LA REGRESSION MULTIPLE

A QUOI CA SERT ?

Décrire

Prédire

Estimer une variable quantitative en fonction de plusieurs variables.

$\longrightarrow$ le mieux est de voir cela concrètement sur un exemple

EXEMPLE

Exprimer un rendement de blé en fonction de doses de fertilisants N, P, K.

Y	X_1	X_2	X_3
rdt	N	P	K
30	80	40	40
50	100	40	40
100	180	100	100
60	100	80	20
70	150	70	120

Ecrire le rendement en fonction de N, P, K :

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_0 + \varepsilon$$

$$\begin{aligned}
Y_1 &= 30 = 80\,\beta_1 + 40\,\beta_2 + 40\,\beta_3 + \beta_0 + \varepsilon_1 \\
Y_2 &= 50 = 100\,\beta_1 + 40\,\beta_2 + 40\,\beta_3 + \beta_0 + \varepsilon_2 \\
Y_3 &= 100 = 180\,\beta_1 + 100\,\beta_2 + 100\,\beta_3 + \beta_0 + \varepsilon_3 \\
Y_4 &= 60 = 100\,\beta_1 + 80\,\beta_2 + 20\,\beta_3 + \beta_0 + \varepsilon_4 \\
Y_5 &= 70 = 150\,\beta_1 + 70\,\beta_2 + 120\,\beta_3 + \beta_0 + \varepsilon_5
\end{aligned}$$

J'ai 5 parcelles sur lesquelles j'ai mesuré le rendement et pour lesquelles je connais la quantité d'engrais apportée.

Je peux écrire ce système de 5 équations d'une autre façon :

$$\begin{pmatrix} Y_1 \\ \\ \vdots \\ \\ Y_5 \end{pmatrix} = \begin{pmatrix} 1 & 80 & 40 & 40 \\ 1 & 100 & 40 & 40 \\ 1 & 180 & 100 & 100 \\ 1 & 100 & 80 & 20 \\ 1 & 150 & 70 & 120 \end{pmatrix} \times \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \\ \vdots \\ \\ \varepsilon_5 \end{pmatrix}$$

Au tableau,

- vérifier qu'en remultipliant, j'ai bien la même chose,
- dire que le système matriciel écrit est de la forme :

$$Y = X\theta + \varepsilon$$

- C'est bien un *modèle linéaire.*

Pour que l'équation

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_0 + \varepsilon$$

soit correcte, il faut :

Estimer les β

Minimiser l'erreur

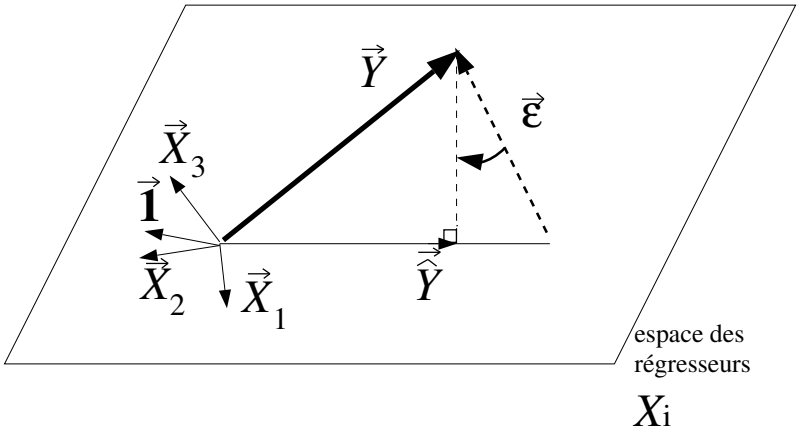
c'est-à-dire $\varepsilon_i = Y_i - \widehat{Y_i}$

$\Longrightarrow$ les moindres carrés
$\Longrightarrow \quad \Sigma \left(Y_i - \widehat{Y_i}\right)^2$
$\Longleftrightarrow \quad \Sigma \left(Y_i - \widehat{\beta_0} - \widehat{\beta_1} X_{1i} - \widehat{\beta_2} X_{2i} - \widehat{\beta_3} X_{3i}\right)^2$
à minimiser

$\sum \varepsilon_i^2$ à minimiser.

Minimiser cette équation, c'est compliqué.

$\longrightarrow$ regardons la géometrie.



$$\Longrightarrow \text{produit scalaire nul}$$

$$\Longrightarrow \overrightarrow{\varepsilon}.\ \overrightarrow{\mathbf{1}} = 0,\ \overrightarrow{\varepsilon}.\ \overrightarrow{X_1} = 0,\ \overrightarrow{\varepsilon}.\ \overrightarrow{X_2} = 0,\ \overrightarrow{\varepsilon}.\ \overrightarrow{X_3} = 0$$

$$X'Y = X'X\,\theta$$

$$\Longrightarrow\ \widehat{\theta} = (X'X)^{-1}\,X'Y$$

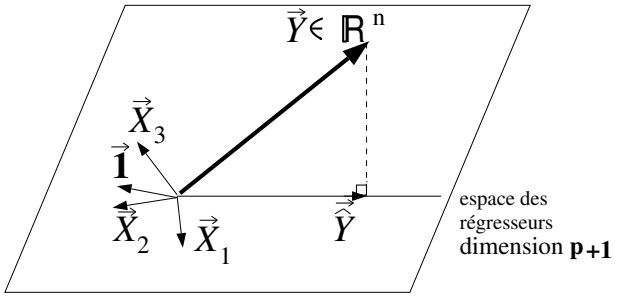
$$\Downarrow$$

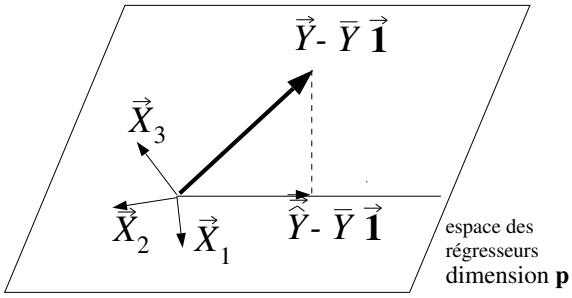
Estimation des paramètres
Tests et contrôles

Bien expliquer que l'on a remplacé l'espace des X_i par un plan pour les commodités du graphique mais c'est faux car les X_i sont des vecteurs libres (ils ne sont pas combinaisons linéaires les uns des autres). Je peux écrire les 4 produits scalaires différents

X	la matrice
X'	sa transposée
$(X'X)^{-1}$	la matrice inverse de $X'X$

au passage, on peut rappeler les définitions de matrice **inverse** et **transposée** et souligner que la multiplication des matrices n'est pas commutative.





On a un triangle rectangle $\Longrightarrow$ Théorème de Pythagore

$$\left\|\overrightarrow{Y} - \overline{Y}\ \overrightarrow{\mathbf{1}}\right\|^2 = \left\|\overrightarrow{\widehat{Y}} - \overline{Y}\ \overrightarrow{\mathbf{1}}\right\|^2 + \left\|\overrightarrow{\varepsilon}\right\|^2$$

$$\Longleftrightarrow \Sigma\left(Y_i - \overline{Y}\right)^2 = \Sigma\left(\widehat{Y}_i - \overline{Y}\right)^2 + \Sigma\left(Y_i - \widehat{Y}_i\right)^2$$

Origine de la variation	Somme de carrés	ddl	CM
régression	$\Sigma\left(\widehat{Y}_i - \overline{Y}\right)^2$	**p**	$\dfrac{\Sigma\left(\widehat{Y}_i - \overline{Y}\right)^2}{\mathbf{p-1}}$
résiduelle	$\Sigma\left(Y_i - \widehat{Y}_i\right)^2$	**n-p-1**	$\dfrac{\Sigma\left(Y_i - \widehat{Y}_i\right)^2}{\mathbf{n-p-1}}$
total	$\Sigma\left(Y_i - \overline{Y}\right)^2$	**n-1**	

Expliquer les manipulations pour enlever $\overrightarrow{\mathbf{1}}$ et avoir ainsi un espace des régresseurs de dimension p.

$$\widehat{\beta_0} = \overline{Y} - \widehat{\beta_1}\overline{X_1} - \widehat{\beta_2}\overline{X_2} - \widehat{\beta_3}\overline{X_3}$$

$$\Longrightarrow Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

$$\Longleftrightarrow Y_i - \overline{Y} = \beta_0 + \beta_1(X_{1i} - \overline{X_1}) + \beta_2(X_{2i} - \overline{X_2}) + \beta_3(X_{3i} - \overline{X_3}) + \varepsilon_i$$

Test et contrôles

1. Test de Fisher de l'analyse de variance
   teste l'hypothèse : $\beta_1 = \beta_2 = \beta_3 = 0$
   avec un risque de première espèce choisi

2. Coefficient de détermination

$$R^2 = \frac{\Sigma\left(\widehat{Y_i} - \overline{Y}\right)^2}{\Sigma\left(Y_i - \overline{Y}\right)^2} = \frac{\text{SC regression}}{\text{SC totale}}$$

3. Ecart-types des estimations des paramètres
   $\Longrightarrow$ Intervalles de confiances

Expliquer le test F

$$F_{cal} = \frac{\text{CM régression}}{\text{CM résiduel}}$$

Si $F_{cal} > F_{table}$, H_0 est rejetée
Seuil choisi : en général 5%

Propriétés des estimateurs

1. $\mathrm{E}\left(\widehat{\beta_i}\right) = \beta_i$

2. Connaissant la variance de chaque estimateur
   $\Longrightarrow$ Intervalle de confiance

3. Si deux régresseurs ne sont pas orthogonaux (dans la matrice X), on aura une covariance entre les estimateurs des coefficients de régression
   $\Longrightarrow$ ne pas interpréter de manière séparée les coefficients de régression

Pas de commentaire

Précautions

- Regarder les graphiques des résidus $\widehat{\varepsilon_i}$ en fonction des $\widehat{Y_i}$

- Les régresseurs sont des variables fixées et connues sans erreur

  $\Longrightarrow$ Attention à ne pas avoir des erreurs trop grandes

Dans l'exemple, cela suppose de connaître assez précisément les doses d'engrais.

Rendement de blé

CORRELATION ANALYSIS

4 'VAR' Variables: RDT N P K

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
RDT	5	62	25.8844	310.0	30.0000	100.0
N	5	122	41.4729	610.0	80.0000	180.0
P	5	66	26.0768	330.0	40.0000	100.0
K	5	64	43.3590	320.0	20.0000	120.0

Pearson Correlation Coefficients / N = 5

	N	P	K
N	1.00000	0.79521	0.85640
P	0.79521	1.00000	0.45991
K	0.85640	0.45991	1.00000

Vérifier le tableau des corrélations
Existent-ils des X_i non orthogonaux ?

OBS	RDT	RESIDU	PREDIT
1	30	-2.58333	32.583
2	50	1.41667	48.583
3	100	-1.58333	101.583
4	60	1.25000	58.750
5	70	1.50000	68.500

$Y_i \quad \widehat{\varepsilon_i} = Y_i - \widehat{Y_i} \quad \widehat{Y_i}$

Commenter le risque d'avoir P et N (c'est–à-dire X_1 et X_2) trop corrélés.

Rendement de blé

Model: MODEL1
Dependent Variable: RDT

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	3	2665.00000	888.33333	59.222	0.0952
Error	1	15.00000	15.00000		
C Total	4	2680.00000			

Root MSE	3.87298	R-square	0.9944
Dep Mean	62.00000	Adj R-sq	0.9776
C.V.	6.24675		

Parameter Estimates

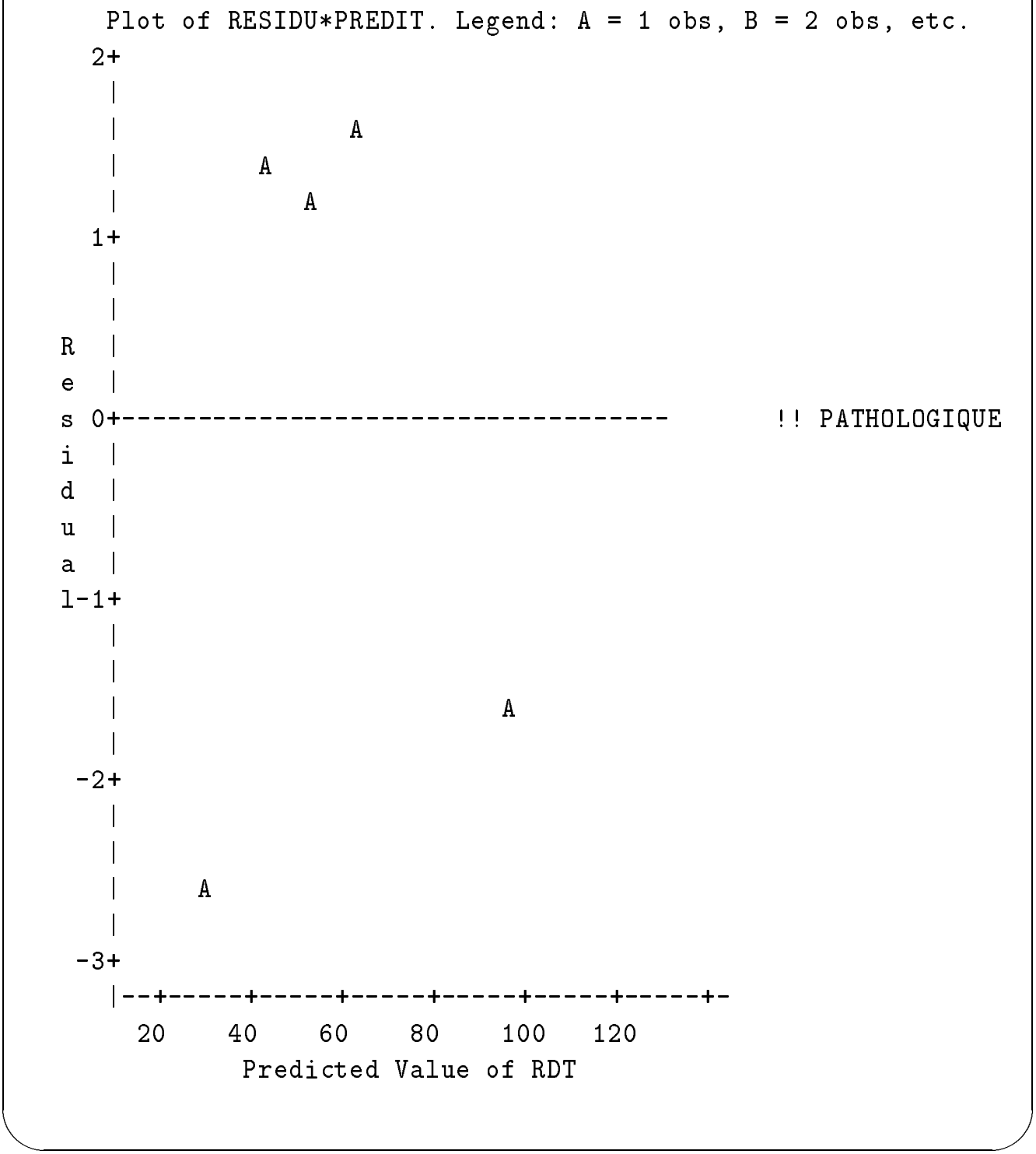
Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob>|T|
INTERCEP	1	-24.083333	7.74551842	-3.109	0.1981
N	1	0.800000	0.18708287	4.276	0.1462
P	1	0.108333	0.17300450	0.626	0.6438
K	1	-0.291667	0.12219065	-2.387	0.2526

$\overrightarrow{\widehat{\beta}}$ Ecart-types

Risque de 1ère espèce trop élevé : rejeter H_0 seulement si on accepte de se tromper trop souvent.

Pas de commentaire.

Rendement de blé



Même s'il n'y a que peu de points.

ETUDIER L'ADAPTATION D'UNE VARIETE DE MOUTARDE A LA SECHERESSE

Notations 34 jours après repiquage (sans irrigation)

- **1 variable expliquée** (dependent variable) :
  Nombre de racines courtes tubérisées ..... RC
  ($\Leftrightarrow$ adaptation de la plante au stress)
- **5 variables explicatives** :
  - longueur de la tige ..... LT
  - potentiel hydrique foliaire ..... HF
  - poids matière sèche des racines ..... PR
  - poids matière sèche des parties aériennes PA
  - nombre de feuilles ..... FE

Question :

Dans quelle mesure les variables explicatives, permettent-elles de connaître (prédire, estimer) la variable expliquée ?

Présentation d'un exemple agronomique d'utilisation de la régression linéaire multiple

31 pieds de moutarde sont soumis un stress hydrique de 34 jours après repiquage.

On veut étudier le nombre de racines courtes tubérisées (qui est un mécanisme d'adaptation à la sécheresse de la moutarde) en fonction de 5 variables mesurées.

TABLEAU DES DONNEES

OBS	RC	LT	HF	PR	PA	FE
1	0.00	29	65	87	43	2
2	0.00	35	65	163	122	2
3	1.10	40	65	175	117	3
4	0.69	25	60	38	49	2
5	0.00	30	30	57	23	1
6	0.00	45	70	270	124	5
7	0.69	40	65	202	78	4
8	1.39	50	70	226	74	3
9	1.61	50	85	525	222	5
10	1.10	55	80	230	92	3
11	3.47	60	155	1109	897	4
12	2.40	80	95	869	628	5
13	1.10	60	60	553	189	8
14	3.00	90	100	903	3022	6
15	3.43	80	145	1216	3049	6
16	1.61	75	85	912	3273	6
17	2.83	60	75	689	443	6
18	1.61	85	85	443	251	5
19	2.20	65	80	643	424	5
20	4.09	60	240	1089	843	6
21	3.09	60	80	825	757	7
22	1.39	70	80	385	1350	5
23	4.22	90	180	1335	728	7
24	4.17	90	175	953	668	3
25	4.57	95	205	1145	696	6
26	3.43	75	305	1129	678	6
27	3.04	70	120	978	529	6
28	3.26	75	70	795	329	6
29	5.40	70	300	1618	1075	7
30	4.16	70	250	1020	881	7
31	3.91	60	280	1020	624	8

Présentation du tableau de données et des statistiques simples sur les 6 variables.

Cela permet de repérer d'éventuelles observations aberrantes (en regardant notamment les minima et les maxima).

Le nombre de racines tubérisées est exprimé en log (pour homogeneiser la variance des observations).

Simple Statistics

Variable	N	Mean	Std Dev	Sum
RC	31	2.35355	1.54563	72.96000
LT	31	62.54839	19.33018	1939
HF	31	123.22581	78.93614	3820
PR	31	696.83871	431.61843	21602
PA	31	718.64516	869.07927	22278
FE	31	5.00000	1.86190	155.00000

Simple Statistics

Variable	Minimum	Maximum	Label
RC	0	5.40000	nbre tuberisations
LT	25.00000	95.00000	longueur tige
HF	30.00000	305.00000	potentiel hydrique
PR	38.00000	1618	pds racines
PA	23.00000	3273	pds part.ariennes
FE	1.00000	8.00000	nbre de feuilles

On voit que les variables PA et PR présentent une gamme très étendue de valeurs observées.

TABLEAU DES CORRELATIONS ENTRE VARIABLES

	RC	LT	HF	PR	PA	FE
RC nbre tuberisations	1.00000	0.72800	0.80342	0.93346	0.35176	0.65189
LT longueur tige	0.72800	1.00000	0.44948	0.75603	0.53304	0.61405
HF potentiel hydrique	0.80342	0.44948	1.00000	0.77919	0.20174	0.52051
PR pds racines	0.93346	0.75603	0.77919	1.00000	0.51263	0.73002
PA pds part.ariennes	0.35176	0.53304	0.20174	0.51263	1.00000	0.38258
FE nbre de feuilles	0.65189	0.61405	0.52051	0.73002	0.38258	1.00000

Avant tout calcul de régression, il est utile d'établir les corrélations linéaires entre variables (surtout entre régresseurs). En effet, si 2 régresseurs ont un coefficient de corrélation très proche de 1, on pourra éliminer l'un des deux dans l'analyse. Sinon cette colinéarité risque de perturber les calculs (probleme d'inversibilité de la matrice des X).

On voit que la variable PR est très liée aux variables LT et HF. Cela implique que si l'on étudie un modèle de régression de RC sur PR, les variables LT et HF seront probablement peu utiles pour expliquer mieux la variation de RC.

Il est important de souligner que les coefficients des variables corrélées ne pourront etre interprétés séparément

MODELE GENERAL

$$\mathrm{RC} = \beta_0 + \beta_1 \mathrm{LT} + \beta_2 \mathrm{HF} + \beta_3 \mathrm{PR} + \beta_4 \mathrm{PA} + \beta_5 \mathrm{FE} + \varepsilon$$

Par le test F, on teste :

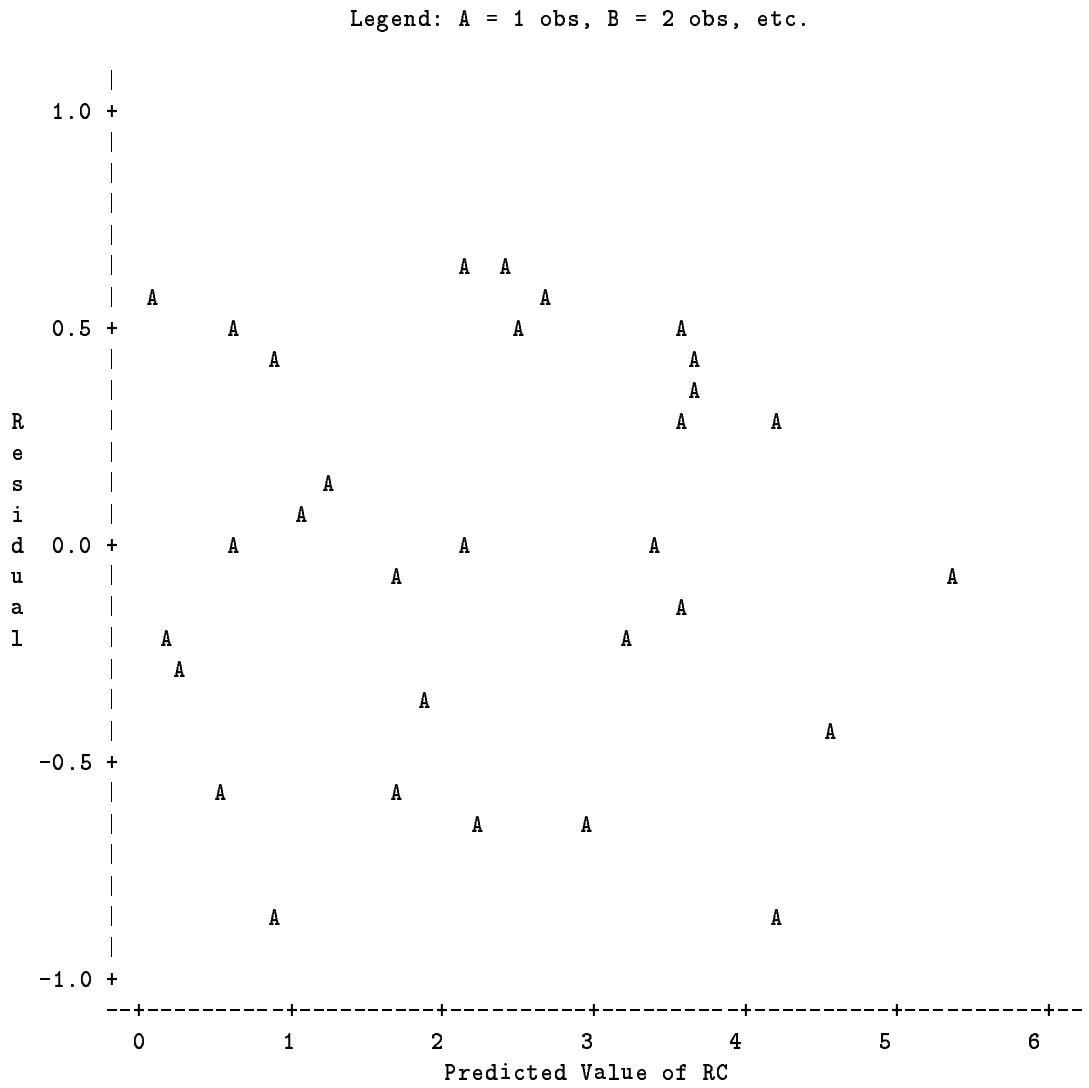
$$H_0 \ : \ \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$

$$\Longleftrightarrow \quad \textit{modèle } M_0 \ : \ \mathrm{RC} = \beta_0 + \varepsilon$$

Si on garde toutes les variables explicatives comme régresseurs, on pose le modèle le plus général.

Par le test F, on pourra tester l'hypothèse nulle selon laquelle il n'existe pas de relations entre la variable expliquée et les régresseurs (dans ce cas, sous H_0, tous les coefficients de la regression sont nuls, on se ramène au modèle M_0).

Plot of RESIDU*PREDIT

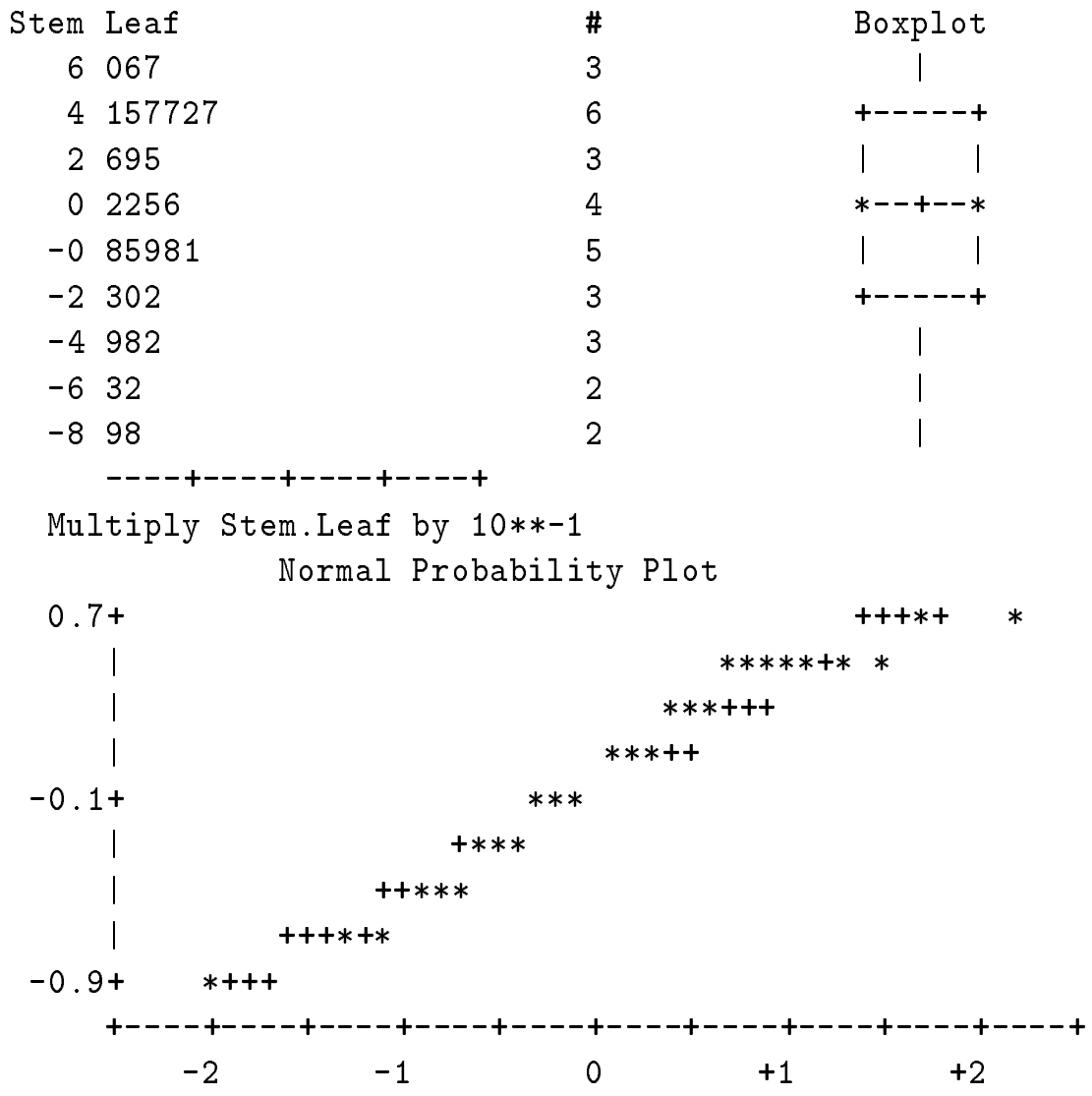


Après avoir fait tourner le programme et avant d'analyser les résultats, il faut jeter un coup d'oeil sur les résidus.

On trace ici le graphique des résidus du modèle M en fonction de RC (variable expliquée).

On ne peut repérer aucune configuration particulière des résidus.

Graphes des résidus



D'autre part, un stem and leaf (histogramme), un box plot et une droite de Henry (qqplot) des résidus semblent suggèrer que ces résidus présentent globalement une répartition Normale.

L'analyse des résultats est donc possible et fiable.

REGRESSION LINEAIRE MULTIPLE EXEMPLE : MOUTARDE ET SECHERESSE

Model: MODEL1
Dependent Variable: RC nbre tuberisations

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	5	65.21663	13.04333	50.539	0.0001
Error	25	6.45208	0.25808		
C Total	30	71.66871			

Root MSE	0.50802	R-square	0.9100
Dep Mean	2.35355	Adj R-sq	0.8920
C.V.	21.58524		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	-0.428901	0.40886893	-1.049	0.3042
LT	1	0.013582	0.00794030	1.711	0.0995
HF	1	0.003234	0.00210214	1.539	0.1365
PR	1	0.002889	0.00058155	4.967	0.0001
PA	1	-0.000285	0.00013538	-2.107	0.0453
FE	1	-0.054721	0.07386775	-0.741	0.4657

$$F = \frac{\text{carré moyen } \mathbf{Régression}}{\text{carré moyen } \mathbf{Résiduel}}$$

Nous avons 0.01% de chances d'être sous H_0 c'est-à-dire d'observer ces résultats expérimentaux s'il n'existe pas de liaison entre la variable expliquée et ces 5 régresseurs.

Ici, la liaison entre RC et LT, HF, PR, PA, FE est très hautement significative (on rejette H_0).

$R^2 = 0.91$, la proportion de variabilité expliquée par les régresseurs est de 91%.

Les tests de nullité de chaque coefficient montrent que seuls PR et PA sont significatifs à p=95% (c'est-à-dire que l'on peut rejeter l'hypothèse de nullité de leur coefficient de régression).

On écrira au tableau le modèle M avec ses valeurs :
RC $= -0.43+0.013$LT$+0.003$HF$+0.002$PR-0.0002PA-0.054FE

```
dm'log;clear;output;clear';
options ps=70;
TITLE 'REGRESSION LINEAIRE MULTIPLE; EXEMPLE : MOUTARDE ET SECHERESSE';
DATA RACINES;
input RC LT HF PR PA FE;
LABEL RC='nbre tuberisations'
      LT='longueur tige'
      HF='potentiel hydrique'
      PR='pds racines'
      PA='pds part.ariennes'
      FE='nbre de feuilles'
;
cards;
0.00 29 65  87   43   2
0.00 35 65  163  122  2
1.10 40 65  175  117  3
0.69 25 60  38   49   2
0.00 30 30  57   23   1
0.00 45 70  270  124  5
0.69 40 65  202  78   4
1.39 50 70  226  74   3
1.61 50 85  525  222  5
1.10 55 80  230  92   3
3.47 60 155 1109 897  4
2.40 80 95  869  628  5
1.10 60 60  553  189  8
3.00 90 100 903  3022 6
3.43 80 145 1216 3049 6
1.61 75 85  912  3273 6
2.83 60 75  689  443  6
1.61 85 85  443  251  5
2.20 65 80  643  424  5
4.09 60 240 1089 843  6
3.09 60 80  825  757  7
1.39 70 80  385  1350 5
```

Selon la matériel disponible, on fera tourner le programme sur SAS (avec un Data Show) ou on pourra montrer aux stagiaires les principales procédures utilisees : proc reg, proc univariate, proc plot, proc means.

```
4.22 90 180 1335 728  7
4.17 90 175 953  668  3
4.57 95 205 1145 696  6
3.43 75 305 1129 678  6
3.04 70 120 978  529  6
3.26 75 70  795  329  6
5.40 70 300 1618 1075 7
4.16 70 250 1020 881  7
3.91 60 280 1020 624  8
;
proc reg corr;
model RC=LT HF PR PA FE;
output out=resid r=residu p=predit;
run;
title'';
proc univariate normal plot data=resid;
var residu;
run;
proc plot;
plot residu*predit;
run;
```