Chapitre 3 Retour au sommaire

STATISTIQUE DESCRIPTIVE
SERIES STATISTIQUES A DEUX VARIABLES

I] VOCABULAIRE

Notion de série double :

L'étude d'une population peut porter sur plusieurs caractères. On peut par exemple étudier deux caractères **quantitatifs** (taille et poids d'un individu, hauteur et diamètre d'un arbre etc.) ou un caractère quantitatif et noter son évolution temporelle, la série obtenue est alors **chronologique.**

Tableaux d'effectifs :

Exemple d'une **série pondérée** : on regroupe les données de la série dans un tableau à deux entrées :

xi	0	1	1	1	2	3	4	4
yi	2	2	3	5	2	2	3	5
ni	3	3	2	4	1	5	1	6

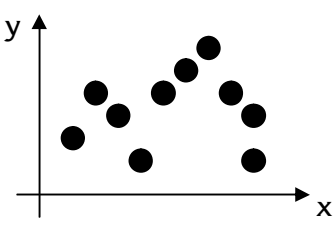
yi \ xi	0	1	2	3	
2	3	3	1	5	
3		2			
5		2			6

Rq : si les variables sont continues, on trouve à la place des valeurs de x ou de y des classes ou des centres de classe.

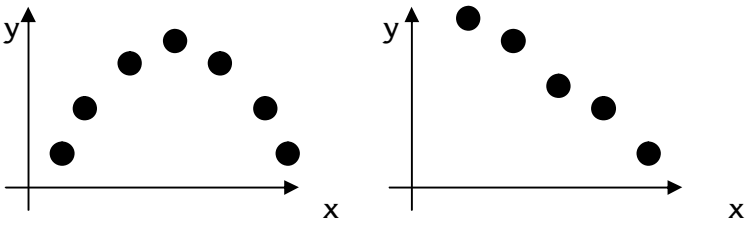
Nuage de points

Si l'on munit le plan d'un repère, on peut associer au couple (xi yi) le point Mi de coordonnées (xi yi). L'ensemble des points Mi constitue le nuage de points représentant la série statistique.

En fonction de la disposition des points, on peut conclure qu'il existe ou qu'il n'existe pas de relation **(corrélation)** entre les points.



Pas de corrélation



Corrélations éventuelles

Point moyen :

On appelle point moyen d'un nuage de points, le point G dont les coordonnées sont respectivement la moyenne des abscisses et la moyenne des ordonnées des points du nuage.

II] AJUSTEMENTS AFFINE :

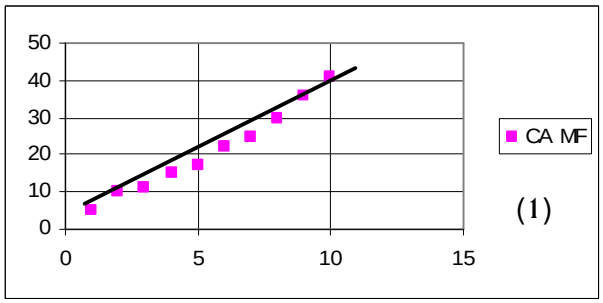
Considérons la série chronologique suivante donnant à la fin de l'année de rang i le CA d'une société en millions de francs.

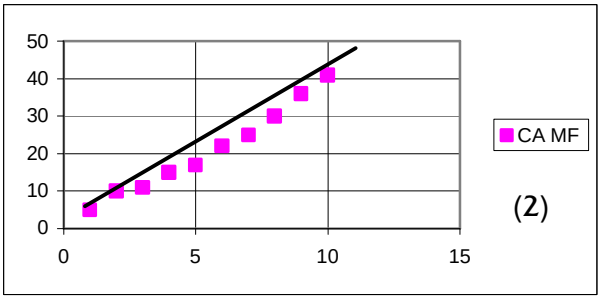
Rang i	1	2	3	4	5	6	7	8	9	10
CA MF	5	10	11	15	17	22	25	30	36	41

Le nuage de points représenté ci dessous à gauche est fortement longiligne, on se rend compte aisément que l'on peut tracer une droite passant au mieux par les points du nuage, on distinguera deux types d'ajustement.

• Graphique à l'oeil ou à l'aide de moyennes. (1)

• Analytique recherche d'une équation de la droite. (2)





1°) Les ajustements graphiques :

a) Ajustement direct à la règle :

Méthode rapide et souvent efficace.

b) Utilisation du point moyen :

On montre par le calcul que pour obtenir le meilleur ajustement affine, il convient de prendre une droite passant par le point moyen G $(\bar{x}\ ;\bar{y})$.

c) Fractionnement du nuage - Méthode de Mayer :

On fractionne la série en deux groupes de points, on calcule le point moyen G_1 de la première moitié du nuage et G_2 pour la seconde moitié, la droite $(G1G_2)$ est **la droite de Mayer**.

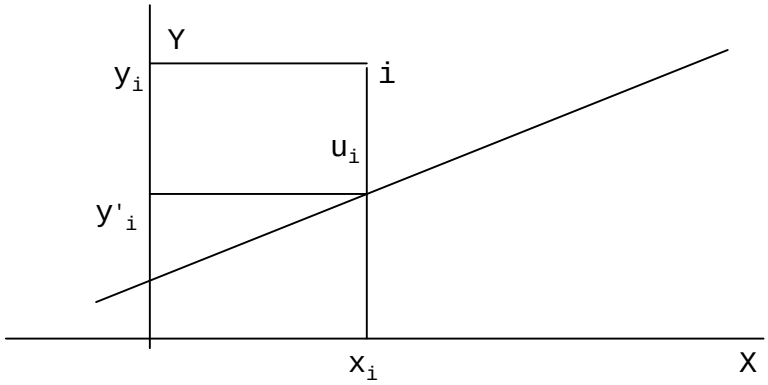
Chacun de ces ajustements affines permet d'effectuer une **projection sur l'avenir.**

2°) Méthode des moindres carrés :

On cherche toujours une droite, d'équation y = **A.x + B** qui approche "au mieux" les données.

- $\mathbf{x_i}$ est la **valeur observée** de la variable explicative x.

- $\mathbf{y_i}$ est la valeur observée de la variable à expliquer y.
- $\mathbf{y'_i = A.x_i + B}$ est la **valeur théorique**, ou **ajustée**, de la variable à expliquer, associée à la valeur observée $\mathbf{x_i}$.
- $\mathbf{u_i = y_i - y'i}$ est l'**erreur d'ajustement** (ou **résidu**), c'est-à-dire l'écart entre la valeur observée et la valeur théorique calculée de la variable à expliquer.



La "meilleure" droite retenue est en fait celle qui rend minimale la somme des carrés des erreurs d'ajustement : Σu_i^2 . On l'appelle **droite des moindres carrés** de y en x ou **droite de régression** de y en x.

On montre que les coefficients A et B, de la droite de régression de y en x s'expriment en fonction des données par :

$$A = \frac{\sum_{i=1}^{n}(xi-\bar{x})(yi-\bar{y})}{\sum_{i=1}^{n}(xi-\bar{x})^2} = \frac{\sum xiyi - n\bar{x}\bar{y}}{\sum xi^2 - n\bar{x}^2}$$

Etant donné que $\bar{y} = A.\bar{x} + B$, cela signifie que la droite passe par le point moyen du nuage $(\bar{x}\ ;\bar{y})$. On a : $V(X) = \frac{\sum_{i=1}^{i=n} x_i^2}{n} - \bar{x}^2$ et $s_x = \sqrt{V(X)}$

3°) Droites de régression :

La régression de y en x donne des rôles différents aux deux variables, on peut renverser le problème et régresser la variable x sur y.

On obtient une droite : x = C y + D, de coefficients C et D, donnés par :

$$C = \frac{\sum(xi-\bar{x})(yi-\bar{y})}{\sum(yi-\bar{y})^2} \qquad \text{et} \qquad \bar{x} = C.\bar{y} + D$$

Les deux droites de régression diffèrent en général, mais sont peu différentes si le nuage est proche de l'alignement.

4°) Lissage :

Lorsque le nuage présente régulièrement des écarts de part et d'autre d'une courbe de tendance (obtenue par ajustement) pour une série chronologique la plupart du temps et

que ces écarts sont « liés » a% des périodes chronologiques, on effectue un lissage en utilisant des coefficients, appelés **corrigés des variations saisonnières.**

a) Méthode des moyennes mobiles :

- On remplace tous les regroupements de n couples consécutifs d'un nuage par un seul dont l'abscisse est la moyenne des abscisses et l'ordonnée est Ici moyenne des ordonnées.

- On ajuste ensuite graphiquement le nouveau nuage lissé.

b) Méthode des moyennes échelonnées :

- On effectue des regroupements non sécants de n couples consécutifs et l'on trace un nouveau nuage.

- On ajuste ensuite graphiquement le nouveau nuage lissé.

c) Méthode des moyennes discontinues:

- Lorsque plusieurs valeurs de y correspondent à une seule valeur de x, on associe à chaque valeur de x la moyenne des différentes valeurs de y.

- On ajuste ensuite graphiquement le nouveau nuage lissé.

5°) Coefficient de corrélation linéaire :

Il est des cas ou aucune des deux variables ne paraît devoir expliquer l'autre (par exemple le taux d'équipement en réfrigérateurs et celui en magnétoscopes).

On s'intéresse alors davantage à mesurer l'intensité de la liaison linéaire éventuelle, qu'à régresser l'une des variables sur l'autre. Pour ce faire, on calcule le **coefficient de corrélation**, noté **R**, entre les deux variables :

$$R = \frac{\sum (xi - \bar{x})(yi - \bar{y})}{\sqrt{\sum (xi - \bar{x})^2} . \sqrt{\sum (yi - \bar{y})^2}}$$

On peut montrer que ce coefficient R est toujours compris entre -1 et 1.

- Si R est proche de 1 : il y a une liaison linéaire marquée, et les deux variables varient dans le même sens.

- Si R est proche de 0 : il n'y a pas de liaison linéaire

- Si R est proche de -1 : il y a une liaison linéaire marquée, et les deux variables varient en sens contraire.

Dans le premier et le dernier cas le nuage montre un bon alignement et les deux droites de régression sont proches. Dans le second cas, le nuage n'a pas de caractère linéaire, et la régression n'est guère justifiée.

TD 1 – APPLICATION DU COURS
SERIES STATISTIQUES A DEUX VARIABLES

Exercice 1

Une entreprise a réalisé un emprunt de 285 000 € pour l'achat d'une nouvelle machine. A la fin de chaque mois, on note dans le tableau ci-dessous le montant des bénéfices cumulés (en milliers d'euros) réalisés depuis l'acquisition de la nouvelle machine.

Rang du mois xi	1	2	3	4	5	6	7	8
Montant des bénéfices cumulés yi	28	40	51	65	78	84	89	98

1°) a) Représenter dans un repère orthonormal le nuage de points associé à la série (x ; y) (On prendra 2 cm pour 1 mois sur l'axe des abscisses et 2 cm pour 10 000 € sur l'axe des ordonnées).

b) Ce nuage permet il d'envisager un ajustement affine ?

c) Donner l'équation y = a x + b de la droite (D) des moindres carrés en utilisant votre calculatrice, puis construire la droite D dans le repère précédent.

d) Calculer la somme S = $\sum_{i=1}^{8}[y_i-(ax_i+b)]^2$ à l'unité près.

e) A partir de quel mois l'emprunt sera-t-il amorti par les bénéfices cumulés réalisés depuis l'acquisition de la nouvelle machine ?

2°) L'expérience d'une évolution linéaire des bénéfices semble trop optimiste. On effectue alors le changement de variable X = $\sqrt{x}$

a) Calculer les valeurs Xi (arrondir au centième).

b) La droite (D ') des moindres carrés de cette nouvelle série (X ; Y) a pour équation y = 39,45 X – 13,80. En déduire une équation de y en fonction de x.
A partir d'un tableau de valeurs, tracer dans le repère précédent la courbe (C) obtenue avec cette expression.
Selon ce nouvel ajustement, à partir de quel mois l'emprunt sera-t-il amorti par les bénéfices cumulés réalisés depuis l'acquisition de la machine ?

3°) Quel est le meilleur ajustement ? Justifiez.

Indications / Réponses

1°) c) $y = 10{,}08\,x + 21{,}25$
d) $S \approx 94$
e) $27^{ème}$ mois

2°) b) $y = 39{,}45\,\sqrt{x} - 13{,}80$, à partir du $58^{ème}$ mois.

3°) Le deuxième ajustement est le meilleur.

Exercice 2

Le tableau suivant donne le total des prestations sociales reçues par les ménages en France de 1988 à 1992 :

Année	1988	1989	1990	1991	1992
Rang x_i de l'année	0	1	2	3	4
Total des prestations en milliards de francs : y_i	1338	1415	1505	1606	1700

Source : INSEE, Tableaux de l'économie française 1993-1994

1°) Représenter le nuage de points associé à la série statistique (x_i, y_i) : le plan est rapporté à un repère orthogonal. Les unités graphiques sont : 2 cm par année sur l'axe des abscisses, 1 cm pour 100 milliards de francs sur l'axe des ordonnées, en commençant la graduation à 1 000 milliards.

2°) a) Calculer à 10^{-3} près par excès, le coefficient de corrélation linéaire de la série (x_i, y_i). En déduire qu'un ajustement affine est justifié.

b) Ecrire une équation de la droite de régression de y en x par la méthode des moindres carrés : on donnera les coefficients à 10^{-1} près par défaut.

c) Estimer le total des prestations sociales reçues par les ménages en 1997.

3°) En supposant que le tendance ainsi constatée se maintienne, à partir de quelle année le total des prestations dépassera-t-il 2 200 milliards ?

TD 2 – UTILISATION DE LA CALCULATRICE
SERIES STATISTIQUES A DEUX VARIABLES (CASIO ©)

Objectifs : utiliser sa calculatrice pour :

- réaliser des ajustements linéaires et non linéaires ,
- déterminer une droite de Mayer,
- déterminer une droite des moindres carrés,
- réaliser une régression exponentielle, logarithmique

I. ENTREE DES DONNEES

Les calculs se font en utilisant les listes dans le mode STAT pour l'enregistrement des données et la lecture des résultats et le mode RUN pour certains calculs.
Remarque : tous les calculs sont à faire avec toute la précision de la machine. Seuls les résultats seront arrondis

Initialisation : utiliser list 1 pour la variable x_i et list 2 pour la variable y_i, les fréquences étant bloquées à 1. List 3 est éventuellement utilisée pour le calcul des valeurs $\hat{y}$, en fonction de x, des ordonnées de la droite d'ajustement et list 4 pour le calcul des résidus.
En cas de problème vérifier les paramètres dans le menu STAT GRPH SET :
Graph type : Scatter
Xlist : list 1
Ylist : list 2
Frequency : 1
Mark Type : ?

Entrée des données : se placer dans le mode STAT. Il est prudent d'effacer les données avec le deuxième menu des STAT : ? DEL-A en se plaçant sur chacune des listes.
Puis dans list 1 et list 2 taper chaque valeur suivie de EXE. Ne faire aucune erreur, ni oubli ni valeur antérieure non effacée et vérifier.

Exemple
Cinéma et télévision (Fractale TES exercice 21 page 24)

Département	Milliers de spectateurs au cinéma	Nombre de téléviseurs
Loire Atlantique	2294	381572
Maine et Loire	1317	240093
Vendée	741	190037
Sarthe	735	188840
Mayenne	346	99650

II. AJUSTEMENT GRAPHIQUE

Dessiner le nuage de points. Calculer les moyennes $\bar{x}$ et $\bar{y}$. Menu CALC 2VAR

$\bar{x} = \frac{\sum_{i=1}^{i=n} x_i}{n} = \frac{5433}{5} \approx 1086$ $\qquad$ $\bar{y} = \frac{\sum_{i=1}^{i=n} y_i}{n} = \frac{1100100}{5} \approx 220038.$

Tracer le centre de gravité G($\bar{x}$, $\bar{y}$) et une droite D passant par G traversant le nuage de points « au plus près ». Lire sur le graphique les coordonnés u et v d'un point M de cette droite (assez éloigné de G), par exemple M(2000, 320 000).

Calculer le coefficient directeur $a = \frac{\bar{y} - v}{\bar{x} - u} = \frac{220038 - 320000}{1086 - 2000} \approx 109.$

Taper STO ALPHA A pour mémoriser a (écran : ans → A).

Puis trouver le coefficient constant sachant que l'équation de la droite passant par G est

$$y - \bar{y} = a\,(x - \bar{x}).$$

soit $b = \bar{y} - a\,\bar{x} = 220038 - a\,1086 = 101664.$

Taper STO ALPHA B pour mémoriser b (écran : ans → B).

La droite a donc pour équation y = 109 x + 101664

III. CALCUL DES RESIDUS

Le coefficient directeur ayant été mémorisé dans la variable A et le coefficient constant dans la variable B, en mode RUN placer les valeurs $\hat{y}$ en fonction de x des ordonnées de la droite d'ajustement d'équation y = ax + b dans list 3 (les valeurs de list 3 permettent aussi de tracer la droite).

Avec la touche OPTN choisir LIST (au commencement taper deux fois sur LIST ; menu LIST et mot List) :

A list 1 + B → list 3

Calculer les résidus et les ranger dans list 4 :

(list 2 –list 3)^2 → list 4

puis calculer leur somme : taper deux fois sur ▶ : valider Sum puis ▶ List :

Sum List 4

Dans cet exemple on trouve la somme des résidus égale à $\sum_{i=1}^{i=n} \left[y_i - (ax_i + b)\right]^2 \approx 2{,}60\ 10^9$.

IV. METHODE DE MAYER

La série étant ordonnée, la partager en deux parties. Ici les deux départements les plus importants d'une part, et les trois autres d'autre part.

Calculer les coordonnées du point $G_1(u, v)$ correspondant au point moyen associé aux points d'abscisses les plus petites et $G_2(s, t)$ aux autres points du nuage.

Une méthode avec la calculatrice Casio : en mode RUN après la touche OPTN recopier list 1 et list 2 dans list 3 et list 4 puis dans list 5 et list 6.

list 1 → list 3
list 2 → list 4
list 1 → list 5
list 2 → list 6

Dans le menu LIST avec la touche DEL supprimer les (ici les trois) derniers éléments de list 3 et list 4 et les (ici les deux) premiers de list 5 et list 6
En mode Run calculer les moyennes (OPTN List ▸ Mean)
Mean (List 3) 1805.5→ALPHA U Mean (List 5) 607.3→ALPHA S
Mean (List 4) 310832→ALPHA V Mean (List 6) 159509→ALPHA T

Calculer le coefficient directeur a = $\frac{v-t}{u-s} = \frac{159509-310832}{607.3-1805.5} = 126{,}3$.
Le conserver dans la variable A : taper →ALPHA A ; l'affichage est Ans→A
126.3.
Puis le coefficient constant sachant que l'équation de la droite passant par exemple par G_1 est y - v = a (x - u) ; soit b = v - a u = 310832 - A×1805.5 = 82811.

Le conserver dans la variable B : → ALPHA B.

La droite de Mayer a donc pour équation y = 126,3 x + 82811

Vérifier sur le graphique que la droite de Mayer passe bien par le point G.

On peut effectuer, comme ci dessus chapitre III, le calcul de la somme des résidus. On trouve alors $\sum_{i=1}^{i=n}\left[y_i-(ax_i+b)\right]^2 \approx 1{,}24\ 10^9$. L'équation de la droite de Mayer est donc une bien meilleure approximation que l'ajustement affine précédent.

V. AJUSTEMENT PAR LA METHODE DES MOINDRES CARRES

La calculatrice donne les principales valeurs remarquables, le coefficient de corrélation, l'équation de la droite de régression $D_{y/x}$. On peut obtenir un graphique et une estimation permettant de prévoir $\hat{y}$ en fonction de x ou $\hat{x}$ en fonction réciproque de y.

a. Calcul des valeurs caractéristiques

Dans le mode STAT utiliser les options CALC 2VAR.
Moyennes :
$\bar{x} = \frac{\sum_{i=1}^{i=n} x_i}{n} = \frac{5433}{5} \approx 1086 \quad \bar{y} = \frac{\sum_{i=1}^{i=n} y_i}{n} = \frac{1100100}{5} \approx 220038$
Variances et écart-types :
Noter les valeurs trouvées sur l'écran de la calculatrice :
V(X) = $\frac{\sum_{i=1}^{i=n} x_i^2}{n} - \bar{x}^2 = \frac{8{,}206\times10^6}{5} - 1086^2 \approx \ldots$; V(Y) = $\frac{\sum_{i=1}^{i=n} y_i^2}{n} \bar{y}^2 = \frac{2{,}849\times10^{11}}{5} - 220038^2$
≈....
$s_x = \sqrt{V(X)} \approx 678{,}6$ $s_y = \sqrt{V(Y)} \approx 92587$ (notés par Casio xsn et ysn et à ne pas confondre avec xsn-1 et ysn-1 où les calculs sont modifiés par une division par n-1.)

Covariance :

$s_{xy} = \frac{\sum_{i=1}^{i=n} x_i y_i}{n} - \bar{x}\,\bar{y} = \frac{1{,}5056\times10^9}{5} - 1086 \times 220038 \approx \ldots.$

Pour le calcul des variances et de la covariance, sans éteindre la calculatrice, quitter les STAT, avec la touche MENU choisir le Mode RUN, la touche VARS puis l'option STAT.
Enfin la variable X et l'écart type xsn x^2 suivi de EXE. La calculatrice affiche s_x^2 :

xsn² 460489,84

à recopier ci dessus dans V(X) = …. ≈ 460 490.
Puis touche EXIT et choisir la variable Y et le calcul de ysn x^2 suivi de EXE. La calculatrice affiche s_y^2 à recopier ci dessus dans V(Y) = …. ≈ $8{,}572\times10^9$.

De même pour le calcul de la covariance utiliser la variable Y puis X, en mémorisant l'effectif n :

Σxy ÷ 5 - $\bar{y}$ × EXIT X $\bar{x}$ EXE
Casio affiche : Σxy ÷ 5 - $\bar{y} \times \bar{x}$ = ≈ 62030747
Soit s_{xy} = …. ≈ $6{,}203\times10^7$

b. droite de régression

Revenir au mode STAT, choisir CALC ou GRPH, puis REG et X.

Explications à fournir

La droite de régression a pour coefficient directeur :

$$a = \frac{s_{xy}}{s_x^2} \approx \frac{6{,}203\times10^7}{678{,}6^2} \approx 134{,}7.$$

La droite passe par le centre de gravité G($\bar{x}$, $\bar{y}$) à représenter sur le graphique ; son équation est :

$$y - \bar{y} = a\,(x - \bar{x}).$$

Le coefficient constant est $b = \bar{y} - a\,\bar{x} \approx 73667$.
La droite de régression de y en x a pour équation : y = 134,7 x + 73677

Mémorisation des coefficients

Après l'exécution des calculs de la régression avec la touche MENU, choisir le mode calcul RUN. Avec la touche VARS choisir F3 les STAT puis encore F3 le mode GRPH.
Le bas de l'écran affiche les coefficients a et b de la droite que l'on peut mémoriser par exemple dans les variables A et B.
Placer le coefficient directeur dans la variable A : F1 → ALPHA A

$a \rightarrow A$

et le coefficient constant dans la variable B : F2 → ALPHA B

$b \rightarrow B$

Il est aussi possible de mémoriser l'équation dans l'éditeur de fonction avec le menu GRPH et, après le calcul de régression REG et X, utiliser la l'option COPY.

c. Coefficient de corrélation

Le coefficient de corrélation r est un nombre compris entre -1 et 1. Le voisinage de 1 ou de -1 indique en principe une bonne corrélation sans que l'on puisse directement en dire plus. On ne peut conclure à la validité d'une corrélation qu'en fonction du contexte statistique.

$$r = \frac{s_{xy}}{s_x s_y} \approx \frac{6{,}203\times10^7}{678{,}6\times92587} \approx 0{,}987.$$

Très bon taux de corrélation. Ce qui montre qu'actuellement en France, contrairement à une idée reçue, le développement du Cinéma va de pair avec celui de la télévision.

Mémorisation du coefficient

Après l'exécution des calculs de la régression avec la touche MENU, choisir le mode calcul RUN. Avec la touche VARS choisir F3 les STAT puis encore F3 le mode GRPH.
Le bas de l'écran affiche les coefficients a, b, c… des courbes. Taper sur F6 ▸ pour obtenir l'écran suivant.
S'affiche le coefficient de corrélation r que l'on peut mémoriser par exemple dans la variable R :

F1 → ALPHA R

$r \rightarrow R$

Estimation

Lorsque r est voisin de 1 ou de –1 on peut effectuer des **estimations** avec, en fonction du contexte, une bonne fiabilité.
Dans le mode RUN avec la touche OPTN choisir les STAT :
taper : 1000 $\hat{y}$ 208372
permet d'estimer à 208 000 le nombre de téléviseur d'un département ou il y a 1 000 milliers d'entrées au cinéma.
De même : 500 000 $\hat{x}$ 3164,9
permet d'estimer 3 200 000 séances de cinéma dans un département ayant 500 000 postes de télévision.

d. Calcul de la somme des résidus

Comme au chapitre III on va placer dans list 3 les valeurs $\hat{y}$ calculées en fonction de x des ordonnées des points de la droite d'ajustement d'équation y = ax + b.

Dans le mode STAT après l'exécution des calculs de la régression revenir à l'écran affichant les liste (avec entre autre la touche EXIT).
Avec les flèches de direction se placer sur le titre list 3 qui apparaît en inversion vidéo.

Pour le coefficient directeur a avec la touche VARS choisir F3 les STAT puis encore F3 le mode GRPH, taper F1 la calculatrice affiche : a

Avec la touche OPTN choisir menu LIST et mot List et taper 1 : a list 1
taper + a list 1 +

Pour le coefficient constant b avec la touche VARS choisir F3 les STAT puis encore F3 le mode GRPH, taper F2 : a list 1 + b

Valider avec EXE le tableau list 3 se remplit automatiquement avec les valeurs $\hat{y}$.

Pour les résidus, carrés des écarts, se placer avec les flèches de direction sur le titre list 4 qui apparaît en inversion vidéo.
taper une parenthèse ouvrante : (
Avec la touche OPTN choisir menu LIST et mot List et taper 2 : (list 2
taper le signe de soustraction - : (list 2 –
enfin F1 mot List et taper 3 : : (list 2 – list 3
fermer la parenthèse, calculer le carré avec la touche x^2 et valider avec EXE : : (list 2 – list 3)^2
list 4 contient alors les résidus.

Se placer sur une case vierge par exemple dans list 5 et calculer leur somme : taper deux fois sur ▶ : valider Sum puis ▶ List et enfin taper 4 et terminer par EXE : Sum List 4

On calcule ainsi la somme des résidus égale à :

$$\sum_{i=1}^{i=n}\left[y_i-(ax_i+b)\right]^2 \approx 1{,}08\ 10^9.$$

La droite $D_{y/x}$ est donc une meilleure approximation que les ajustements précédents.
Par définition il n'est pas possible de trouver une droite admettant un résidu moindre.

VII. REGRESSION AVEC CHANGEMENT DE VARIABLE

a. fonction inverse : valeur d'une voiture d'occasion

Voici, en 1999, la cote Argus d'un type de voiture d'occasion :

Année de mise en circulation	1998	1997	1996	1995	1994	1993	1992
Cote Argus en milliers de francs	58	48	38	32	24	19	15

Le but de ce problème est d'estimer le prix d'une voiture de ce type non cotée mise en circulation en 1990.

On note x l'âge de la voiture (en années) et y la cote Argus (en milliers de francs).

1°) **a)** Représenter graphiquement la série (x_i, y_i).
b) Déterminer une équation de la droite de régression de y en x et tracer cette droite. (Effectuer les calculs avec trois chiffres significatifs).
c) Calculer le coefficient de corrélation linéaire r_1 de y en x.
d) Peut-on estimer la valeur d'une voiture mise en circulation en 1990 ? Expliquer.

2°) Les spécialistes pensent qu'on aura un meilleur ajustement en remplaçant les sept valeurs y_i par les valeurs $z_i = \frac{1}{y_i}$.
a) Présenter dans un tableau la série double (x_i, z_i), i variant de 1 à 7.
b) Calculer le coefficient de corrélation linéaire r_2 de z en x. Comparer r_1 et r_2.
c) Déterminer, à l'aide de la calculatrice, l'équation de la droite de régression de z en x sous la forme z = mx + p (m et p étant arrondis à 10^{-6} près).
d) À l'aide de cette équation peut-on estimer la valeur d'une voiture mise en circulation en 1990 ? Donner le résultat et expliquer.

Indications de correction :
Dans la première partie introduire en mode STAT x_i dans list 1 et y_i dans list 2.
La séquence CALC REG X permet de trouver l'équation $y = -7{,}17\,x + 62{,}1$.
Bien que le coefficient de corrélation $r_1 = -0{,}989$ soit très proche de - 1 pour $x = 9$ une valeur négative, ce qui est absurde, permet de rejeter la méthode dans ce cas.

Pour la deuxième partie conserver list 1 pour la variable x_i ;
Transférer la variable y_i de list 2 vers list 3 (list 2 → list 3) :
avec les flèches de direction se placer sur le titre list 3 qui apparaît en inversion vidéo.
Après la touche OPTN choisir LIST (au commencement taper deux fois sur LIST : menu LIST et mot List) ; taper 2 et valider avec EXE.

Puis faire le changement de variable z_i dans list 2 (en calculant 1/ list 3 → list 2) ;
avec les flèches de direction se placer sur le titre list 2 qui apparaît en inversion vidéo.
Taper 1 ÷ puis sur F1 après le mot List taper 3. Valider 1 ÷ list 3 avec EXE.

En tapant deux fois sur EXIT, le retour au mode CALC REG X permet de trouver le coefficient $r_2 = 0{,}973$ et l'équation $z = m\,x + p$ avec $m = 8{,}115\ 10^{-3}$ et $p = 4{,}197\ 10^{-3}$.
Le coefficient r_2 est en principe moins pertinent que r_1 mais son voisinage de 1 indique une bonne corrélation :
pour $x = 9$ on trouve $z = 0{,}0772$ donc $y = 12{,}9$ permet d'estimer la valeur d'une voiture de 1990 à 13 000 francs.

b. fonction racine : distance de freinage

Au cours d'une séance d'essai un pilote d'automobile doit, quand il reçoit un signal sonore dans son casque, arrêter le plus rapidement possible son véhicule.
Au moment du top sonore, on mesure la vitesse de l'automobile puis la distance nécessaire pour arrêter le véhicule.
Pour six expériences, on a obtenu les résultats suivants :

Vitesse v_i en km/h	21	43	62	77	98	115
Distance d'arrêt y_i en m	8	20	33	55	102	137

1. Calculer, à l'aide d'une calculatrice, le coefficient de corrélation linéaire r_1 de y en v.

2. Les spécialistes pensent qu'on aura un meilleur ajustement en remplaçant les six valeurs v_i par les valeurs $x_i = v_i^2$
Présenter dans un tableau la série double (x_i, y_i), i variant de 1 à 6.
Calculer le coefficient de corrélation linéaire r_2 de y en x. Comparer r_1 et r_2.

3. Dans un repère orthogonal construire le nuage de points associé à cette nouvelle série double.
les x_i en abscisses avec 1 cm pour 1000,
les y_i. en ordonnées avec 1 cm pour 10.

4. a. Déterminer, à l'aide de la calculatrice, l'équation de la droite de régression de y en x sous la forme $y = mx + p$ (m et p étant arrondis à 10^{-2} près).
Tracer cette droite dans le repère précédent.

b. Quelle est la distance d'arrêt estimée correspondant à une vitesse de 150 km/h ?
c. À l'aide de cette équation, déterminer la valeur estimée de x correspondant à une distance d'arrêt de 180 m, puis la vitesse correspondante du véhicule.
d. Le manuel du code de la route donne, pour calculer la distance d'arrêt (en mètres), la méthode suivante :
«Prendre le carré de la vitesse exprimée en dizaines de kilomètres par heure.»
Comparer le résultat obtenu au **c**. à celui que l'on obtiendrait par cette méthode.

Indications de correction :
Pour la première question introduire en mode STAT v_i dans list 1 et y_i dans list 2. La séquence CALC REG X permet de trouver $r_1 = 0,96$.
L'équation d'ajustement, non demandée, $y = 1,4\ v - 37,6$ donne des valeurs négatives pour les vitesse inférieur à 27 km/h et des distances de freinage sous évaluées pour les grandes vitesses n'est pas satisfaisante.

Pour la suite conserver list 2 pour la variable y_i ;
Transférer la variable x_i de list 1 vers list 3 (list 1 $\rightarrow$ list 3) :
avec les flèches de direction se placer sur le titre list 3 qui apparaît en inversion vidéo.
Après la touche OPTN choisir LIST (au commencement taper deux fois sur LIST : menu LIST et mot List) ; taper 1 et valider avec EXE.

Puis faire le changement de variable x_i dans list 1 (en calculant le carré : list 3^2 $\rightarrow$ list 1) :
avec les flèches de direction se placer sur le titre list 1 qui apparaît en inversion vidéo.
Taper F1 pour List et taper 3. Calculer le carré avec la touche x^2 , valider list 3^2 avec EXE.

En tapant deux fois sur EXIT, le retour au mode CALC REG X permet de calculer $r_2 = 0,996$ ce qui est une très bonne corrélation.
L'équation $y = m\ x + p$ avec $m = 0,0104$ et $p = -1,09$ donne la distance de freinage en fonction de la vitesse : $.y = 0,0104\ v^2 - 1,09$.
Pour $v = 150$ on peut estimer la distance de freinage à 232 mètre.
On peut prévoir qu'une voiture s'arrêtant sur 180 m ferait du 132 km/h.
La prévention routière utilise la fonction $y = 0,01\ v^2$ qui donne une très bonne approximation et permet de prévoir 225 m de freinage pour 150 km/h ou 134 km/h pour 180 m de freinage.

VII. REGRESSION EXPONENTIELLE

Lors d'une épidémie on a relevé toute les semaines x_i le nombre de cas y_i.

x_i	1	2	3	4
y_i	94	221	446	1050

1. Représenter le nuage de points dans un repère convenable. Un ajustement affine paraît-il justifié ?
2. On pose $z_i = \ln y_i$ (ln désigne le logarithme népérien). Calculer l'équation de la droite de régression de z en x par la méthode des moindres carrés.
3. Trouver une relation entre x et y de la forme $y = a\ b^x$.
4. Combien de malades peut-on prévoir pour la cinquième semaine ?

Indications de correction :

Question **2.** : introduire x_i dans list 1 et y_i dans list 3.
Puis faire le changement de variable z_i (ln list 3 $\rightarrow$ list 2) :
Avec les flèches de direction se placer sur le titre list 2 qui apparaît en inversion vidéo.
Calculer le logarithme avec la touche ln, avec OPTN choisir List et taper 3., valider ln list 3 avec EXE.

En tapant deux fois sur EXIT, le mode CALC REG X permet de trouver l'équation ln y = b x + a avec b = 0,7941 et a = 3,764.

Question **3.** : transférer la variable y_i de list 3 vers list 2 (list 3 → list 2) :
Avec les flèches de direction se placer sur le titre list 2. Avec la touche F1 choisir List, taper 3 et valider avec EXE.
De nouveau le mode CALC REG EXP permet de trouver :

$$\alpha = e^{3,76} = 43,12 \text{ et } \beta = e^{0,794} = 2,213.$$

Question **4.** On a une bonne corrélation avec r = 0,9992, on peut donc prévoir pour la cinquième semaine $\alpha\beta^5 = 2287$ malades.

VIII. REGRESSION LOGARITHMIQUE

La marge brute d'autofinancement (M.B.A.) d'une entreprise de 1996 à 2001 en pourcentage de son chiffre d'affaire est donnée par le tableau suivant où x_i représente le rang de l'année et y_i la M.B.A. en pourcentage :

x_i	1	2	3	4	5	6
y_i	8,13	8,51	8,79	9	9,15	9,31

1. Représenter le nuage de points dans un repère convenable.
2. On pose $z_i = e^{yi}$. Calculer l'équation de la droite de régression de z en x par la méthode des moindres carrés.
3. En déduire une relation entre x et y.
4. En quelle année la marge brute d'autofinancement devrait dépasser 10 % ?

Indications de correction :

Introduire x_i dans list 1 et y_i dans list 3.
Faire le changement de variable la variable z_i (e^x list 3 → list 2) :
Avec les flèches de direction se placer sur le titre list 2 qui apparaît en inversion vidéo.
Frapper la touche e^x pour l'exponentielle, dans le menu OPTN choisir List et taper 3.
Valider e^x list 3 avec EXE.

Le mode CALC REG X permet de trouver l'équation $e^y = a\,x + b$ avec a = 1518 et b = 1933 ; r = 0,9996 indiquant une très bonne corrélation.
On a donc la relation $e^y = 1518\,x + 1933$ soit $y = \ln(1518\,x + 1933)$.
Cet ajustement permet de prévoir 10% en 2014 au bout de 19 ans.

Utilisation du programme de la calculatrice
Le programme d'ajustement par régression logarithmique effectue un ajustement linéaire entre les variables $t_i = \ln(x_i)$ et y_i.

Transférer la variable y_i de list 3 vers list 2 : se placer sur le titre list 2 et valider List 3.

De nouveau le mode CALC REG LOG permet de trouver :
$y = a + b \ln x$ avec a = 8,094 et b = 0,6575.
On a alors la relation $y = 0,6575 \ln x + 8,094$.
Avec une moins bonne corrélation on prévoit 10% en 2009 au bout de 14 ans.
Ces prévisions à trop long terme ne sont certainement pas fiables.