D11 : SERIES STATISTIQUES A DEUX VARIABLES

Que disent le programme et les recommandations pédagogiques ?

Le programme : **Statistique descriptive**

Séries statistiques à deux variables : nuage de points ; ajustement affine (méthode des moindres carrés) ; ajustements, qui par un changement de variable, se ramènent à un ajustement affine ; régression, coefficient de corrélation.

Les recommandations pédagogiques : Des situations de la vie économique, des sciences et techniques seront exploitées pour des études de régression. On distinguera variable explicative et variable expliquée. On veillera à attirer l'attention des étudiants sur l'étude des résidus (on vérifiera que leur somme est nulle). La représentation graphique des résidus permettra de vérifier le bien-fondé du modèle d'ajustement envisagé : cette représentation ne doit laisser apparaître aucune tendance. On pourra déterminer le coefficient de détermination et on en donnera une interprétation.

Précisions de l'inspection.

Dans le programme et les recommandations pédagogiques qui l'accompagnent seul le mot « régression » peut poser problème et donner lieu à des interprétations différentes. Faut-il s'en tenir à la recherche de la courbe d'ajustement (dite aussi courbe de régression ou courbe d'estimation), au calcul des résidus et à leur interprétation, au calcul de l'estimation de la variable expliquée pour une valeur donnée de la variable explicative ? Ou bien faut-il aller plus loin, préciser les hypothèses du modèle linéaire et aborder les problèmes d'inférence statistique qui lui sont liés ? L'ambiguïté du mot « régression » a posé question à l'équipe du GRES qui a sollicité l'avis de l'inspection.

A l'occasion d'une réunion de travail, le GRES a invité Monsieur l'Inspecteur principal PACULL. Concernant le mot « régression » figurant dans le libellé du programme D11, Monsieur PACULL a précisé que l'idée qui avait prévalu lors de l'élaboration du programme était celle correspondant à la première interprétation. Il a insisté sur le fait que dans l'étude de séries statistiques à deux variables, les variables sont statistiques et non aléatoires, le hasard n'intervient pas, on n'a pas une situation probabiliste. Dès lors il n'est pas possible de faire de l'inférence, en faire constituerait une erreur.

Avec les moyens de calcul dont on dispose aujourd'hui, que faire sur les séries statistiques à deux variables ?

Compte tenu des moyens de calcul dont disposent les étudiants, il serait désuet de leur demander de construire le traditionnel tableau de calcul pour déterminer par la méthode des moindres carrés le coefficient de corrélation linéaire, le coefficient directeur et l'ordonnée à l'origine de la droite d'ajustement (ceci ne veut évidemment pas dire qu'il ne faille pas traiter manuellement un ou deux exemples d'école).

Tous possèdent des calculatrices qui fournissent directement ces résultats. Ils ne comprendraient pas que l'on n'utilise pas ces possibilités qui soulagent des calculs longs et fastidieux que nécessite la méthode des moindres carrés.
En outre toutes les calculatrices de la dernière génération fournissent non seulement les résultats pour les ajustement affines mais aussi pour les ajustements qui peuvent se ramener au modèle linéaire ou multilinéaire : ajustement exponentiel, ajustement puissance et ajustements polynomiaux (jusqu'au degré 4 pour certaines). De plus elles proposent un mini tableur (mode liste) qui permet de faire très simplement les opérations élémentaires de calcul matriciel et avec lequel le calcul des résidus devient facile. Dès maintenant certaines calculatrices vont jusqu'à fournir la série des résidus. On n'est plus limité par le nombre des observations.

Les présupposés théoriques et les pièges à éviter.

Concernant les résidus, les résultats théoriques utilisés sont établis dans le cas du modèle linéaire :

- Leur somme est nulle.
- Ils se répartissent à peu près équitablement entre résidus positifs et résidus négatifs et sans qu'aucune tendance n'apparaisse.

Etablis dans le cadre du modèle linéaire ces résultats ne s'appliquent que dans le cas d'un ajustement affine.

- Dans le cas de l'ajustement exponentiel d'une série (x_i,y_i), l'ajustement affine s'applique à la série transformée (x_i,z_i) avec $z_i = \ln(y_i)$ et bien entendu les résultats concernant les résidus s'appliquent à cette série mais pas à la série (x_i,y_i).
  En particulier la somme des résidus $e_i = y_i - \hat{y}_1$ de la série (x_i,y_i) n'est pas nulle. Certaines calculatrices fournissent ces résidus.
- Dans le cas de l'ajustement puissance d'une série (x_i,y_i), l'ajustement affine s'applique à la série transformée (u_i,z_i) avec $u_i = \ln(x_i)$ et $z_i = \ln(y_i)$ et les résultats concernant les résidus s'appliquent à cette série mais pas à la série (x_i,y_i).
  En particulier la somme des résidus $e_i = y_i - \hat{y}_i$ de la série (x_i,y_i) n'est pas nulle.

  Certaines calculatrices fournissent ces résidus.

La plupart des calculatrices fournissent directement un coefficient de corrélation linéaire r dans le cas d'un ajustement exponentiel ou d'un ajustement puissance. Il faut bien comprendre que :

- Dans le cas d'un ajustement exponentiel, le coefficient de corrélation linéaire fourni par la calculatrice est celui correspondant à l'ajustement affine de la série transformée (x_i,z_i). C'est le coefficient de corrélation linéaire entre les variables *X* et *Z*.
- Dans le cas d'un ajustement puissance, le coefficient de corrélation linéaire fourni par la calculatrice est celui correspondant à l'ajustement affine de la série transformée (u_i,z_i). C'est le coefficient de corrélation linéaire entre les variables *U* et *Z*.

Activité - Etude de la croissance d'une tige de tomate. (D'après un document de biologie)

On mesure l'allongement *X* de la tige d'une tomate, exprimé en mm/j, en fonction de la température diurne *T*, exprimée en °C. Le tableau suivant fournit le relevé des valeurs du couple de variables statistiques (*T*, *X*).

t_i	5	7	10	13	15	18	20	22	25	28	30
x_i	1	2	3	6	8	10	11	15	17	20	23

1- Construire le nuage de points représentant cette série statistique double. Que suggère l'examen du nuage ?

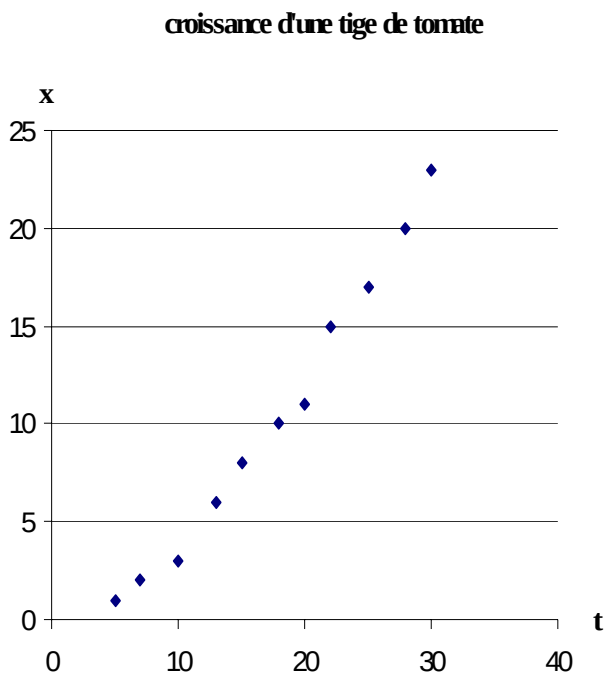
2- En utilisant une calculatrice :
a. donner une équation de la droite d'ajustement de *X* en *T*, obtenue par la méthode des moindres carrés.
b. construire la droite.
c. donner le coefficient de corrélation linéaire r_1 entre *X* et *T*.
d. donner le coefficient de détermination r_1^2 et interpréter ce coefficient.

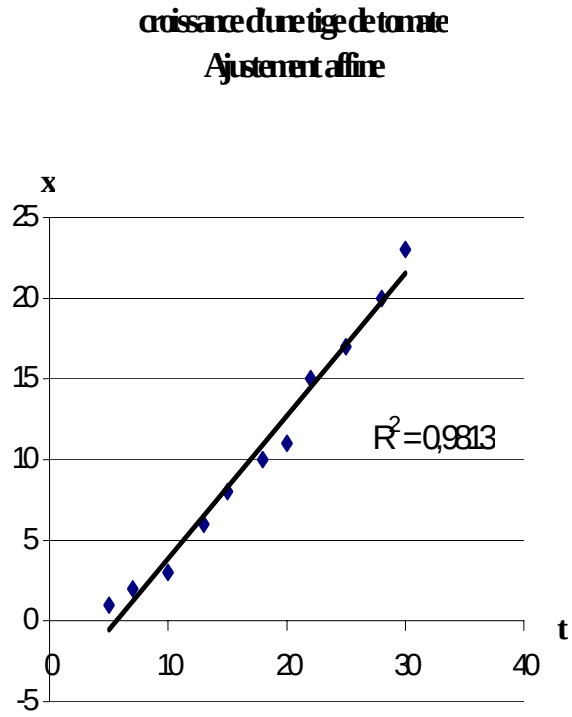
3- *a*. Calculer les écarts résiduels $e_i = x_i - \hat{x}_i$ où $\hat{x}_i$ est la valeur estimée correspondant à t_i, calculée en utilisant l'équation de la droite d'ajustement obtenue à la question 2- *a*.
b. Représenter les résidus en fonction de la variable explicative puis de la variable expliquée.
c. Que peut-on dire de la répartition des résidus et de l'ajustement affine ? Quelle remarque peut-on faire sur les coefficients de corrélation et de détermination entre *X* et *T* ?

4- **Ajustement exponentiel**. On pose $z = \ln x$:
a. Calculer les z_i. Construire le nuage de points représentant la série (t_i,z_i).
b. En utilisant une calculatrice : donner une équation de la droite d'ajustement de *Z* en *T*, obtenue par la méthode des moindres carrés.
c. Calculer les écarts résiduels $e'_i = z_i - \hat{z}_i$ où $\hat{z}_i$ est la valeur estimée correspondant à t_i, calculée en utilisant l'équation de la droite d'ajustement obtenue à la question 4- *b*.
***d*.** Représenter les résidus en fonction de la variable explicative *T*. Que peut-on dire de la répartition des résidus ? L'ajustement affine de la série (t_i,z_i) est-il envisageable ?

5- **Ajustement puissance**. On pose $u = \ln t$:
a. Calculer les u_i. Construire le nuage de points représentant la série (u_i,z_i).
b. En utilisant une calculatrice : donner une équation de la droite d'ajustement de *Z* en *U*, obtenue par la méthode des moindres carrés.
c. Calculer les écarts résiduels $e'_i = z_i - \hat{z}_i$ où $\hat{z}_i$ est la valeur estimée correspondant à u_i, calculée en utilisant l'équation de la droite d'ajustement obtenue à la question 5- *b*.
d. Représenter les résidus en fonction de la variable explicative *U*. Que peut-on dire de la répartition des résidus ? L'ajustement affine de la série (u_i,z_i) est-il envisageable ?
e. donner les coefficients de corrélation linéaire et de détermination entre les variables statistiques *Z* et *U* . Que peut-on dire de l'ajustement puissance ?
f. Déduire de l'équation de la droite d'ajustement de *Z* en *U* une relation de la forme $\hat{x} = f(t)$ liant $\hat{x}$ et *t*. Construire la courbe d'équation $\hat{x} = f(t)$ dans le même repère que le nuage de points représentant la série (t_i,x_i).

Eléments de correction.





Nuage de points représentant la série statistique (t_i,x_i).

Au vu du nuage de points on peut légitimement envisager un ajustement affine de la série (t_i,x_i)..

Remarque :

Le choix des unités, pour le repère dans lequel on construit le nuage de points, n'est pas neutre. Si celles-ci sont convenablement choisies on constate, dans l'exemple traité, que le nuage est incurvé vers le haut, ce qui conduit à penser que le modèle linéaire n'est peut-être pas le plus approprié. Mais un choix inadéquat des unités peut écraser le nuage, masquer cette courbure et induire que le seul modèle pertinent est l'ajustement affine. **Seul le diagramme des résidus va permettre de démasquer cette erreur**, en faisant apparaître que les résidus ont une distribution tendancieuse.

Ajustement affine de la série (t_i,x_i).

Résultats calculatrice : a = 0,883 *résultats numériques à 10^{-3} près*

b = -4,941

Equation de la droite d'ajustement : $\hat{x} = 0{,}883.t - 4{,}941$

Coefficient de corrélation linéaire : r_1 = 0,991

Coefficient de détermination : r_1^2 = 0,981

Interprétation du coefficient de détermination : 98% de la variabilité totale de X est expliquée par l'ajustement affine.

L'ajustement par une droite d'équation $\hat{x} = a.t + b$ semble, de ce point de vue, parfaitement justifié.

Toutefois l'étude de l'ajustement affine serait incomplète sans celle des résidus

ableau de calcul des résidus :

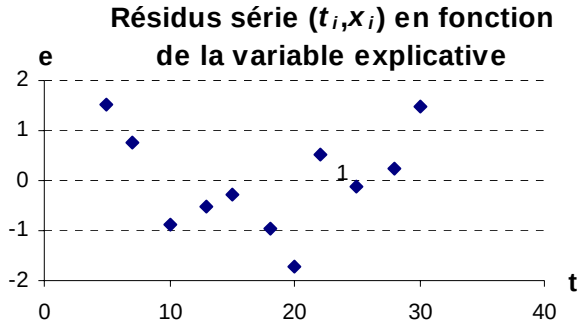
t_i	x_i	résidus série (t_i,x_i)	$z_i = \ln(x_i)$	résidus série (t_i,z_i)	$u_i = \ln(t_i)$	résidus série (u_i,z_i)
5	1	1,5	0	-0,54	1,609	-0,04
7	2	0,8	0,693	-0,08	1,946	0,06
10	3	-0,9	1,099	-0,03	2,303	-0,15
13	6	-0,5	1,792	0,32	2,565	0,08
15	8	-0,3	2,079	0,37	2,708	0,12
18	10	-0,9	2,303	0,25	2,89	0,02
20	11	-1,7	2,398	0,11	2,996	-0,06
22	15	0,5	2,708	0,19	3,091	0,08
25	17	-0,1	2,833	-0,04	3,219	-0,02
28	20	0,2	2,996	-0,23	3,332	-0,05
30	23	1,5	3,135	-0,32	3,401	-0,03
		Somme des résidus 0		Somme des résidus 0		Somme des résidus 0

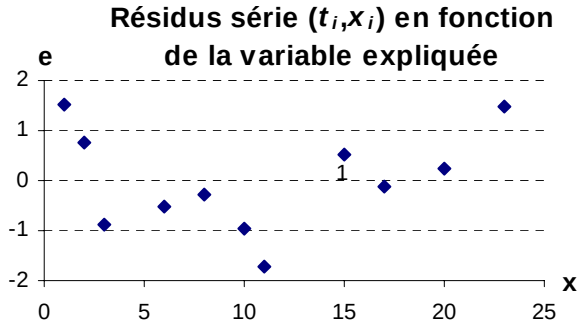
Etude des résidus : $e_i = x_i - \hat{x}_i$

$= x_i - (a.t_i + b)$

Calcul des résidus voir tableau ci-dessus.
On vérifie que la somme des résidus est nulle aux arrondis près.
L'examen des diagrammes des résidus montre que les points qui les représentent sont mal répartis par rapport aux axes des abscisses. Les résidus sont négatifs au centre positifs ailleurs.

Représentations des résidus :





Cela conduit à chercher un modèle plus pertinent pour traduire la relation entre *X* et *Y*.

La forme des nuages, représentant respectivement la série et les résidus, suggère qu'une courbe d'équation soit $\hat{y} = b.a^x$ soit $\hat{y} = b.x^a$ peut fournir un meilleur modèle d'ajustement.

Remarque : l'ajustement affine n'est manifestement pas le modèle le mieux adapté et pourtant coefficients de corrélation linéaire et de détermination prennent des valeurs très proches de 1 (respectivement r_1 = 0,99 et r_1^2 = 0,98). Cet exemple montre que ces coefficients ne permettent pas de juger de la pertinence du modèle retenu.
Le coefficient de détermination permet de mesurer la qualité d'un ajustement affine mais il ne permet pas de juger de sa validité.

Ajustement exponentiel.

$$\hat{x} = b.a^t \Leftrightarrow \ln \hat{x} = \ln b + t.\ln a$$

$$\Leftrightarrow \begin{cases} \ln \hat{x} = z,\ \ln a = A,\ \ln b = B \\ z = A.t + B \end{cases}$$

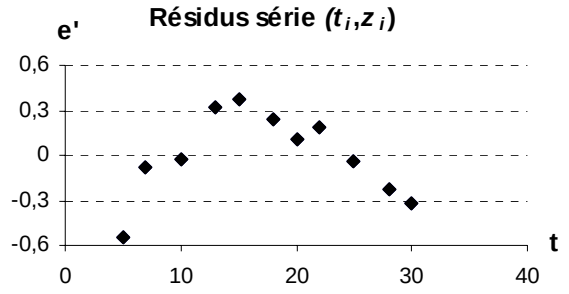
Si les points de coordonnées (t_i,x_i) sont ajustés par la courbe d'équation $\hat{x} = b.a^t$ alors les points de coordonnées (t_i,z_i) sont ajustés par la droite d'équation $\hat{z} = A.t + B$ et réciproquement.
Pour répondre à la question

- on calcule les valeurs $z_i = \ln(x_i)$, on construit le nuage de points représentant la série (t_i,z_i),
- selon l'aspect du nuage de points on envisage ou non un ajustement affine de la série (t_i,z_i).
- on effectue l'ajustement affine de la série (t_i,z_i)
  Résultats calculatrice : A = 0,117
  B = -0,042
  Equation de la droite d'ajustement : $\hat{z} = 0{,}117.t - 0{,}042$
- on calcule les résidus (voir tableau de calcul) et on les représente

Etude des résidus :

$$e'_i = z_i - \hat{z}_i$$
$$= z_i - (A.t_i + B)$$

On vérifie que la somme des résidus est nulle aux arrondis près.



Le diagramme des résidus montre que le modèle exponentiel n'est pas adapté.
On ne retient pas l'ajustement exponentiel.

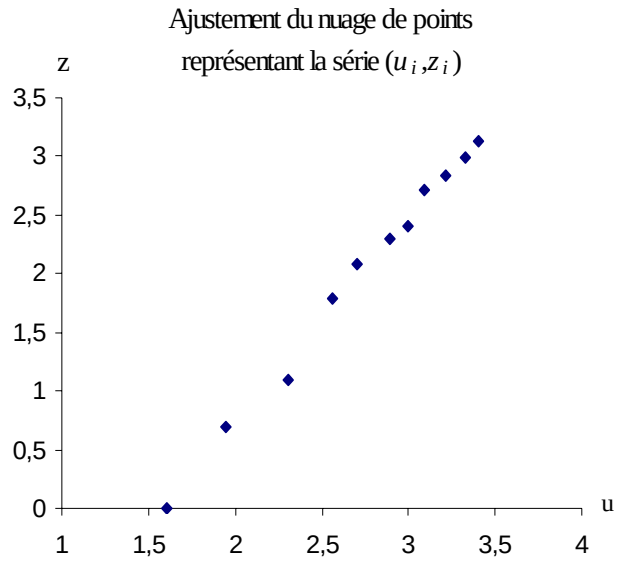
Ajustement puissance.

$$\hat{x} = b.t^a \Leftrightarrow \ln \hat{x} = \ln b + a.\ln t$$

$$\Leftrightarrow \begin{cases} \ln t = u,\ \ln \hat{x} = \hat{z},\ \ln b = B \\ \hat{z} = a.u + B \end{cases}$$

Si les points de coordonnées (t_i,x_i) sont ajustés par la courbe d'équation $\hat{x} = b.t^a$ alors les points de coordonnées (u_i,z_i) sont ajustés par la droite d'équation $\hat{z} = a.u + B$ et réciproquement.
Pour répondre à la question

- on calcule les $u_i = \ln(t_i)$ (voir tableau de calcul), on construit le nuage de points représentant la série (u_i,z_i),
- selon l'aspect du nuage de points on envisage ou non un ajustement affine de la série (u_i,z_i).



Les points du nuage sont approximativement alignés. L'ajustement par une droite d'équation $\hat{z} = a.u + B$ est tout à fait envisageable.

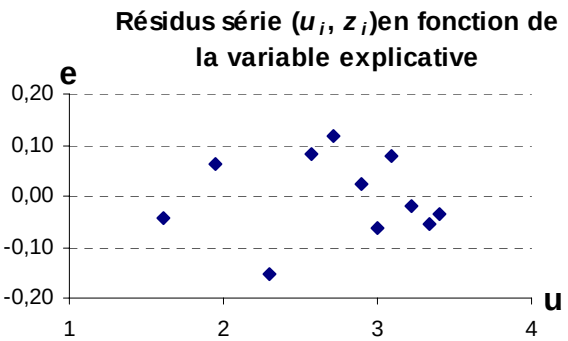
Résultats calculatrice : a = 1,745
B = -2,765
Equation de la droite d'ajustement : $\hat{z} = 1{,}745.u - 2{,}765$

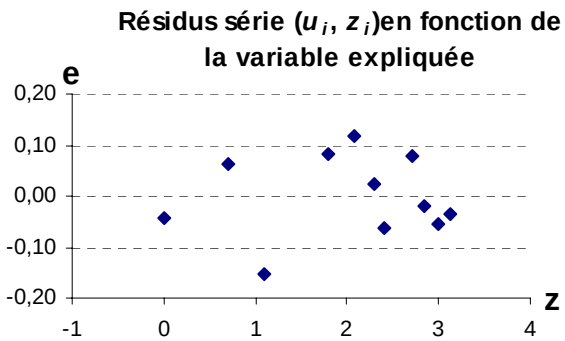
Pour étudier la validité du modèle on calcule les résidus, on construit leurs représentations graphiques en fonction de la variable explicative ou de la variable expliquée.

Etude des résidus.

Résidus : $e_i'' = z_i - \hat{z}_i$
$= z_i - (a.u_i + B)$

Calcul des résidus voir tableau. On vérifie que la somme des résidus est nulle aux arrondis près.
Représentation des résidus :





A première vue, il semble que les résidus soient convenablement répartis de part et d'autre des axes des abscisses et ne fassent pas apparaître de tendance. Toutefois un examen plus attentif montre que cette apparence tient au seul résidu –0,15 correspondant au point de coordonnées u_i = 2,303 z_i = 1,099. Les nuages des résidus, privés de ce point, font apparaître nettement une tendance.

L'ajustement puissance, lui non plus, ne semble pas très bien adapté.

Dans le cadre du D11.
Toutefois des trois ajustements au programme du D 11, c'est celui pour lequel les résidus sont les mieux répartis (aucune tendance perceptible) et, en restant dans le cadre du programme, c'est celui que nous retiendrons.
Nous allons maintenant évaluer la qualité de cet ajustement et pour ceci calculer le coefficient de détermination.
Résultats calculatrice : Coefficient de corrélation linéaire entre *U* et *Z* : r_3 = 0,997

Coefficient de détermination entre *U* et *Z* : r_3^2 = 0,994

Plus de 99% de la variabilité totale de *Z* est expliquée par l'ajustement affine.

Conclusion.
Des trois ajustements au programme du D11, l'ajustement puissance est le plus approprié pour établir la relation exprimant l'allongement d'une tige de tomate en fonction de la température.

Equation de la courbe d'ajustement .
Une équation de la droite d'ajustement des moindres carrées de z en u est :

$\hat{z} = 1{,}745.u - 2{,}765$ *avec* $u = \ln t$ *et* $\hat{z} = \ln \hat{x}$

on remplace *u* et $\hat{z}$ par leur expression en fonction de *t* et $\hat{x}$: $\ln \hat{x} = 1{,}745.\ln t - 2{,}765$

propriété de la fonction logarithme népérien : pour tout réel strictement positif *x*, pour tout rationnel α, $\alpha \ln x = \ln x^{\alpha}$: $\ln \hat{x} = \ln t^{1{,}745} - 2{,}765$

passage aux exponentielles : $e^{\ln \hat{x}} = e^{\ln t^{1{,}745} - 2{,}765}$

propriété de la fonction exponentielle : pour tout réel strictement positif *x*, $e^{\ln x} = x$: $\hat{x} = e^{\ln t^{1{,}745} - 2{,}765}$

propriété de la fonction exponentielle (propriété des puissances d'un nombre) $e^{a+b} = e^a.e^b$: $\hat{x} = e^{\ln t^{1{,}745}}.e^{-2{,}765}$

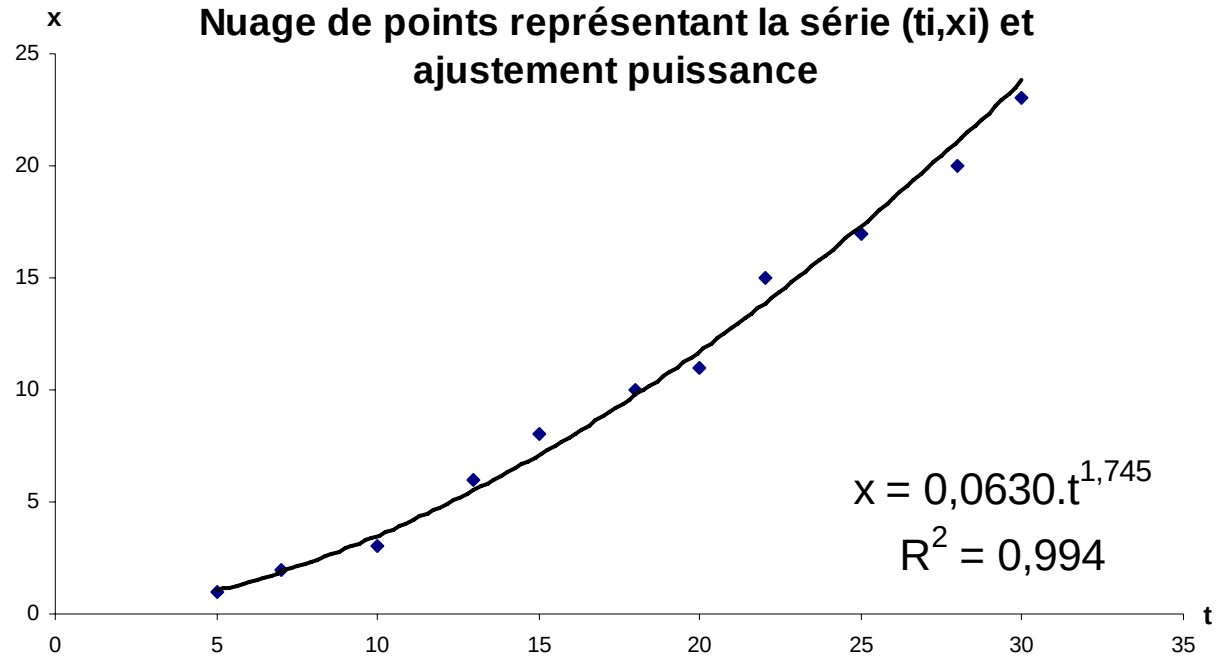
propriété de la fonction exponentielle : pour tout réel strictement positif *x*, $e^{\ln x} = x$: $\hat{x} = e^{-2{,}765}.t^{1{,}745}$ or

$e^{-2{,}765} = 0{,}063$

$$\hat{x} = 0.063 \times t^{1{,}745}$$

Tableau de valeurs

t	5	10	15	20	25	30
$\hat{x}$	1,0	3,5	7,1	11,7	17,3	23,8



Au-delà du D11.

Les calculatrices proposent des ajustements par des polynômes, alors pourquoi ne pas aller un peu plus loin, au moins entre profs ?

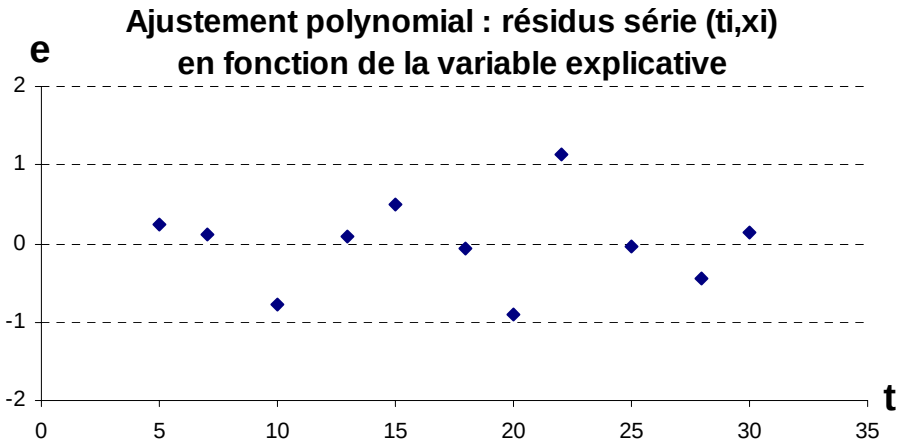
Ajustement par un polynôme du second degré .

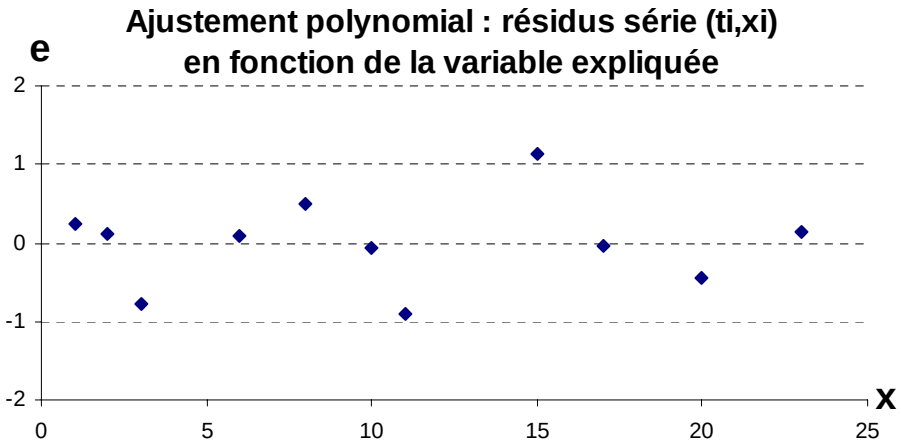
$\hat{x} = a.t^2 + b.t + c$

Résultats calculatrice : $a = 1.41.10^{-2}$, b = 0,391, c = -1,54

Calcul des résidus :

ti	5	7	10	13	15	18	20	22	25	28	30
xi	1	2	3	6	8	10	11	15	17	20	23
résidus	0,23	0,11	-0,78	0,08	0,51	-0,06	-0,91	1,13	-0,03	-0,44	0,15





Des quatre ajustements étudiés, l'ajustement quadratique est celui qui conduit à la meilleure distribution des résidus (voir représentations graphiques) ce qui ne signifie pas qu'il n'y en ait pas de plus pertinent.

C'est celui que l'on retient.

Pour évaluer la qualité de cet ajustement on calcule le coefficient de détermination en utilisant sa définition : $r^2 = \dfrac{\sum(\hat{x}_i - \bar{x})^2}{\sum(x_i - \bar{x})^2}$

A vos calculatrices !
Si on effectue le calcul du coefficient de détermination avec la calculatrice CASIO *fx-6910G,* celle-ci fournit le résultat avec 9 décimales : 0,994137663.
On retient $r^2 = 0{,}994$
99,4 % de la variabilité totale de X est expliquée par l'ajustement par un polynôme du second degré.

Voici, toujours avec la calculatrice CASIO *fx-6910G*, la liste des données et des commandes à saisir pour calculer les coefficients de l'ajustement polynomial et le coefficient de détermination :

En mode STAT : t_i en liste 1, x_i en liste 2, puis CALC SET 2Var X : List1 2Var Y : List2 2Var F : 1 QUIT CALC REG X^2.

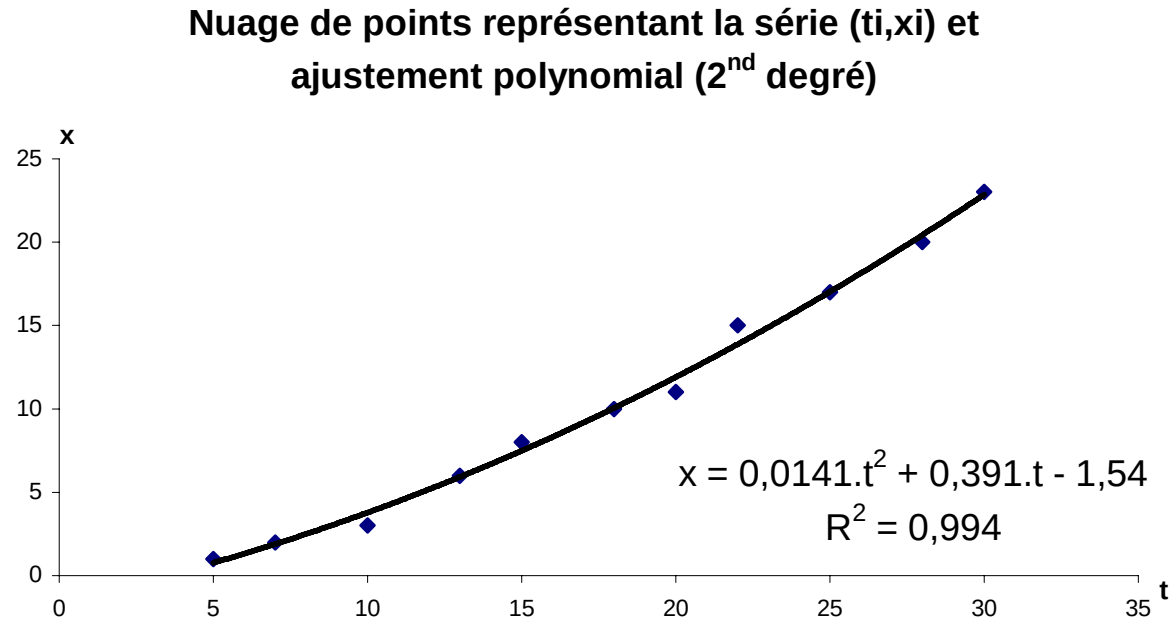
En mode RUN : OPTN LIST Sum ((VARS STAT GRPH a OPTN LIST List 1 ^ 2 + VARS STAT GRPH b OPTN LIST List 1 + VARS STAT GRPH c – OPTN LIST Mean(List 2)) ^ 2) / Sum ((List 2 – Mean(List 2)) ^ 2)

Ce qui donne à l'écran de la calculatrice :
Sum((aList 1^2+bList 1+c–Mean(List 2))^2)/Sum((List 2– Mean(List 2))^2)

Certaines calculatrices (TI 83) donnent le coefficient de détermination dans le cas d'un ajustement quadratique.

Avec le tableur EXCEL pour obtenir ces résultats on peut utiliser :

- soit le grapheur : après avoir représenté la série à deux variables par un nuage de points, on sélectionne ce nuage en cliquant sur un de ses points, puis ou on fait un clic droit ou on rentre dans le menu « graphique » et on utilise la commande « Ajouter une courbe de tendance ». Une boite de dialogue se présente avec deux onglets, l'onglet « Type » permet de choisir l'ajustement, l'onglet « Option » permet d'afficher sur le graphique l'équation de la courbe d'ajustement et le coefficient de détermination.
- soit l'outil d'analyse « Régression linéaire » de l'utilitaire d'analyse.



-

= - = - = - = - = - = - = - = -