Université Mohamed Khider Biskra

Faculté des SE et SNV

Département de SNV

Module: Biostatistiques

1ère année Master

2019/2020

TP 3

Régression | Plan

Plan

- Définitions et Principes
- Régression linéaire simple
- Exercice

Régression | Définitions et Principes

Le but de la régression simple (resp. multiple) est d'expliquer une variable *Y* à l'aide d'une variable X (resp. plusieurs variables $X_1, X_2, \ldots, X_p$).

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 $X_1, X_2, \ldots, X_p$).**
- **De prédire la variable Y si la variable X (ou les variables $X_1, X_2, \ldots, X_p$) est connue.**

Remarque:

La variable Y est appelée variable dépendante, ou variable à expliquer et les variables X_j (j=1,...,p) sont appelées variables indépendantes, ou variables explicatives.

Régression | Définitions et Principes

C'est- à-dire, dans la régression simple (resp. multiple) on cherche d'une fonction *f* telle que

$$Y = f(X) + \varepsilon, \ (resp.\ Y = f(X_1, X_2, \dots, X_p) + \varepsilon),$$

où ε est une variable aléatoire (résidus).

Régression linéaire simple et multiple:

Si la fonction *f* est **affinée** (la relation est linéaire) on parle sur **la régression linéaire**. Alors

- **Dans le cas de la régression linéaire simple :**

$$f(X) = aX + b.$$

- **Dans le cas de la régression linéaire multiple :**

$$f(X_1, X_2, \dots, X_p) = a_0 + a_1X_1 + a_2X_2 + \cdots + a_pX_p.$$

Régression | Définitions et Principes

Exemples sur l'application de la régression

- **Étude de la température (Y) en fonction de l'altitude (X).**
- **Étude du poids (Y) en fonction de la taille (X) ou l'inverse.**
- **Étude du nombre de morts (Y) d'une maladie en fonction du nombre des infectés (X).**
- **Étude de la taille (Y) en fonction du poids (X_1) et de l'âge (X_2) .**

Modèle de la régression linéaire simple:

Soit un échantillon de n individus. Pour chaque individu, on a les observations qu'elles sont les valeurs des réalisations des **variables quantitatives X, Y** respectivement.
Pour n observations, on peut écrire le modèle de régression linéaire simple sous la forme

$$y_i = ax_i + b + \varepsilon_i, \text{ pour i} = 1, \dots, \text{n.}$$

On suppose que

-) $E(\varepsilon_i) = 0,$

-) $Cov(\varepsilon_i, \varepsilon_j) = 0, \forall i \neq j$ et $Var(\varepsilon_i) = \sigma^2$, i=1,…, n,

-) De plus, $\varepsilon \sim N(o, \sigma^2).$ (hypothèse pour les inférences)

Représentation graphique

Une étude de régression simple débute toujours par un tracé du nuage des n points de coordonnées (x_i, y_i) , i = 1, ..., n. Cette première représentation permet de savoir, à priori, si le modèle linéaire d'équation $Y = aX + b + \varepsilon$ est pertinent (le nuage des points est distribué sous une forme linéaire) et d'où les données peuvent être représentées sous forme d'une droite (dite droite de régression), qu'on peut l'ajouter au nuage de points.

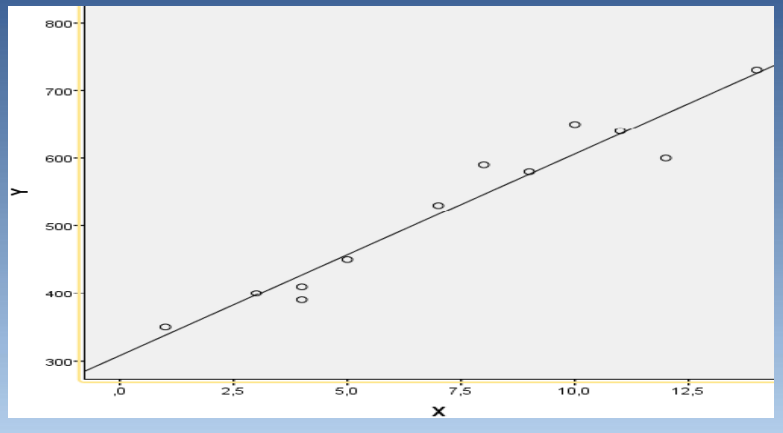


Figure 1. Nuage de points représentant la relation entre X et Y et la droite de régression.

Estimation des paramètres

Pour établir l'équation de la droite de régression, il convient de déterminer les valeurs de **a** (pente de la droite) et **b** (l'ordonnée à l'origine) dans l'équation.

Les estimateurs de ces deux valeurs sont donnés à l'aide de la méthode des moindre carrés par :

$$\hat{a} = \frac{Cov(X,Y)}{Var(X)}, \text{ où } Cov(X,Y) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) - \bar{X}\bar{Y} \text{ et } Var(X) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^{\,2}\right) - (\bar{X})^2 ;$$

$$\hat{b} = \bar{Y} - \hat{a}\bar{X}.$$

Une fois les coefficients de la droite estimés, on calcule

- Pour chaque individu :

- **la valeur prédite ou ajustée de *Y* par le modèle**

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

- **le résidu de l'observation *i*. C'est l'écart entre la valeur de *Y* observée sur l'individu *i* et la valeur prédite**

$$\hat{\varepsilon}_i = y_i - \hat{y}_i$$

- **La somme des carrés des résidus**

$$SCE = \sum_{i=1}^{n} \hat{\varepsilon}_i^{\,2}.$$

- **La variance est estimée par la variation résiduelle :**

$$\hat{\sigma}^2 = \frac{SCE}{n-2}.$$

Qualité et validation du modèle

a- Décomposition de la variation totale:

La variation totale de Y se décompose comme suit:

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{Y})^2}_{SCT}=\underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{SCR}+\underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{Y})^2}_{SCE},$$

où

- **SCT: Variation de Y ou variation totale.**
- **SCR: Variation des résidus (variation qui n'est pas expliquée par le modèle de régression).**
- **SCE: Variation de la régression ou variation expliquée par la régression.**

b- Coefficients de corrélation

Le coefficient de corrélation nous donne des informations sur l'existence d'une relation linéaire entre les deux variables considérées, il est donné par

$$\rho = Cor(X,Y) = \frac{Cov(X,Y)}{\sigma_X \times \sigma_Y},$$

où

$$\sigma_X{}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{X})^2 \; ;$$
$$\sigma_Y{}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{Y})^2.$$

Remarques:

- **Le coefficient de corrélation est toujours compris entre -1 et +1.**
- **Le signe positif indique que les deux variables varient dans le même sens, tandis que le signe négatif indique que les deux variables varient en sens inverse.**

➢ **Plus $|\rho|$ est près de 1, plus la corrélation est grande donc le modèle linéaire décrit bien le phénomène étudiée.**

➢ **Si ρ est nul ou proche de zéro, il n'ya pas de dépendance linéaire entre les deux variables.**

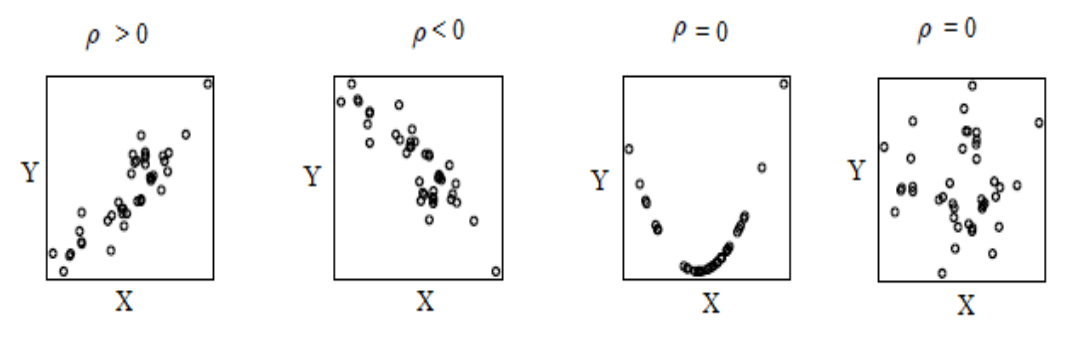


Figure 2. Nuages de points et coefficients de corrélation.

c- Coefficient de détermination

Afin d'avoir une idée globale de la qualité de l'ajustement linéaire, on définit $R^2\ (0 \leq R^2 \leq 1)$ le coefficient de détermination qui est le carré du coefficient de corrélation par:

$$R^2 = \frac{SCE}{SCT}.$$

Il mesure la part de la variation totale de Y expliquée par le modèle de régression sur X.

Remarque: *Plus le* R^2 *est près de 1, plus le modèle est adéquate et le contraire est vrai.*

d- Test sur le modèle (Test de Fisher)

Pour valider le modèle (Sous l'hypothèse de normalité des erreurs), on fait un test sur la pente, donc on pose les deux hypothèses suivantes :

$$\begin{cases} H_0 : a = 0 \\ H_1 : a \neq 0 \end{cases}$$

La statistique associée à ce test est la suivante

$$F = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{Y})^2/1}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/n-2} = \frac{SCE/1}{SCR/n-2} \sim f(1, n-2)$$

où $f(1, n-2)$ désigne une loi de Fisher de degrés de liberté n_1=1 et n_2=n-2. Ainsi, pour un risque α on décide que

- **Si $f_c > f(\alpha, 1, n-2)$, on peut rejeter H_0 alors le modèle est valide,**

- **Si $f_c \leq f(\alpha, 1, n-2)$, on ne peut pas rejeter H_0 le modèle donc n'est pas valide,**

dont f_c est la réalisation de la statistique F et $f(\alpha, 1, n-2)$ est le quantile d'ordre (1-α) de la loi de Fisher de degrés de liberté 1 et (n-2).

Régression | Régression linéaire simple

Remarque : ***Les résultats du test de Fisher associés au modèle de la régression linéaire simple peuvent être présentés dans un tableau (dite table d'analyse de variance) sous la forme suivante***

Source de Variation	ddl	Sommes des carrés (SC)	Carrés Moyens (CM)	f_c	f_α
Régression	1	SCE	$\text{CME} = \frac{\text{SCE}}{1}$	$f_c = \frac{\text{CME}}{\text{CMR}}$	$f_\alpha = f(\alpha, 1, n-2)$
Résidu	n-2	SCR	$\text{CMR} = \frac{\text{SCR}}{n-2}$		
Total	n-1				

Tab 1. Table d'analyse de variance.

Exercice: (exercice 13 page 27 polycopié du cours) *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 (µmol), nous avons réalisée des expériences dont la biomasse moyenne (Y) ainsi que la concertation du l'Azote (X) en question sont données dans le tableau suivant*

Régression | Régression linéaire simple

Concentration μmol	0	100	200	400	600
Biomasse mg	305	378	458	540	565

On donne : $\sum x_i = 1300$; $\sum y_i = 2246$; $\sum x_i^2 = 570000$; $\sum y_i^2 = 1056498$; $\sum x_i y_i = 684400$;

1. ***Présenter graphiquement le nuage des points. Que peut-on conclure sur le modèle proposer ?***
2. ***Calculer les estimations des paramètres a et b et donner la droite de régression.***
3. ***Calculer le coefficient de corrélation linéaire. Que peut-on conclure ?***
4. ***Pour un seuil de risque α=5%, le modèle proposé est-il pertinent ?***
5. ***Quelle Biomasse prévoyez-vous à une concentration 500 µmol ?***

Dans cet exercice on veut étudier:

la biomasse (Y) d'une plante en fonction de la concentration de l'Azote (X) .

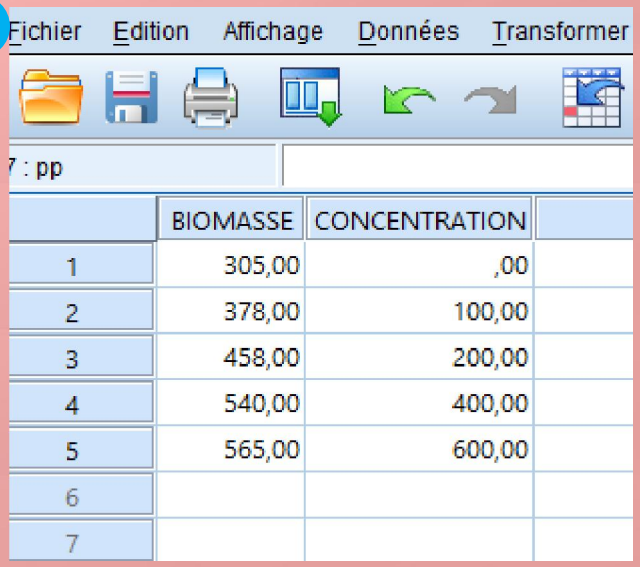
Variable dépendante

Variable indépendante

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.
(Comme vous avez vu précédemment).

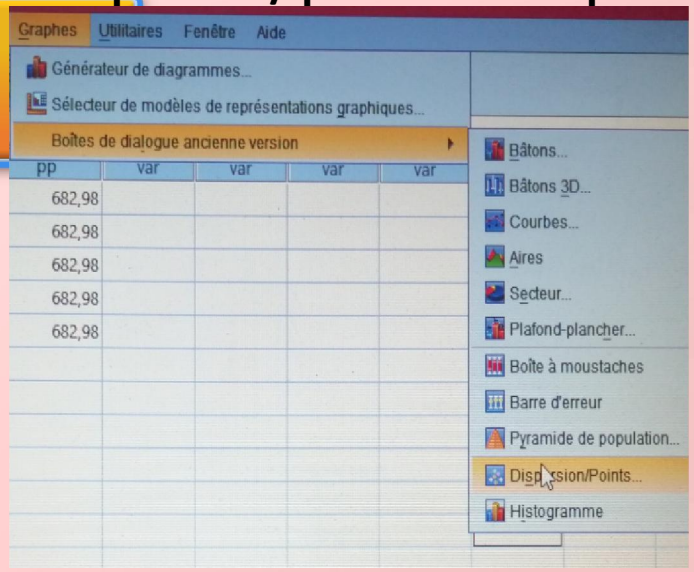


	BIOMASSE	CONCENTRATION
1	305,00	,00
2	378,00	100,00
3	458,00	200,00
4	540,00	400,00
5	565,00	600,00
6		
7		

***Figure 3*. Saisie des données sous SPSS pour une régression linéaire simple.**

Etape 2. Présentation graphique données (Nuage de points (question 1))

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



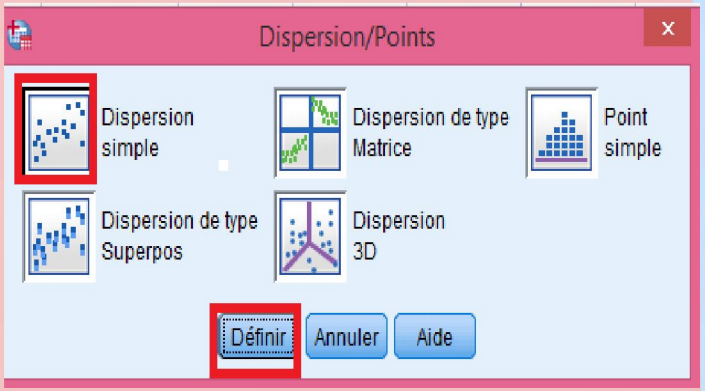


Figure 4 . Présentation graphique des données (Partie a).

Régression | Régression linéaire simple

Pla

Régr
simp

Exercice

b- *Dans la boîte de dialogue qui va apparaître (Figure 5) insérez la variable dépendante dans la case Axe des Y et la variable indépendante dans la case Axe des X puis cliquez sur OK.*

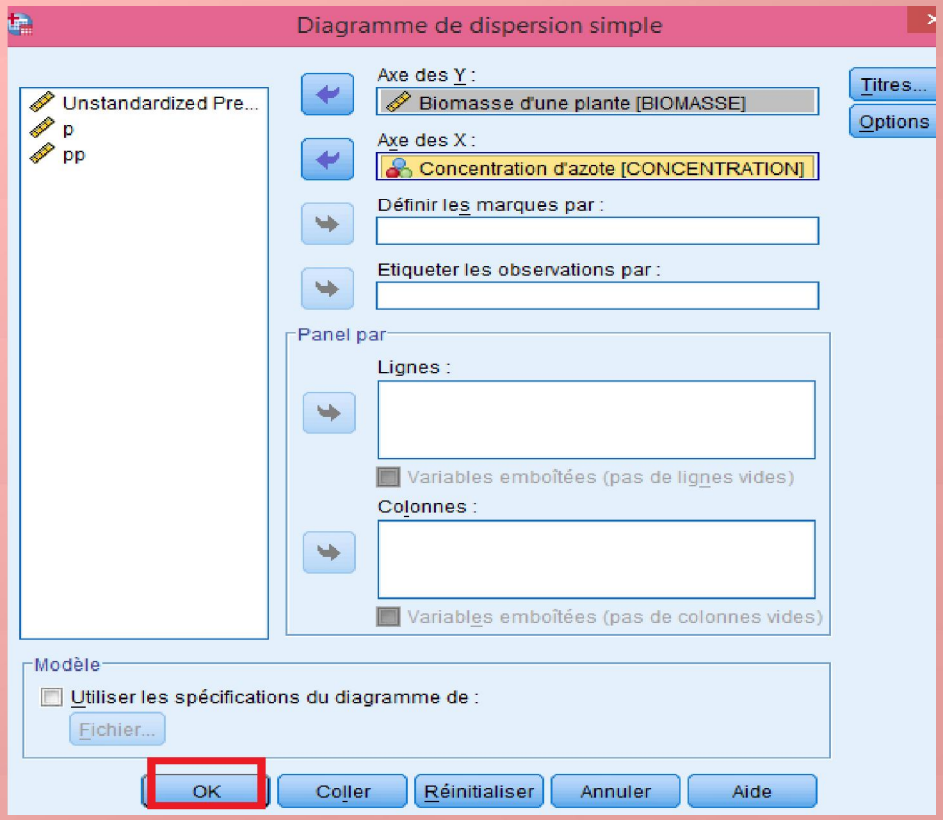


Figure 5. Présentation graphique des données (Partie b).

Régression | Régression linéaire simple

Plan

Défi
Princ

Régres
simple

Exercice

Une fois que vous cliquez sur OK, vous obtiendrez la Figure 6.

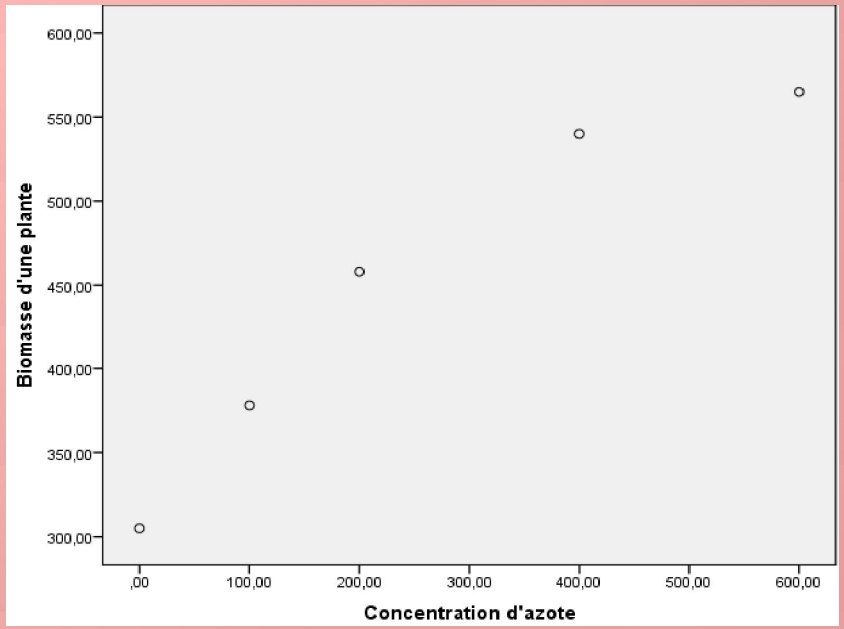


Figure 6. Nuage de points présente la Biomasse d'une Plante en fonction de la Concentration d'azote.

Au vue du graphique (Figure 6), il semble que le modèle linéaire est adéquat pour l'explication de Y en fonction de X (car le nuage des points est distribué sous une forme linéaire).

Régression | Régression linéaire simple

Remarque : *Vous pouvez ajouter la droite de régression au nuage de points en cliquant deux fois sur le nuage de points, puis cliquez sur Eléments → ajouter une courbe d'ajustement au total, vous obtiendrez alors ce qui suit (voir Figure 7).*

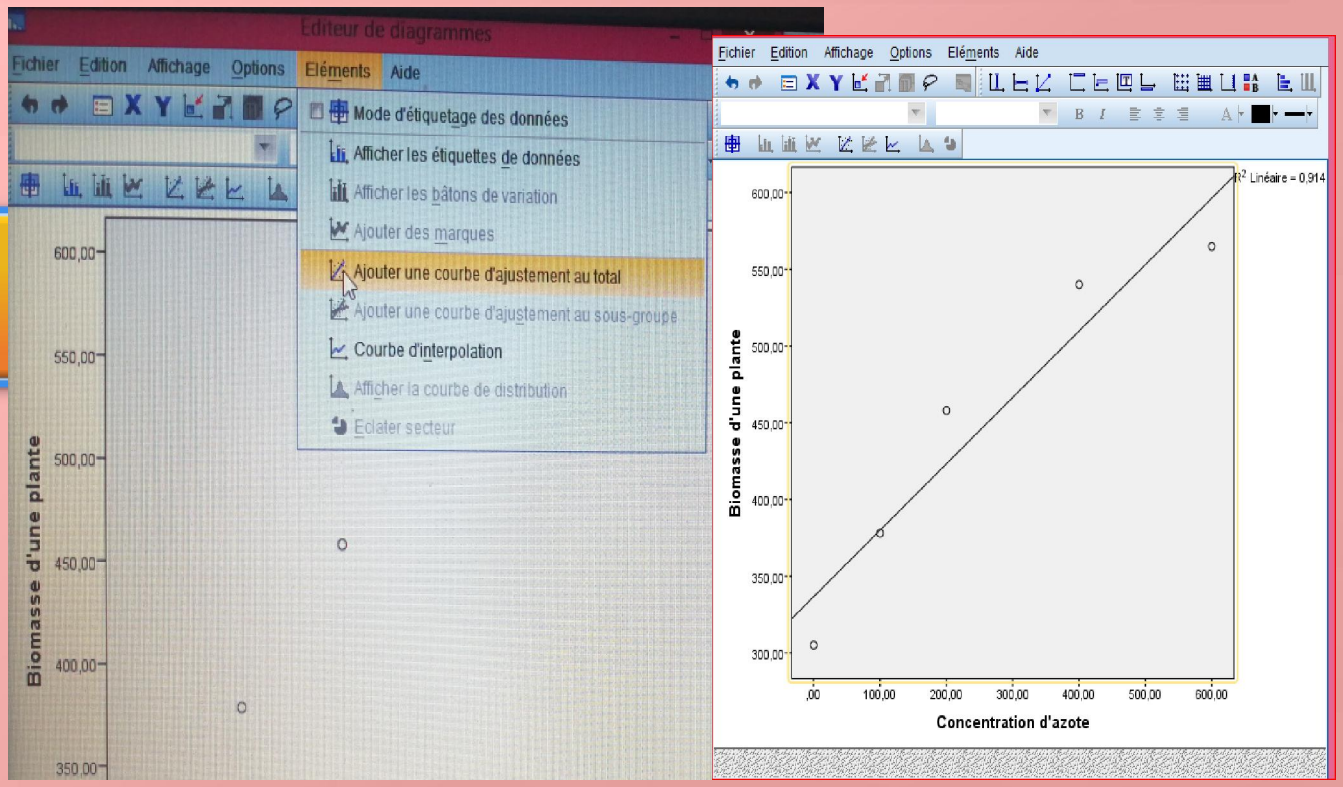


Figure 7. Nuage de points et droite de régression.

Régression | Régression linéaire simple

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 (voir Figure 8)

Analyse → Régression → Linéaire.

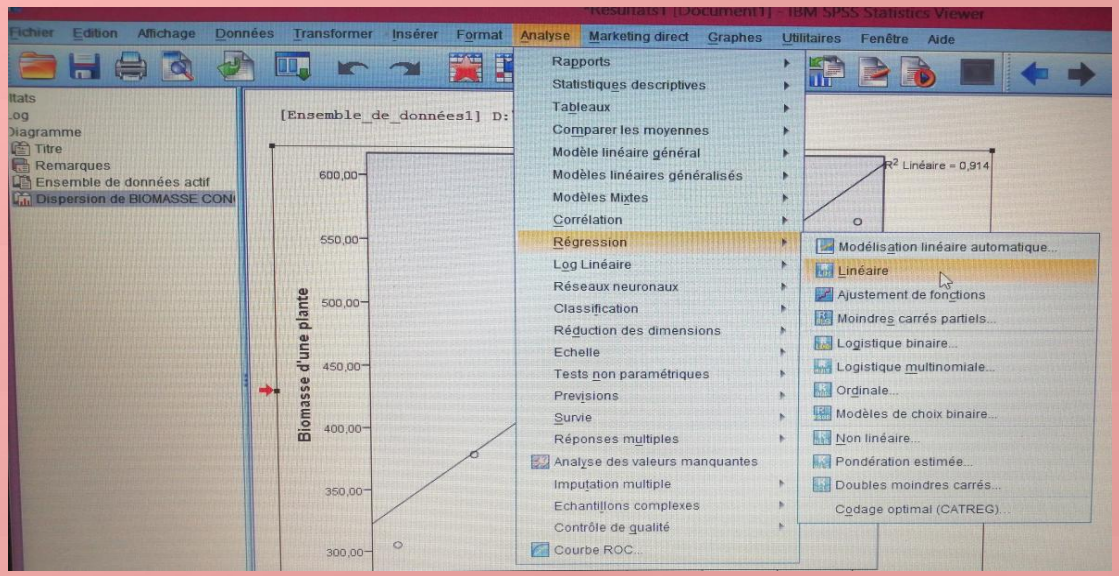


Figure 8. Procédures de la réalisation de la régression linéaire simple (Partie 1).

Régression | Régression linéaire simple

2- Dans la boîte de dialogue de la figure 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**, la variable indépendante (variable explicative) dans la case **Variables indépendantes** puis cliquez sur OK .

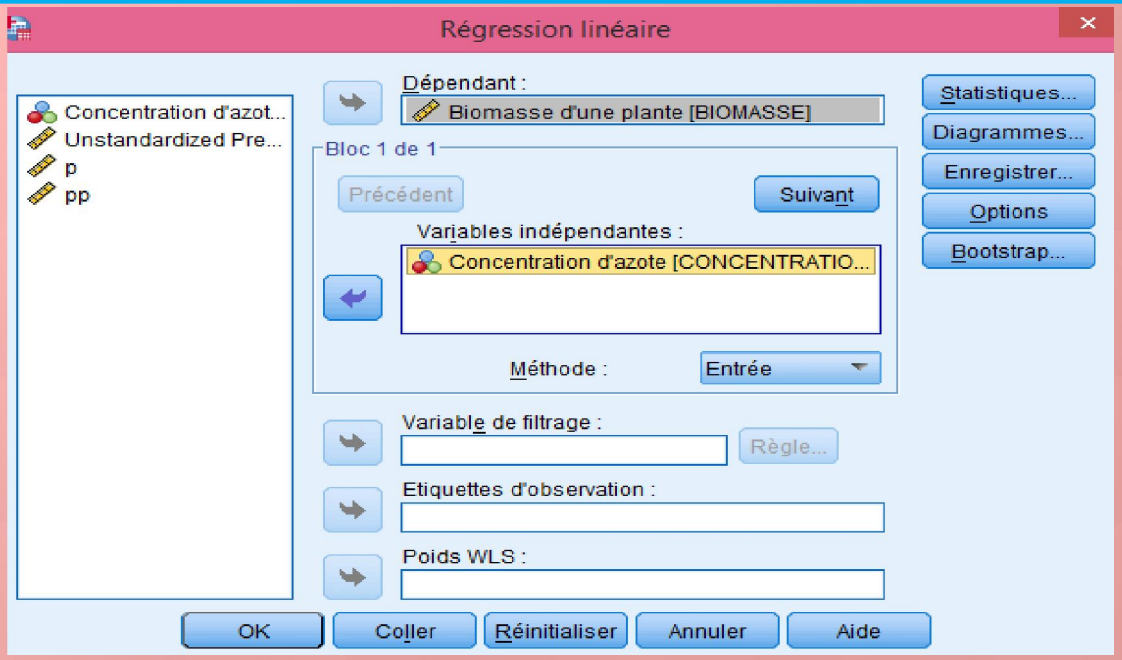


Figure 9. Procédures de la réalisation de la régression linéaire simple (Partie 2).

Régression | Régression linéaire simple

Pla

Défi
Prin

Régre
simpl

Exercice

L'application de ces étapes sur les données de l'exercice permet d'obtenir les résultats (4 tableaux) qui sont présentés dans la figure 10. Ces résultats contiennent des réponses sur les questions 2, 3 et 4.

Variables introduites/supprimées[a]

Modèle	Variables introduites	Variables supprimées	Méthode
1	Concentration d'azote[b]	.	Entrée

a. Variable dépendante : Biomasse d'une plante

b. Toutes variables requises saisies.

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,956[a]	,914	,885	37,01895

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

ANOVA[a]

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

a. Variable dépendante : Biomasse d'une plante

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

Coefficients[a]

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

Figure 10. Résultats de la régression linéaire simple.

Régression | Régression linéaire simple

➢ **Le premier tableau récapitule Les variables explicatives prises en compte dans le modèle.**

Ici, il n'y a qu'une seule variable puisque nous travaillons sur une régression simple.

Variables introduites/supprimées[a]

Modèle	Variables introduites	Variables supprimées	Méthode
1	Concentration d'azote[b]	.	Entrée

a. Variable dépendante : Biomasse d'une plante

b. Toutes variables requises saisies.

➢ **Le deuxième tableau donne deux valeurs importantes dans le modèle de régression :**

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,956[a]	,914	,885	37,01895

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

R=0.956 (coefficient de corrélation) et R^2 =0.914 (coefficient de détermination).

- Le coefficient de corrélation est presque égal à 1, ce qui indique qu'il y a une forte liaison linéaire entre X et Y.

- Le coefficient de détermination égale à 0.914, ce qui indique que 91.4% de la variation totale de Y est expliquée par le modèle de régression sur X.

Régression | Régression linéaire simple

Le troisième tableau (c'est la table d'analyse de Variance) indique si le modèle est valide ou non. Il donne :

ANOVA[a]

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

a. Variable dépendante : Biomasse d'une plante

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

-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 risque α=5% que le modèle obtenu est pertinent (valide) car Sig < 0,05, 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 : $\hat{Y} = \hat{a}X + \hat{b}$ **, où** $\hat{a}$ **et** $\hat{b}$ **sont donnés dans le quatrième tableau.**

Régression | Régression linéaire simple

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: $\hat{Y} = \hat{a}X + \hat{b}$ **, où** $\hat{a}$ **et** $\hat{b}$ **sont donnés dans le quatrième tableau.**

Coefficients[a]

Modèle		Coefficients non standardisés		Coefficients standardisés		
		A	Erreur standard	Bêta	t	Sig.
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

➢ **Le quatrième tableau donne alors les deux valeurs suivantes :**

- **336 .638 qui est la constante ou l'ordonnée à l'origine** $\hat{b}$ **,**
- **0.433 qui est la pente** $\hat{a}$ **.**

Alors

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

Pour la dernière question, il suffit de substituer la valeur de la concentration d'azote (X=500 µmol) dans la dernière équation pour trouver la valeur prédite $(\hat{Y})$ **de la biomasse de la plante.**