre d'années d'études de 26 familles (enfant, mère et père).

I) Nous proposons premièrement le modèle linéaire suivant :

$$y_i = aM_i + bP_i + \varepsilon_i, \quad \varepsilon \sim N\left(0, \sigma_\epsilon^2\right), \qquad i = 1, 2, ..., 26.$$

On vous donne les sommes suivantes:

$$\sum M_i^2 = 288, \qquad \sum P_i^2 = 202, \qquad \sum M_i P_i = 144, \qquad \sum y_i M_i = 184, \qquad \sum y_i P_i = 158.$$

1) Ecrire ce modèle sous forme matricielle : $Y = XA + \varepsilon$, en précisant Y, X et A.

2) Donnez les matrices (X^tX) et X^tY à l'aide des valeurs numériques des sommes ci-dessus.

3) Calculez l'estimateur $\widehat{A}$ de A.

4) Calculez $\widehat{\sigma}_\epsilon^2$, sachant que $\sum \widehat{\varepsilon}_i^{\,2} = 36$.

5) Calculez les variances estimes de chacun des coefficients : $\widehat{\sigma}_a^2$ et $\widehat{\sigma}_b^2$.

6) Tester la signification des paramètres a et b à 90%.

7) Donner un intervalle de confiance à 90% des paramètres a et b.

8) Le nombre d'années d'études des mères à un effet sur celui de leur enfants ?

9) Le nombre d'années d'études des pères à un effet sur celui de leur enfants ?

10) Quel est le nombre d'années d'études prévu pour un enfant, sachant que le nombre d'années d'études de ces parents a été 17 ans pour le père et 15 ans pour la mère ?

II) On décide d'ajuster ces données par un modèle linéaire avec constante:

$$y_i = c + aM_i + bP_i + \varepsilon_i, \quad \varepsilon \sim N\left(0, \sigma_\epsilon^2\right), \qquad i = 1, 2, ..., 26$$

et soient les résultats empiriques :

$$X^tY = \begin{pmatrix} 232 \\ 2691 \\ 2650 \end{pmatrix}, \quad X^tX = \begin{pmatrix} 26 & 187 & 178 \\ 187 & 2425 & 1955 \\ 178 & 1955 & 2260 \end{pmatrix} \quad \text{et} \quad \left(X^tX\right)^{-1} = \begin{pmatrix} 0.09 & -0.0004 & -0.004 \\ -0.0004 & 0.0015 & -0.001 \\ -0.004 & -0.001 & 0.0016 \end{pmatrix}.$$

1) Ecrire ce modèle sous forme matricielle : $Y = XA + \varepsilon$, en précisant Y, X et A.

2) Que représente le paramètre c dans cette dexième partie ?

3) Calculer l'estimateur $\widehat{A}$ de A.

4) Calculez $\widehat{\sigma}_\epsilon^2$, sachant que $\sum \widehat{\varepsilon}_i^{\,2} = 34$.

5) Le nombre d'années d'études des mères à un effet sur celui de leur enfants ?

6) Le nombre d'années d'études des pères à un effet sur celui de leur enfants ?

7) Tester l'hypothèse $H_0 : c = 12$.

8) Quel est le nombre d'années d'études prévu pour un enfant, sachant que le nombre d'années d'études de ces parents a été 17 ans pour le père et 15 ans pour la mère ?

Département d'Informatique **Module : Analyse des Données**

Interrogation (08/05/2018)

Exercice (10 pts): Considérons une régression multiple avec constante et deux variables explicatives :

$$y_i = c + ax_{1i} + bx_{2i} + \varepsilon_i, \quad \varepsilon \sim N\left(0, \sigma_\epsilon^2\right), \qquad i = 1, 2, ..., 26$$

et soient les résultats empiriques :

$$X^tY = \begin{pmatrix} 232 \\ 2691 \\ 2650 \end{pmatrix}, \quad X^tX = \begin{pmatrix} 26 & 187 & 178 \\ 187 & 2425 & 1955 \\ 178 & 1955 & 2260 \end{pmatrix} \text{ et } \left(X^tX\right)^{-1} = \begin{pmatrix} 0.09 & -0.0004 & -0.004 \\ -0.0004 & 0.0015 & -0.001 \\ -0.004 & -0.001 & 0.0016 \end{pmatrix}.$$

1) Ecrire le modèle sous forme matricielle : $Y = XA + \varepsilon$, en précisant Y, X et A.

2) Calculer l'estimateur $\widehat{A}$ de A.

3) Calculer $\widehat{\sigma}_\epsilon^2$, sachant que $\sum \widehat{\varepsilon_i}^2 = 34$.

4) La variable x_1 à un effet sur Y ?

5) Donner un intervalle de confiance du paramètre b.

6) Tester l'hypothèse $H_0 : c = 11$?

7) Tester l'hypothèse $H_0 : a + b = 2$?

8) Donner une estimation de la prévision de y_h sachant que $x_{1h} = 10$ et $x_{2h} = 20$.

Département d'Informatique **Module : Analyse des Données**

Interrogation (08/05/2018)

Exercice (10 pts): Considérons une régression multiple avec constante et deux variables explicatives :

$$y_i = c + ax_{1i} + bx_{2i} + \varepsilon_i, \quad \varepsilon \sim N\left(0, \sigma_\epsilon^2\right), \qquad i = 1, 2, ..., 26$$

et soient les résultats empiriques :

$$X^tY = \begin{pmatrix} 232 \\ 2691 \\ 2650 \end{pmatrix}, \quad X^tX = \begin{pmatrix} 26 & 187 & 178 \\ 187 & 2425 & 1955 \\ 178 & 1955 & 2260 \end{pmatrix} \text{ et } \left(X^tX\right)^{-1} = \begin{pmatrix} 0.09 & -0.0004 & -0.004 \\ -0.0004 & 0.0015 & -0.001 \\ -0.004 & -0.001 & 0.0016 \end{pmatrix}.$$

1) Ecrire le modèle sous forme matricielle : $Y = XA + \varepsilon$, en précisant Y, X et A.

2) Calculer l'estimateur $\widehat{A}$ de A.

3) Calculer $\widehat{\sigma}_\epsilon^2$, sachant que $\sum \widehat{\varepsilon_i}^2 = 34$.

4) La variable x_1 à un effet sur Y ?

5) Donner un intervalle de confiance du paramètre b.

6) Tester l'hypothèse $H_0 : c = 11$?

7) Tester l'hypothèse $H_0 : a + b = 2$?

8) Donner une estimation de la prévision de y_h sachant que $x_{1h} = 10$ et $x_{2h} = 20$.

Université de Biskra **Département d'Informatique**
Module : Analyse des Données **Master 1 - SIOD**

EMD - S2 - (21/05/2018 - 11h00-12h30)

Exercice 1 (06 pts): L'étude statistique ci-dessous porte sur les poids respectifs des Pères (P) et de leur Fils (F). Voici les résultats numériques que nous avons obtenus :

$$\sum P_i = 800, \qquad \sum P_i^2 = 53418, \qquad \sum P_iF_i = 54107, \qquad \sum F_i = 811, \qquad \sum F_i^2 = 54849.$$

1) Calculer les estimateurs des paramètres du modèle du poids des Fils en fonction du poids des Pères:

$$F_j = a_1 P_j + b_1 + \varepsilon_j, \qquad j = 1, ..., 12$$

2) Calculer les estimateurs des paramètres du modèle du poids des Pères en fonction du poids des Fils:

$$P_j = a_2 F_j + b_2 + \varepsilon_j, \qquad j = 1, ..., 12$$

3) Montrer que le produit des pentes $\hat{a}_1$ et $\hat{a}_2$ des deux droites est égal au coefficient de détermination.

Exercice 2 (08 pts): Considérons la régression multiple sans constante suivante :

$$Y_j = aX_{1j} + bX_{2j} + \varepsilon_j, \quad j = 1, ..., 25 \qquad \varepsilon \sim N\left(0, \sigma_\epsilon^2\right)$$

et soient les résultats empiriques :

$$X^tY = \begin{pmatrix} 28,2 \\ 15.7 \end{pmatrix}, \quad X^tX = \begin{pmatrix} 20 & 5 \\ 5 & 15 \end{pmatrix}, \quad \left(X^tX\right)^{-1} = \frac{1}{55}\begin{pmatrix} 3 & -1 \\ -1 & 4 \end{pmatrix} \quad et \quad \sum_{i=1}^{n} \widehat{\varepsilon}_i^2 = 15,7.$$

1) Donner une estimation des paramètres a *et* b du modèle.

2) Calculer les variances estimes : $\hat{\sigma}_\epsilon^2, \quad \hat{\sigma}_a^2 \quad et \quad \hat{\sigma}_b^2$.

3) La variable X_2 à un effet sur Y ?

4) Donner un intervalle de confiance du paramètre a.

5) Tester l'hypothèse $H_0 : a + b = 2$?

6) Donner une estimation de la prévision de Y_h sachant que $X_{1h} = 1,4$ et $X_{2h} = 2,2$.

Exercice 3 (06 pts): Considérons les résultats d'AV (2) :

Variation	ddl	SC	MC	F
Fac. A :				
Fac. B :	5		112.5	
Résiduelle:		1250.3		
Total:	47	2018		

1) Compléter le tableau.

2) S'agit-il d'une AV (2) avec ou sans répétitions ? Justifier.

3) Combien de niveaux dans le facteur A, même chose pour B ? et quelle est la taille n de l'échantillon ?

Corrigé Type de l'Examen (1) du Module : Analyse des Données –Master 1 - SIOD - 2018

Exercice 1 (06 pts): On a : $\bar{F} = 67.583, \quad \bar{P} = 66.667$.
1) Modèle du poids des Fils en fonction du poids des Pères: $F_j = a_1 P_j + b_1 + \varepsilon_j$,

$$\hat{a}_1 = \frac{\sum P_i F_i - n\bar{F}\bar{P}}{\sum P_i^2 - n\bar{P}^2} = 0.48 \text{ et } \hat{b}_1 = \bar{F} - \hat{a}_1 \bar{P} = 35.58 \qquad (2\text{pts})$$

2) Modèle du poids des Pères en fonction du poids des Fils: $P_j = a_2 F_j + b_2 + \varepsilon_j$,

$$\hat{a}_2 = \frac{\sum P_i F_i - n\bar{P}\bar{F}}{\sum F_i^2 - n\bar{F}^2} = 1.02 \text{ et } \hat{b}_2 = \bar{P} - \hat{a}_2 \bar{F} = -2.27 \qquad (2\text{pts})$$

3) Coefficient de détermination : $D = \frac{S_{FP}^2}{S_F^2 S_P^2} = \frac{S_{FP}}{S_F^2}\frac{S_{PF}}{S_P^2} = \hat{a}_1 \times \hat{a}_2 = 0.48 \times 1.02 = 48.96\%$. $(2pts)$

Exercice 2 (08 pts): Considérons la Reg Mult : $Y_j = aX_{1j} + bX_{2j} + \varepsilon_j$
1) Estimation des paramètres a et b :

$$\hat{A} = \left(X^t X\right)^{-1} X^t Y = \begin{pmatrix} \hat{a} \\ \hat{b} \end{pmatrix} = \frac{1}{55}\begin{pmatrix} 3 & -1 \\ -1 & 4 \end{pmatrix}\begin{pmatrix} 28,2 \\ 15.7 \end{pmatrix} = \begin{pmatrix} 1.24 \\ 0.63 \end{pmatrix} \qquad (2\text{pts})$$

2) Variances estimes :

$$\hat{\sigma}_\epsilon^2 = \frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{n-2} = \frac{15.7}{23} = 0.68, \qquad (0.5\text{pt})$$
$$Var(A) = \hat{\sigma}_\epsilon^2 \left(X^t X\right)^{-1} = \begin{pmatrix} 0.04 & -0.012 \\ -0.012 & 0.05 \end{pmatrix}, \text{ donc } \hat{\sigma}_a^2 = 0.04 \ \ et \ \ \hat{\sigma}_b^2 = 0.05 \ (1.5\text{pt})$$

3) La variable X_2 à un effet sur Y ? Il s'agit de tester $H_0 : b = 0$?

$$\text{En effet, } T_b = \frac{\hat{b}}{\hat{\sigma}_b} = \frac{0.63}{\sqrt{0.05}} = 2.817 > t_{.975}(23) = 2.07 \qquad (0.5\text{pt})$$

Alors H_0 est rejeté et a variable X_2 à un effet sur Y à 95%. $(0.5pt)$
4) Intervalle de confiance du paramètre a : $\quad a = \hat{a} \pm \hat{\sigma}_a t_{.975}(23) \rightarrow a = 1.24 \pm 0.41$. $(1pt)$

5) Test de l'hypothèse $H_0 : a + b = 2$

$$T_{a+b} = \frac{(a+b) - \left(\hat{a} + \hat{b}\right)}{\hat{\sigma}_{a+b}} = \frac{2 - 1.87}{\sqrt{0.04 + 0.05 - 2(0.012)}} = 0.506 < t_{.975}(23) = 2.07 \qquad (1\text{pt})$$

Alors H_0 est accepté et $(a + b = 2)$ à 95%. $(0.5pt)$
6) Prévision de Y_h : $\quad \hat{Y}_h = 1.24X_{1h} + 0.63X_{2h} = 3.12$. $(0.5pt)$

Exercice 3 (06 pts): **1)** Tableau d'AV (2): $(2pt)$

Tableau d'AV (2)

Variation	ddl	SC	MC	F
Fac. A	7	205.2	29.31	0.819
Fac. B	5	562.5	112.5	3.14
Résiduelle	35	1250.3	35.72	
Total	47	2018		

Université de Biskra **Département d'Informatique**

Module : Analyse des Données **Master 1 - SIOD**

EMD - S2 - (31/05/2018 - 10h00-11h30)

Exercice 1 (07 pts): L'étude statistique ci-dessous porte sur les poids respectifs des Pères (P) et de leur Fils (F). Voici les résultats numériques que nous avons obtenus :

$$\bar{F} = 67.583, \quad \bar{P} = 66.667 \qquad \sum \left(P_i - \bar{P}\right)^2 = 84.133, \qquad \sum \left(P_i F_i - \bar{P}\bar{F}\right) = 40.33, \qquad \sum \left(F_i - \bar{F}\right)^2 = 39.457$$

1) Calculer les estimateurs des paramètres du modèle du poids des Pères en fonction du poids des Fils:

$$P_j = a_1 F_j + b_1 + \varepsilon_j, \qquad j = 1, ..., 12$$

2) Calculer les estimateurs des paramètres du modèle du poids des Fils en fonction du poids des Pères:

$$F_j = a_2 P_j + b_2 + \varepsilon_j, \qquad j = 1, ..., 12$$

3) Montrer que le produit des pentes $\hat{a}_1$ et $\hat{a}_2$ des deux droites est égal au coefficient de détermination.

Exercice 2 (07 pts): Considérons le modèle estimé d'une régression multiple :

$$\hat{y}_j = \underset{\hat{a}}{0.65}\ \hat{x}_j \ + \ \underset{\hat{b}}{0.42}\ \hat{z}_j, \qquad j = 1, ..., 26.$$

Nous avons obtenu les résultats suivants:

$$Var\left(\hat{a}\right) = 0.84, \qquad Var\left(\hat{b}\right) = 0.15, \qquad Cov\left(\hat{a}, \hat{b}\right) = -0.03, \qquad \sum \hat{\varepsilon}_j^2 = 2.4.$$

1) Ecrire le modèle sous forme matricielle et construire la matrice inverse $\left(X^t X\right)^{-1}$.

2) Sous hypothèse de normailté des erreurs. Tester à 90% les hypothèses suivantes :

$$i)\ H_0 : a = 0\ ? \quad ii)\ H_0 : b = 0.5\ ? \quad iii)\ H_0 : a + b = 2\ ?$$

3) Donner un intervalle de confiance à 90% des paramètres a et b.

Exercice 3 (06 pts): Dans le but de déterminer si un produit C est efficace par rapport aux produits $A\ et\ B$. Nous obtenons les résultats empiriques suivants:

A:	1.0	0.2	1.3	0.5		
B:	2.7	3.1	2.6	2.8	3.2	0.7
C:	2.3	2.7	1.9	3.2		

Sachant que l'hypothèse nulle d'équivalence des produits est rejetée et que la somme des carrés des erreurs est $11,02$. Testez l'efficacité du nouveau produit C comparée au deux autres.

Corrigé Type de l'Examen (2)du Module : Analyse des Données –Master 1 - SIOD - 2018

Exercice 1 (07 pts): On a : $\bar{F} = 67.583, \quad \bar{P} = 66.667$.
1) Modèle du poids des Pères en fonction du poids des Fils: $P_j = a_1 F_j + b_1 + \varepsilon_j$,

$$\hat{a}_1 = \frac{\sum\left(P_i F_i - \bar{P}\bar{F}\right)}{\sum\left(F_i - \bar{F}\right)^2} = 1.02 \text{ et } \hat{b}_1 = \bar{P} - \hat{a}_1 \bar{F} = -2.27 \qquad (2\text{pts})$$

2) Modèle du poids des Fils en fonction du poids des Pères: $F_j = a_2 P_j + b_2 + \varepsilon_j$,

$$\hat{a}_2 = \frac{\sum\left(P_i F_i - \bar{P}\bar{F}\right)}{\sum\left(P_i - \bar{P}\right)^2} = 0.48 \text{ et } \hat{b}_2 = \bar{F} - \hat{a}_2 \bar{P} = 35.58 \qquad (2\text{pts})$$

3) Coefficient de détermination : $D = \frac{S_{FP}^2}{S_F^2 S_P^2} = \frac{S_{FP}}{S_F^2}\frac{S_{PF}}{S_P^2} = \hat{a}_1 \times \hat{a}_2 = 0.48 \times 1.02 = 48.96\%$. $(3pts)$

Exercice 2 (07 pts): Considérons la Reg Mult : $Y_j = aX_{1j} + bX_{2j} + \varepsilon_j$.
1) Forme matricielle et matrice $\left(X^t X\right)^{-1}$;

$$Y = XA + \varepsilon \rightarrow \begin{pmatrix} y_1 \\ \vdots \\ y_{26} \end{pmatrix} = \begin{pmatrix} x_1 & z_1 \\ \vdots & \vdots \\ x_{26} & z_{26} \end{pmatrix} + \begin{pmatrix} a \\ b \end{pmatrix} \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_{26} \end{pmatrix} \qquad (1\text{pt})$$

$$\hat{\sigma}_\epsilon^2 = \frac{\sum_{i=1}^n \widehat{\varepsilon}_i^2}{n-2} = 0.1 \qquad \left(X^t X\right)^{-1} = \frac{Var\left(A\right)}{\hat{\sigma}_\epsilon^2} = \begin{pmatrix} 8.4 & -0.3 \\ -0.3 & 1.5 \end{pmatrix} \qquad (1\text{pt})$$

2) Tests :
i) $H_0 : a = 0$? on a

$$T_a = \frac{\hat{a}}{\hat{\sigma}_a} = 0.77 \lessdot t_{.95}\left(24\right) = 1.71$$

Alors H_0 est acceptée ($a = 0$ à 90%) $(01pt)$
ii) $H_0 : b = 0.5$? on a

$$T_b = \frac{\hat{b} - 0.5}{\hat{\sigma}_b} = 0.2 \lessdot t_{.95}\left(24\right) = 1.71$$

Alors H_0 est acceptée ($b = 0.5$ à 90%) $(01pt)$
iii) $H_0 : a + b = 2$? on a

$$T_{a+b} = T_{a+b} = \frac{(a+b) - 2}{\hat{\sigma}_{a+b}} = 0.96 \lessdot t_{.95}\left(24\right) = 1.71$$

Alors H_0 est accepté et $(a + b = 2)$ à 90%. $(02pt)$
3) IC à 90% de a et de b :

$$a = \hat{a} \pm \hat{\sigma}_a t_{.95}\left(24\right) \rightarrow a = 0.65 \pm 1.56. \qquad (0.5\text{pt})$$

$$b = \hat{b} \pm \hat{\sigma}_b t_{.95}\left(24\right) \rightarrow b = 0.42 \pm 0.66. \qquad (0.5\text{pt})$$

Exercice 3 (06 pts): La variance résiduelle est $MCR = \hat{\sigma}_\epsilon^2 = 1$
1) Test A vs C : on a

$$T_{AC} = \frac{\left|\bar{y}_A - \bar{y}_C\right|}{\hat{\sigma}_\epsilon \sqrt{1/n_A + 1/n_C}} = 2.5 \lessdot t_{.99}\left(11\right) = 2.718$$

Alors H_0 est accepté [illegible] à 95% $(02pt)$

Université de Biskra
Département de Mathématiques

2017/2018
1^er A. Master

Examen de Rattrapage (Module : Séries temporelles)

Exercice 1 (08 pts) : Répondre par **Vrai** ou **Faux** sans justification:

1) 41 est le $5^{ème}$ élément de la sérié : 3, 8, 16, 17.

2) 101 est le $12^{ème}$ élément de la sérié : 4, 10, 18, 26.

3) Si $X_t = (a+bt)^2 + \varepsilon_t$, alors, $\left(1 - L^2\right) X_t$ est un modèle stationnaire.

4) Si $X_t = (a+bt)^2 + \varepsilon_t$, alors, $(1-L)^2 X_t$ est un modèle stationnaire.

5) Le modèle $X_t - X_{t-2} = \varepsilon_t - \varepsilon_{t-1}$ est une marche aléatoire.

6) Si $X_t = 6 + X_{t-1} + \varepsilon_t$, alors $E[X_t] = 6$.

7) Le modèle : $X_t = 2t + (1 - 0.5L)\,\varepsilon_{t-1}$ est inversible.

8) $(1 - \alpha L) X_t = \left(1 - \frac{1}{\alpha} L\right) \varepsilon_t$ est un bruit blanc ($\alpha < 1$).

Exercice 2 (06 pts) : Soient $X_0, X_1, X_2, ...$ des variables aléatoires indépendantes et de même loi:

$$P(X_k = 0) = P(X_k = 1) = \frac{1}{2}, \qquad k = 0, 1, 2, ...$$

Considérons le modèle:

$$Y_t = \frac{X_t}{2} + \frac{X_{t+1}}{2^2} + \frac{X_{t+2}}{2^3} + ...$$

Calculer $E(Y_t)$ et $\gamma_Y(h)$. Ce modèle est-il stationnaire?

Exercice 3 (06 pts) : Soit L un opérateur de retard ($LX_t = X_{t-1}$) et $\alpha \in IR$:

1) Montre que, *si* $|\alpha| < 1$, alors

$$\frac{1}{1-\alpha L} = (1-\alpha L)^{-1} = \sum_{k=0}^{\infty} (\alpha L)^k .$$

2) Montre que, *si* $|\alpha| > 1$, alors

$$\frac{1}{1-\alpha L} = (1-\alpha L)^{-1} = -\sum_{k=1}^{\infty} (\alpha L)^k .$$

3) Calculer l'inverse des filtres suivants si possible:

$$(1 - 0.2L), \quad (1 + 2L), \quad \left(1 - \frac{2}{5}L\right)\left(1 + \frac{5}{2}L\right), \quad \left(1 + \frac{1}{2}L\right)\left(1 + \frac{3}{2}L\right)^2 .$$

4) Considérons le modèle : $Y_t - 2Y_{t-1} = X_t + \frac{1}{2} X_{t-1}$.

Ecrire Y_t en fonction de X puis X_t en fonction de Y.

Bonne chance

Examen de Rattrapage (Module : Analyse des Données)

Exercice 1 (14 pts) : La distribution suivante montre une relation linéaire simple :

$$y_j = ax_j + b + \varepsilon_j, \quad j = 1, ..., 10 \text{ et } \varepsilon \sim N\left(0, \sigma_\epsilon^2\right)$$

entre le revenu mensuelle $X\left(\times 10^4 DA\right)$ est la consommation $Y\left(\times 10^4 DA\right)$ de dix familles :

x	y	x^2	y^2	xy	$\hat{a}x$	$\hat{y} = \hat{a}x + \hat{b}$	$\hat{\varepsilon}$	$\hat{\varepsilon}^2$
2	1							
3	2.5							
3.5	2.5							
4	3							
5	3.5							
6	3.5							
7	7							
9	6.5							
11.5	8							
14	10							

1) Compléter les trois premières colonnes du tableau. En déduire les valeurs de :

$$\bar{x}, \quad \bar{y}, \quad S_x^2, \quad S_y^2, \quad S_{xy}$$

2) Compléter le reste du tableau et proposez des estimateurs pour les paramètres du modèle.

3) Donner une estimation de la variance résiduelle $\hat{\sigma}_\epsilon^2$.

4) Vérifier que $\hat{\sigma}_a^2 = 0.004$ et que $\hat{\sigma}_b^2 = 0.248$.

5) Etudier la signification des paramètres a et b du modèle.

6) Donner des intervalles de confiance des paramètres a et b du modèle.

7) Calculer la valeur du coefficient de détermination D du modèle.

8) Testez la validité globale du modèle.

9) Sachant qu'une famille à un revenu qui vaut $6.5 \times 10^4 DA$, quelle est la valeur prédite de sa consommation mensuelle? Même question pour un revenu de $18 \times 10^4 DA$?

Exercice 2 (06 pts) : L'équation estimée ci-dessous a été obtenue par les MCO en utilisant des données trimestrielle sur la période de 2014 à 2017 (bornes inclues):

$$\widehat{Y}_t = \underbrace{4,23}_{(1.3)} + \underbrace{1,24}_{(0.05)} X_{t1} - \underbrace{2,18}_{(0.7)} X_{t2} + \underbrace{0,44}_{(0.15)} X_{t3}.$$

Les chiffres entre parenthèses sont les écarts-types estimés. La $SCE\,(Erreur)$ et la $SCR\,(régression)$ étaient respectivement $19,16$ et $108,48$. Calculez le coefficient de détermination et testez la significativité de chaque coefficient.

Bonne chance

Exercice 1 **(08 pts) : Vrai** ou **Faux**

1) $\rightarrow F$, 2) $\rightarrow F$, 3) $\rightarrow F$, 4) $\rightarrow V$, 5) $\rightarrow V$, 6) $\rightarrow F$, 7) $\rightarrow F$, 8) $\rightarrow V$.

Exercice 2 **(06 pts) :** 1) On a $E(X_t) = \frac{1}{2}$, donc

$$E(Y_t) = \frac{1}{2}\left(\frac{1}{2} + \frac{1}{2^2} + ... + \frac{1}{2^k} + ...\right) = \frac{1}{2}\frac{1}{2}\frac{1}{1-(1/2)} = \frac{1}{2}$$

2) De même,

$$\begin{aligned} \gamma_Y(h) &= cov\left(\frac{X_t}{2} + \frac{X_{t+1}}{2^2} + ... + \frac{X_{t+h}}{2^{h-1}} + \frac{X_{t+h+1}}{2^h}... \quad ; \quad \frac{X_{t+h}}{2} + \frac{X_{t+h+1}}{2^2} + ...\right) \\ &= \frac{1}{2^h} var(X_{t+h}) + \frac{1}{2^{h+2}} var(X_{t+h+1}) + ... = \frac{1}{2^h}. \end{aligned}$$

En déduire que le modèle est stationnaire.

Exercice 3 **(06 pts) :** 1) Si $|\alpha| < 1$, alors

$$\sum_{k=0}^{\infty} (\alpha L)^k = \lim_{n\to\infty} \sum_{k=0}^{n} (\alpha L)^k = \frac{1-(\alpha L)^{n+1}}{1-\alpha L} = (1-\alpha L)^{-1} = \frac{1}{1-\alpha L}$$

2) Si $|\alpha| > 1$, alors on pose $\beta = 1/\alpha$ tq $(|\beta| < 1)$:

$$\begin{aligned} \frac{1}{1-\alpha L} &= \frac{1}{1-(\beta L^{-1})^{-1}} = -\beta L^{-1}\frac{1}{1-\beta L^{-1}} = -\beta L^{-1}\sum_{k=1}^{\infty}\left(\beta L^{-1}\right)^k \\ &= -\sum_{k=1}^{\infty}\left(\beta L^{-1}\right)^{k+1} = -\sum_{k=1}^{\infty}(\alpha L)^{-k}. \end{aligned}$$

3) Calcul de l'inverse des filtres :

$$\begin{aligned} (1-0.2L)^{-1} &= \frac{1}{1-\frac{1}{5}L} = \sum_{k=0}^{\infty}\left(\tfrac{1}{5}L\right)^k = 1 + \tfrac{1}{5}L + \tfrac{1}{5^2}L^2 + ... \\ (1+2L)^{-1} &= \frac{1}{1-2L} = -\sum_{k=1}^{\infty}(2L)^{-k} = -\tfrac{1}{2}L^{-1} - \tfrac{1}{2^2}L^{-2} - \tfrac{1}{2^3}L^{-3} + ... \end{aligned}$$

$$\left(1-\tfrac{2}{5}L\right)^{-1}\left(1+\tfrac{5}{2}L\right)^{-1} = \left(1 + \tfrac{2}{5}L + \left(\tfrac{2}{5}\right)^2 L^2 + ...\right)\left(\tfrac{2}{5}L^{-1} - \tfrac{2^2}{5^2}L^{-2} + ...\right)$$

$$\left(1+\frac{1}{2}L\right)\left(1+\frac{3}{2}L\right)^2 = \left(1 - \tfrac{1}{2}L + \tfrac{1}{2^2}L^2 - ...\right)\left(-\tfrac{2}{3}L^{-1} + \tfrac{2^2}{3^2}L^{-2} - ...\right)^2$$

4-a) Y_t en fonction de X :

$$\begin{aligned} Y_t &= (1-2L)^{-1}\left(1+\frac{1}{2}L\right)X_t = \left(1+\frac{1}{2}L\right)\left(\tfrac{1}{2}L^{-1} - \tfrac{1}{2^2}L^{-2} + \tfrac{1}{2^3}L^{-3} - ...\right)X_t \\ &= \left(\tfrac{1}{4} + \tfrac{3}{8}L^{-1} - \tfrac{3}{16}L^{-2} + ...\right)X_t = \tfrac{1}{4}X_t + \tfrac{3}{8}X_{t+1} - \tfrac{3}{16}X_{t+2} + ... \end{aligned}$$

4-b) X_t en fonction de Y :

$$\begin{aligned} X_t &= \left(1+\frac{1}{2}L\right)^{-1}(1-2L)\,Y_t = (1-2L)\left(1 - \tfrac{1}{2}L + \tfrac{1}{2^2}L^2 - ...\right)Y_t \\ &= \left(1 - \tfrac{5}{2}L + \tfrac{3}{2}L^2 - ...\right)Y_t = Y_t - \tfrac{5}{2}Y_{t-1} + Y_{t-2} - .... \end{aligned}$$

Université de Biskra **Département d'Informatique**
Module : Analyse des Données **Master 1 - SIOD**

TD N 01 (Régression Linéaire Simple)

Exercice 1: L'étude statistique ci-dessous porte sur les poids respectifs des Pères (P) et de leur Fils (F). Voici les résultats numériques que nous avons obtenus :

$$\sum P_i = 800, \qquad \sum P_i^2 = 53418, \qquad \sum P_iF_i = 54107, \qquad \sum F_i = 811, \qquad \sum F_i^2 = 54849.$$

1) Calculer les estimateurs des paramètres du modèle du poids des Fils en fonction du poids des Pères:

$$F_j = a_1 P_j + b_1 + \varepsilon_j, \qquad j = 1, ..., 12$$

2) Calculer les estimateurs des paramètres du modèle du poids des Pères en fonction du poids des Fils:

$$P_j = a_2 F_j + b_2 + \varepsilon_j, \qquad j = 1, ..., 12$$

3) Montrer que le produit des pentes $\hat{a}_1$ et $\hat{a}_2$ des deux droites est égal au coefficient de détermination.

Exercice 2. La distribution suivante montre une relation linéaire simple entre l'évolution du taux de croissance économique Y(en %) et le prix de pétrole X(en \$) pendant 10 ans :

$$y_j = ax_j + b + \varepsilon_j, \quad j = 1, ..., 10 \text{ et } \varepsilon \sim N\left(0, \sigma_\epsilon^2\right).$$

	x	y	$(x-\overline{x})$	$(x-\overline{x})(y-\overline{y})$	$(x-\overline{x})^2$	$(y-\overline{y})^2$	$\hat{y}$	$\hat{\varepsilon}$	$\hat{\varepsilon}^2$
	111	1.9							
	120	4.4							
	133	5.6							
	105	4.1							
	101	3.2							
	95	-3							
	82	-2							
	70	0							
	66	-2.9							
	54	-3							
$\sum =$									

1) Donner la représentation graphique des données.
2) Completer le tableau et estimer les paramètres du modèle.
3) Calculer la valeur de $\hat{\sigma}_\epsilon^2$.
4) Calculer les quantités : $\hat{\sigma}_a^2$ *et* $\hat{\sigma}_b^2$ (i.e., estimateurs des variances de a et b resp.).
5) Testez la signification des paramètres du modèle au niveau 95%.
6) Déterminer l'intervalle de confiance à 95% des paramètres a et b.
7) Testez la validité du modèle.
8) Quelle quantité mesure la bonne explication de la variable à expliquer par la variable explicative? Que vaut-elle ici ? Est-ce une valeur satisfaisante ?
9) Donnez le taux de croissance lorsque le prix du pétrole est 130\$?
10) Considérons le cas du prix du pétrole 100\$, estimer la valeur correspondante du taux de croissance.

Université de Biskra | **Département d'Informatique**
Module : Analyse des Données | **Master 1 - SIOD**

Interrogation (2019)

Exercice 1. Dans une étude de régression simple:

$$y_j = ax_j + b + \varepsilon_j, \quad j = 1, ..., n \text{ et } \varepsilon \sim N\left(0, \sigma_\epsilon^2\right)$$

Nous avons obtenu le tableau d'AV pour $n = 19$ observations, comme suit:

Variation	ddl	SC	MC	F
Factorielle				
Résiduelle		1053		
Total		5654		

1) Compléter le tableau et tester la signification locale et globale du modèle à 95%.
2) Calculer le coefficient de détermination et $\hat{\sigma}_\epsilon^2$ du modèle.

Université de Biskra | **Département d'Informatique**
Module : Analyse des Données | **Master 1 - SIOD**

Interrogation (2019)

Exercice 1. Dans une étude de régression simple:

$$y_j = ax_j + b + \varepsilon_j, \quad j = 1, ..., n \text{ et } \varepsilon \sim N\left(0, \sigma_\epsilon^2\right)$$

Nous avons obtenu le tableau d'AV pour $n = 19$ observations, comme suit:

Variation	ddl	SC	MC	F
Factorielle	1	4601	4601	$=\frac{4601}{61.94} = 74.282$
Résiduelle	17	1053	$=\frac{1053}{17} = 61.941$	
Total	18	5654		

1) Compléter le tableau et tester la signification locale et globale du modèle à 95%.
2) Calculer le coefficient de détermination et $\hat{\sigma}_\epsilon^2$ du modèle.