INSEE
DIRECTION GÉNÉRALE
INSTITUT NATIONAL
DE LA STATISTIQUE
ET DES ÉTUDES
ÉCONOMIQUES

Direction des Statistiques Démographiques et Sociales
Unité Méthodes Statistiques
Division Méthodologie d'Elaboration et d'Analyse de Données
Timbre F410

LA REGRESSION LINEAIRE SOUS SAS®

Document n° F 9605

révision Mars 1997

Josiane CONFAIS & Monique LE GUEN

Série des documents de travail

de la Direction des Statistiques Démographiques et Sociales

TIMBRE - 18, BLD ADOLPHE PINARD - 75675 PARIS CEDEX 14 - TEL.

1. Introduction : Où se place la régression linéaire

La **régression linéaire** se classe parmi les méthodes d'analyses multivariées qui traitent des données quantitatives.

C'est une méthode d'investigation sur données d'observations, où l'objectif fondamental est de rechercher une liaison linéaire entre une variable Y quantitative et une ou plusieurs variables X également quantitatives.

C'est la méthode la plus utilisée pour deux raisons majeures :

- c'est une **méthode ancienne**
- c'est l'**outil de base** de la plupart des modélisations plus sophistiquées comme la régression logistique, le modèle linéaire généralisé, etc. et les méthodes de traitement des séries temporelles.

A l'aide du tableau suivant on peut repérer les méthodes les plus courantes d'analyses statistiques et les procédures SAS utiles pour rechercher des liaisons, selon le type[1] (nominal, ordinal, intervalle, ratio) des variables Y et X.

	X intervalle/ratio	X ordinales/nominales	
Y intervalle/ratio	Régression linéaire ***PROC REG***	Analyse de la variance ***PROC ANOVA***	Modèles linéaires généralisés ⇦ ***PROC GLM***
Y ordinale/nominale	*Si Y est ordinale ou à 2 modalités* Régression logistique ***PROC LOGISTIC***	Analyses de tableaux de contingence ***PROC FREQ*** Régression logistique ***PROC LOGISTIC***	traitements des variables catégorielles[1] ⇦ ***PROC CATMOD***

Tableau : Procédures SAS adaptées selon le type (nominal, ordinal, intervalle, ratio) des variables

[1]Pour le lecteur peu familiarisé avec la terminologie des variables SAS voir Annexe 1 dans les Annexes Générales.

2. Ajustement affine ou Régression Simple

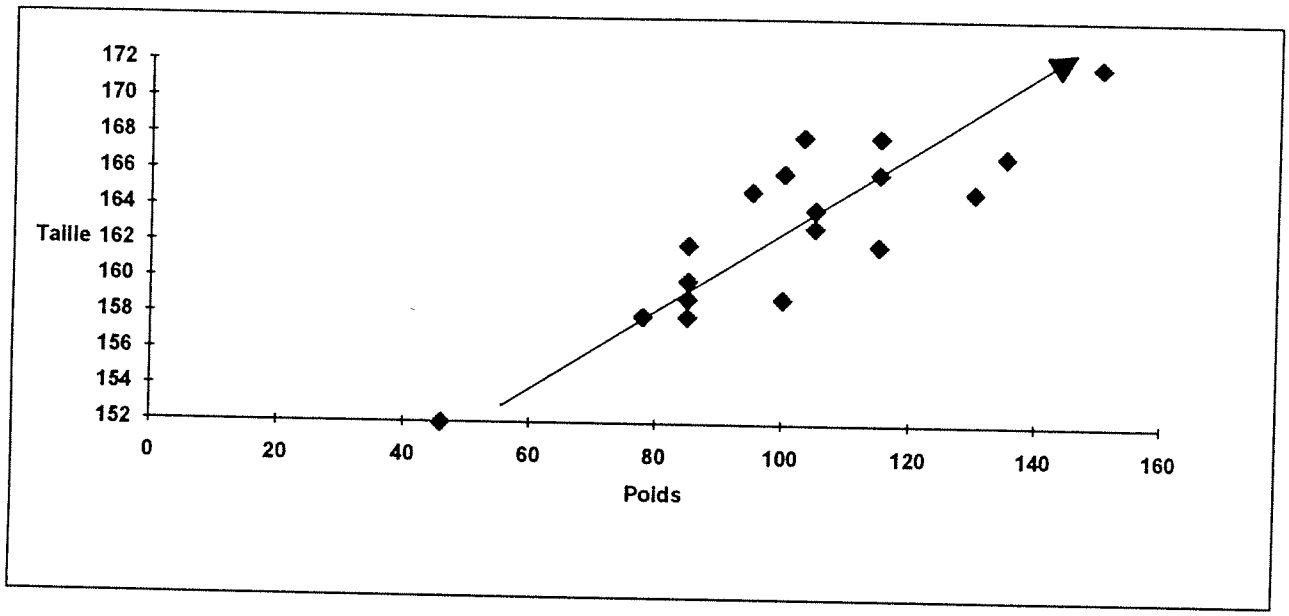
Exemple

Soient les 2 mensurations poids (variable X) et taille (variable Y) relevées sur un échantillon de 20 individus.

identifiant	poids (X)	taille (Y)
1	46	152
2	78	158
3	85	160
4	85	162
5	85	158
6	85	159
7	95	165
8	95	165
9	100	166
10	100	159
11	100	166
12	103	168
13	105	163
14	105	164
15	115	168
16	115	166
17	115	162
18	130	165
19	135	167
20	150	172

Tableau des données

Par une représentation graphique du nuage de points d'abscisse le **poids** et d'ordonnée la **taille** on voit qu'il existe une **relation linéaire** entre ces deux variables.



*Graphique Taille*Poids*

Les points du nuage sont approximativement alignés sur une droite (y=ax+b) à une erreur près.

$$\textbf{Taille}=\beta_0 + \beta_1 \textbf{ Poids} + \textbf{erreur}$$

La variable Taille (Y) est appelée la variable "**réponse**".

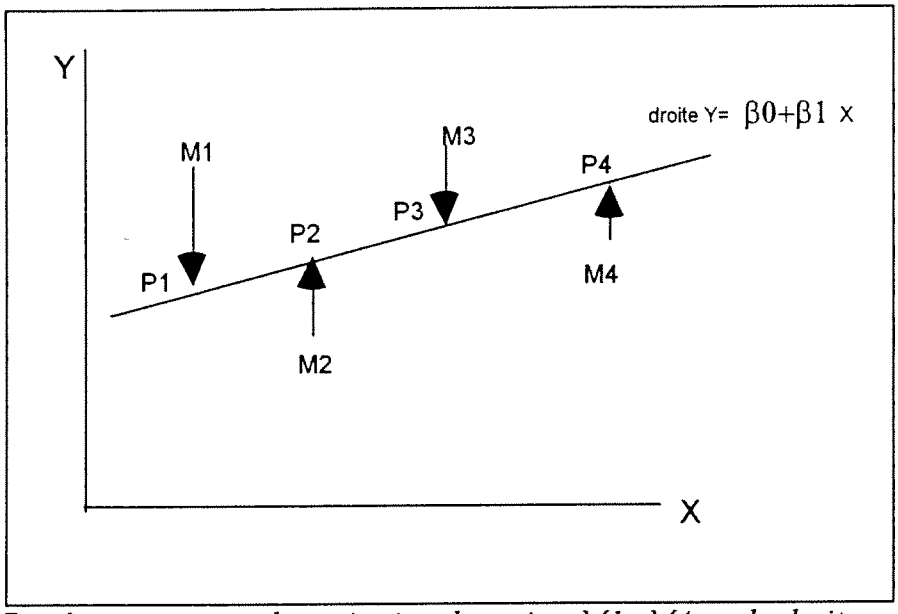
La variable Poids (X) est la variable "**régresseur**".

β_0 est l'ordonnée à l'origine.

β_1 est la pente de la droite d'ajustement.

2.1. Comment trouver la droite qui passe "au plus près" de tous les points?

Pour trouver la droite qui passe "au plus près" de tous les points il faut se donner un **critère d'ajustement**.



Graphique montrant la projection des points M1...M4 sur la droite.

On projette les points M1 à M4 parallèlement à l'axe des Y. Sur la droite on obtient les points P1 à P4, cf. le graphique ci-dessus.

Le critère retenu pour déterminer la droite D passant au plus près de tous les points sera tel que :

La somme des carrés des écarts (SCE) des points observés à la droite solution soit minimum.

La droite solution sera appelée **droite de régression de Y sur X**.

Le critère est le **"critère des Moindres Carrés Ordinaires"** (MCO).

Note :
Les écarts sont calculés en projetant les points M **parallèlement à l'axe des Y**.
On pourrait aussi projeter les points M parallèlement à l'axe des X, on aurait alors une autre droite solution (régression de X sur Y). Y et X ne jouent pas le même rôle.

On pourrait aussi projeter les points M perpendiculairement à la droite solution. Y et X joueraient dans ce cas le même rôle. C'est cette situation que l'on rencontre dans une Analyse en Composantes Principales, ou dans une régression orthogonale.

2.2. Méthode d'estimation des paramètres β_0 et β_1

La Somme des Carrés des Ecarts est donnée par :

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

La valeur de cette fonction S est minimum lorsque les dérivées de S par rapport à β_0 *et* β_1 s'annulent. La solution est obtenue en résolvant le système :

$$\frac{\partial S}{\partial \beta_0} = 0 \quad \text{et} \quad \frac{\partial S}{\partial \beta_1} = 0$$

Les dérivées par rapport à β_0 et β_1 sont :

$$\frac{\partial S}{\partial \beta_0} = -2\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)$$

$$\frac{\partial S}{\partial \beta_1} = -2\sum_{i=1}^{n} X_i (Y_i - \beta_0 - \beta_1 X_i)$$

Ces dérivées s'annulent pour deux valeurs b_0 et b_1 solutions des 2 équations à 2 inconnues :

équation 1 :

$$\sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i) = 0$$

équation 2 :

$$\sum_{i=1}^{n} X_i (Y_i - b_0 - b_1 X_i) = 0$$

Ce système de 2 équations à 2 inconnues déterminent les **équations normales.**

Pourquoi le terme équations normales ?.....
voir Annexe 1 de la partie I page 38, "Aparté linguistique du terme *Normal*, comment s'y retrouver".

Développons ces 2 équations normales :

• l'équation 1 donne :

$$\sum Y_i - nb_0 - b_1 \sum X_i = 0 \quad soit \quad en \quad divisant \quad par \quad n$$
$$\bar{Y} = b_0 + b_1 \bar{X}$$

Remarque:

La droite solution passe par le centre de gravité du nuage : $(\bar{X}, \bar{Y})$ ou $\left(\frac{\sum_i X_i}{n}, \frac{\sum_i Y_i}{n} \right)$

• L'équation 2 donne

$$\sum Y_i X_i - b_0 \sum X_i - b_1 \sum X_i^2 = 0$$

dans laquelle on remplace b_0

$$\sum Y_i X_i - (\bar{Y} - b_1 \bar{X}) \sum X_i - b_1 \sum X_i^2 = 0$$

Solution :

$$b_1 = \frac{\sum X_i Y_i - (\sum X_i \sum Y_i)/n}{\sum X_i^2 - (\sum X_i)^2 / n}$$

en divisant numérateur et dénominateur par n on retrouve les expressions de la covariance et de la variance empiriques :

formule n° 1

$$b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{Cov(X,Y)}{Var(X)}$$

Les points qui sont sur la droite de régression auront pour ordonnée:

$$\hat{Y} = b_0 + b_1 X$$

$\hat{Y}$ est l'**estimation** de Y obtenue à partir de l'équation de régression.
$\hat{Y}$ *se prononce Y chapeau.*

b_0 et b_1 sont les **estimateurs** des moindres carrés de β_0 et β_1.

2.3. Décomposition de l'écart entre Y_i et la moyenne de Y

En un point d'observation (X_i, Y_i) on décompose l'écart entre Y_i et la moyenne des Y en ajoutant puis retranchant $\hat{Y}_i$ la valeur estimée de Y par la droite de régression. Cette procédure fait apparaître une somme de deux écarts :

$$(Y_i - \overline{Y}) = (Y_i - \hat{Y}_i + \hat{Y}_i - \overline{Y})$$

$$(Y_i - \overline{Y}) = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \overline{Y})$$

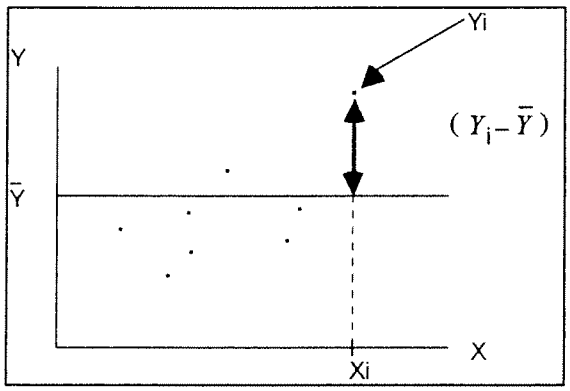
$$\Downarrow \qquad \Downarrow \qquad \Downarrow$$

cf. graph1 graph2 graph3

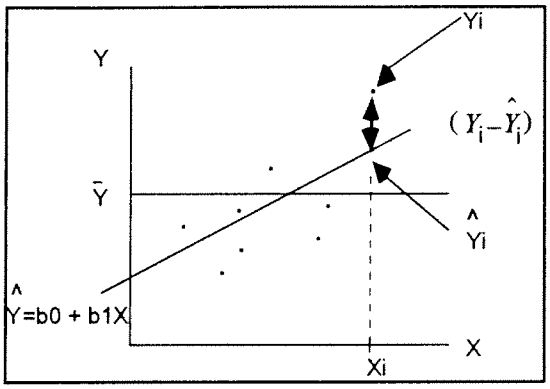
Ainsi l'écart total $(Y_i - \overline{Y})$ peut être vu comme la somme de deux composantes :

- un écart entre Y_i et $\hat{Y}_i$ la valeur estimée par le modèle
- un écart entre $\hat{Y}_i$ la valeur estimée par le modèle et la moyenne $\overline{Y}$.

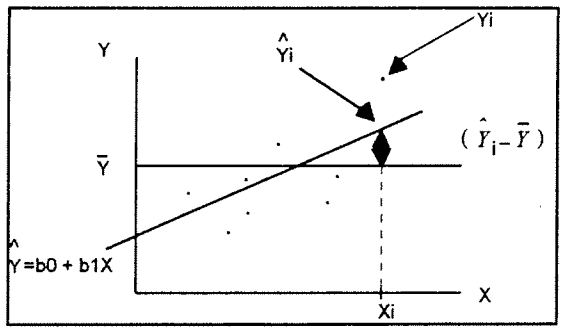
Les 4 graphiques suivants montrent l'explication géométrique de cette décomposition.
Cet artifice de décomposition aura un intérêt fondamental dans l'analyse de la variance abordée au paragraphe suivant.



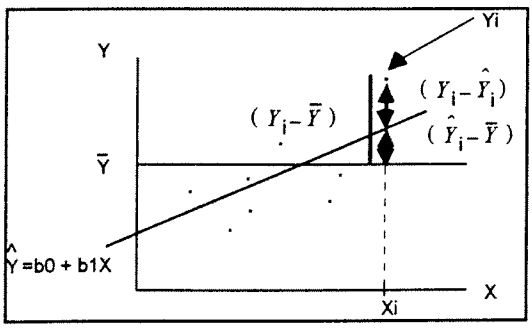
Graphique 1 : Ecart $(Y_i - \overline{Y})$
le graphique 1 montre l'écart entre Y_i et $\overline{Y}$



Graphique 2 : Ecart $(Y_i - \hat{Y}_i)$
le graphique 2 montre l'écart entre Y_i et Y estimé



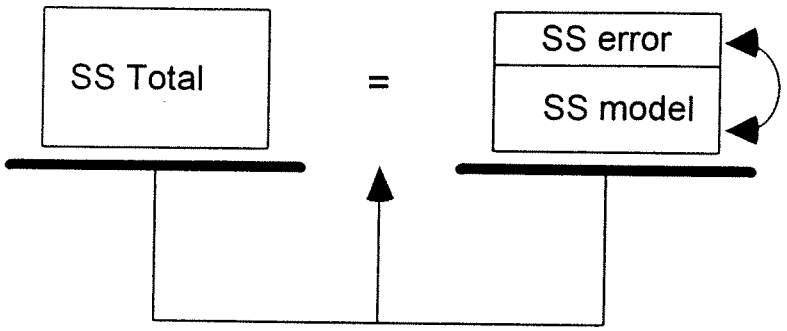
Graphique 3 : Ecart $(\hat{Y}_i - \overline{Y})$
le graphique 3 montre l'écart entre Y estimé et $\overline{Y}$



Graphique 4 :
Décomposition de l' écart total
$Y_i - \hat{Y}_i = (Y_i - \overline{Y}) - (\hat{Y}_i - \overline{Y})$

4.6. Ce qu'il faut retenir des 'SS'

Décomposition des SS : Sum of Squares



Lorsqu'on introduit une nouvelle variable dans un modèle :

SS model augmente de la même quantité que

SS error décroît

Donc le coefficient de détermination **R-square augmente toujours**.

Cependant on n'améliore pas nécessairement la précision de l'estimation de Y.
En effet SS Error décroît mais s^2= MS Error=SS error/(n-p-1) peut croître donc augmenter la largeur de l'intervalle de confiance de Y estimé qui est proportionnel à "s", cf. Partie I page 30.

A la limite si le nombre de variables p + 1 (le 1 correspond à la variable constante X_0) est égal au nombre d'observations (n), l'équation de régression passera exactement par tous les points du nuage, l'ajustement sera parfait.
Dans ce cas SSerror vaut 0, et le coefficient de détermination R^2 vaut 1.

Ce n'est plus de la Statistique mais de la résolution d'équations !

Modèles "*parcimonieux*"
Les statisticiens parlent de modèles "*parcimonieux*", pour signifier qu'un modèle doit comporter un nombre limité de variables par rapport au nombre d'observations, si on veut que le modèle ait une portée prévisionnelle et/ou explicative.

5. Quand les résultats d'une régression ne sont pas forcément pertinents

5.1. Exemples en régression simple.

5.1.1. Une même valeur pour des situations différentes

Ces exemples sont empruntés à Tomassone, Lesquoy, Millier (1986).
Soient les 5 couples de 16 observations (X,Ya),(X,Yb),(X,Yc),(X,Yd),(Xe,Ye) sur lesquels on effectue 5 régressions linéaires.

OBS	X	Ya	Yb	Yc	Yd	Xe	Ye
1	7	5,535	0,113	7,399	3,864	13,715	5,654
2	8	9,942	3,770	8,546	4,942	13,715	7,072
3	9	4,249	7,426	8,468	7,504	13,715	8,491
4	10	8,656	8,792	9,616	8,581	13,715	9,909
5	12	10,737	12,688	10,685	12,221	13,715	9,909
6	13	15,144	12,889	10,607	8,842	13,715	9,909
7	14	13,939	14,253	10,529	9,919	13,715	11,327
8	14	9,450	16,545	11,754	15,860	13,715	11,327
9	15	7,124	15,620	11,676	13,967	13,715	12,746
10	17	13,693	17,206	12,745	19,092	13,715	12,746
11	18	18,100	16,281	13,893	17,198	13,715	12,746
12	19	11,285	17,647	12,590	12,334	13,715	14,164
13	19	21,365	14,211	15,040	19,761	13,715	15,582
14	20	15,692	15,577	13,737	16,382	13,715	15,582
15	21	18,977	14,652	14,884	18,945	13,715	17,001
16	23	17,690	13,947	29,431	12,187	33,281	27,435

Les estimations des Y sont les suivantes :

OBS	Ya_est	Yb_est	Yc_est	Yd_est	Ye_est
1	6,18	6,18	6,18	6,18	11,61
2	6,99	6,99	6,99	6,99	11,61
3	7,80	7,80	7,80	7,80	11,61
4	8,61	8,61	8,61	8,61	11,61
5	10,22	10,23	10,22	10,22	11,61
6	11,03	11,03	11,03	11,03	11,61
7	11,84	11,84	11,84	11,84	11,61
8	11,84	11,84	11,84	11,84	11,61
9	12,65	12,65	12,65	12,65	11,61
10	14,27	14,27	14,27	14,27	11,61
11	15,07	15,08	15,08	15,08	11,61
12	15,88	15,89	15,89	15,89	11,61
13	15,88	15,89	15,89	15,89	11,61
14	16,69	16,69	16,69	16,69	11,61
15	17,50	17,50	17,50	17,50	11,61
16	19,12	19,12	19,12	19,12	27,43

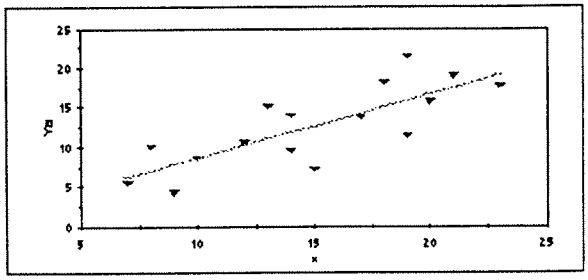
Les résultats des 5 régressions, voir page suivante, sont identiques, mêmes estimations, mêmes statistiques pour R^2=0.617, b_0=0.520, b_1=0.809, mêmes erreurs-types sur les coefficients, et pourtant les situations sont bien différentes.

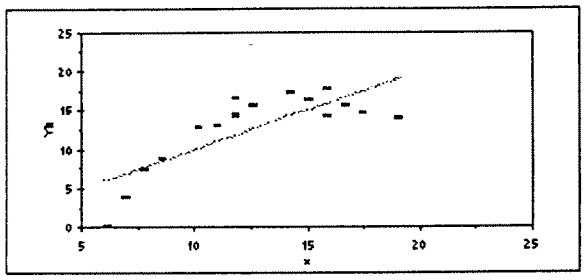
Analyses des résultats

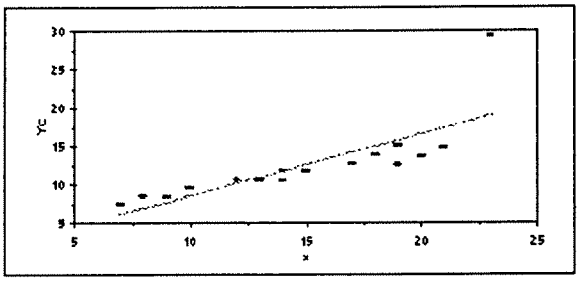
- Sur le 1er graphique on peut voir que le modèle semble bien **adapté**.
- Sur le 2ème graphique le modèle linéaire est **inadapté**, un modèle quadratique de la forme $Y = \beta X^2$ serait préférable. Dans ce cas la liaison entre Y et X^2 serait linéaire.
- Sur le 3ème graphique un point est **suspect** et entraîne la droite de régression vers le haut.
- Sur le 4ème graphique la variance des erreurs varie. Il y a un phénomène **d'hétéroscédasticité** (variance de Y sachant X non constante).
- Sur le 5ème graphique le **plan expérimental** défini par les valeurs de Xe est particulièrement mauvais.

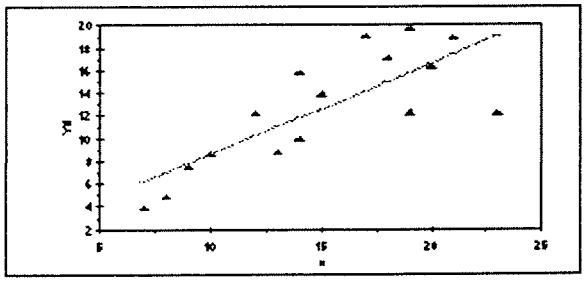
LES PIEGES DE LA REGRESSION

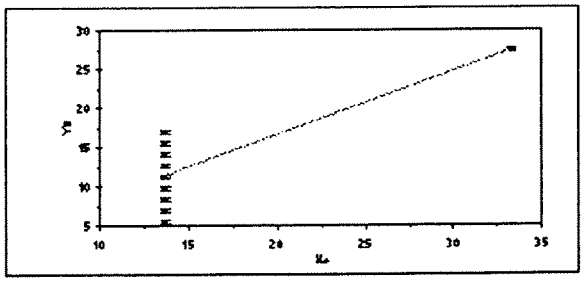
5 situations différentes mais 5 analyses de régression identiques











Résultats Y=b0 + b1X	
Nombre d'observations	16
Degrés de liberté	14
R carré	0,617
Moyenne de Y	12,600
Erreur std de l'est. de Y=RMSE	3,226
Moyenne de X	14,937
Constante b0	0,520
Coefficient b1	0,809
Erreur std du coef. b1	0,170

5.2. Exemple en régression multiple.

Cet exemple est emprunté à D.Ladiray (1990)[7].
On cherche à expliquer le taux d'urbanisation, variable URBA, en fonction de 10 variables régresseurs, POP87 à ESPER.

Tableau des données :

OBS	URBA	POP87	NAT	MORT	ACCR	DOUB	FERTI	MORTI	AGE15	AGE65	ESPER	ACCR_CAL
1	81	0.4	32	5	2.8	25	4.6	32.0	41	2	67	2.7
2	53	0.7	20	9	1.1	63	2.5	12.0	25	11	74	1.1
3	79	0.6	47	8	4.0	18	7.4	59.0	50	3	64	3.9
4	68	17.0	46	13	3.3	21	7.2	80.0	49	4	62	3.3
5	90	4.4	23	7	1.7	41	3.1	12.3	33	9	75	1.6
6	60	3.7	45	8	3.7	19	7.4	54.0	51	3	67	3.7
7	80	1.9	34	3	3.2	22	4.4	19.0	40	1	72	3.1
8	80	3.3	30	8	2.2	32	3.8	52.0	38	5	65	2.2
9	9	1.3	47	14	3.3	21	7.1	117.0	44	3	52	3.3
10	86	0.3	34	4	3.0	23	5.6	42.0	34	2	69	3.0
11	72	14.8	39	7	3.1	22	6.9	79.0	37	2	63	3.2
12	49	11.3	47	9	3.8	18	7.2	59.0	49	4	63	3.8
13	46	51.4	30	9	2.1	33	4.0	92.0	36	4	62	2.1
14	81	1.4	30	4	2.6	27	5.9	38.0	30	1	68	2.6
15	15	6.5	53	19	3.4	20	7.8	137.0	49	3	47	3.4
16	40	2.4	47	17	3.0	23	7.3	135.0	48	3	48	3.0
17	16	14.2	48	22	2.6	27	7.6	182.0	46	4	39	2.6
18	13	107.1	44	17	2.7	26	6.2	140.0	44	4	50	2.7
19	5	1.5	38	18	2.0	34	5.5	142.0	40	3	46	2.0
20	25	800.3	33	12	2.1	33	4.3	101.0	38	4	55	2.1
21	51	50.4	45	13	3.2	21	6.3	113.0	44	3	57	3.2
22	26	0.2	48	10	3.8	18	7.1	68.0	45	2	51	3.8
23	7	17.8	42	17	2.5	28	6.1	112.0	41	3	52	2.5
24	28	104.6	44	15	2.9	24	6.6	125.0	45	4	50	2.9
25	22	16.3	25	7	1.8	38	3.7	29.8	35	4	70	1.8
26	64	0.2	30	4	2.6	26	3.6	12.0	38	3	62	2.6
27	24	38.8	34	13	2.1	33	4.4	103.0	39	4	53	2.1
28	12	0.7	48	23	2.5	28	5.8	183.0	35	3	40	2.5
29	22	174.9	31	10	2.1	33	4.2	88.0	40	3	58	2.1
30	11	6.5	39	18	2.1	33	4.7	160.0	35	3	43	2.1
31	16	3.8	41	16	2.5	28	5.8	122.0	43	3	50	2.5
32	32	16.1	31	7	2.4	28	3.9	30.0	39	4	67	2.4
33	40	61.5	35	7	2.8	25	4.7	50.0	41	3	65	2.8
34	100	2.6	17	5	1.1	61	1.6	9.3	24	5	71	1.2
35	17	53.6	29	8	2.1	33	3.5	57.0	36	3	63	2.1
36	19	62.2	34	8	2.6	27	4.5	55.0	40	4	63	2.6
37	32	1062.0	21	8	1.3	53	2.4	61.0	28	5	66	1.3
38	92	5.6	14	5	0.9	77	1.6	7.5	24	7	75	0.9
39	76	122.2	12	6	0.6	124	1.8	5.5	22	10	77	0.6
40	64	21.4	30	5	2.5	28	4.0	33.0	39	4	65	2.5
41	65	42.1	20	6	1.4	51	2.1	30.0	31	4	67	1.4
42	97	0.4	23	6	1.7	41	3.7	12.0	34	8	68	1.7
43	51	2.0	37	11	2.6	26	5.1	53.0	42	3	62	2.6
44	67	19.6	17	5	1.2	59	1.8	8.9	30	5	73	1.2

[7] Ladiray D.(1990) *Autopsie d'un résultat: L'exemple des procédures Forecast, X11, Cluster.* Club SAS 1990

Dans la matrice X(44,10) des observations, 2 valeurs de la variable NAT (taux de natalité) pour OBS=11 et OBS=30 sont légèrement modifiées (39 est remplacé par 40)

Les régressions effectuées avant et après modifications donnent les résultats suivants :

Résultat 1 (valeur 39) Avant	**Résultat 2** (valeur 40) Après
URBA =	URBA=
25.541	20.689
-0.026 POP87	-0.026 POP87
-6.661 NAT	**-4.047 NAT**
+2.681 MORT	**-0.005 MORT**
+64.506 ACCR	**+39.832 ACCR**
+0.019 DOUB	+0.015 DOUB
+7.834 FERTI	+7.307 FERTI
+0.101 MORTI	+0.128 MORTI
-1.132 AGE15	-1.157 AGE15
+2.709 AGE65	+2.848 AGE65
+0.910 ESPER	+0.969 ESPER

Les résultats "Avant" et "Après" sont particulièrement instables pour les estimations des coefficients des 3 variables, NAT (taux de natalité) MORT (taux de mortalité) et ACCR (taux d'accroissement de la population).

Explication

Ces 3 variables **ne sont pas indépendantes**, elles sont liées entre elles par une relation quasi-linéaire qui est sensiblement ACCR= (NAT-MORT) / 10, voir sur la page précédente la variable la dernière colonne *ACCR_CAL*, calculée avec la formule exacte.

Lors de l'inversion de la matrice X'X, il y a une valeur propre qui est presque nulle.

Conséquence une légère perturbation des données entraîne de grands changements dans les estimations.

Conclusion

Des multicolinéarités rendent les résultats instables.

6. Conditions d'utilisation de la régression, les diagnostics, quelques remèdes

Les différents exemples présentés au paragraphe précédent ont montré l'importance des analyses et diagnostics effectués avant et après les premiers traitements.

Rappel : Pour élaborer la régression on a supposé, en pure théorie, que le modèle linéaire postulé est correct, puis on a dû faire des hypothèses ou suppositions, *a priori* sur les erreurs.

-les erreurs sont de moyenne nulle ce qui est vérifié par construction si la constante β_0 existe dans le modèle.
-les erreurs sont de variance constante
-les erreurs sont indépendantes
-les erreurs suivent une distribution normale

Ces **suppositions** sont nécessaires pour définir les tests.

Mais les erreurs sont inconnues et le resteront toujours. Elles peuvent cependant être approchées par l'**examen des résidus**.

Après examen des résidus, on peut conclure : les suppositions semblent ou ne semblent pas être violées.
Ce qui ne signifie pas que les suppositions soient correctes. Cela veut dire que sur la base des données que l'on a étudiées, on n'a aucune raison de dire que les suppositions sont fausses.

6.1. Modèle Inadapté

C'est par l'examen des résidus que l'on peut vérifier si le modèle postulé est vraisemblablement correct, ou s'il est **inadapté**.
Les résidus contiennent à la fois des erreurs de mesure et des erreurs de spécification du modèle, comme des variables omises, ou des liaisons non linéaires.
On peut avoir des tests tout à fait satisfaisants, des précisions sur les estimateurs des paramètres excellentes, alors que l'on a un modèle inadapté à l'étude.

En dehors de certaines visualisations il n'y a que le bon sens et la **pré-connaissance du problème** qui permettent de repérer l'inadéquation du modèle aux données.

6.2. *L'influence de certaines données, les données aberrantes*

Certaines données anormales peuvent fausser complètement les résultats. Les visualisations graphiques permettent parfois de les identifier. On peut alors être amené à retirer ces points anormaux des analyses.

Le livre de Belsley, Kuh et Welsh a popularisé une méthode rigoureuse de recherche des observations influentes.

L'option INFLUENCE de PROC SAS permet cette analyse.

6.3. *Multicolinéarité ou Corrélation entre les Régresseurs*

C'est le gros PROBLEME

C'est également le trio "BKW" qui a proposé des indicateurs de détection de multicolinéarités.

Les options TOL,VIF et COLLIN/COLLINOINT de Proc REG sont des aides aux diagnostics de multicolinéarité.

Tous ces compléments à la régression, représentations graphiques, indicateurs techniques de BKW etc, nécessitent de faire appel aux nombreuses options de Proc REG, qui seront présentées dans la partie "Compléments sur la régression avec PROC REG" par Josiane Confais.

1. Que feriez-vous à sa place?

A-berrantes ?

A-normales ?

A-typiques ?

2. Quand la variable à expliquer n'est pas forcément expliquée

Une fois réglées toutes les difficultés reste un dernier point et non des moindres, **l'interprétation.**

2.1. Etendue du vocabulaire et taille des pieds .

Prenons un exemple concret: Parmi des enfants, on effectue une enquête permettant de mesurer l'étendue du vocabulaire et la taille de leurs pieds[1].
La corrélation entre ces deux variables est nettement significative!

Le **bon sens** permet d'éviter d'en tirer des conclusions aberrantes. Sous cette corrélation se cache l'influence de la variable AGE.

2.2. Taux de criminalité et taux de fréquentation dans les églises.

Autre exemple : Dans un Etat des U.S.A on a corrélé sur les 20 dernières années, le taux de criminalité et le taux de fréquentation dans les églises.
Là aussi la corrélation obtenue est très élevée, mais le bon sens ne vient que peu en aide.

La variable **cachée** est l'immigration italienne et irlandaise.

A la lumière de ces deux exemples, gardons-nous de toutes interprétations hâtives.

Avoir toujours à l'esprit que sous une corrélation peut se cacher

l'effet d'une autre variable, ou d'un autre facteur.

Remarque: On peut cependant utiliser le modèle identifié, s'il est correct, dans un but de **prévision** mais surtout pas dans un but de **"control"** (action sur les variables explicatives dans l'espoir d'agir sur Y).

Sinon, on pourrait augmenter l'intelligence de nos enfants en augmentant la taille de leurs pieds!
Que penseraient alors les chinoises?

[1]Exemples empruntés à R. Astier et G. Oppenheim, Professeurs de Statistique- DEA (1985) de Statistiques et Modèles Université d'ORSAY.

2.3. *La causalité*

La liaison entre 2 variables X et Y peut se rencontrer dans 5 situations:

- X cause Y
- Y cause X
- X et Y inter-agissent l'une sur l'autre
- X et Y évoluent ensemble sous l'effet d'une même variable
- X et Y sont liées par hasard

Le **problème** de la causalité est une notion qui ne peut être validée par "l'outil" régression.
C'est d'ailleurs un débat historique qui est encore d'actualité.

La régression ne permet pas de déduire une quelconque relation de cause à effet de X sur Y et/ou de Y sur X. Il faut d'autres pratiques méthodologiques pour expliquer la causalité.

L'erreur qui perdure dans la littérature, est de donner le nom de variable dépendante ou variable **expliquée** à Y et de variables indépendantes ou variables **explicatives** à X, ce qui amène à déduire logiquement qu'il existe une idée de cause à effet entre X et Y, ce qui est totalement faux.

Corrélation n'est pas Causalité

3. Qualité de l'information apportée par les données

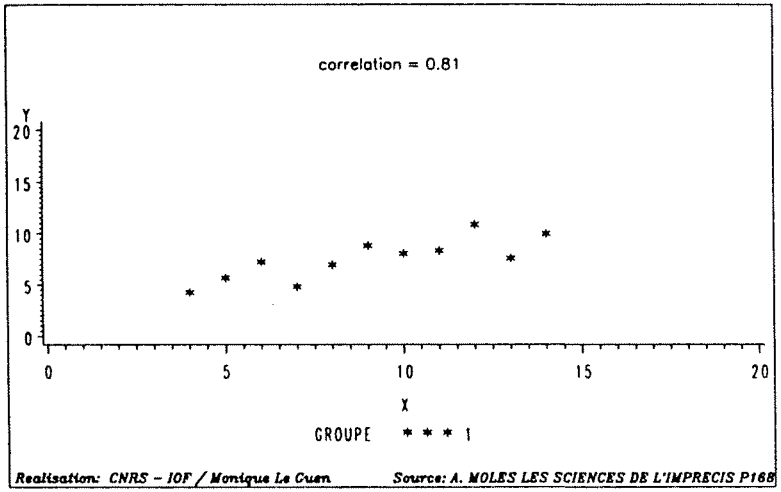
La qualité de l'information apportée par les données (observations) intervient dans la validité et la robustesse d'un modèle de régression. Mais cette qualité n'est pas appréhendable par l'analyse statistique.
Ce sont des connaissances *externes* à la statistique mais *internes* à l'étude qui doivent intervenir.
Ces connaissances sont aussi indispensables pour déterminer le plan d'échantillonnage.
Cette étape qui se situe au niveau de la collecte des données et donc en amont de l'analyse statistique des données mériterait à elle seule un long développement.

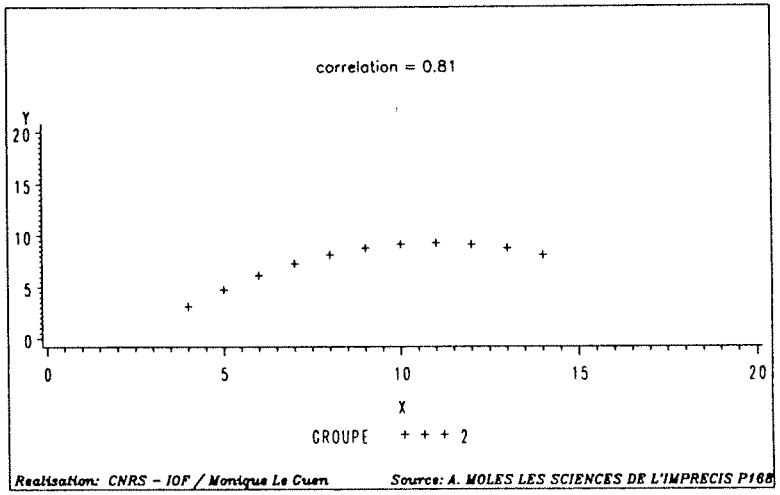
4. Comment les chiffres peuvent masquer la réalité

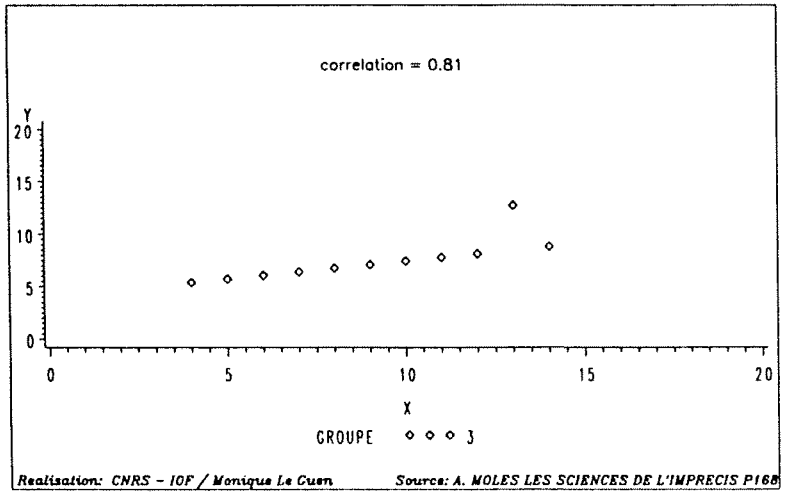
On montre sur les graphiques de la page suivante comment des situations totalement différentes peuvent cependant donner la même valeur à une statistique: la **corrélation**.
Ces graphiques révèlent l'importance de la visualisation des données comme pratique parallèle des calculs statistiques.

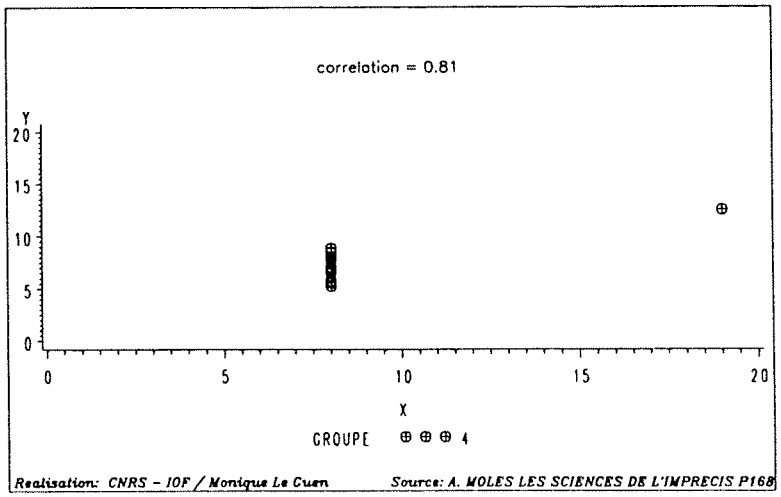
omment les chiffres peuvent masquer la realite

pour ces 4 graphes le coefficient de correlation est identique









5. Stratégies à développer

Pour vous aider à synthétiser l'ensemble des stratégies à développer dans le cadre d'une analyse de régression nous proposons le schéma de la page suivante: "*Méthodologie de la régression*".

Conclusion

Les méthodes de régression sont des méthodes très **puissantes**, mais qui doivent être utilisées avec beaucoup de **discernement** et de **précaution.**

En toute honnêteté il ne faut se contenter d'un seul modèle et d'une seule procédure REG, il faut en tester plusieurs.

C'est un travail d'explorateur.

Alors, pourquoi ne pas utiliser l'analyse exploratoire multidimensionnelle pour s'aider à rechercher un "bon" modèle?

Porte ouverte à l'analyse de données et ses visualisations.

Dans ce domaine, il existe d'autres techniques de régression dérivées de l'analyse des données comme la régression par Boule et par l'Analyse des Correspondances. Pour plus d'information nous vous renvoyons à l'auteur Pierre Cazes [2].

[2] CAZES Pierre (1976) Régression par Boule et par l'Analyse des correspondances, RSA Vol XXIV n°4, pp5-22.

Méthodologie de la REGRESSION