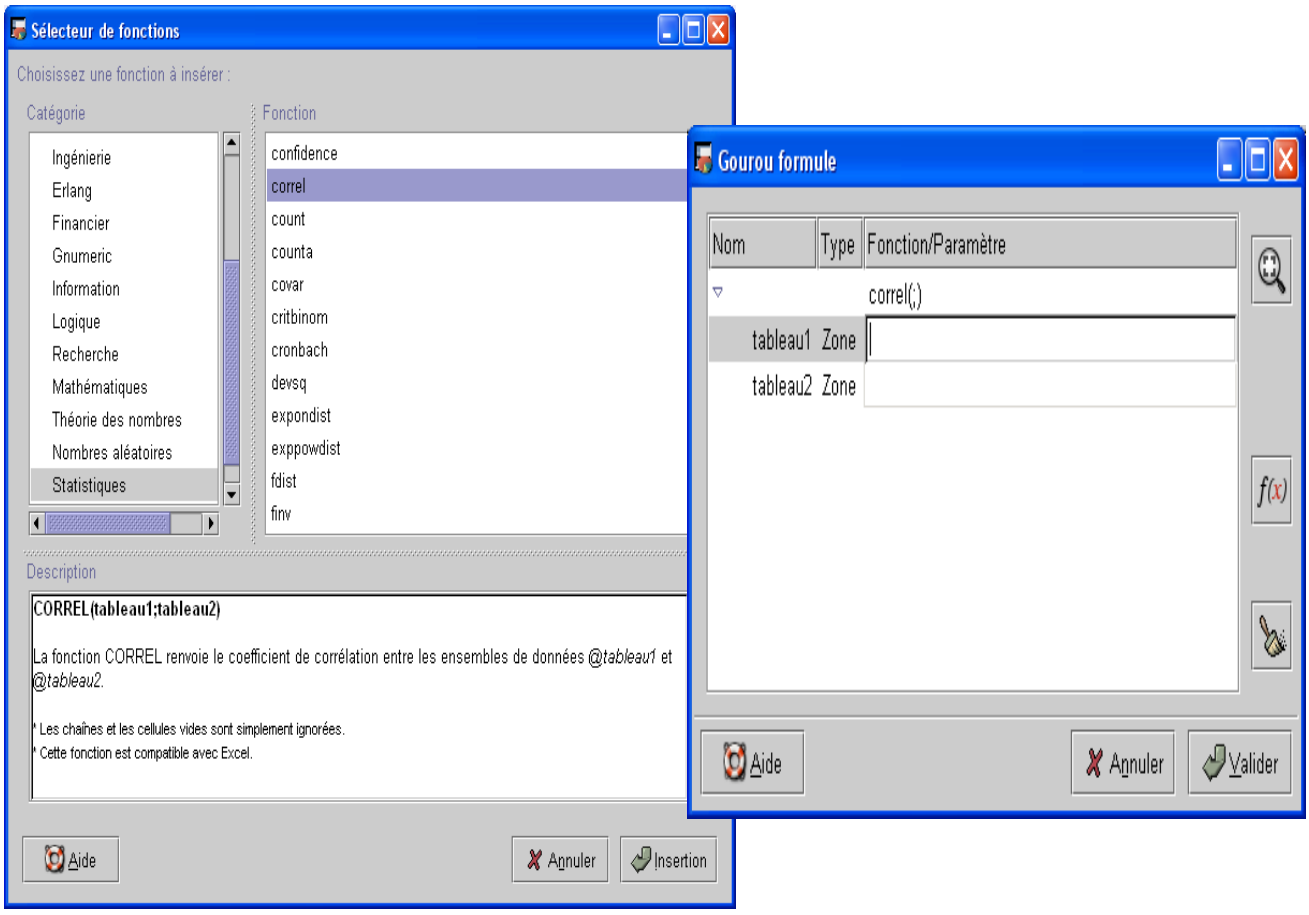
TP 4 : Statistique descriptive bidimensionnelle

A. PHILIPPE

Exercice 1 Corrélation

La corrélation entre deux variables X et Y mesure le lien linéaire entre deux variables . La fonction **correl** calcule la corrélation entre deux variables :



On considère deux vecteurs $X=(X_n), n\in\{1,...100\}$ et $Y=(Y_n), n\in\{1,...100\}$.

(1) Vérifier que si $X_n=n$ et $Y_n=2X_n$ alors le coefficient de corrélation entre X et Y vaut 1

(2) Vérifier que si $X_n=n$ et $Y_n=-2X_n$ alors le coefficient de corrélation entre X et Y vaut -1

(3) Calculer le coefficient de corrélation entre X et Y si $X_n=n$ et $Y_n=\sqrt{X_n}$

(4) Calculer le coefficient de corrélation entre X et Y si $X_n=n$ et $Y_n=X_n^2$

Exercice 2 coefficient de corrélation (suite)

(1) Simuler deux échantillons $X=(X_n), n\in\{1,...100\}$ et $Y=(Y_n), n\in\{1,...100\}$ indépendants et de même loi normale de moyenne 0 et de variance 1.

(2) Tracer le nuage de points $(X_n, Y_n), n\in\{1,...100\}$

(3) Calculer le coefficient de corrélation entre les deux échantillons aléatoires X et Y.

(4) Commenter.

(5) Construire la série $Z=(Z_n), n\in\{1,...100\}$ définie par $Z_1=X_1$ et pour j>1 $Z_j=a\,Z_{j-1}+X_j$ où a=0,85

(6) Tracer le nuage de points $(Z_n, Z_{n+1}), n\in\{1,...99\}$ et calculer le coefficient de corrélation entre $(Z_1,\dots,Z_{99})$ et $(Z_2,\dots,Z_{100})$

(7) Commenter.

Exercice 3 Le bon usage du coefficient de corrélation

Le fichier Anscombre.gnumeric contient 5 séries de points.

(1) Récupérer le fichier de données à l'adresse suivante

http://www.math.sciences.univ-nantes.fr/~philippe/lecture/Anscombre.gnumeric

(2) Calculer pour chacune des cinq séries

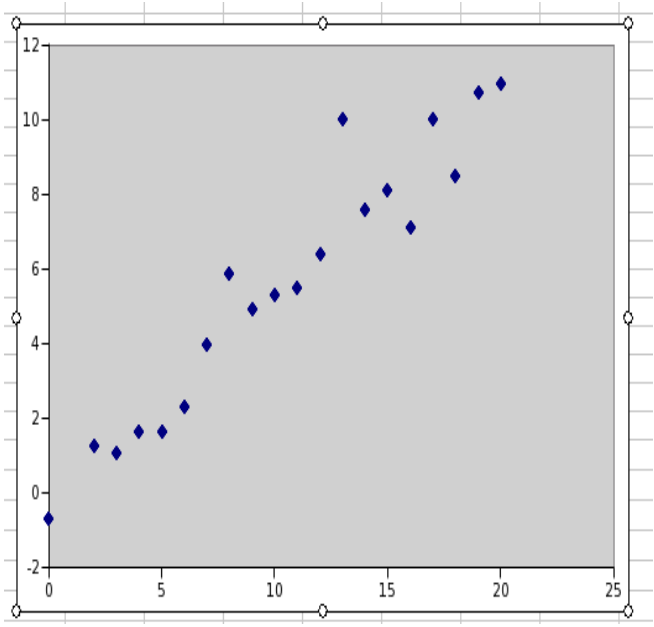
- la moyenne et la variance de la variable X
- la moyenne et la variance de la variable Y
- la corrélation entre X et Y

(3) Comparer les résultats obtenus

(4) Tracer les 5 nuages de points $(X_n, Y_n), n \in \{1,...100\}$

(5) Commenter et conclure

Régression linéaire



Lorsque l'on observe un nuage de points $(X_n, Y_n), n \in \{1,...100\}$ tel que

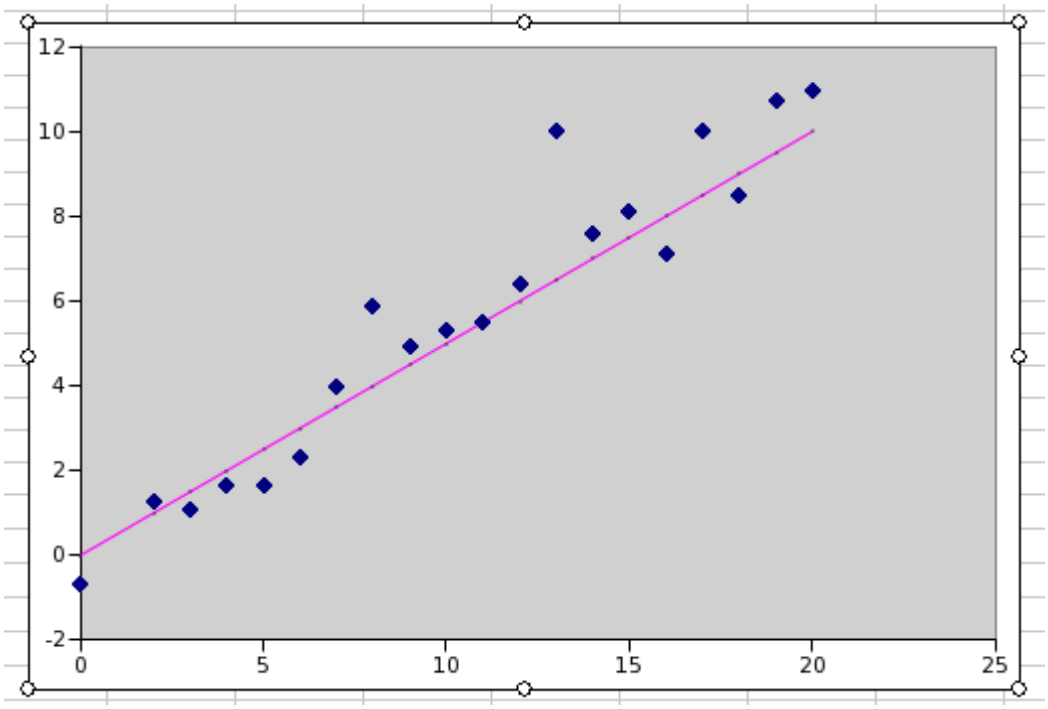
- l'allure est proche de celle d'une droite
- la valeur absolue du coefficient de corrélation est proche de 1,

  ici on obtient 0,95

On peut modéliser le nuage de points par une droite.

On cherche la droite $y = ax + b$ qui approche le mieux le nuage de points pour le critère des moindres carrés c'est à dire on cherche les coefficients (a, b) qui minimisent

$$\sum_{j=1}^{n} (Y_j - aX_j - b)^2$$



Exercice 4 Modélisation de la distance de freinage

Le fichier car.gnumeric contient pour n véhicules

- la distance de freinage
- la vitesse

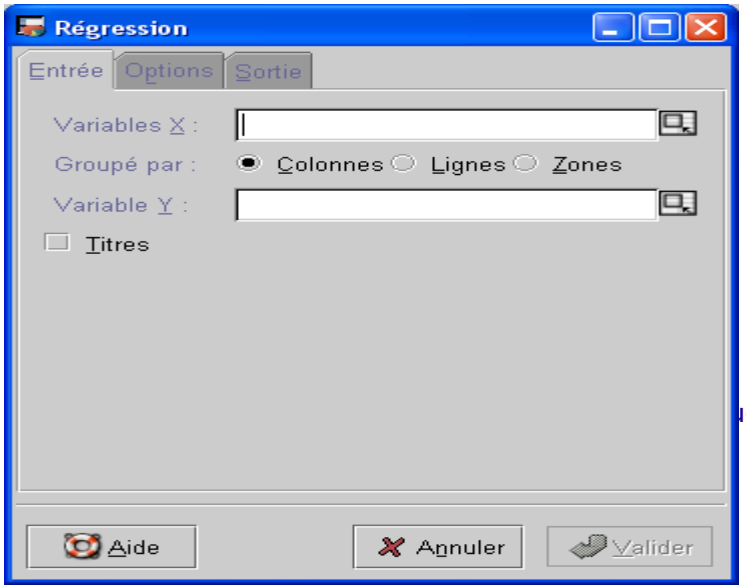
Dans cet exercice, on veut prévoir la distance de freinage à partir de la vitesse.

(1) Recupérer le fichier de données à l'adresse suivante
http://www.math.sciences.univ-nantes.fr/~philippe/lecture/car.gnumeric

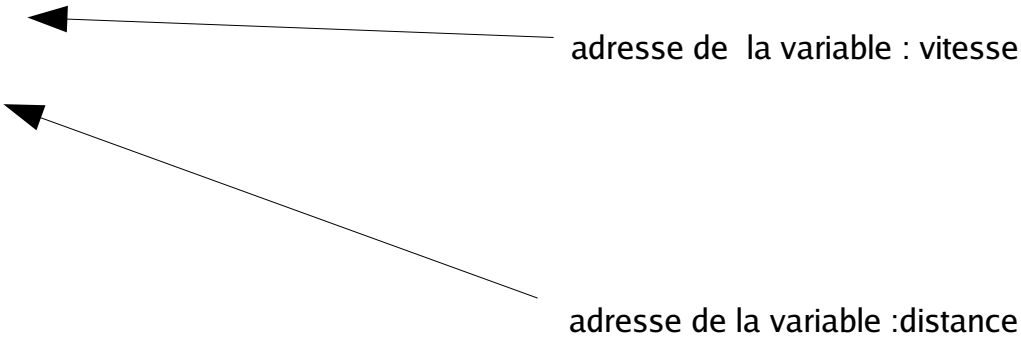
(2) Tracer le nuage de points $(vitesse_n, distance_n), n \in \{1,...50\}$

(3) Calculer le coefficient de corrélation entre les variables vitesse et distance.

(4) Commenter les résultats. Le modèle linéaire $distance = a\, vitesse + b$ est -il pertinent ?



(5) Calculer les coefficients a et b en utilisant la fonction régression accessible via le menu outils / analyse statistique / régression



(6) Représenter sur un même graphique, la droite obtenue et le nuage de points

(7) Pour chacun des points du nuages, on peut calculer l'erreur commise en utilisant le modèle linéaire, c'est à dire $e_j = distance_j - a\ vitesse_j - b$. Calculer puis tracer la valeur absolue des erreurs $e_j,\ j = 1,...50$.

(8) Calculer la moyenne et la variance de l'échantillon $e_j,\ j = 1,...50$

(9) On veut maintenant prévoir la distance de freinage par $dist_{prév} = a\,v + b$ où v est la vitesse. Calculer et tracer la prévision pour v=1,.....,40.

Un autre modèle.

(10)On exprime la distance en fonction de la vitesse et du carré de la vitesse. Construire une colonne contenant le carré de la vitesse

	A	B	C	D	E	F
1		vitesse2	vitesse	distance		
2	1	16	4	2		
3	2	16	4	10		
4	3	49	7	4		
5	4	49	7	22		
6	5	64	8	16		
	6	81	9	10		
	7	100	10	18		

(11)En utilisant la fonction régression, calculer les coefficients $a_{1,} a_{2,} b$ qui minimisent

$$\sum_{j=1}^{n} (distance_j - a_1\ vitesse_j - a_2\ vitesse_j^2 - b)^2$$

(12)Superposer sur le nuage de points et la parabole $a_1\ vitesse_j + a_2\ vitesse_j^2 + b$

(13)Calculer les erreurs $e_j' = distance_j - a_1\ vitesse_j - a_2\ vitesse_j^2 - b$

(14)Calculer la moyenne et la variance de l'échantillon $e'_j,\ j = 1, ...50$

(15)Représenter sur un même graphique les séries $(|e_j|,\ j = 1..50)$ et $(|e_j'|,\ j = 1..50)$

(16)Conclure