TP : 02

Ce TP vise d'affirmé l'existence d'une relation entre deux variables, on utilisant la corrélation ensuite déterminé la relation entre les deux variable on pratiquant la régression simple.

Rappel sur la corrélation entre deux variables.

La corrélation de Pearson ou de Bravais – Pearson est une mesure de la liaison linéaire existant entre deux variables quantitatives normales.

$$r_{xy} = \frac{\mathrm{Cov.xy}}{\mathrm{Sx} * \mathrm{Sy}} = \frac{\sum \mathrm{xy} - \frac{\sum \mathrm{x} \sum \mathrm{y}}{\mathrm{n}}}{\sqrt{\mathrm{SCEx} * \mathrm{SCEy}}}$$

Toutes les valeurs de $r_{x,y}$ sont comprises entre -1 et +1. r = -1 ou +1 si tous les points du diagramme de dispersion sont situés sur une ligne droite et r = 0 lorsque le nuage de dispersion ne montre aucune tendance de relation entre les deux variables. Aussi, si r est positif ⇒ les deux variables augmentent au même temps et si r est négatif ⇒ l'une des variables augmente quand l'autre diminue.

<u>Application numerique</u> : On désire vérifier la corrélation entre la taille (en cm) et le poids (en kg) des enfants de 2 ans sur un échantillon de 15 individus.

Taille (X)	82.9	83.4	82.4	82.1	84.8	86.7	84.0	89.0	85.0	85.4	87.7	87.7	86.4	86.4	86.9
Poids (Y)	8.7	9.2	9.5	10.1	10.4	10.5	10.8	11.0	11.5	11.6	12.4	13.6	13.8	13.9	14.6

Calcul de $r_{x,y}$

XY	721.23	767.28	782.8	829.21	881.92	910.35	907.2	979	977.5	990.64	1087.5	1192.3

$\sum x = 1280.8$	$\sum y = 171.6$
SCEx = 62.1	SCEy = 47.9

$\sum xy = 14689.4$

$$r_{xy} = \frac{\mathrm{Cov.xy}}{\mathrm{Sx}*\mathrm{Sy}} = \frac{\sum \mathrm{xy} - \frac{\sum \mathrm{x} \sum \mathrm{y}}{\mathrm{n}}}{\sqrt{\mathrm{SCEx} * \mathrm{SCEy}}} = \frac{14689.4 - \frac{1280.8*171.6}{15}}{\sqrt{62.1 * 47.9}} = \frac{37}{54.5} = 0.68$$

Test de signification du coefficient de corrélation de Pearson

on estime r et on vérifie si l'estimation obtenue est suffisamment distante de 0 pour rejeter l'hypothèse d'indépendance (r = 0). H0 : r = 0 Les variables ne sont pas corrélées

:

$$t_r = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$ et comparer cette valeur à $t_{1-\alpha/2}$ pour un ddl = n-2.

Si $t_r > t_{1-\alpha/2} \Rightarrow$ H0 est rejetée $\Rightarrow$ la liaison est significative.

Si $t_r \leq t_{1-\alpha/2} \Rightarrow$ H0 est acceptée $\Rightarrow$ la liaison n’est pas significative.

Régression linéaire simple

On parle de régression linéaire simple lorsqu’on désire calculer une fonction du premier degré liant les variables « x » et « y ». Cette fonction linéaire est de la forme :

$$\boldsymbol{y = ax + b}$$

Elle correspond à l’équation d’une ligne droite qui traverse au mieux le nuage de points et permet de calculer une valeur estimée « ŷ » pour chaque point de l’axe des x correspondant à la variable prédictive. Cette droite porte le nom de droite d’estimation ou droite de régression de « y » en « x ». Cette droite est obtenue par la méthode des moindres carrés dont le principe consiste à choisir la pente « a » et l’ordonnée à l’origine « b » de la droite qui minimisent la somme des carrés des erreurs. L’erreur est représentée par l’écart entre la valeur observée « y_i » et la valeur prédite par la droite « $\hat{y}_i$ ». Cet écart « $e_i = y_i - \hat{y}_i$ » est appelé également « résidu ». La méthode des moindres carrés consiste à réduire au maximum la somme des carrés des écarts (résidus) :

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2.$$

$b = \bar{y} - a\bar{x}$ et $a = \frac{S_{xy}}{S_x^2}$ avec $S_{xy} = \frac{n\sum xy - (\sum x)(\sum y)}{n(n-1)}$

Application numérique :

Hauteur	Diamètre	Hauteur	Diamètre	Hauteur	Diamètre	Hauteur	Diamètre
10	65	14	72	19	82	23	92
11	68	15	75	20	86	24	98
12	67	16	76	20	84	24	101
12	69	17	80	21	85	25	105
13	69	18	82	22	88	25	109

Déterminons l’équation de régression linéaire qui permet de prédire le diamètre du tronc à partir de la hauteur des arbres et calculons le coefficient de corrélation linéaire « r ».

L’estimation des coefficients « a » et « b » de la droite de régression et le calcul du coefficient de corrélation « r » nécessitent les calculs suivants :4.84 12.78

N = 20 arbres $\sum x = 361$ $\sum x^2 = 6985$ $\bar{X}$ = 18.05 $S_x = 4.97$ $S_x^2 = 24.7$

$\sum y = 1653$ $\sum y^2 = 139889$ $\bar{y}$ = 82.65 $S_y = 13.11$ $S_y^2 = 171.87$ $\sum xy = 31033$

$S_{xy} = \frac{n \sum xy - (\sum x)(\sum y)}{n(n-1)} = \frac{20 * 31033 - 361 * 1653}{20(19)} = \frac{620660 - 596733}{380} = 62.96$

$r = \frac{S_{xy}}{S_x * S_y} = \frac{62.96}{65.15} = 0.96$

La pente de la droite de régression « a » est égale à :

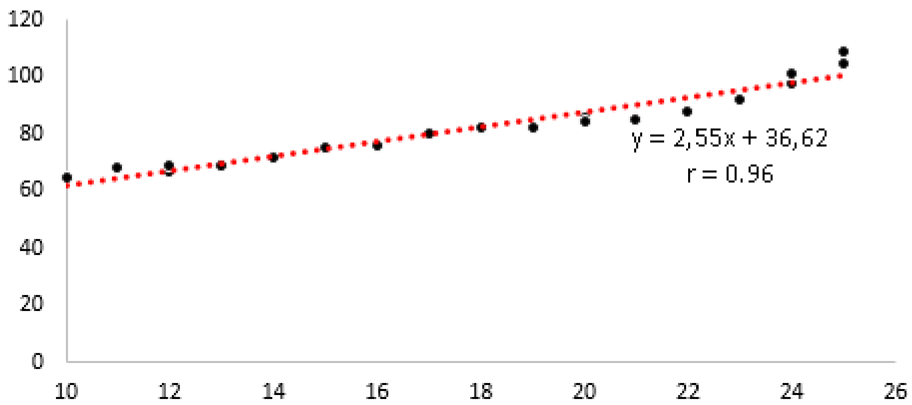
$a = \frac{S_{xy}}{S_x^2} = \frac{62.96}{24.70} = 2.55$

L’ordonnée à l’origine « b » est égale à :

$b = \bar{y} - a\bar{X} = 82.65 - 2.55\,(18.05) = 36.62$

On aura ainsi l’équation de régression suivante :

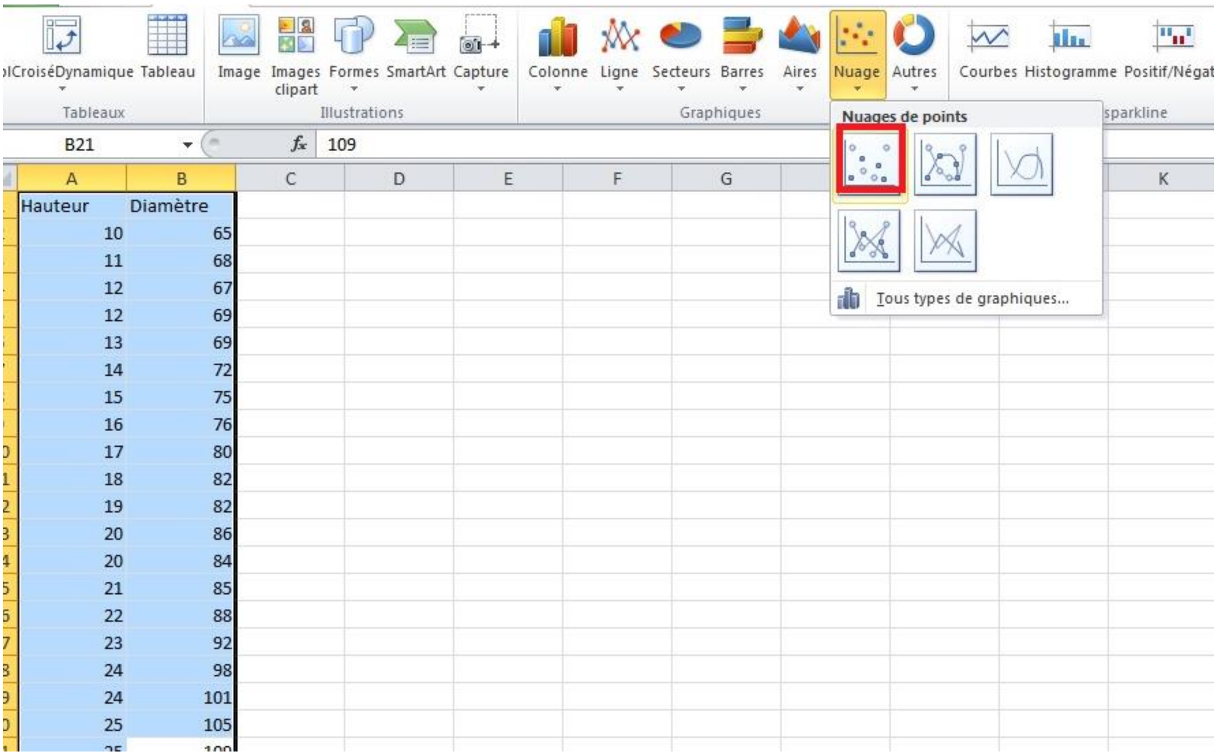
$y = 2.55\,x + 36.62$



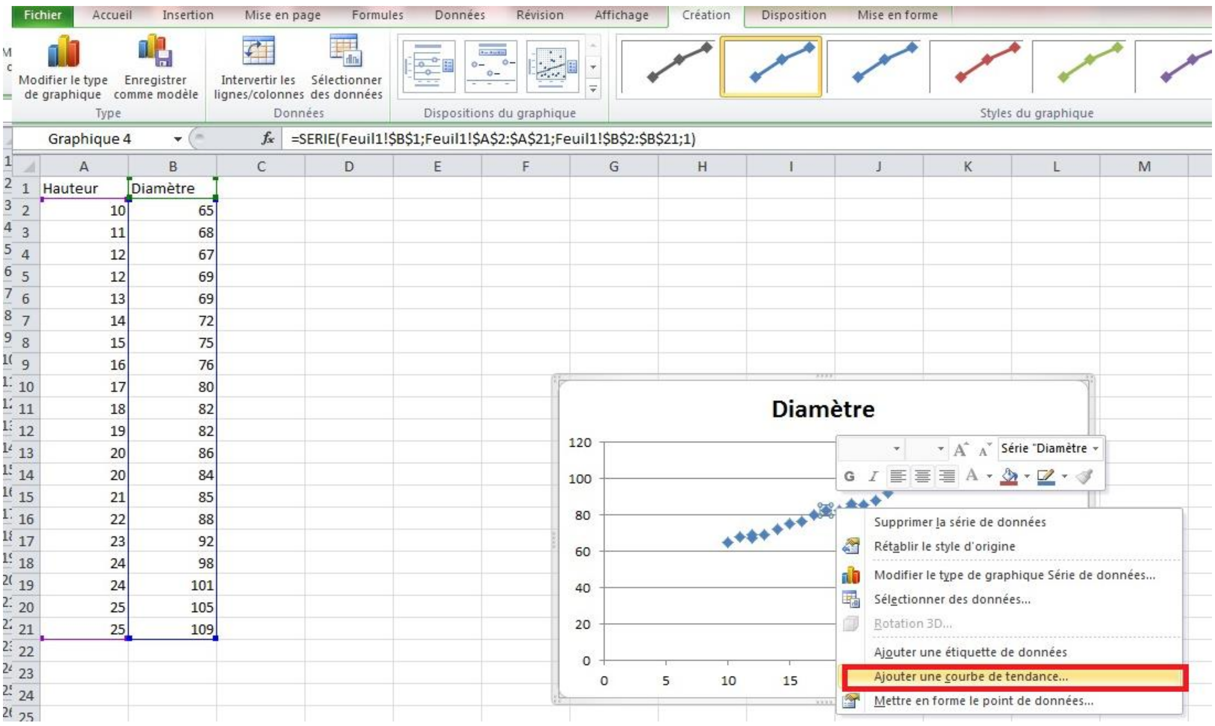
Application TP sur Excel :

Une étude sur la relation entre deux caractéristiques dendrométriques (Hauteur des arbres en mètre et diamètre du tronc en cm.) d’un peuplement forestier a été menée sur un échantillon aléatoire de 20 arbres et les résultats sont les suivants :

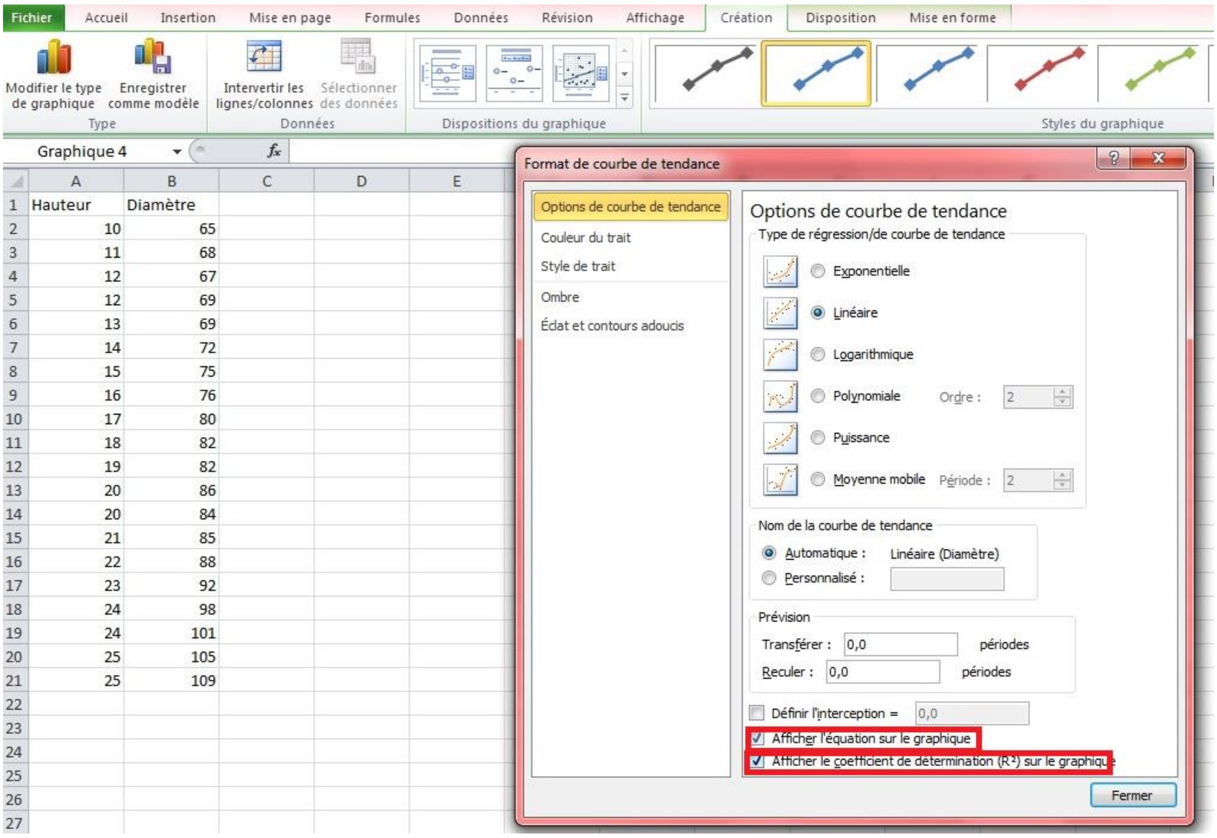
Entrer les données des hauteurs et des diamètres des arbres/ **sélectionner** les plages de données / **Insertion/ Nuage de points / Nuage de points avec marqueur uniquement**



Cliqué sur un des point du graphe avec le bouton de droite de la souris/ **Ajouté une courbe de tendance**



Dans la fenêtre suivante /coché -1 **Afficher l'équation sur le graphe** -2 **Afficher le coefficient de détermination (R^2)**


Fichier
Accueil
Insertion
Mise en page
Formules
Données
Révision
Affichage
Création
Disposition
Mise en forme
Modifier le type de graphique
Enregistrer comme modèle
Type
Intervertir les lignes/colonnes
Sélectionner des données
Données
Dispositions du graphique
Styles du graphique
Graphique 4
A
B
C
D
E
Hauteur
Diamètre
10 65
11 68
12 67
12 69
13 69
14 72
15 75
16 76
17 80
18 82
19 82
20 86
20 84
21 85
22 88
23 92
24 98
24 101
25 105
25 109
Format de courbe de tendance
Options de courbe de tendance
Couleur du trait
Style de trait
Ombre
Éclat et contours adoucis
Options de courbe de tendance
Type de régression/de courbe de tendance
Exponentielle
Linéaire
Logarithmique
Polynomiale
Ordre : 2
Puissance
Moyenne mobile
Période : 2
Nom de la courbe de tendance
Automatique : Linéaire (Diamètre)
Personnalisé :
Prévision
Transférer : 0,0 périodes
Reculer : 0,0 périodes
Définir l'interception = 0,0
Afficher l'équation sur le graphique
Afficher le coefficient de détermination (R²) sur le graphique
Fermer