

République Algérienne Démocratique et Populaire  
Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

**UNIVERSITÉ MOHAMED KHIDER, BISKRA**

FACULTÉ des SCIENCES EXACTES et des SCIENCES de la NATURE et de la VIE

**DÉPARTEMENT DE SNV**



Polycopié du TP :

**Biostatistiques**

Régression linéaire sous SPSS

(TP 4)

Préparé par

ROUBL. A

2020/2021

# Table des matières

<b>Table des matières</b>	<b>2</b>
<b>Liste des figures</b>	<b>3</b>
<b>1 Régression linéaire sous SPSS</b>	<b>1</b>
1.1 Rappel . . . . .	1
1.2 Régression linéaire simple . . . . .	2
1.2.1 Modèle de régression linéaire simple . . . . .	2
1.2.2 Régression linéaire simple sous SPSS . . . . .	3
1.3 Régression linéaire multiple . . . . .	15
1.3.1 Modèle de régression linéaire multiple . . . . .	15
1.3.2 Régression linéaire multiple sous SPSS . . . . .	16

# Table des figures

1.1	Saisie des données sous SPSS. . . . .	4
1.2	Présentation graphique des données (Etape 2 partie a). . . . .	5
1.3	Présentation graphique des données (Etape2 partie b). . . . .	6
1.4	Nuage de points présente la Biomasse en fonction de la Concentration d'azote. . . . .	6
1.5	Ajout du droite de régression au nuage de points. . . . .	7
1.6	Ajout du droite de régression au nuage de points. . . . .	8
1.7	Nuage de points et droite de régression. . . . .	8
1.8	Procédures de la réalisation de la régression linéaire simple (Partie 1). . . . .	9
1.9	Procédures de la réalisation de la régression linéaire simple (Partie 2). . . . .	10
1.10	Résultats de la régression linéaire simple. . . . .	11
1.11	Tableau récapitule les variables explicatives introduites dans le modèle. . . . .	12
1.12	Tableau des coefficients de corrélation et de détermination. . . . .	12
1.13	Table de L'ANOVA. . . . .	14
1.14	Tableau des coefficients de la régression linéaire simple. . . . .	14
1.15	Saisie des données. . . . .	18
1.16	Réalisation de la régression linéaire multiple sous SPSS (Partie 1). . . . .	19
1.17	Réalisation de la régression linéaire multiple sous SPSS (Partie 2). . . . .	20
1.18	Résultats de la régression linéaire multiple sous SPSS. . . . .	21
1.19	Procédure de la réalisation de la régression linéaire multiple par étapes (en choisissant la méthode "Pas à Pas"). . . . .	23

# Chapitre 1

## Régression linéaire sous SPSS

Dans ce chapitre, on va présenter un rappel sur la notion de la régression linéaire. Puis, à travers d'un exemple d'application, on va donner les principales étapes à suivre pour faire une régression linéaire à l'aide du logiciel SPSS.

### 1.1 Rappel

Le but de la régression simple (respectivement. multiple) est d'expliquer une variable  $Y$  à l'aide d'une variable  $X$  (resp. plusieurs variables  $X_1, X_2, \dots, X_p$ ). Ou d'une autre façon la régression permet :

- De trouver (modéliser) la relation entre la variable  $Y$  et la variable  $X$  (ou entre la variable  $Y$  et plusieurs variables  $X_1, X_2, \dots, X_p$ ).
- De prédire les valeurs de la variable  $Y$  à partir des valeurs de  $X$  (si la variable  $X$  (ou les variables  $X_j, j = 1, 2, \dots, p$ ) est connue).

**Remarque 1.1** *La variable  $Y$  est appelée **variable dépendante**, ou **variable à expliquer** et les variables  $X_j$  ( $j = 1, 2, \dots, p$ ) sont appelées **variables indépendantes**, ou **variables explicatives**. C'est-à-dire, dans la régression simple (respectivement. multiple) on cherche d'une fonction  $f$  telle que*

$$Y = f(X) + \epsilon, (\text{resp. } Y = f(X_1, X_2, \dots, X_p) + \epsilon),$$

où  $\epsilon$  est une variable aléatoire (résidus).

**Exemple 1.1** - Étude de la température ( $Y$ ) en fonction de l'altitude ( $X$ ).

- Étude de la taille ( $Y$ ) en fonction du poids ( $X$ ) ou l'inverse.
- Étude du nombre de morts d'une maladie ( $Y$ ) en fonction du nombre des infectés ( $X$ ).
- Étude de la taille ( $Y$ ) en fonction du poids ( $X_1$ ) et de l'âge ( $X_2$ ).
- Étude du poids de l'enfant à la naissance ( $Y$ ) en fonction du poids de la mère ( $X_1$ ), de son âge ( $X_2$ ) et de sa taille ( $X_3$ ).

Si la fonction  $f$  est affiné (la relation est linéaire) on parle sur la régression linéaire. Alors

### Cas de régression linéaire simple

$$f(x) = a + bX;$$

### Cas de régression linéaire multiple

$$f(x_1, x_2, \dots, x_p) = a + b_1X_1 + b_2X_2 + \dots + b_pX_p.$$

## 1.2 Régression linéaire simple

### 1.2.1 Modèle de régression linéaire simple

Soit un échantillon de  $n$  individus. Pour chaque individu, on a les observations  $x_i$  et  $y_i$ ,  $i = 1, 2, \dots, n$  qu'elles sont les valeurs des réalisations des variables quantitatives  $X$  et  $Y$  respectivement.

Le modèle de régression linéaire simple est de la forme suivante :

$$Y = a + bX + \epsilon. \tag{1.1}$$

Pour la  $i^{\text{ème}}$  observation, on peut réécrire le modèle (1.1) sous la forme

$$y_i = a + bx_i + \varepsilon_i, \text{ pour } i = \overline{1, n}. \quad (1.2)$$

Les hypothèses relatives à ce modèle sont

i)  $E(\varepsilon_i) = 0$ ;

ii)  $\text{var}(\varepsilon_i) = \sigma^2 < \infty \quad \forall i = \overline{1, n}$ ;

iii)  $\text{cov}(\varepsilon_i, \varepsilon_j) = 0 \quad \forall i \neq j$ .

De plus, une hypothèse complémentaire pour les inférences : les variables aléatoires  $\varepsilon_i$  sont normalement distribuées, alors  $\varepsilon_i \sim N(0, \sigma^2)$ .

## 1.2.2 Régression linéaire simple sous SPSS

**Exemple 1.2** (*Exercice 13 sérieTP3*) (*Régression linéaire simple*)

Dans le cadre de travaux de recherche sur la Biomasse (mg), d'un certain type de plante, en fonction de la concentration de l'Azote  $\text{NH}_4^+$  ( $\mu \text{ mol}$ ), nous avons réalisée des expériences dont la biomasse moyenne ( $Y$ ) ainsi que la concentration du l'Azote ( $X$ ) en question sont données dans le tableau ci-dessus :

Concentration $\mu \text{mol}$	0	100	200	400	600
Biomasse mg	305	378	458	540	565

Afin de modéliser ces données, nous avons proposé le modèle linéaire suivant :

$$Y = a + bX + \varepsilon.$$

**Questions :**

1. Pour un seuil de risque  $\alpha = 5\%$ , le modèle proposé est-il pertinent ?
2. Donner les estimations des paramètres  $a$  et  $b$  et donner la droite de régression.

3. Donner le coefficient de corrélation linéaire. Que peut-on conclure ?

4. Quelle Biomasse prévoyez-vous à une concentration 500  $\mu$  mol ?

Avant de répondre aux questions, on peut **présenter graphiquement le nuage des points**  $(x_i, y_i)$  pour faire une idée préliminaire sur la distribution de ces points et si le modèle linéaire (modèle proposé) peut décrire les données.

Dans cet exemple on veut étudier

La biomasse ( $Y$ ) d'une plante en fonction de la concentration de l'Azote  $\text{NH}_4^+$  ( $X$ ), alors

- La variable dépendante : la biomasse ;
- La variable indépendante : la concentration de l'Azote  $\text{NH}_4^+$ .

Pour répondre aux questions de cet exercice sous SPSS, il faut suivre les étapes suivantes

**Etape 1.** Saisie des données

Entrez les données dans SPSS, dont vous avez deux variables quantitatives  $Y$  et  $X$  à définir séparément dans SPSS (voir figure 1.1).

The screenshot shows the IBM SPSS Statistics data entry window for 'Série 3 TP3.sav'. The data table is as follows:

	BIOMASSE	CONCENTRATION	HauteurH	DiamètreD	Absorbance Y	ConcentrationX	HauteurNeige	Altitude	Rugosité	Pente	Orientation	Latitude
1	305,00	,00	9,2073	,1999	,000	0	95	2768	252	22	324	876021
2	378,00	100,00	9,6794	,3012	,205	20	150	4208	333	29	308	876015
3	458,00	200,00	10,8049	,3791	,331	40	4	4045	62	5	349	876016
4	540,00	400,00	13,4637	,6005	,515	60	0	4572	85	8	14	876013
5	565,00	600,00	14,1540	,6570	,584	80	0	4614	115	10	63	876010
6	-	-	-	-	,671	100	80	4321	176	16	130	876007
7	-	-	-	-	-	-	95	3886	72	6	199	876003
8	-	-	-	-	-	-	20	4206	57	5	32	876001
9	-	-	-	-	-	-	90	4192	266	23	197	875998
10	-	-	-	-	-	-	10	4051	69	6	113	875995
11	-	-	-	-	-	-	10	3746	62	5	149	875992
12	-	-	-	-	-	-	50	3789	42	3	218	875985
13	-	-	-	-	-	-	45	3771	44	4	53	875984
14	-	-	-	-	-	-	60	3796	48	4	101	875983
15	-	-	-	-	-	-	55	3885	77	7	332	875981
16	-	-	-	-	-	-	3	4295	113	10	18	875978
17	-	-	-	-	-	-	33	4467	147	13	50	875976
18	-	-	-	-	-	-	0	4764	12	1	276	875973
19	-	-	-	-	-	-	35	4313	38	3	350	875970
20	-	-	-	-	-	-	45	4387	40	3	46	875967
21	-	-	-	-	-	-	-	-	-	-	-	-

FIG. 1.1 – Saisie des données sous SPSS.

**Remarque 1.2** *Il faut sauvegarder votre fichier.*

**Etape 2.** Présentation graphique des données (Nuage de points)

a- Allez à Barre de menus → Graphes → Boîtes de dialogue ancienne version  
 puis cliquez sur Dispersion/points → Dispersion simple → Définir

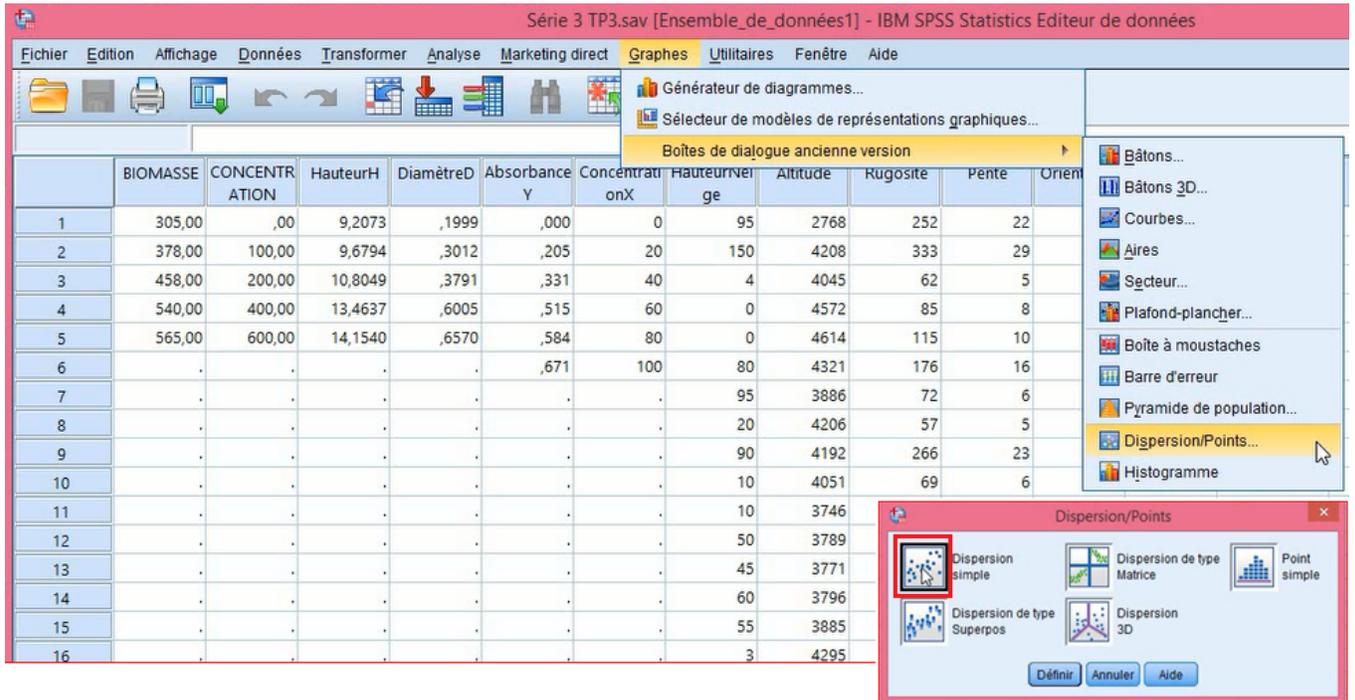


FIG. 1.2 – Présentation graphique des données (Etape 2 partie a).

b- Dans la boîte de dialogue qui va apparaître (figure 1.3) insérez **la variable dépendante** dans la case **Axe des Y** et **la variable indépendante** dans la case **Axe des X** puis cliquez sur **OK**.

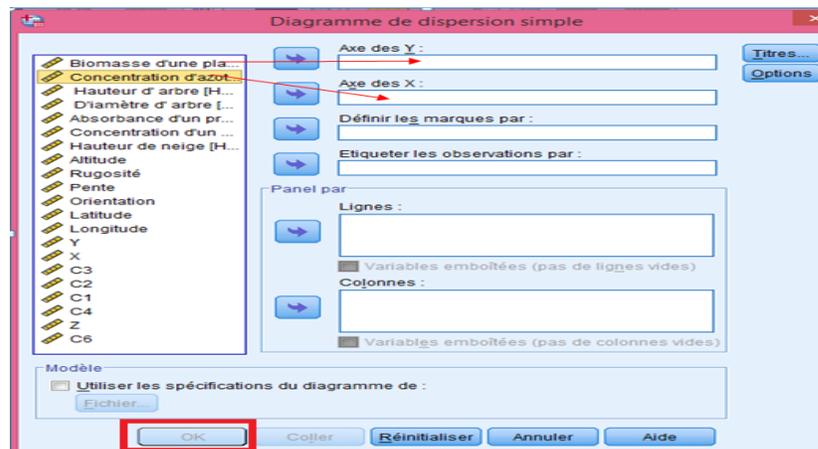


FIG. 1.3 – Présentation graphique des données (Etape2 partie b).

une fois que vous exécutez l'étape 2, vous obtiendrez la figure 1.4.

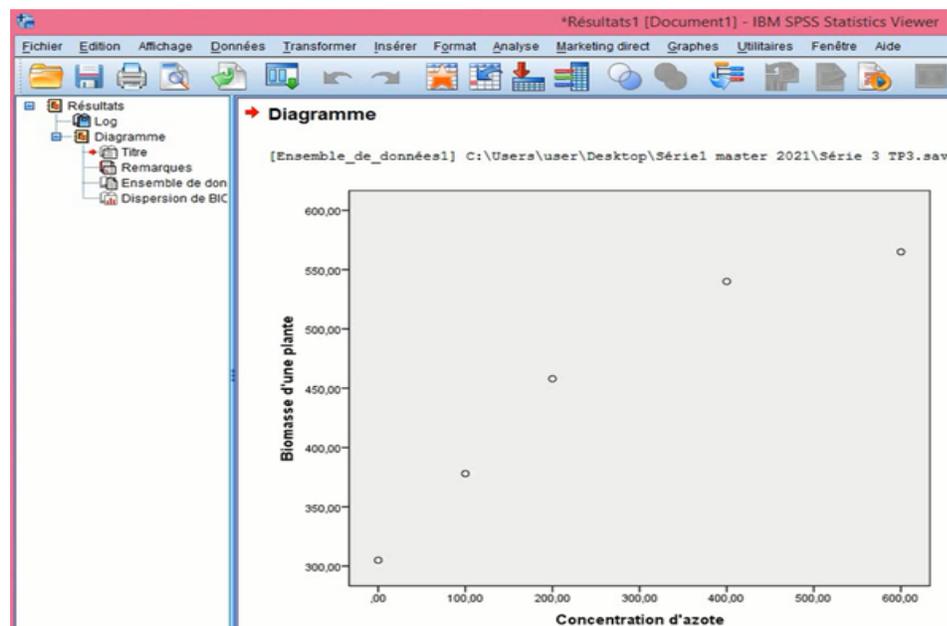


FIG. 1.4 – Nuage de points présente la Biomasse en fonction de la Concentration d'azote.

Au vue du graphique (figure 1.4), il semble que le modèle linéaire est adéquat pour l'explication de biomasse en fonction de la concentration de l'Azote (car le nuage des points est distribué sous une forme linéaire).

**Remarque 1.3** Vous pouvez ajouter la droite de régression au nuage de points en suivant ce qui suit

- 1- cliquez deux fois sur le nuage de points une fenêtre va s'ouvrir appelée "**Editeur de diagrammes**";
- 2- sur la fenêtre "**Editeur de diagrammes**" cliquez sur **Eléments** → **ajouter une courbe d'ajustement au total** (voir figure 1.5);
- 3- une autre fenêtre va s'ouvrir appelée "**propriété**" fermez cette fenêtre puis fermez la fenêtre d'Editeur de diagrammes (voir figure 1.6).

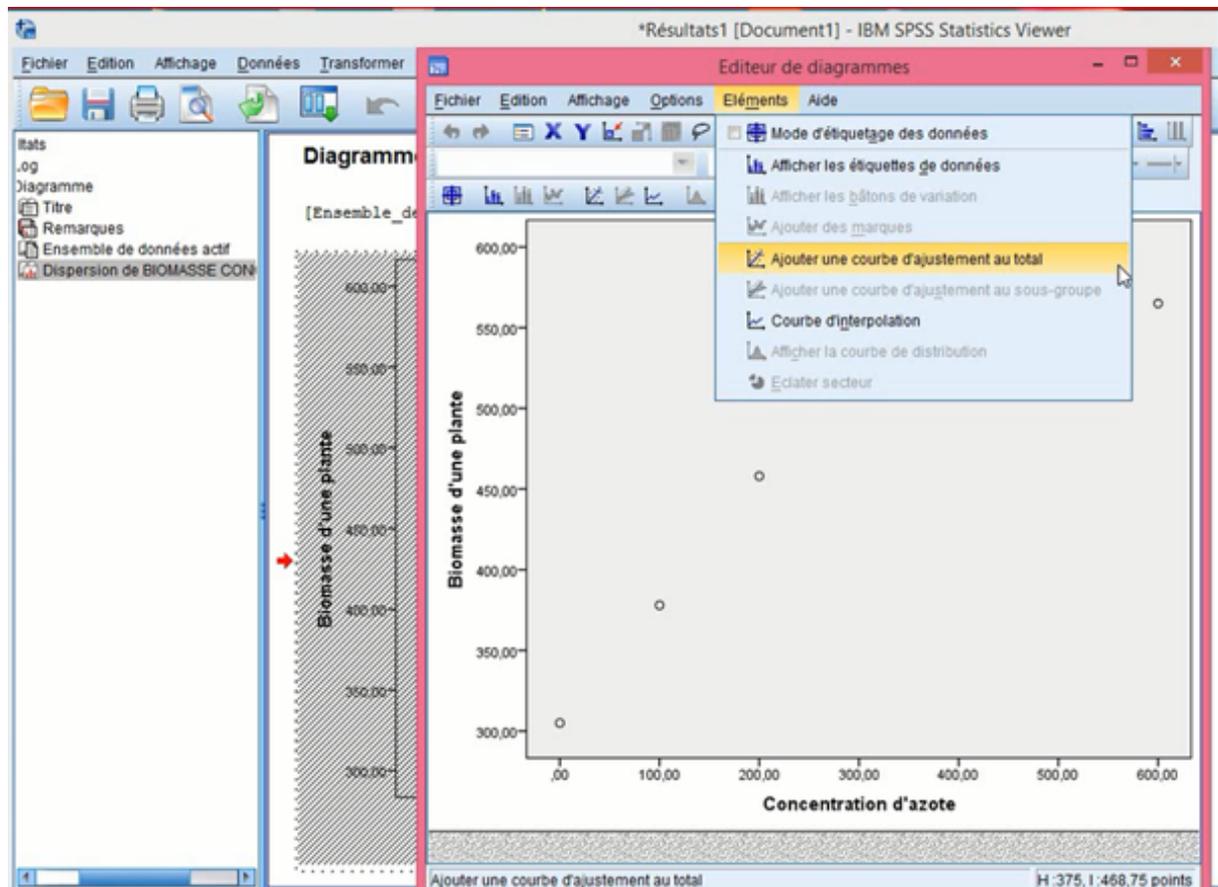


FIG. 1.5 – Ajout du droite de régression au nuage de points.

En suivant les étapes citées, vous obtiendrez alors la figure 1.7.

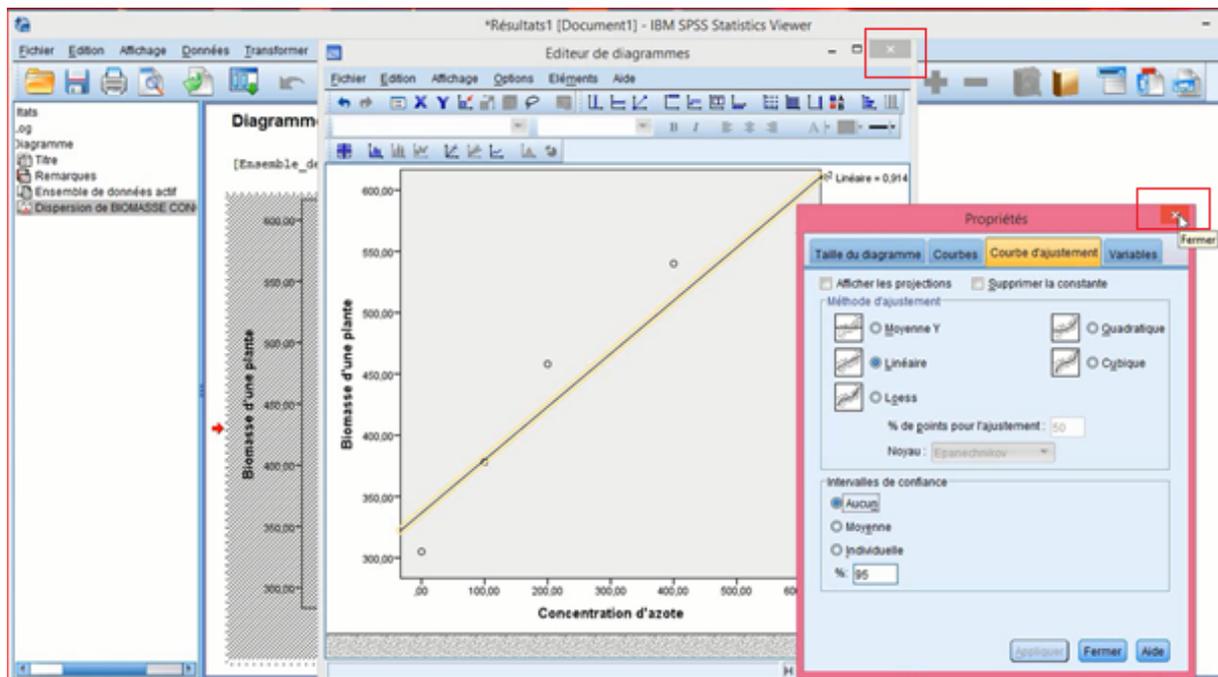


FIG. 1.6 – Ajout du droite de régression au nuage de points.

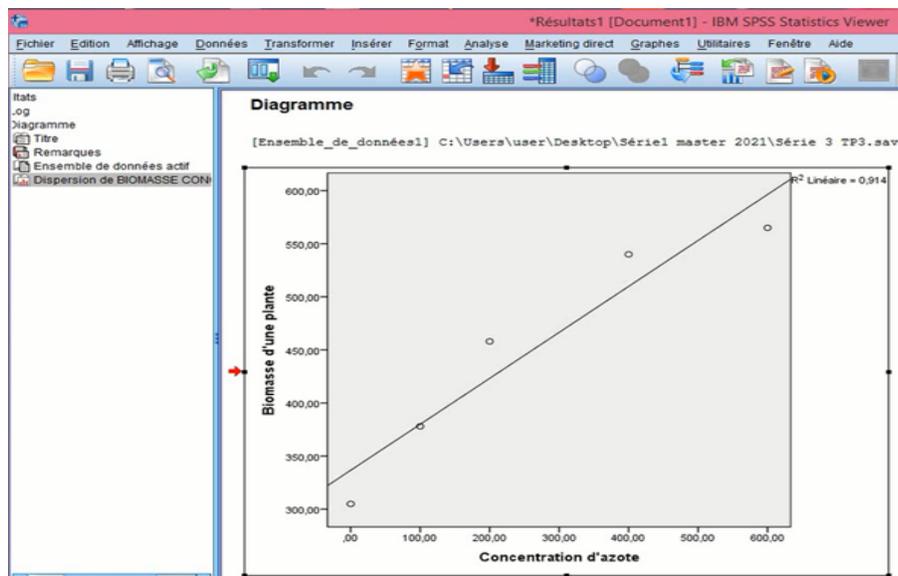


FIG. 1.7 – Nuage de points et droite de régression.

**Etape 3.** Réalisation de la régression linéaire sous SPSS

Pour obtenir une régression linéaire simple il faut suivre ces étapes

1- Sélectionnez sur la barre de menu

Analyse → Régression → Linéaire

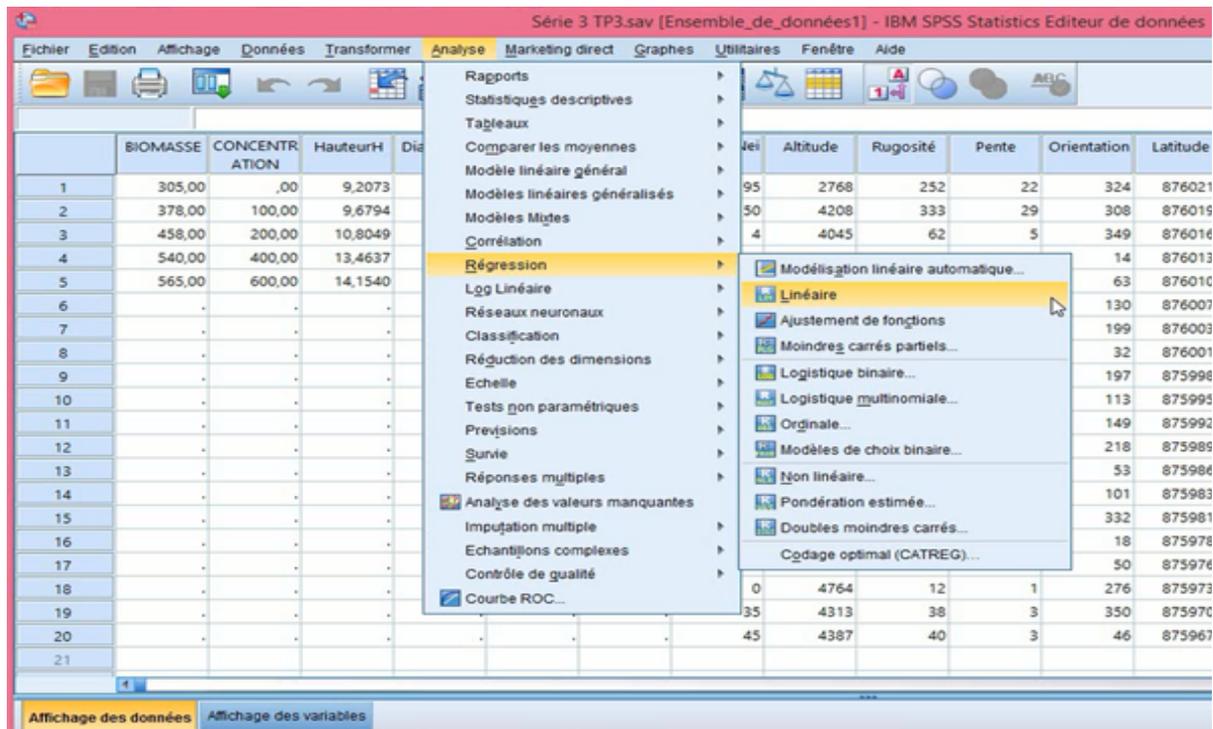


FIG. 1.8 – Procédures de la réalisation de la régression linéaire simple (Partie 1).

2- Dans la boîte de dialogue de la figure 1.9 apparaît : sélectionnez, dans la liste des variables, les deux variables que vous souhaitez à analyser, et mettez, en cliquant sur les flèches, la **variable dépendante** (variable à expliquer) dans la case **Dépendant** et la **variable indépendante** (variable explicative) dans la case **Variables indépendantes**, puis cliquez sur **OK**.

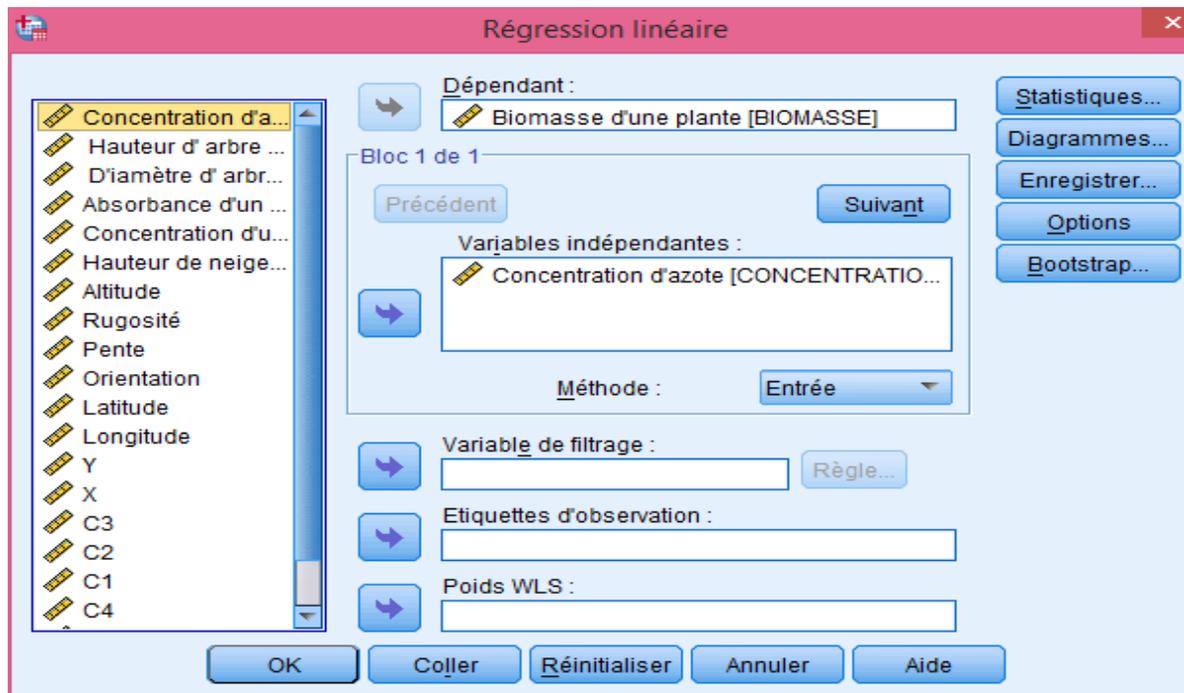


FIG. 1.9 – Procédures de la réalisation de la régression linéaire simple (Partie 2).

L'application de ces étapes sur les données de l'exercice permet d'obtenir les résultats (4 tableaux) qui sont présentés dans la figure 1.10. Ces résultats contiennent des réponses sur les questions 1, 2, 3 et 4.

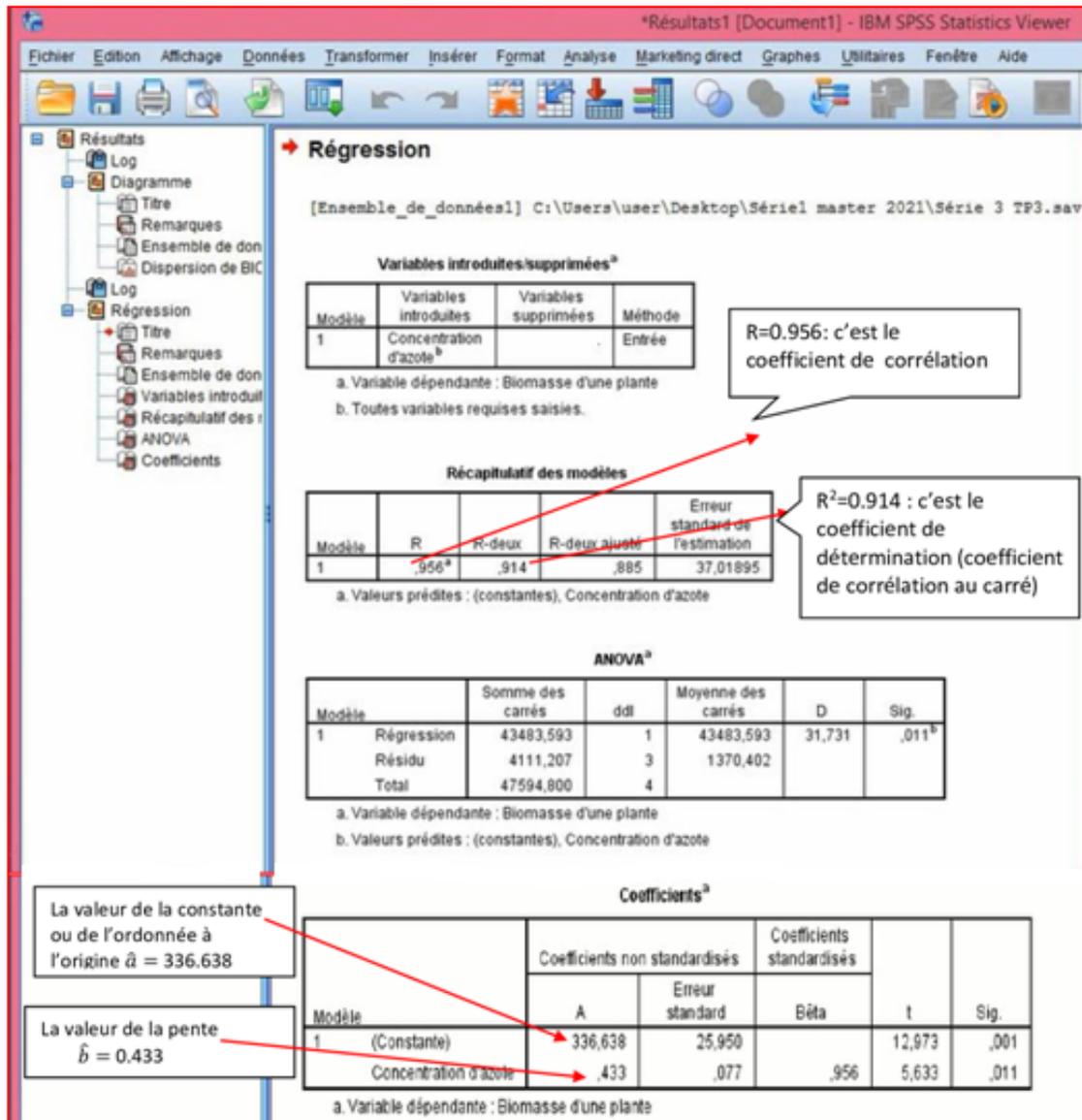


FIG. 1.10 – Résultats de la régression linéaire simple.

**Variables introduites/supprimées<sup>a</sup>**

Modèle	Variables introduites	Variables supprimées	Méthode
1	Concentration d'azote <sup>b</sup>	.	Entrée

a. Variable dépendante : Biomasse d'une plante

b. Toutes variables requises saisies.

FIG. 1.11 – Tableau récapitule les variables explicatives introduites dans le modèle.

**Récapitulatif des modèles**

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,956 <sup>a</sup>	,914	,885	37,01895

a. Valeurs prédites : (constantes), Concentration d'azote

FIG. 1.12 – Tableau des coefficients de corrélation et de détermination.

### Interprétation des Résultats obtenus

**Premier tableau** (figure 1.11) : ce tableau récapitule les variables explicatives prises en compte dans le modèle. Ici, il n'y a qu'une seule variable dans la case "variables introduites" (Concentration de d'Azote) tandis qu'il n'ya pas des variables supprimées puisque nous travaillons sur une régression simple.

**Deuxième tableau** (figure 1.12) : il donne deux valeurs importantes dans le modèle de régression :

- Le coefficient de corrélation :  $R = 0.956$ . Donc, le coefficient de corrélation est presque égal à 1, ce qui indique qu'il y a une forte liaison linéaire entre la biomasse ( $Y$ ) et la concentration de l'azote ( $X$ ). Le signe positif de  $R$  indique que les deux variables varient dans le même sens.
- Le coefficient de détermination :  $R^2 = 0.914$ , ce qui indique que 91.4% de la variation totale de  $Y$  est expliquée par le modèle de régression sur  $X$ .

**Troisième tableau** (figure 1.13) : c'est la table d'analyse de la variance, il indique si le modèle est valide ou non à partir d'un test sur la pente de la droite de régression. C'est à dire, il nous permet de répondre au test d'hypothèses suivant :

$$\begin{cases} H_0 : b = 0 \text{ ((le modèle n'est pas valide)}; \\ H_1 : b \neq 0 \text{ (le modèle est valide)}. \end{cases}$$

Alors, à l'aide de la valeur de signification (Sig) et au seuil de risque  $\alpha$ , on décide de

$$\begin{cases} \text{ne pas rejeter } H_0 & \text{si } \alpha < \text{Sig}; \\ \text{rejeter } H_0 & \text{si } \alpha \geq \text{Sig}. \end{cases}$$

Ce tableau donne les valeurs suivantes

- La somme des carrés des résidus (variation qui n'est pas expliquée par le modèle de régression)

$$SCR = 4111.207.$$

- La somme des carrés de régression (la variation expliquée par la régression)

$$SCE = 43483.593.$$

- La réalisation de la statistique de Fisher

$$f_c = 31.731.$$

- La valeur de signification :  $Sig = 0.011$ . Il résulte de cette valeur et pour un seuil de risque  $\alpha = 5\%$  que le modèle obtenu est pertinent (valide) car  $\alpha > Sig$  ( $0.05 > 0.011$ ), c'est-à-dire, il existe une relation linéaire statistiquement significative entre la Biomasse de la plante et la concentration de l'azote donnée par l'équation :  $\hat{Y} = \hat{a} + \hat{b}X$ , où  $\hat{a}$  et  $\hat{b}$  sont donnés dans le quatrième tableau.

**ANOVA<sup>a</sup>**

Modèle		Somme des carrés	ddl	Moyenne des carrés	D	Sig.
1	Régression	43483,593	1	43483,593	31,731	,011 <sup>b</sup>
	Résidu	4111,207	3	1370,402		
	Total	47594,800	4			

a. Variable dépendante : Biomasse d'une plante

b. Valeurs prédites : (constantes), Concentration d'azote

FIG. 1.13 – Table de L'ANOVA.

**Quatrième tableau** (figure 1.14) : il donne les estimations des paramètres  $a$  et  $b$  (coefficients de la droite de régression) dits "**A**" dans SPSS (Partie encadrée) suivantes :

- La valeur de la constante ou l'ordonnée à l'origine  $\hat{a} = 336.638$  ;
- La valeur de la pente  $\hat{b} = 0.433$ .

Alors l'équation de la droite de régression sera donnée par

$$\hat{Y} = 336.638 + 0.433X.$$

**Coefficients<sup>a</sup>**

Modèle		Coefficients non standardisés		Coefficients standardisés	t	Sig.
		A	Erreur standard	Bêta		
1	(Constante)	336,638	25,950		12,973	,001
	Concentration d'azote	,433	,077	,956	5,633	,011

a. Variable dépendante : Biomasse d'une plante

FIG. 1.14 – Tableau des coefficients de la régression linéaire simple.

Pour la dernière question, on veut prédire, sur la base de notre modèle, la biomasse ( $Y$ ) pour une concentration d'azote égale à  $500 \mu\text{mol}$ , il suffit alors de substituer la valeur de la concentration d'azote ( $X = 500 \mu\text{mol}$ ) dans la dernière équation pour trouver **la valeur**

prédite ( $\hat{Y}$ ) de la biomasse de la plante

$$\hat{Y} = 336.638 + 0.433(500) = 553.138.$$

## 1.3 Régression linéaire multiple

Les idées utilisées en régression multiple sont les mêmes que celles qu'on a vu en régression linéaire simple. La régression linéaire multiple diffère de la régression linéaire simple par le nombre de variables explicatives présentes dans le modèle.

### 1.3.1 Modèle de régression linéaire multiple

Le modèle de régression linéaire multiple est de la forme suivante :

$$Y = a + b_1X_1 + b_2X_2 + \dots + b_pX_p + \varepsilon. \quad (1.3)$$

Pour la  $i^{\text{ème}}$  observation, le modèle (1.3) peut être représenté de la manière suivante

$$Y_i = a + b_1x_{1i} + b_2x_{2i} + \dots + b_px_{pi} + \varepsilon_i, \quad i = \overline{1, n}; \quad (1.4)$$

où :

$Y_i$  est la valeur de la variable dépendante  $Y$  (variable quantitative),

$x_{1i}, x_{2i}, \dots, x_{pi}$  sont les valeurs de  $p$  autres variables quantitatives (variables explicatives)

$X_1, \dots, X_p$ , pour  $i = 1, \dots, n$ .

les  $\varepsilon_i$  sont les termes des erreurs.

Les hypothèses relatives à ce modèle sont

i)  $E(\varepsilon_i) = 0$ ;

ii)  $var(\varepsilon_i) = \sigma^2 < \infty \quad \forall i = \overline{1, n}$ ;

iii)  $cov(\varepsilon_i, \varepsilon_j) = 0 \quad \forall i \neq j$ ;

iv) Les termes  $x_j$  ( $j = \overline{1, p}$ ) étant déterministes;

de plus, une hypothèse complémentaire pour les inférences : les variables aléatoires  $\varepsilon_i$  sont normalement distribuées, alors  $\varepsilon_i \sim N(0, \sigma^2)$ .

On peut aussi écrire le modèle (1.4) sous sa forme matricielle :

$$Y = Xb + \varepsilon;$$

où

- $Y$  est un vecteur aléatoire de dimension  $n$ ,
- $X$  est une matrice de taille  $n \times (p + 1)$  connue,
- $\beta$  est le vecteur de dimension  $(p + 1)$  des **paramètres inconnus** du modèle,
- $\varepsilon$  est le vecteur de dimension  $n$  des erreurs.

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{p1} \\ 1 & x_{12} & x_{22} & \dots & x_{p2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{pn} \end{bmatrix}, b = \begin{bmatrix} a \\ b_1 \\ \vdots \\ b_p \end{bmatrix},$$

$$\text{et } \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

### 1.3.2 Régression linéaire multiple sous SPSS

**Exemple 1.3** On étudie l'influence des heures de travail ( $X_1$ ) et du capital utilisé ( $X_2$ ) sur la production industrielle ( $Y$ ) ou d'une autre façon, on cherche à établir une relation entre la

production, les heures de travail et le capital utilisé. Pour cela, on dispose des observations de 9 entreprises résumées dans le tableau ci-dessous

Entreprise ( $i$ )	Travail (heures) ( $x_{1i}$ )	Capital (machines/heures) ( $x_{2i}$ )	Production (100 tonnes) ( $Y_i$ )
1	1100	300	60
2	1200	400	120
3	1430	420	190
4	1500	400	250
5	1520	510	300
6	1620	590	360
7	1800	600	380
8	1820	630	430
9	1800	610	440

On suppose que la production ( $Y$ ) est expliquée par un modèle de régression linéaire multiple avec deux variables explicatives, le travail ( $X_1$ ) et le capital ( $X_2$ ), c'est à dire par le modèle

$$Y = a + b_1X_1 + b_2X_2 + \varepsilon.$$

**Questions :**

1. Donner les estimations des paramètres du modèle proposé.
2. Pour un seuil de risque  $\alpha = 5\%$ , le modèle proposé est-il pertinent ?
3. Tester l'hypothèse nulle  $H_0 : b_j = 0$  contre l'alternative  $H_1 : b_j \neq 0$  pour  $j = 1; 2$ .
4. Tester l'hypothèse nulle  $H_0 : a = 0$  contre l'alternative  $H_1 : a \neq 0$ .

La méthode de la régression linéaire multiple se réalise sous le logiciel SPSS en générale en effectuant presque les mêmes étapes de la régression linéaire simple :

**Etape 1.** Saisie des données

Dans cet exercice on a

- Une variable dépendante ou variable à expliquée qui est la production ;
- Deux variables explicatives qu'elles sont : le travail ( $X_1$ ) et le capital ( $X_2$ ).

Entrez les données dans SPSS, dont vous avez 3 variables quantitatives à définir séparément dans SPSS (voir figure 1.15).

	Production	HeursTravail	Capital	var								
1	60	1100	300									
2	120	1200	400									
3	190	1430	420									
4	250	1500	400									
5	300	1520	510									
6	360	1620	590									
7	380	1800	600									
8	430	1820	630									
9	440	1800	610									
10												
11												
12												
13												
14												
15												
16												
17												
18												
19												
20												
21												

FIG. 1.15 – Saisie des données.

**Etape 2.** Réalisation de la régression linéaire multiple sous SPSS

Pour obtenir une régression linéaire multiple il faut suivre ces étapes

1- Sélectionnez sur la barre de menu

*Analyse → Régression → Linéaire*

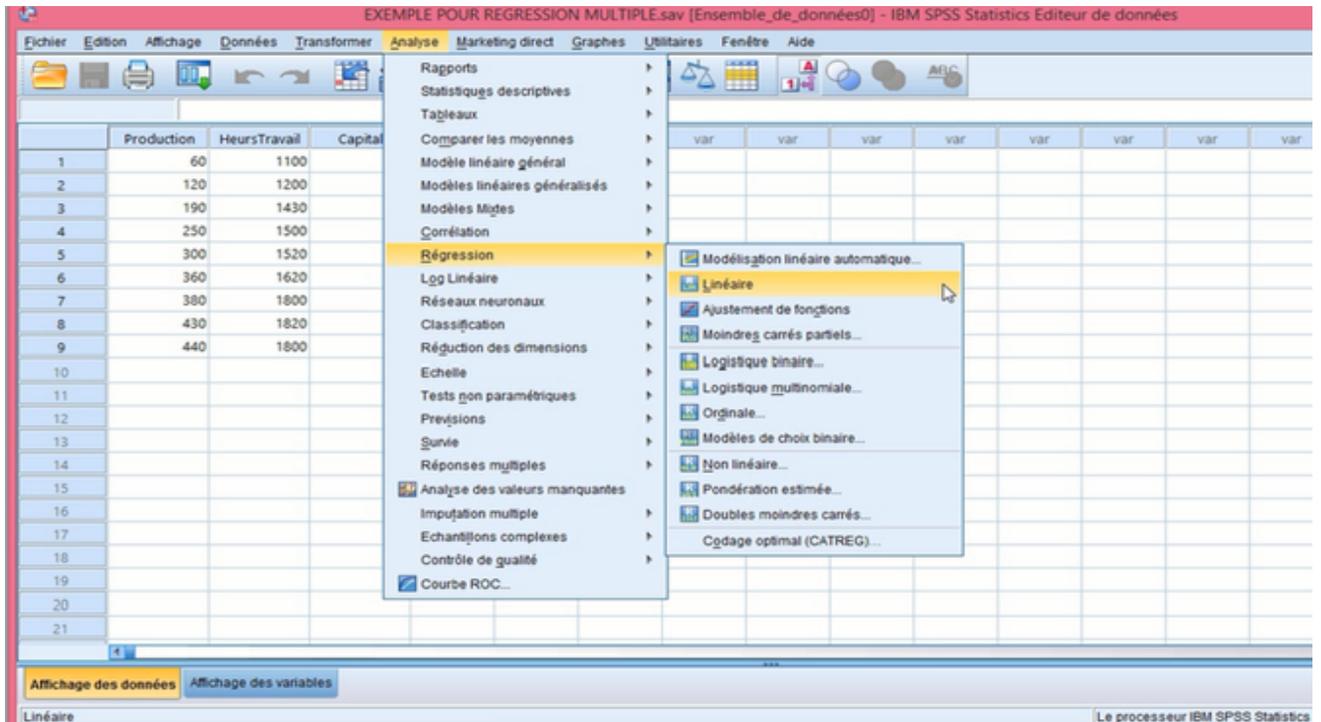


FIG. 1.16 – Réalisation de la régression linéaire multiple sous SPSS (Partie 1).

2- Dans la boîte de dialogue de la figure 1.17 apparaît, il faut réaliser ce qui suit :

- a- Transférez dans la liste des variables la **variable dépendante** (variable à expliquer) dans la case **Dépendant** ;
- b- Transférez **les deux variables indépendantes** (variables explicatives) dans la case **Variables indépendantes** ;
- c- Dans la case "**Méthode**" laissez la méthode par défaut c'est à dire "**Entrée**". Cette méthode (Entrée) est choisie parmi **d'autres méthodes de sélection des variables** selon notre but dans l'exercice.
- d- Cliquez sur OK

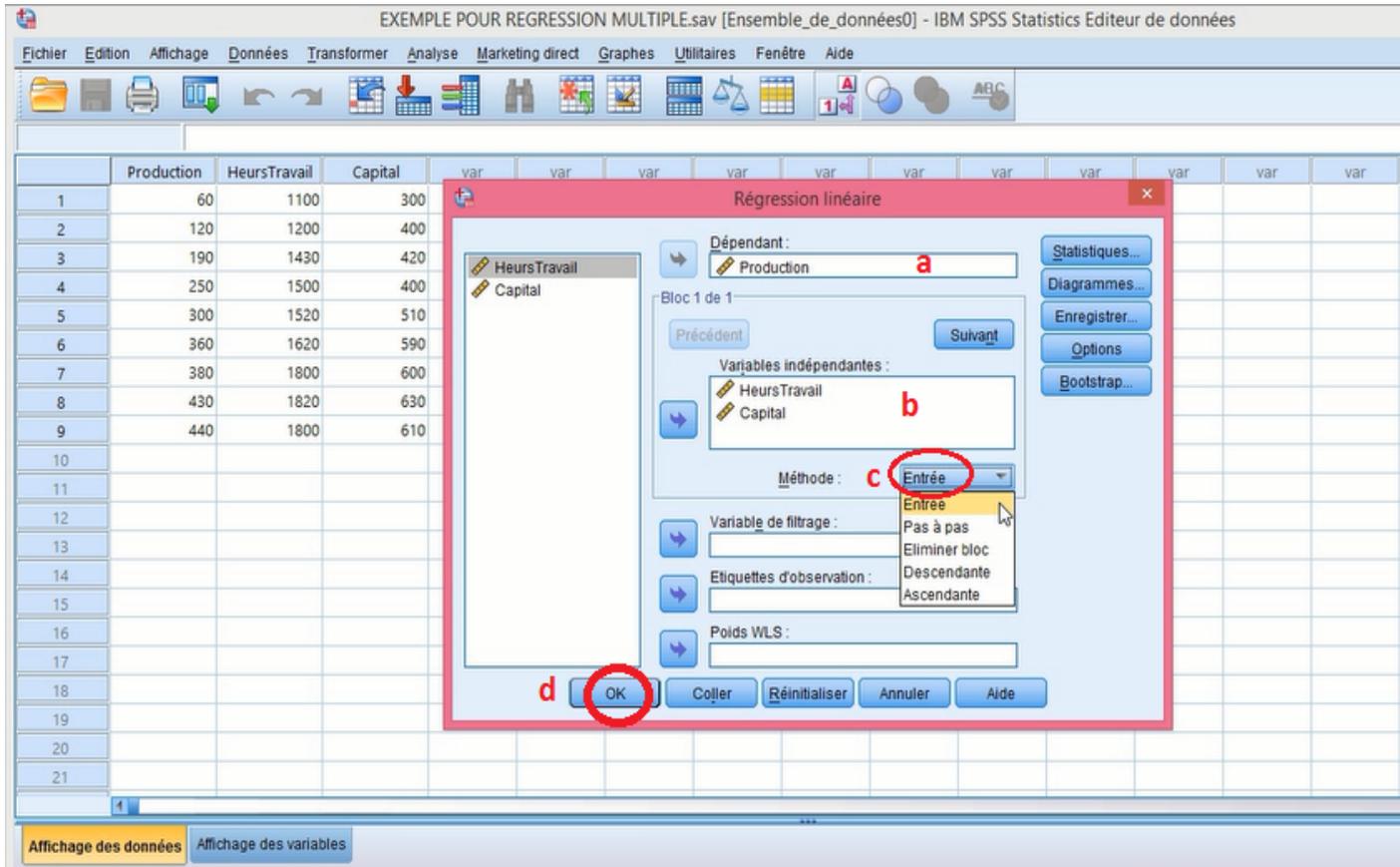


FIG. 1.17 – Réalisation de la régression linéaire multiple sous SPSS (Partie 2).

L'application de ces étapes sur les données de l'exemple permet d'obtenir les résultats (4 tableaux) qui sont présentés dans la figure 1.18. Ces résultats contiennent des réponses sur les questions posées.

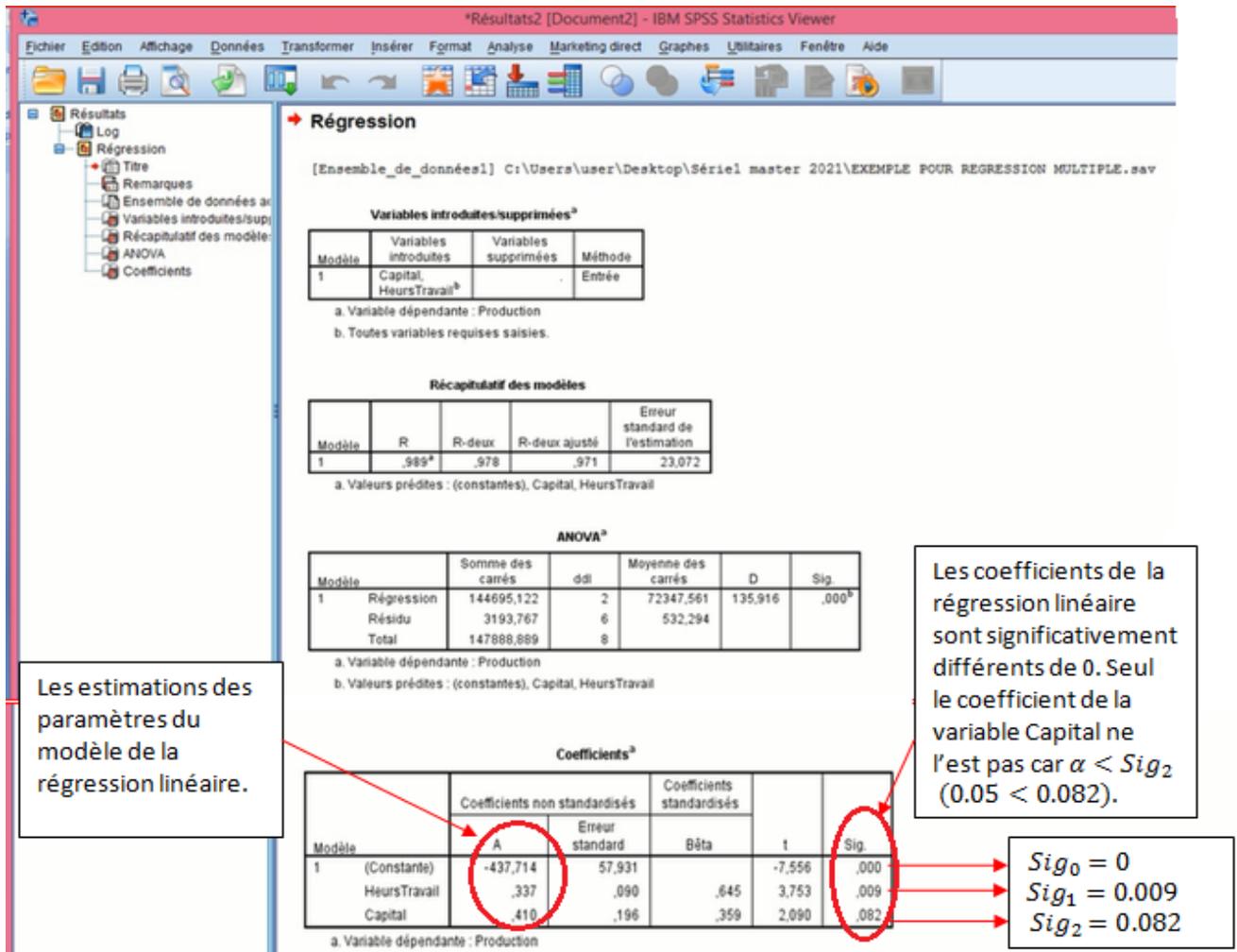


FIG. 1.18 – Résultats de la régression linéaire multiple sous SPSS.

### Interprétation des résultats obtenus

- Le premier tableau de la figure 1.18 présente les variables introduites dans le modèle de régression (le capital et l'heur de travail) et les variables qui sont exclues de l'entrée dans le modèle : ici la méthode "entrée" n'exclut pas les variables.
- Du deuxième tableau de la figure 1.18 on constate que les deux variables prises en compte (ou le modèle) expliquent 97.8% ( $R^2 = 0.978$ ) de la variance de la variable production.
- Pour la première question : les estimations des paramètres  $a$ ,  $b_1$  et  $b_2$  sont données dans le dernier tableau de la figure 1.18, alors
  - La valeur de la constante  $\hat{a} = -437.714$  ;
  - La valeur du coefficient de la variable Travail ( $X_1$ ) :  $\hat{b}_1 = 0.337$ .

- La valeur du coefficient de la variable Capital ( $X_2$ ) :  $\hat{b}_2 = 0.41$ .

Alors le modèle de la régression linéaire multiple sera donné par

$$\hat{Y} = -437.714 + 0.337X_1 + 0.41X_2.$$

**Remarque 1.4** *A la base d'un modèle de régression linéaire multiple, on peut prédire la valeur de la variable  $Y$  si on a des valeurs données pour les variables explicatives, donc il suffit de substituer ces valeurs dans l'équation du modèle obtenu.*

- Dans la deuxième question, on veut réaliser le test suivant (dit Test de significativité globale du modèle)

$$\begin{cases} H_0 : b_1 = b_2 = 0 \\ H_1 : \exists j/b_j \neq 0 (j = 1, 2) \end{cases} ;$$

et à l'aide de la valeur de signification (Sig) et au seuil de risque  $\alpha$ , on décide de

$$\begin{cases} \text{ne pas rejeter } H_0 & \text{si } \alpha < \text{Sig}; \\ \text{rejeter } H_0 & \text{si } \alpha \geq \text{Sig}. \end{cases}$$

Alors, de la table de l'ANOVA (troisième tableau de la figure 1.18) et au seuil de risque  $\alpha = 5\%$ , on constate qu'on peut rejeter  $H_0$  car  $\alpha > \text{Sig}$  ( $0.05 > 0$ ), donc le modèle obtenu est pertinent (valide).

- Les questions 3 et 4 concernant les tests de nullité de chacun des paramètres du modèle de la régression (Tests de significativité des paramètres du modèle). La réponse à ces questions se trouve dans la colonne signification du quatrième tableau de la figure 1.18, donc au seuil de risque  $\alpha = 5\%$ , on constate que

- pour le test d'hypothèse nulle ( $H_0 : a = 0$ ), on rejette  $H_0$  car  $\alpha > \text{Sig}_0$  ( $0.05 > 0$ ) (voir la première ligne du tableau 4), alors  $a$  est significativement différente de 0.

- pour le test d'hypothèse nulle ( $H_0 : b_1 = 0$ ), on rejette  $H_0$  car  $\alpha > Sig_1$  ( $0.05 > 0.009$ ) (voir la deuxième ligne du tableau 4), alors  $b_1$  est significativement différente de 0..
- pour le test d'hypothèse nulle ( $H_0 : b_2 = 0$ ), on ne peut pas rejeter  $H_0$  car  $\alpha < Sig_2$  ( $0.05 < 0.082$ ) (voir la troisième ligne du tableau 4) alors  $b_2$  est significativement égale à 0. Cela veut dire que la variable  $X_2$  (Capital) ne contribue pas significativement à la régression (elle n'a pas une influence significative sur la variable production) donc on peut l'exclure du modèle de la régression linéaire multiple et on fait à nouveau une autre régression en fonction d'une seule variable explicative ( $X_1$ ).

**Remarque 1.5** Pour choisir automatiquement les variables explicatives ayant une influence sur  $Y$ , en choisissant une option "*pas à pas*" dans "*Méthode*".

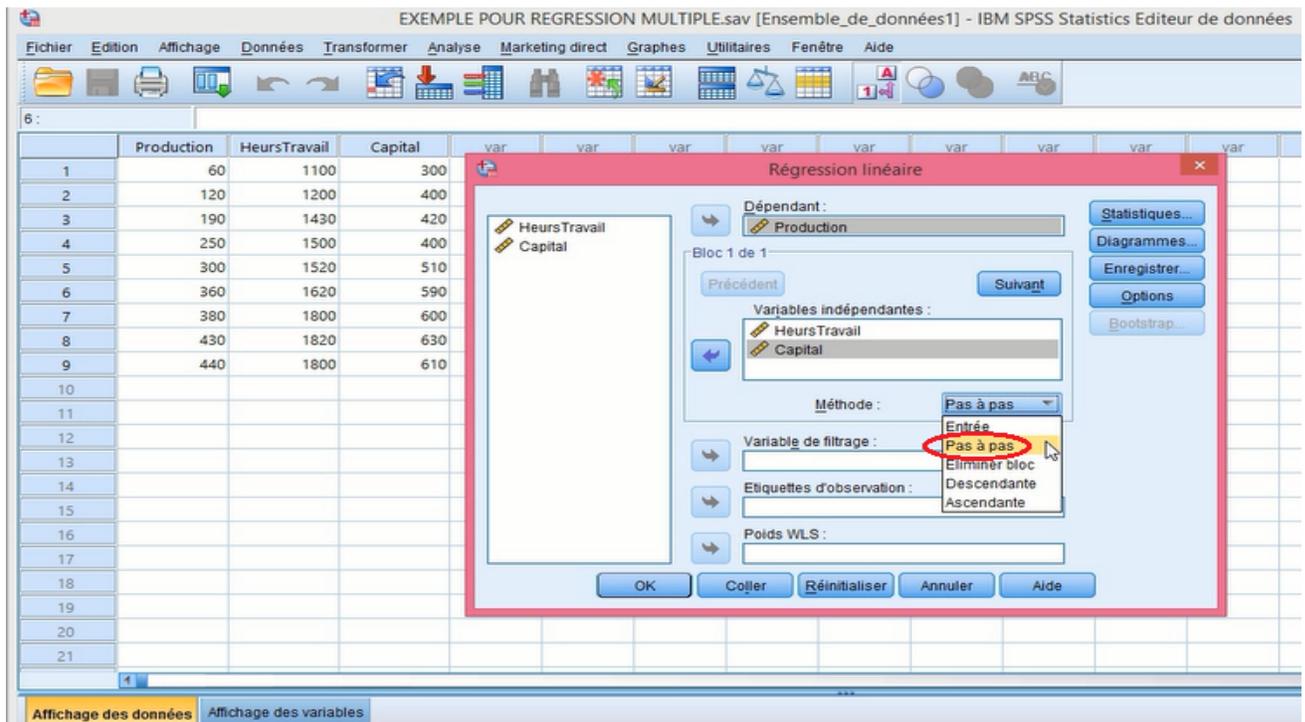


FIG. 1.19 – Procédure de la réalisation de la régression linéaire multiple par étapes (en choisissant la méthode "Pas à Pas").

### Méthodes de sélection des variables de régression linéaire

La sélection d'une méthode vous permet de spécifier la manière dont les variables indépendantes sont entrées dans l'analyse. En utilisant différentes méthodes, vous pouvez construire

divers modèles de régression à partir du même groupe de variables.

**Méthode "Entrée" (par défaut)** : Méthode qui introduit toutes les variables indépendantes simultanément. À utiliser si on veut déterminer l'équation de la droite de régression avec toutes les variables indépendantes.

Les autres méthodes sont des méthodes hiérarchiques. Seulement à utiliser si on pense qu'une des variables est plus importante que les autres.

**Méthode "Pas à pas"** : les variables indépendantes sont ajoutées à l'équation une par une et peuvent être enlevées subséquemment si elles ne contribuent plus significativement à la régression. Le processus s'arrête lorsqu'aucune variable ne peut plus être introduite ou éliminée.

**Méthode "Éliminer bloc"** : toutes les variables dans un bloc sont supprimées en une seule étape.

**Méthode "Descendante"** : toutes les variables sont entrées initialement dans l'équation et sont ensuite éliminées une à une. La variable ayant la plus petite corrélation avec la variable dépendante est d'abord étudiée pour l'élimination. Si elle est éliminée par le modèle, la prochaine variable avec le plus petit coefficient de corrélation est étudiée, jusqu'à ce qu'aucune variable ne satisfasse plus au critère d'élimination.

**Méthode "Ascendante"** : les variables sont introduites séquentiellement une par une. Si la première variable est introduite dans l'équation, la variable explicative ne figurant pas dans l'équation et présentant la plus forte corrélation partielle est considérée ensuite. La procédure s'arrête lorsqu'il ne reste plus de variables satisfaisant le critère d'introduction.