1 Formules mathématiques de la droite de régression linéaire

Définitions.
Soit X et Y deux séries statistiques (numériques) ayant même effectif n :

$$\boxed{X = (x_1, x_2, x_3, \dots, x_n) \quad ; \quad Y = (y_1, y_2, y_3, \dots, y_n)}$$

Le **nuage de points** d'abscisses dans X et d'ordonnées dans Y est l'ensemble des points $M(x_i, y_y)$ du plan rapporté à un repère orthonormé, $i \in [\![1, n]\!]$. C'est un ensemble de n points du plan.

La droite approchant le mieux le nuage de points, au sens des moindres carrés, est la **droite de régression linéaire**. C'est la droite d'équation

$$Y = a.X + b$$

$$\text{avec} \qquad a = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{\sum_{i=1}^{n}(x_i - \overline{X})^2} \quad \text{et} \quad b = \overline{Y} - a.\overline{X}$$

où $\overline{X}$ et $\overline{Y}$ désignent les moyennes des séries X et Y.
Le **coefficient de corrélation** est défini comme :

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{X})^2}.\sqrt{\sum_{i=1}^{n}(y_i - \overline{Y})^2}}$$

Alors $r \in [-1, 1]$ et la droite de régression linéaire approche d'autant mieux le nuage de point que $|r|$ est proche de 1.

• En notant pour deux série statistiques $X = (x_1, x_2, x_3, \dots, x_n)$ et $Y = (y_1, y_2, y_3, \dots, y_n)$, de même effectif n :

- La **variance** de X, $V(X) = \overline{(X - \overline{X})^2}$, soit :

$$\boxed{V(X) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2}$$

- L'**écart-type** de X, $\sigma_X = \sqrt{V(X)}$, soit :

$$\boxed{\sigma_X = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})^2}}$$

- La **covariance** de X et Y, $cov(X, Y) = \overline{(X - \overline{X})(Y - \overline{Y})}$, soit :

$$\boxed{cov(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}$$

la **droite de régression linéaire** $Y = a.X + b$, et son **coefficient de corrélation** sont alors obtenus par les formules :

$$\boxed{a = \frac{cov(X, Y)}{V(X)} \quad ; \quad b = \overline{Y} - a.\overline{X}} \qquad \boxed{r = \frac{cov(X, Y)}{\sigma_X \cdot \sigma_Y}}$$

2 Exemple pratique

En laboratoire on a mesuré l'évolution de la concentration d'un réactif lors d'une réaction chimique. Les différents temps de mesure et concentrations mesurées figurent dans le tableau ci-contre.
On souhaite modéliser l'évolution de la concentration de ce réactif, et pour cela nous allons confronter deux modèles cinétiques, celui d'une réaction chimique d'ordre 1 avec celui d'une réaction chimique d'ordre 2.
Dans ces modèles la vitesse d'accroissement $C'(t)$ au temps t de la concentration d'un produit est proportionnelle respectivement à la concentration $C(t)$ au temps t ou à son carré $C(t)^2$, c'est à dire :

Temps (s)	Concentration (mol/l)
0	34,97
3	17,52
5	13,23
7	10,13
10	8,16
12	6,55
15	5,78
18	5,22
20	4,66
23	4,15
25	3,86

Ordre 1 : $\frac{d}{dt}C(t) = -\lambda\, C(t)$ Ordre 2 : $\frac{d}{dt}C(t) = -\lambda\, C(t)^2$

On admettra ici (et l'on montrera en cours de Mathématiques) que les solutions à ces équations différentielles sont de la forme (respectivement à l'ordre 1 et à l'ordre 2) :

$$C(t) = C_0 \exp(-\lambda t) \quad ; \quad C(t) = \frac{C_0}{1 - \lambda C_0 t}$$

pour C_0 une contante réelle.

À l'aide d'une régression linéaire, nous allons déterminer quel modèle, réaction chimique d'ordre 1 ou d'ordre 2, décrit le mieux l'évolution de la concentration mesurée.

Pour cela, on remarque d'abord que :

• Si la réaction chimique est d'ordre 1 :

$$C(t) = C_0 \exp(-\lambda t)$$

alors en posant $Y(t) = \ln(C(t))$, la fonction $Y(t)$ est affine :

$$Y(t) = -\lambda \cdot t + \ln(C_0).$$

• Si la réaction chimique est d'ordre 2, alors en posant $Z(t) = 1/C(t)$, la fonction $Z(t)$ est affine :

$$Z(t) = -\lambda \cdot t + \frac{1}{C_0}.$$

1. Dans un tableur (calc, open office), saisir sur deux lignes les séries statistiques des temps de mesure, et des concentrations mesurées :

	A	B	C	D	E	F	G	H	I	J	K	L
1	**T**	0	3	5	7	10	12	15	18	20	23	25
2	**C**	34,97	17,52	13,23	10,13	8,16	6,55	5,78	5,22	4,66	4,15	3,86

2. Ajouter deux nouvelles lignes avec les valeurs de $Y = \ln(C)$ et $Z = 1/C$.

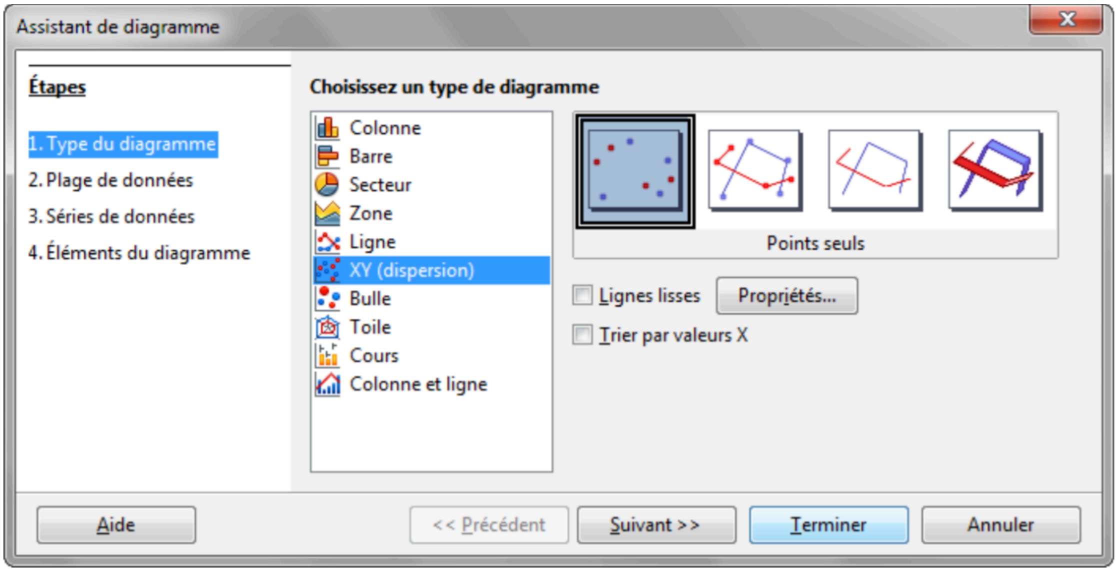
Pour cela taper dans la cellule B3 la commande = `LN(B2)`, valider, puis copier la cellule B3 dans les cellules C3 à L3. Faire de même ligne 4, en tapant dans B4 la comande = `1/B2`, *etc.*

	A	B	C	D	E	F	G	H	I	J	K	L
1	**T**	0	3	5	7	10	12	15	18	20	23	25
2	**C**	34,97	17,52	13,23	10,13	8,16	6,55	5,78	5,22	4,66	4,15	3,86
3	**Y=ln(C)**	3,5545	2,863	2,582	2,316	2,099	1,879	1,7544	1,652	1,539	1,4231	1,3507
4	**Z=1/C**	0,0286	0,057	0,076	0,099	0,123	0,153	0,173	0,192	0,2146	0,241	0,2591

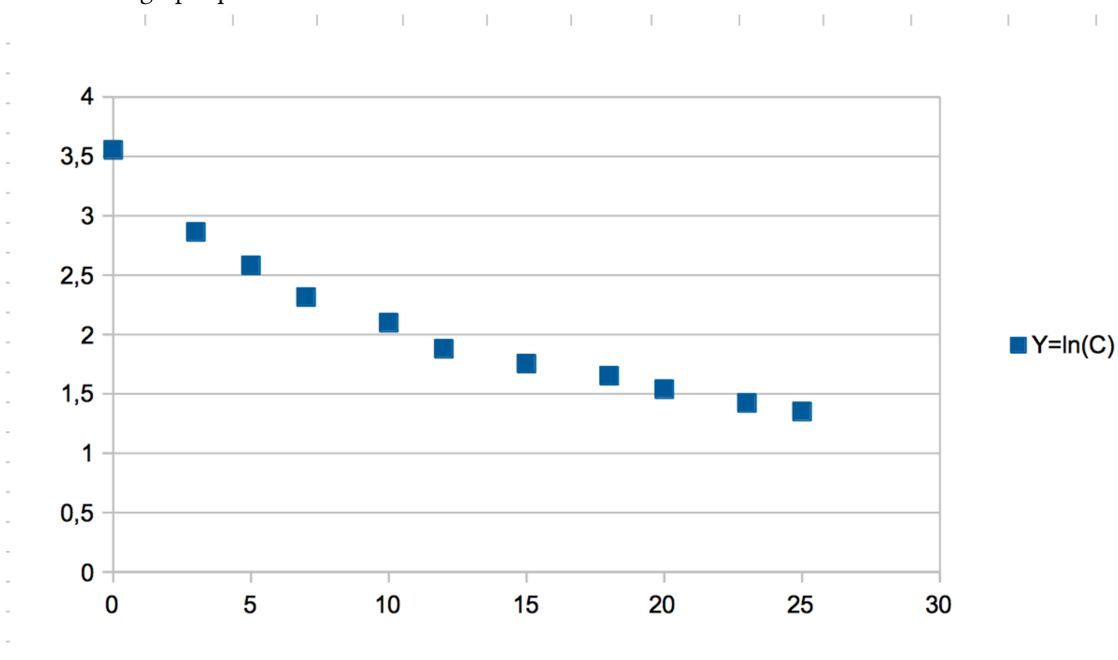
3. À l'aide de l'outil graphique, insérer un graphique contenant le nuage de point des séries (T, Y). Pour cela :

– Barre de menu **Insertion** >> **Diagramme**

Une nouvelle fenêtre s'ouvre pour sélectionner le type de diagramme.
– Sélectionner **XY (dispersion)** >> **Points seuls** >> **Suivant**



– Sélectionner **Séries de données** et **ajouter** puis définir les plages de données X (=T) et Y (=Y). Puis cliquer sur **Terminer**. On obtient le graphique :

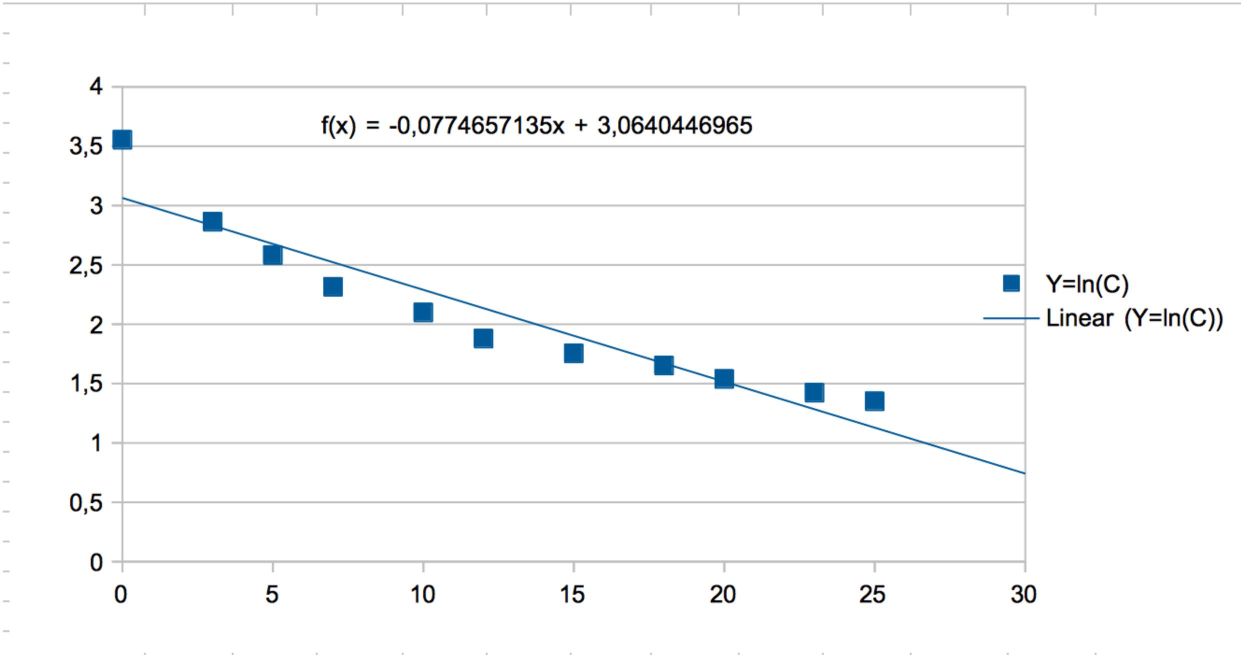


4. Insérer la droite de régression linéaire sur le graphique ; pour cela, le graphique étant toujours sélectionné :
– Menu **Insertion** >> **Courbe de tendance** >> **Type de régression** : cocher **"régression linéaire**". Cliquer sur OK.
Remarque : deux boutons en bas de fenêtre peuvent être cochés pour afficher sur le graphe R^2 ou l'équation de la droite.

Le **coefficient de détermination** R^2 est le rapport entre la somme des carrés des écarts à la moyenne des valeurs prédites $(\widehat{y_i})_{1\leqslant i\leqslant n}$ et des valeurs mesurées $(y_i)_{1\leqslant i\leqslant n}$:

$$R^2 = \frac{\sum_{i=1}^n \left(\widehat{y_i} - \overline{Y}\right)^2}{\sum_{i=1}^n \left(y_i - \overline{Y}\right)^2}$$

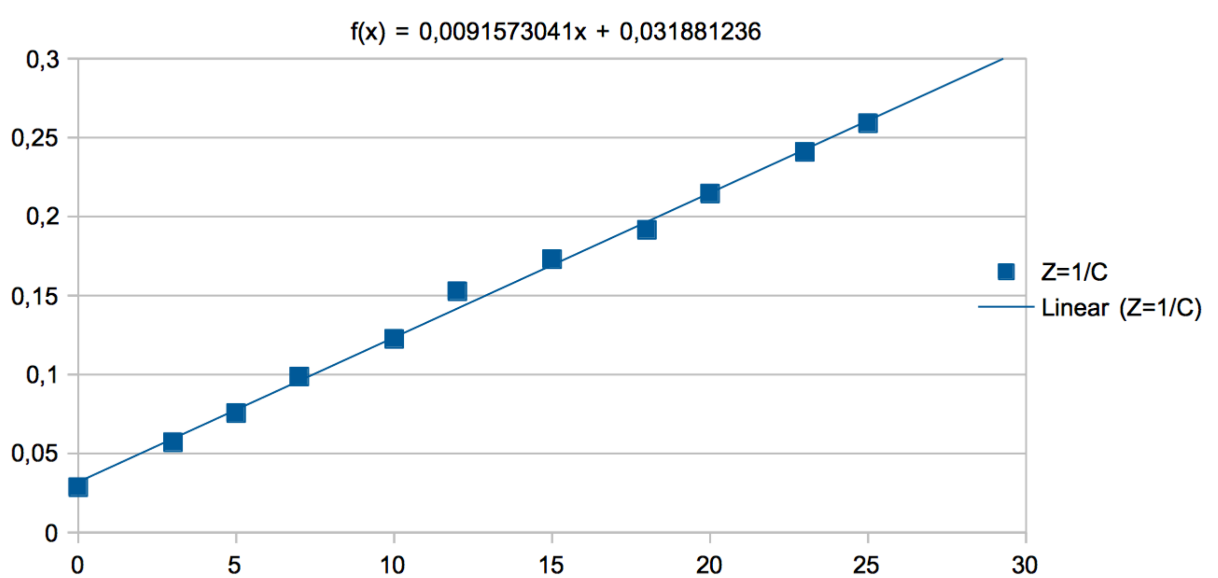
C'est un nombre entre 0 et 1 : la corrélation est d'autant meilleure que R^2 est proche de 1.
Tout comme le coefficient de corrélation linéaire, c'est un estimateur de la qualité de la régression.



5. Obtenir le coefficient de corrélation (**Insertion** >> **fonction** >> **Statistiques** >> **Coefficient de corrélation**), la pente a de la droite de régression linéaire (**Insertion** >> **fonction** >> **Statistiques** >> **pente**) ainsi que son ordonnée à l'origine $b = \overline{Y} - a\overline{T}$:

R:	-0,94659
pente	-0,07747
MOY(T)	12,5455
MOY(Y)	2,0922
b	3,06404

6. Effectuer les mêmes action avec le nuage de points (T, Z) :

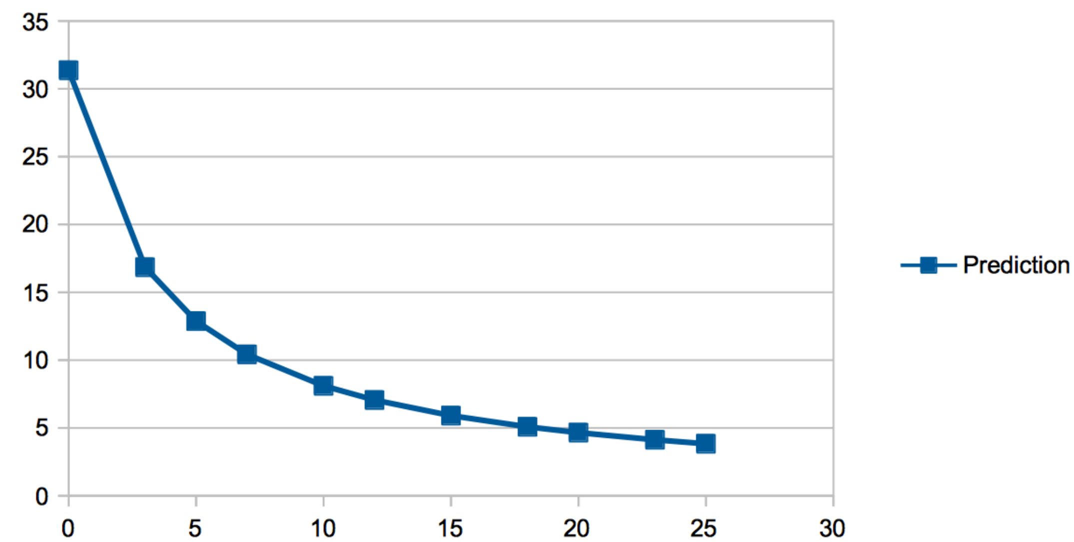


R:	0,99835098
pente	0,0091573
MOY(T)	12,5454545
MOY(Y)	0,14676378
b	0,031881236

7. En déduire que le modèle à retenir est celui d'une réaction chimique d'ordre 2. L'évolution de la concentration est alors donnée dans ce modèle par :

$$\frac{1}{C(t)} = 0,0091573 \cdot t + 0,031881236 \implies \boxed{C(t) = \frac{1}{0,0091573 \cdot t + 0,031881236}}$$

8. En déduire les concentrations prédites par ce modèle :



et le tracé de l'évolution des concentrations aux temps mesurés :

Insertion >> **Diagramme**

– Sélectionner **XY (dispersion)** >> **Points et lignes** >> **Suivant**

– Sélectionner **Séries de données** et **ajouter** puis définir les plages de données X (=T) et Y (Prévisions). Cliquer sur **Terminer**.