RÉGRESSION LINÉAIRE

Plan du Chapitre 5

1. La régression linéaire
2. Colinéarité
3. Sélection de régresseurs
4. Ordonnée à l'origine nulle
5. Une heuristique
6. Conclusion

La régression linéaire

Quand et pourquoi l'utilise-t-on ?

$\longrightarrow$ pour expliquer/décrire

$\longrightarrow$ pour prédire

une variable quantitative en fonction de plusieurs variables **quantitatives**

Ecriture du modèle

$$\mathrm{Y_n} = \theta_0 + \left(\sum_{\mathrm{p}=1}^{\mathrm{P}-1} \theta_\mathrm{p} \quad \mathrm{Z}\underset{\downarrow}{\mathrm{p_n}} \right) + \varepsilon_\mathrm{n}$$

$$\mathrm{X}\mathrm{p_n}$$

Écriture du modèle linéaire dans le cas de la régression

$$Y_n = \theta_0 + \left(\sum_{p=1}^{P-1} Zp_n \ . \theta_p \right) + \varepsilon_n$$

$\theta_0 + \left(\sum_{p=1}^{P-1} Zp_n \ .\theta_p \right) + \varepsilon_n$ = modèle de l'espérance + erreur

Y_n = variable à expliquer

avec N = nombre d'observations

P-1 = nombre total de régresseurs

P = nombre de paramètres

en écriture matricielle

$$Y_n = X.\theta + \varepsilon$$

Étude du rendement de blé (Y) en fonction des doses de fertilisants AZ, PH et PO (Z1, Z2, Z3)

RENDEMENT	AZ	PH	PO
30	80	40	40
50	100	40	40
100	180	100	100
60	100	80	20
70	150	70	120

On peut écrire le rendement en fonction de AZ, PH et PO

$$\mathrm{Y_n} = \theta_0 + \mathrm{Z1_n} \ .\theta_1 + \mathrm{Z2_n} \ .\theta_2 + \mathrm{Z3_n} \ .\theta_3 + \varepsilon_\mathrm{n}$$

On peut écrire ce système sous forme de 5 équations

$$\begin{array}{lcccccccccc}
\mathrm{Y}_1 = & 30 & = & \theta_0+ & 80 & \theta_1+ & 40 & \theta_2+ & 40 & \theta_3+ & \varepsilon_1 \\
\mathrm{Y}_2 = & 50 & = & \theta_0+ & 100 & \theta_1+ & 40 & \theta_2+ & 40 & \theta_3+ & \varepsilon_2 \\
\mathrm{Y}_3 = & 100 & = & \theta_0+ & 180 & \theta_1+ & 100 & \theta_2+ & 100 & \theta_3+ & \varepsilon_3 \\
\mathrm{Y}_4 = & 60 & = & \theta_0+ & 100 & \theta_1+ & 80 & \theta_2+ & 20 & \theta_3+ & \varepsilon_4 \\
\mathrm{Y}_5 = & 70 & = & \theta_0+ & 150 & \theta_1+ & 70 & \theta_2+ & 120 & \theta_3+ & \varepsilon_5
\end{array}$$

Soit $\mathrm{Y} = \mathrm{X}.\theta + \varepsilon$

On retrouve bien l'écriture générale d'un modèle linéaire

Test de qualité d'un modèle

$M_1 \ : \ Y_n = \theta_0 + \left(\sum_{p=1}^{P-1} Zp_n \ .\theta_p \right) + \varepsilon_n$	P	paramètres
$\downarrow \quad H_0 \ : \ \theta_1 = \cdots = \theta_p = 0$ $\quad H_1 \ : \ \exists p; \ \theta_p \neq 0$		
$M_0 \ : \ Y_n = \theta_0 + \varepsilon_n$	1	paramètre

On teste la qualité du modèle complet M_1 par rapport au modèle le plus simple qui puisse exister M_0 par un **test intrinsèque F**

On utilise le rapport :

$$\frac{\frac{SCE_{M_0} - SCE_{M_1}}{P-1}}{\frac{SCE_{M_1}}{N-P}} \text{ qui suit une loi } F(P-1, N-P)$$

Estimation des paramètres

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

Commandes SAS

exemple : rendement de blé

```
data rendem;
input RDT AZ PH PO;
cards;
  30    80    40    40
  50   100    40    40
 100   180   100   100
  60   100    80    20
  70   150    70   120
;
proc reg ;
model RDT= AZ PH PO / ADJRSQ;
run;
```

Sortie SAS

Model : MODEL1 = M_1 Dependent variable : RDT

Analysis of variance

Source	DF	Sum of Squares	Mean Square	F value	Prob>F
Model	3	2665.000	888.333	59.22	0.0952
Error	1	15.000	15.000		
C Total	4	2680.000			
Root MSE		3.87298	R-square	0.9944	
Dep MEAN		62.00000	Adj R-sq	0.9776	

Source	DF	Sum of Squares	Mean Square	F value
Model	P-1	$SCE_{M_0} - SCE_{M_1}$	$\frac{SCE_{M_0} - SCE_{M_1}}{P-1}$	$\frac{SCE_{M_0} - SCE_{M_1} / \ P-1}{SCE_{M_1} / \ N-P}$
Error	N-P	SCE_{M_1}	$\frac{SCE_{M_1}}{N-P}$	
C Total	N-1	SCE_{M_0}		

$$R^2 = \frac{SCE_{M_0} - SCE_{M_1}}{SCE_{M_0}}$$

Root MSE = $\sqrt{\frac{SCE_{M_1}}{N-P}} = \sqrt{\hat{\sigma}^2}$, estimateur de l'écart type des ε

Adj R-sq = $1 - (1 - R^2)\frac{(N-1)}{(N-P)}$

Étudie de l'adaptation d'une variété de moutarde à la sécheresse

Mesures 34 jours après repiquage (sans irrigation).

— 1 variable expliquée :

Nombre de racines courtes tubérisées	RC
(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 ?

Les données

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

Écriture du modèle

$$M_1 \ : \ \mu = \theta_0 + \theta_1 \mathrm{LT} + \theta_2 \mathrm{HF} + \theta_3 \mathrm{PR} + \theta_4 \mathrm{PA} + \theta_5 \mathrm{FE} \qquad \text{P=6}$$

$$\downarrow \quad H_0 \ : \ \theta_1 = \theta_2 = \theta_3 = \theta_4 = \theta_5 = 0$$

$$\downarrow \quad H_1 \ : \ \exists p; \theta_p \neq 0$$

$$M_0 \ : \ \mu = \theta_0 \qquad \text{P=1}$$

Sortie SAS

Régression linéaire multiple

Exemple : Moutarde et sécheresse

Model : MODEL1 = M_1 Dependent variable : RC Nbre de tubérisations

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			

Sortie SAS : estimation des paramètres

Model : MODEL1
dependent variable RC Nbre tubérisations

Parameter Estimates

1	2	3	4	5	6
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

Effet de chacun des régresseurs (par ex : FE)

$M_1 : \mu = \theta_0 + \theta_1 LT + \theta_2 HF + \theta_3 PR + \theta_4 PA + \theta_5 FE$ P=5

$H_0 : \theta_5 = 0$

$\downarrow$

$H_1 : \theta_5 \neq 0$

$M_{0_5} : \mu = \theta_0 + \theta_1 LT + \theta_2 HF + \theta_3 PR + \theta_4 PA$ P=4

$$F : \frac{\left(SCE_{M_{0_5}} - SCE_{M_1}\right)/1}{SCE_{M_1}/N-P} = T^2$$

Intervalle de confiance des θ_{P}

$$\widehat{\theta}_{\mathrm{p}} \sim \mathcal{N}\left(\theta_{\mathrm{p}}, \sigma^2 \ (\mathrm{X}'\mathrm{X})^{-1}_{\mathrm{pp}}\right)$$

On montre que :

$$\frac{\widehat{\theta}_{\mathrm{p}} - \theta_{\mathrm{p}}}{\left\{\sqrt{\widehat{\mathrm{Var}}\left(\widehat{\theta}_{\mathrm{p}}\right)}\right.} \sim \mathrm{t}(\mathrm{N}-\mathrm{P})$$

Parameter Estimates
Standard Error
(colonne 4) de la sortie SAS

d'où l'intervalle de confiance où seuil de α

$$\widehat{\theta}_{\mathrm{p}} - \mathrm{t}_{1-\alpha/2,\mathrm{N}-\mathrm{P}} \times \sqrt{\widehat{\mathrm{Var}}\left(\widehat{\theta}_{\mathrm{p}}\right)} < \theta_{\mathrm{p}} < \widehat{\theta}_{\mathrm{p}} + \mathrm{t}_{1-\alpha/2,\mathrm{N}-\mathrm{P}} \times \sqrt{\widehat{\mathrm{Var}}\left(\widehat{\theta}_{\mathrm{p}}\right)}$$

Exemple :

Paramètre de FE : $-0.202 < \theta_{\mathrm{FE}} < 0.96$ Intervalle de confiance à 95%

Paramètre de PR : $0.0017 < \theta_{\mathrm{PR}} < 0.0041$ Intervalle de confiance à 95%

Région de confiance simultanée des θ cas de 2 régresseurs

Notation

$$v_1^2 = \sum_{\mathrm{n}} (\mathrm{Z}_{1\mathrm{n}} - \mathrm{Z}_{1.})^2$$

$$v_2^2 = \sum_{\mathrm{n}} (\mathrm{Z}_{2\mathrm{n}} - \mathrm{Z}_{2.})^2$$

$$\rho = \sum_{\mathrm{n}} \frac{(\mathrm{Z}_{1\mathrm{n}} - \mathrm{Z}_{1.})(\mathrm{Z}_{2\mathrm{n}} - \mathrm{Z}_{2.})}{\sqrt{v_1^2} \qquad \sqrt{v_2^2}}$$

alors

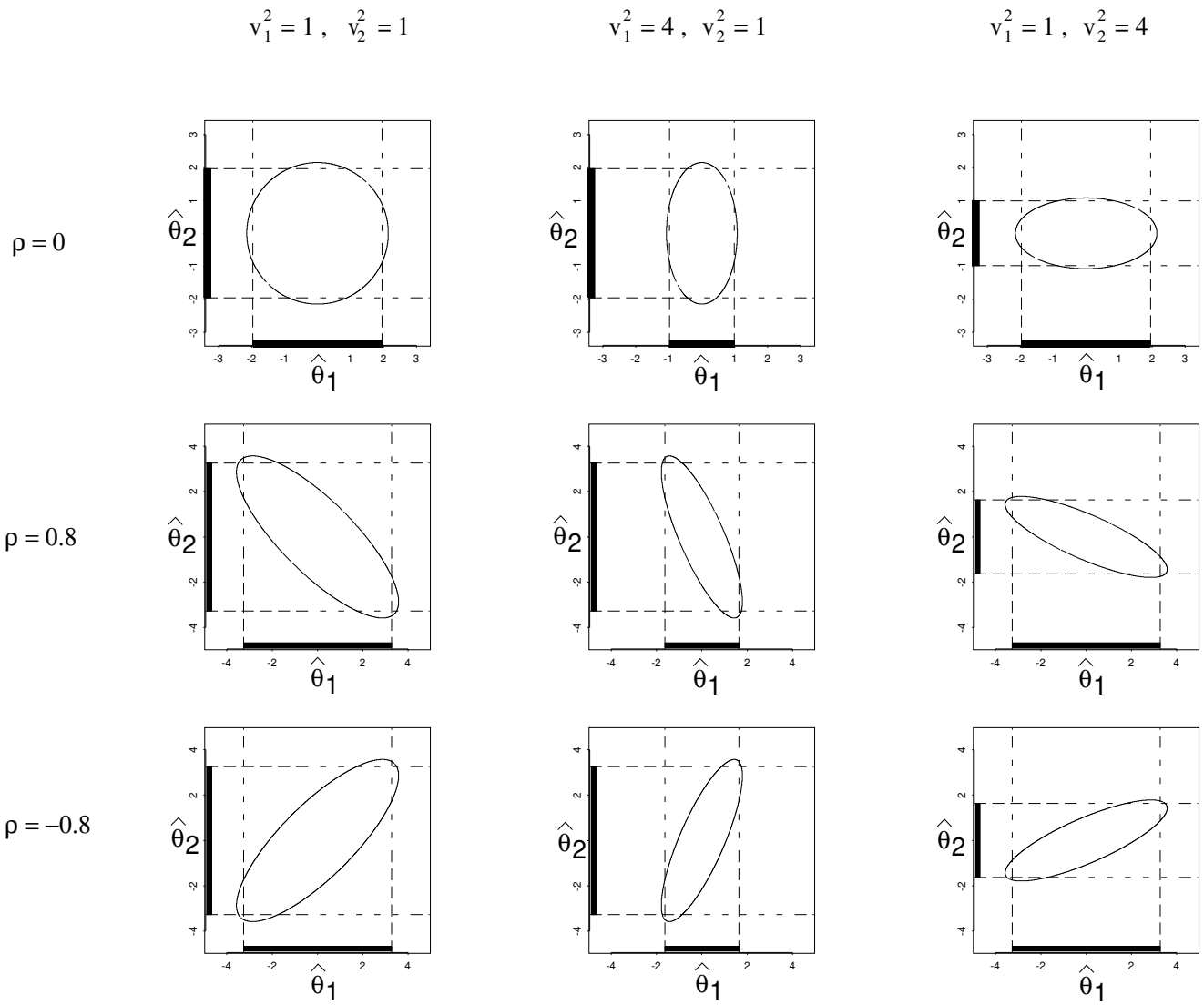
$$\begin{pmatrix} \widehat{\theta}_1 \\ \widehat{\theta}_2 \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, \quad \sigma^2 \quad \begin{pmatrix} \frac{1}{v_1^2(1-\rho^2)} & \frac{-\rho}{v_1 v_2(1-\rho^2)} \\ \frac{-\rho}{v_1 v_2(1-\rho^2)} & \frac{1}{v_2^2(1-\rho^2)} \end{pmatrix} \right]$$

La région de confiance simultanée (lorsque N-P est grand) de (θ_1, θ_2) est la région telle que

$$\frac{1}{\widehat{\sigma^2}} \left[v_1^2 \left(\widehat{\theta}_1 - \theta_1 \right)^2 + v_2^2 \left(\widehat{\theta}_2 - \theta_2 \right)^2 + 2\rho v_1 v_2 \left(\widehat{\theta}_1 - \theta_1 \right) \left(\widehat{\theta}_2 - \theta_2 \right) \right]$$

$$\leq \chi_\alpha^2(2)$$

où $\leq \chi_\alpha^2(2)$ est la quantile d'un χ^2 à 2 degrés de liberté


$v_1^2 = 1$, $v_2^2 = 1$
$v_1^2 = 4$, $v_2^2 = 1$
$v_1^2 = 1$, $v_2^2 = 4$
$\rho = 0$
$\rho = 0.8$
$\rho = -0.8$
$\hat{\theta}_1$
$\hat{\theta}_2$

Colinéarité vraie

Exemple :

Niveau Scolaire

$\mu_n = \theta_0 + \theta_1 Z_{1n} + \theta_2 Z_{2n} + \theta_3 Z_{3n} + \theta_4 Z_{4n}$

μ_n = note moyenne espérée au lycée

Z_{1n} = note moyenne au collège

Z_{2n} = note moyenne en math au lycée

Z_{3n} = note moyenne en philo au lycée

Z_{4n} = notes moyennes au lycée ($Z_{2n} + Z_{3n}$)

Colinéarité vraie = modèle réductible
$\Rightarrow (X'X)$ n'est plus inversible

Remède : Rendre le modèle irréductible (par exemple en supprimant Z_{4n})

Conséquences de la quasi colinéarité

— tend à augmenter la variance des $\widehat{Y}$

en particulier pour les jeux de données n'appartenant pas à l'échantillon

— tend à augmenter la variance des $\widehat{\theta}$

$\Rightarrow$ diminue la puissance des tests sur les paramètres

$\Rightarrow$ risque d'estimation incorrecte des paramètres (ordre de grandeur ou même signe incorrect)

$\Rightarrow$ risque de mauvaise interprétation des paramètres

Comment détecter la quasi colinéarité ?

Regarder la matrice de corrélation des régresseurs

$\Rightarrow$ ne révèle que les corrélations 2 à 2

Utiliser le VIF (Variance Inflation Factor) :

$$\mathrm{M}_1 \ : \ \mu_n = \theta_0 + \sum_{p=1}^{P-1} \mathrm{Z}p_n \ .\theta_p$$

$$\downarrow \quad \begin{array}{l} \mathrm{H}_0 \ : \ \theta_j = 0 \\ \mathrm{H}_1 \ : \ \theta_j \neq 0 \end{array}$$

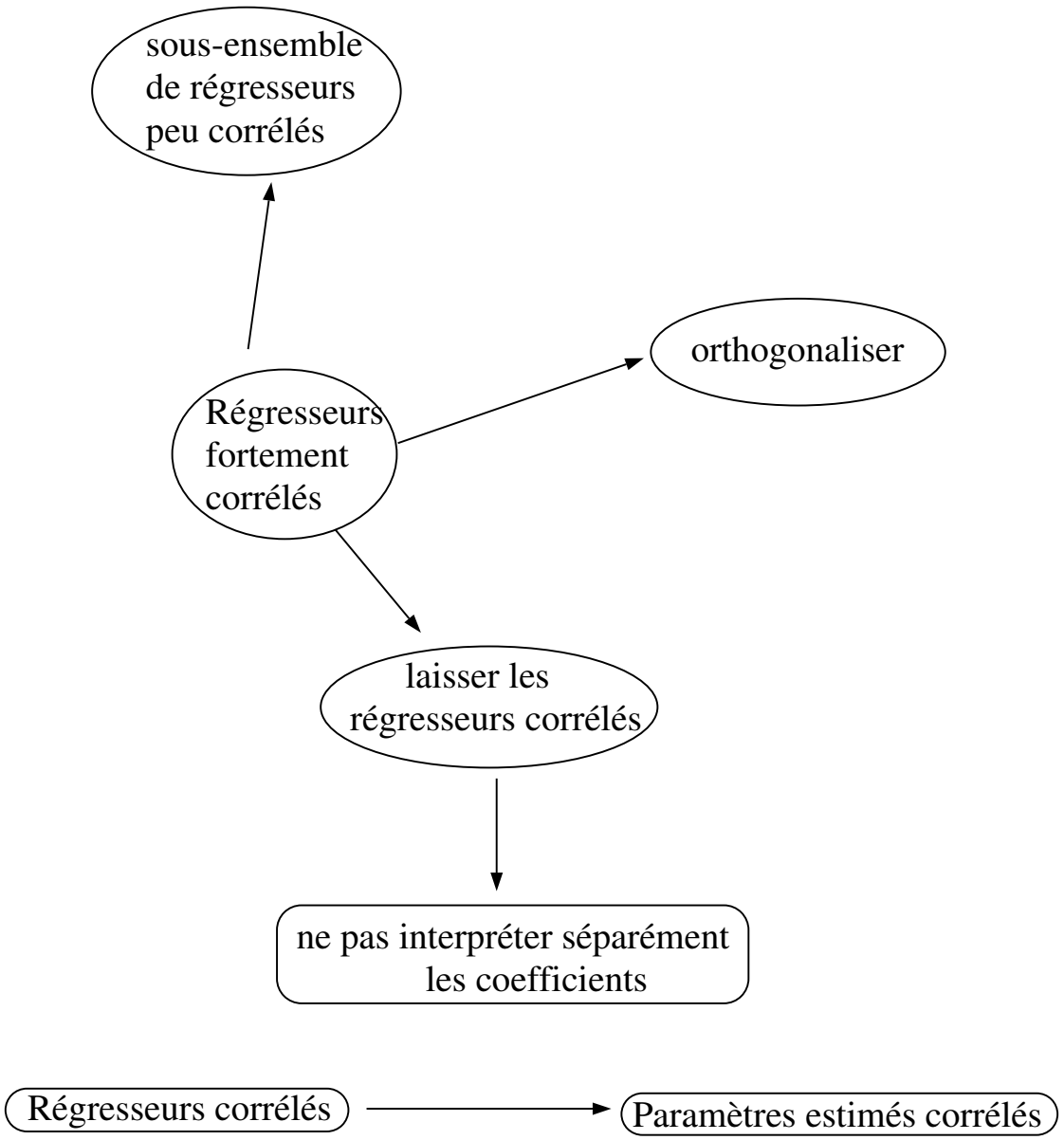
$$\mathrm{M}_{0_j} \ : \ \mu_n = \theta_0 + \sum_{p=1}^{j-1} \mathrm{Z}p_n \ .\theta_p + \sum_{p=j+1}^{P} \mathrm{Z}p_n \ .\theta_p$$

$$\mathrm{R}_j^2 = \frac{\mathrm{SCE}_{\mathrm{M}_{0_j}} - \mathrm{SCE}_{\mathrm{M}_1}}{\mathrm{SCE}_{\mathrm{M}_1}}$$

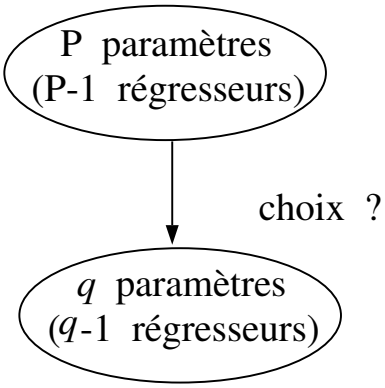
et $\mathrm{VIF}_j = \frac{1}{1-\mathrm{R}_j^2}$

Heuristique : Si $\mathrm{VIF}_j > 10$, se méfier .....

Comment remédier à la colinéarité

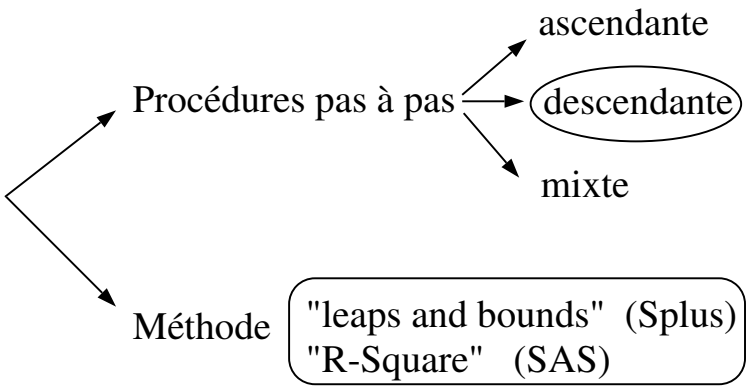


Sélection de Régresseurs



choix délicat

techniques = approches descriptives



Techniques de sélection de sous-ensemble de régresseurs

Procédures “Stepwise” : régresseurs introduits ou supprimés l’un après l’autre.

3 VARIANTES

— Sélection ascendante : “forward selection”

— Sélection descendante : “backward selection”

— Sélection progressive “stepwise” : sélection ascendante avec élimination possible de variables déjà introduites

Régression descendante "backward sélection"

1. On pose le modèle général (avec tous les régresseurs)

2. A chaque étape :

   calcul de la statistique F correspondant au retrait de chaque variable

3. On enlève du modèle le régresseur associé au F le plus faible

— Arrêt de la procédure quand le F du régresseur enlevé reste significatif (pour un niveau α fixé)

Exemple : Moutarde et sécheresse

Commandes SAS

```
Proc reg ;
model : RC = LT HF PR PA / selection = BACKWARD ;
run ;
```

Sortie SAS

Backward Elimination Procedure for Dependent Variable RC

Step 0 All Variables Entered R-square =0.90997350 C(p) = 6.00000000

	DF	Sum of Squares	Mean Square	F value	Prob>F
Regression	5	65.21663	13.04333	50.54	0.0001
Error	25	6.45208	0.25808		
C Total	30	71.66871			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	1	-0.428901	0.40886893	0.28399178	1.100	0.3042
LT	1	0.013582	0.00794030	0.75515641	2.927	0.0995
HF	1	0.003234	0.00210214	0.61090571	2.3685	0.1365
PR	1	0.002889	0.00058155	6.36800672	24.671	0.0001
PA	1	-0.000285	0.00013538	1.14615152	4.44	0.0453
FE	1	-0.054721	0.07386775	0.14163284	0.55	0.4657

Step 1 Variable Removed R-square = 0.90799728 C(p)=4.54878726
FE

	DF	Sum of Squares	Mean Square	F value	Prob>F
Regression	4	65.07499361	16.2687484	64.15	0.0001
Error	26	6.59371607	0.25360446		
C Total	30	71.66871			

Parameter Estimates

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	-0.55945487	0.36572318	0.59344714	2.34	0.1382
LT	0.01291499	0.00782027	0.69167359	2.73	0.1107
HF	0.00336084	0.00207692	0.66407217	2.62	0.1177
PR	0.00271614	0.00052819	6.70628811	26.44	0.0001
PA	-0.00028062	0.00013406	1.11123573	4.38	0.0462

Step 2 Variable Removed R-square = 0.89873142 C(p)=5.12187934
HF

	DF	Sum of Squares	Mean Square	F value	Prob>F
Regression	3	64.41092144	21.47030715	79.87	0.0001
Error	27	7.25778824	0.26880697		
C Total	30	71.66870968			

Parameter Estimates

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	-0.34453496	0.35081919	0.25926293	0.96	0.3348
LT	0.00946080	0.00774552	0.40104674	1.49	0.2325
PR	0.00338087	0.00034183	26.29562764	97.82	0.0001
PA	0.00034732	0.00013133	1.88002585	6.99	0.0135

Bounds on condition number: 2.501807, 19.15523

Step 3 Variable Removed R-square = 0.89313558 C(p)=4.67582217
LT

	DF	Sum of Squares	Mean Square	F value	Prob>F
Regression	2	64.00987470	32.00493735	117.01	0.0001
Error	28	7.65883498	0.27352982		
C Total	30	71.66870968			

Parameter Estimates

Variable	Parameter Estimate	Standard Error	Type II Sum of Squares	F	Prob>F
INTERCEP	0.02402802	0.18052453	0.00484583	0.02	0.8951
PR	0.00365835	0.00025766	55.14186285	201.59	0.0001
PA	0.00030579	0.00012796	1.56202320	6.71	0.0238

Bounds on condition number: 1.356468, 5.425871

All variables in the model are significant at the 0.1000 level.

Sous-modèle sélectionné :
RC = 0.024 + 0.0037 PR - 0.0003 PA

Un critère de qualité d'un sous-modèle : la statistique de Mallows (1976) : C_q

Modèle M_1 à P paramètres à N observations

(P-1 régresseurs)

On cherche le modèle sous-emboîté M_1' à q paramètres

(q-1 régresseurs)

$$C_q = \frac{SCE_{M_1'}}{CM_{M_1}} + (2q - N)$$

$SCE_{M_1'}$: somme des carrés des écarts résiduelle du modèle M_1'

CM_{M_1} : carré moyen des écarts résiduel du modèle M_1

$C_q > q \Rightarrow$ modèle sous paramétré

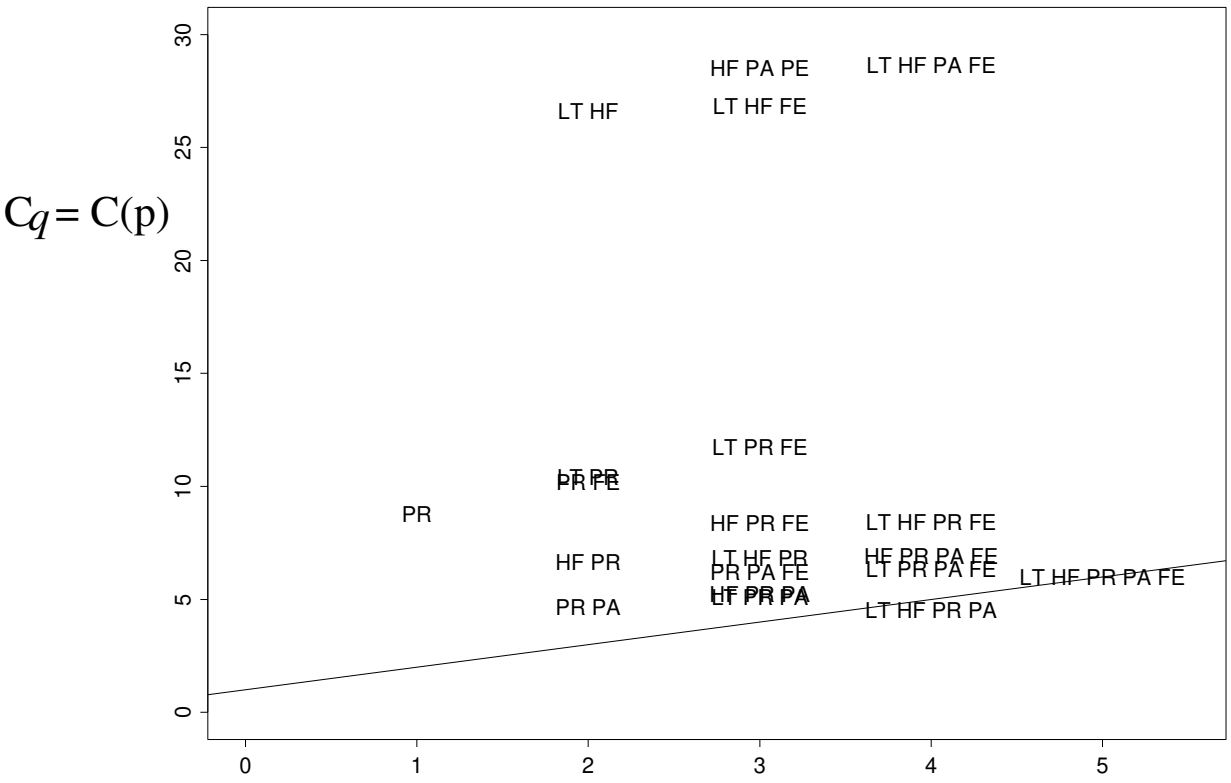
$C_q < q \Rightarrow$ modèle sur paramétré ($\exists$ colinéarité)

Sortie SAS

Option sélection = RSQUARE dans model de proc reg.

Number in Model	R-Square	C(p)	Variables in Model
1	0.87134053	8.7822	PR
1	0.64547979	71.44885	HF
1	0.52998616	103.52097	LT
1	0.42495726	132.68707	FE
1	0.12373617	216.33496	PA
2	0.89313558	4.67582	PR PA
2	0.88607174	6.63742	HF PR
2	0.87321073	10.20887	PR FE
2	0.87249925	10.40645	LT PR
2	0.81416081	26.60680	LT HF
2	0.72039147	52.64617	HF FE
2	0.68298489	63.03383	HF PA
2	0.59735658	86.81247	LT FE
2	0.53182615	105.01001	LT PA
2	0.43723197	131.27843	PA FE
3	0.89873142	5.12188	LT PR PA
3	0.89834630	5.22883	HF PR PA
3	0.89482888	6.20560	PR PA FE
3	0.89249211	6.85451	LT HF PR
3	0.88695651	8.39172	HF PR FE
3	0.87484597	11.75477	LT PR FE
3	0.82050150	26.84602	LT HF FE
3	0.81442384	28.53376	LT HF PA
3	0.73227218	51.34695	HF PA FE
3	0.60152955	87.65365	LT PA FE
4	0.90799728	4.54879	LT HF PR PA
4	0.90144948	6.36709	LT PR PA FE
4	0.89943673	6.92602	HF PR PA FE
4	0.89398114	8.44101	LT HF PR FE
4	0.82112012	28.67423	LT HF PA FE
5	0.90997350	6.00000	LT HF PR PA FE

Exemple : Moutarde et sécheresse



Comment choisir un bon modèle ?

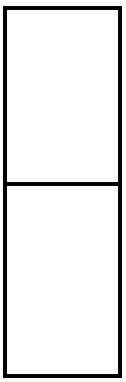
L'heuristique proposée avec le C(p) de Mallow est une bonne technique de choix

MAIS il existe

- d'autres critères (AIC, BIC)
- des techniques utilisant le calcul intensif par ordinateur

$\Rightarrow$ par exemple :

La validation croisée avec un **critère** de choix basé sur le modèle qui prédit le mieux les observations (MSEP)



$\longrightarrow$ partie servant à estimer les paramètres pour tous les modèles possibles

$\longrightarrow$ partie servant à calculer **le critère** $\mathrm{E}\left(\mathrm{Y}-\widehat{\mathrm{Y}}\right)^2$ pour tous les modèles estimés

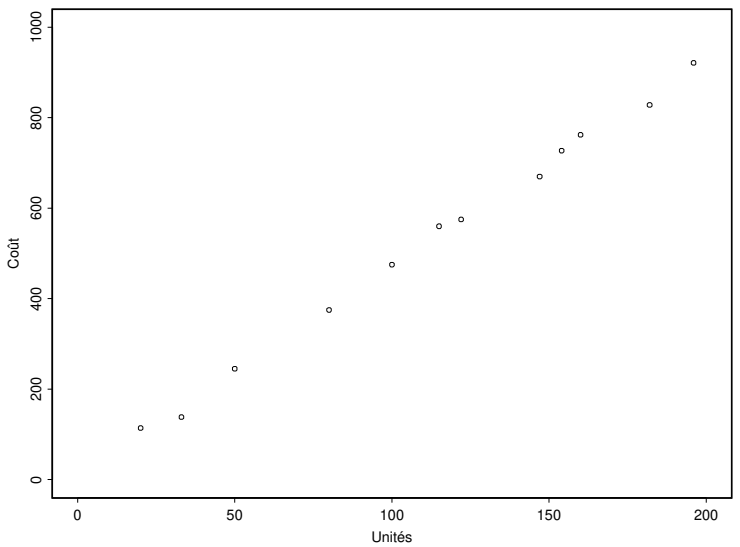
Ordonnée à l'origine nulle

Exemple : Les ateliers

— Y = coût de production
— Z = nombre d'unités produites
— N = 12 ateliers

Les données

atelier	unités	couts ($)
1	20	114
2	196	921
3	115	560
4	50	245
5	122	575
6	100	475
7	33	138
8	154	727
9	80	375
10	147	670
11	182	828
12	160	762



Commandes SAS

(avec intercept)

```
proc reg ;
model : couts = unites /ADJRSQ ;
run ;
```

(sans intercept)

```
proc reg ;
model : couts = unites / NOINT ADJRSQ ;
run ;
```

Sortie SAS

(avec intercept)

$M_1 : \mu_n = \theta_0 + \theta_1 Z_n$

$H_0 : \theta_1 = 0$

$\downarrow$

$H_1 : \theta_1 \neq 0$

$M_0 : \mu_n = \theta_0$

Source	DF	Sum of Squares	Mean Square	F value	Prob>F
Model	1	789526.87232	789526.87232	3530.777	0.0001
Error	10	2236.12768	223.61277		
C Total	11	791763.00000			

Root MSE	14.95369	R-square	0.9972
Dep MEAN	532.50000	Adj R-sq	0.9969
C.V.	2.80820		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0 : Parameter=0	Prob>| T |
INTERCEP	1	9.753739	9.79944674	0.995	0.3430
UNITE	1	4.615861	0.07768149	59.420	0.0001

Sortie SAS

(sans intercept)

$M_1 : \mu_n = \theta_1 Z_n$

$H_0 : \theta_1 = 0$

$\downarrow$ $H_1 : \theta_1 \neq 0$

$M_0 : \mu_n = 0$

Source	DF	Sum of Squares	Mean Square	F value	Prob>F
Model	1	4191980.3407	4191980.3407	18762.480	0.0001
Error	11	2457.65933	223.42358		
U Total	12	4194438			
Root MSE		14.94736	R-square	0.9994	
Dep MEAN		532.50000	Adj R-sq	0.9994	
C.V.		2.80702			

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0 : Parameter=0	Prob>| T |
UNITE	1	4.685274	0.03420502	136.976	0.0201

Une Heuristique

Ou la démarche pour obtenir à partir de P variables explicatives un jeu de P' variables

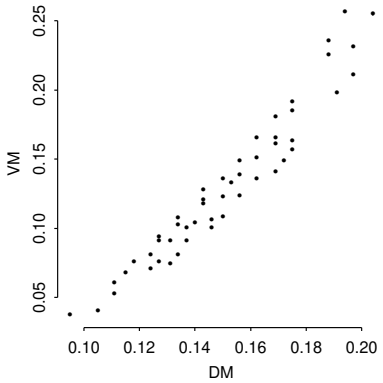
Exemple : Volume de bois

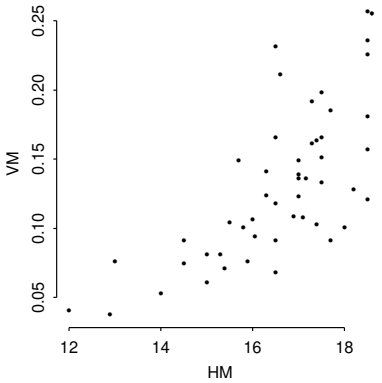
Exemple : l'estimation du volume d'un arbre **(VM)** à partir de sa hauteur **(HM)** et de son diamètre à hauteur de poitrine **(DM)**

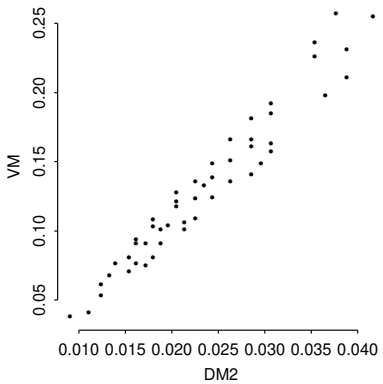
Les données

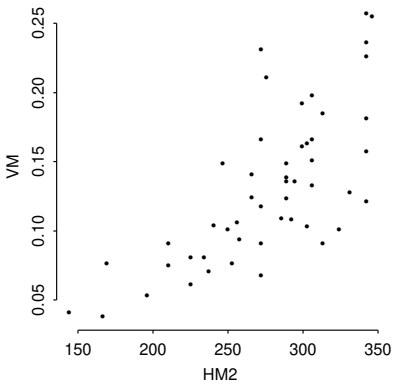
	DM	HM	VM		DM	HM	VM
1	0.095	12.9	0.037	28	0.153	17.5	0.132
2	0.105	12	0.04	29	0.156	17	0.138
3	0.111	14	0.052	30	0.156	17	0.148
4	0.111	15	0.06	31	0.156	16.3	0.123
5	0.115	16.5	0.067	32	0.162	17.16	0.135
6	0.118	15.9	0.075	33	0.162	17.5	0.165
7	0.124	15.4	0.07	34	0.162	17.5	0.15
8	0.124	15.3	0.08	35	0.169	18.5	0.18
9	0.127	16.5	0.09	36	0.169	17.3	0.16
10	0.127	13	0.075	37	0.169	16.3	0.14
11	0.127	16.05	0.093	38	0.169	16.5	0.165
12	0.131	14.5	0.074	39	0.172	15.7	0.148
13	0.134	17.4	0.102	40	0.175	18.5	0.156
14	0.134	17.1	0.107	41	0.175	17.7	0.184
15	0.137	18	0.1	42	0.175	17.3	0.191
16	0.134	15	0.08	43	0.175	17.4	0.162
17	0.137	14.5	0.09	44	0.188	18.5	0.225
18	0.131	17.7	0.09	45	0.188	18.5	0.235
19	0.14	15.5	0.103	46	0.191	17.5	0.197
20	0.143	18.2	0.127	47	0.194	18.5	0.256
21	0.143	18.5	0.12	48	0.197	16.5	0.23
22	0.143	16.5	0.117	49	0.197	16.6	0.21
23	0.146	15.8	0.1	50	0.204	18.6	0.254
24	0.146	16	0.105				
25	0.15	17	0.122				
26	0.15	17	0.135				
27	0.15	16.9	0.108				

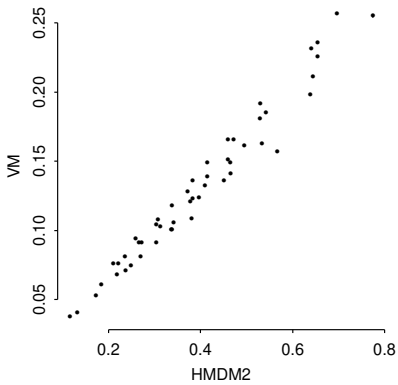
Première étape

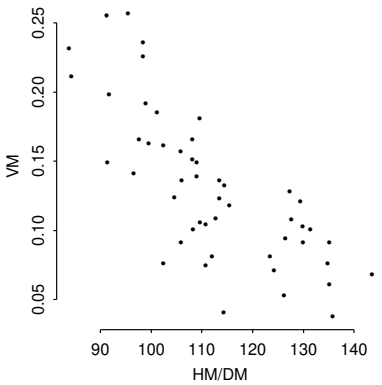




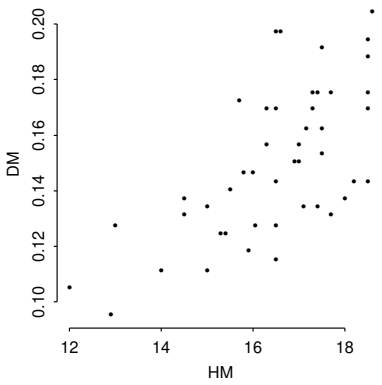


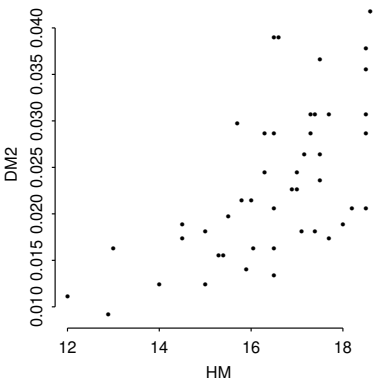


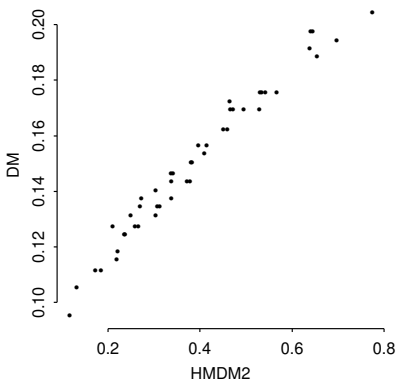


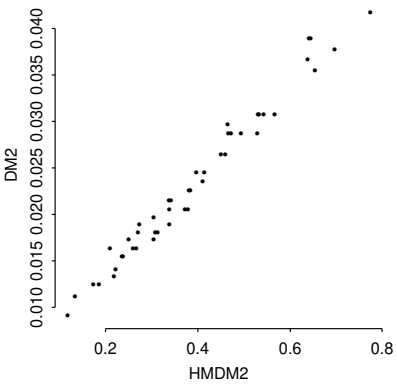


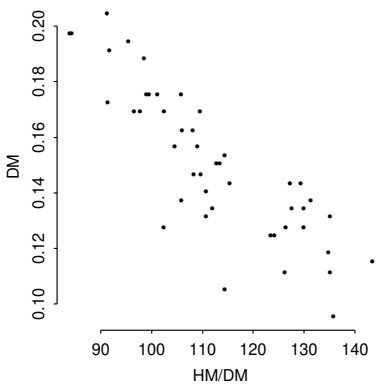
Deuxième étape

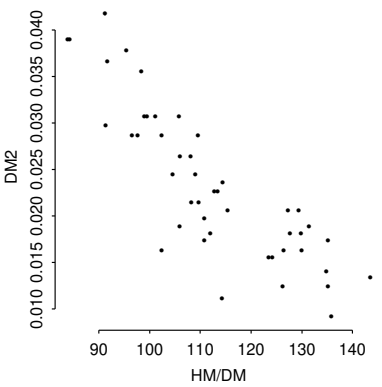












Troisième étape

(1) $R^2 = 0,9442 \quad R^2 adj = 0,9430$

Variable	DF	Parameter Estimate	Standard Error	T for H0 : Parameter=0	Prob>| T |
INTERCEP	1	-0.025508	0.00576662	-4.423	0.0001
DM2	1	6.629485	0.23262280	28.499	0.0001

(2) $R^2 = 0,9574 \quad R^2 adj = 0,9536$

Variable	DF	Parameter Estimate	Standard Error	T for H0 : Parameter=0	Prob>| T |
INTERCEP	1	-0.099118	0.01996072	-4.966	0.0001
DM2	1	5.978778	0.26704178	22.389	0.0001
HM	1	0.005367	0.00140721	3.814	0.0004

(3) $R^2 = 0,9605 \quad R^2 adj = 0,9580$

Variable	DF	Parameter Estimate	Standard Error	T for H0 : Parameter=0	Prob>| T |
INTERCEP	1	-0.153987	0.0345914	-4.452	0.0001
DM2	1	7.659445	0.91456016	8.375	0.0001
HM	1	0.001165	0.00258455	0.451	0.6542
HM/DM	1	0.000760	0.00039658	1.917	0.0615

(4) $R^2 = 0,9604 \quad R^2 adj = 0,9587$

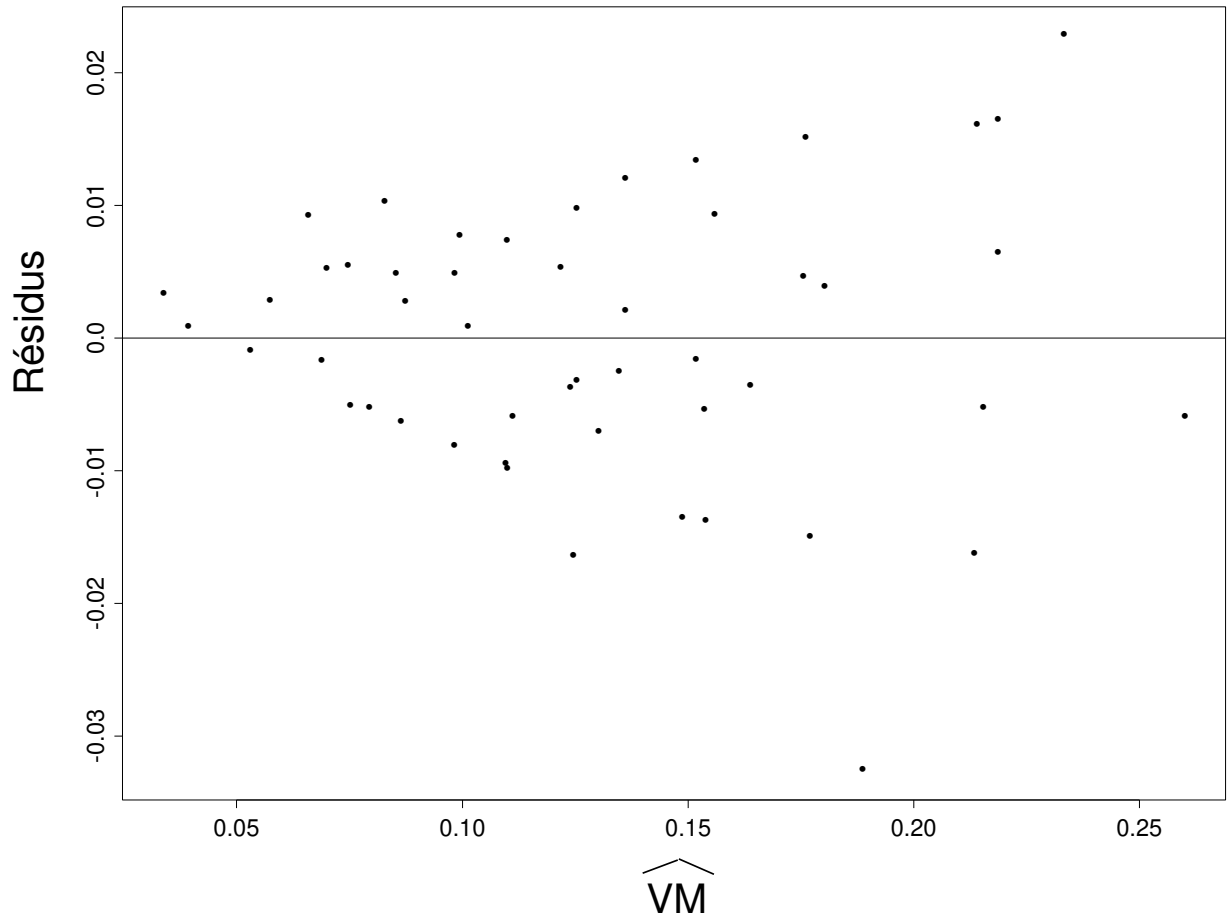
Variable	DF	Parameter Estimate	Standard Error	T for H0 : Parameter=0	Prob>| T |
INTERCEP	1	-0.160453	0.03121066	-5.141	0.0001
DM2	1	8.034431	0.37713097	21.304	0.0001
HM/DM	1	0.000912	0.00020825	4.378	0.0001

(5) $R^2 = 0,9557 \quad R^2 adj = 0,9650$

Variable	DF	Parameter Estimate	Standard Error	T for H0 : Parameter=0	Prob>| T |
INTERCEP	1	-0.006284	0.00398355	-1.577	0.1213
HMDM2	1	0.344002	0.00935428	36.775	0.0001

Quatrième étape

Modèle : **VM** = a + b **HMDM2**



Cinquième étape

Model : MODEL1

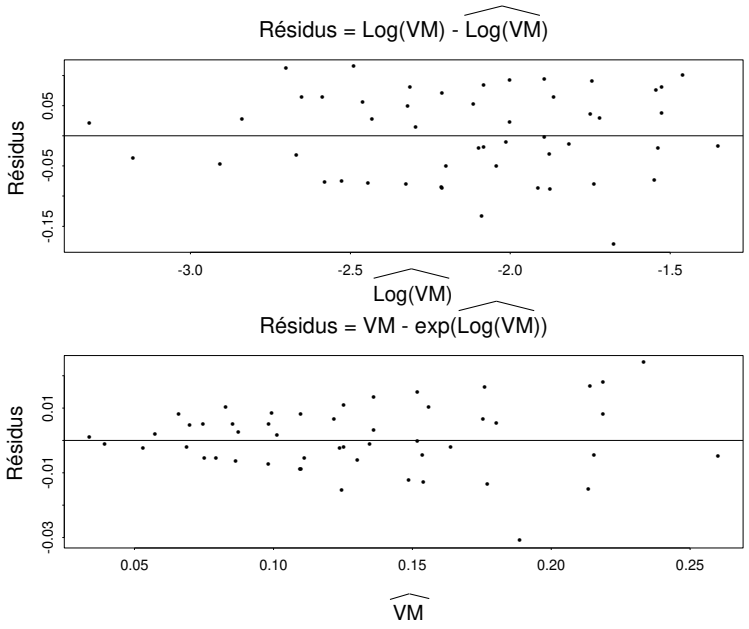
Dependent Variable : LOGVM

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F value	Prob>F
Model	2	9.81589	4.90795	931.117	0.0001
Error	47	0.24774	0.00527		
C Total	49	10.06363			
Root MSE		0.07260	R-square	0.9754	
Dep MEAN		-2.13373	Adj R-sq	0.9743	
C.V.		-3.40258			

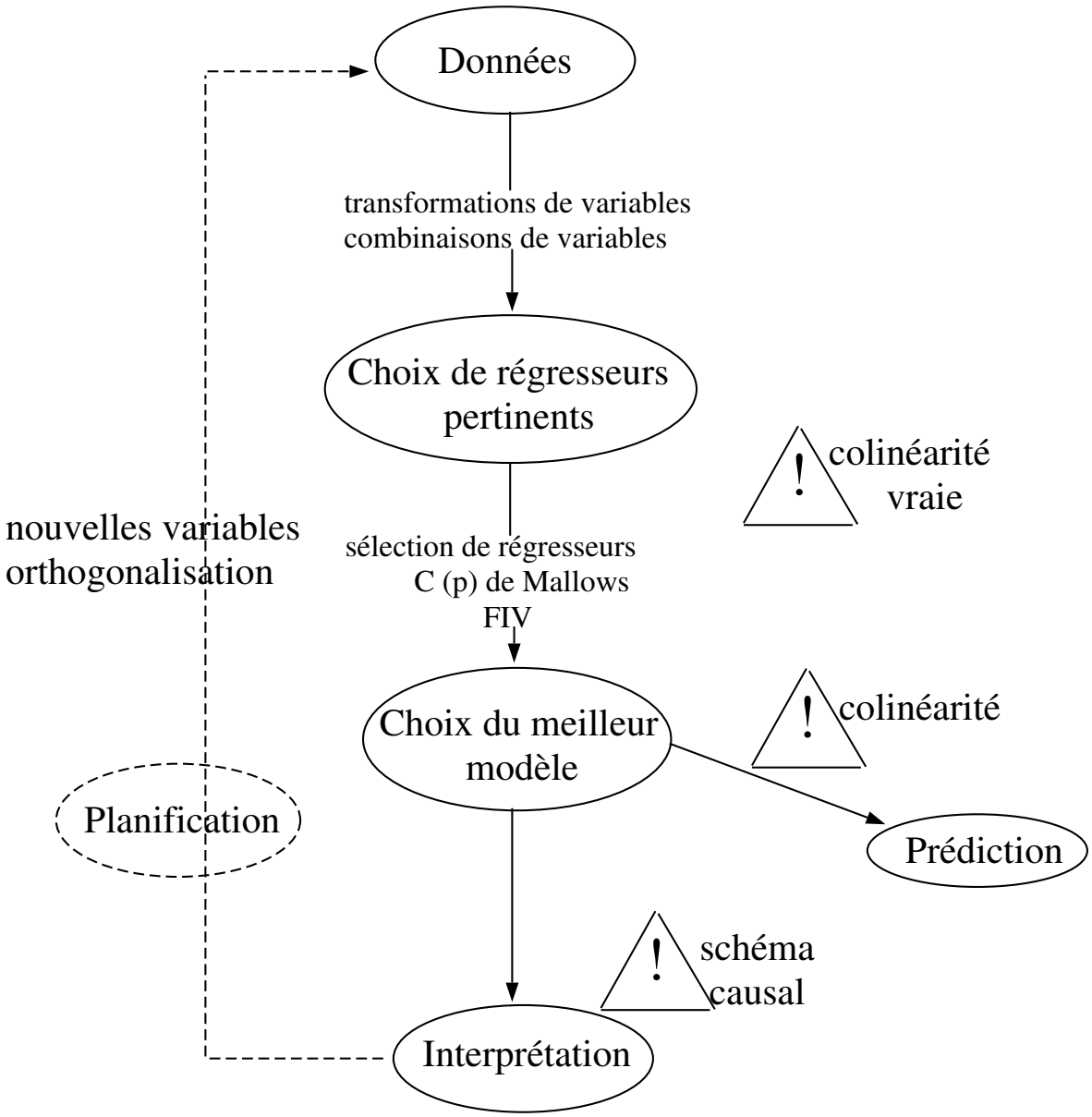
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0 : Parameter=0	Prob>F<| T |
INTERCEP	1	-0.966086	0.53221938	-1.815	0.0759
LOGDM	1	2.089999	0.08085075	25.850	0.0001
LOGHM	1	1.004970	0.14744389	6.816	0.0001

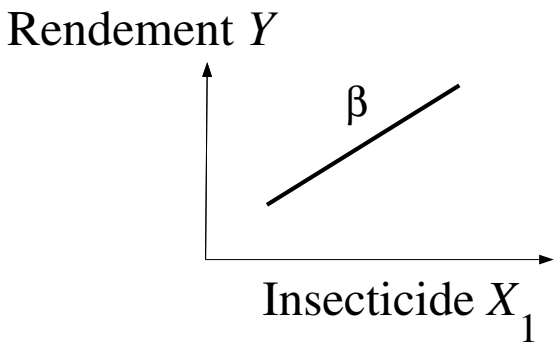


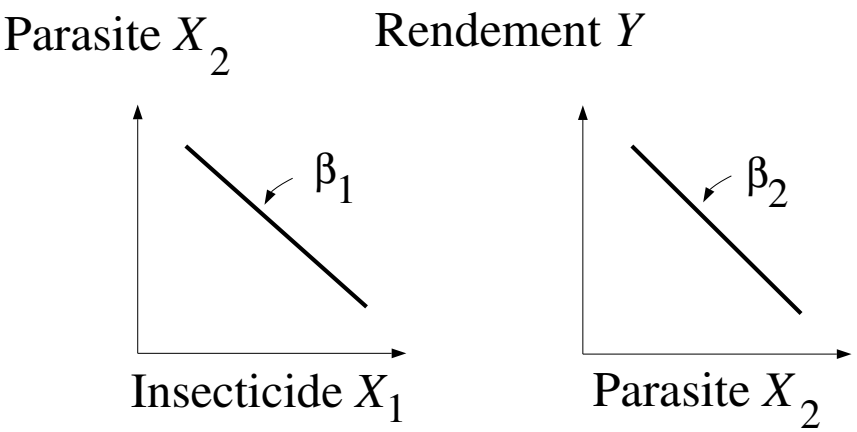
Régression linéaire

Conclusion



Exemple de schéma causal :
Attaque parasitaire d'une culture végétale





$$Y = \alpha + \beta X_1$$
$$\beta = \beta_1 \beta_2$$