

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 ?

→ pour expliquer/décrire

→ pour prédire

une variable quantitative en fonction de plusieurs variables
quantitatives

Ecriture du modèle

$$Y_n = \theta_0 + \left(\sum_{p=1}^{p-1} \theta_p Z_{pn} \right) + \varepsilon_n$$

\downarrow
 X_{pn}

On appelle régresseur une variable théoriquement contrôlée

→ Insister sur les variables **quantitatives**

Écriture du modèle linéaire dans le cas de la régression

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

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

Y_n = variable à expliquer

avec N = nombre d'observations

$P-1$ = nombre total de régresseurs

P = nombre de paramètres

en écriture matricielle

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

Insister sur la linéarité en fonction des paramètres

Rappeler que le nombre total de paramètres $P \leq (N-1)$

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

$$Y_n = \theta_0 + Z1_n \cdot \theta_1 + Z2_n \cdot \theta_2 + Z3_n \cdot \theta_3 + \varepsilon_n$$

On peut écrire ce système sous forme de 5 équations

$$\begin{aligned} Y_1 = 30 &= \theta_0 + 80 \theta_1 + 40 \theta_2 + 40 \theta_3 + \varepsilon_1 \\ Y_2 = 50 &= \theta_0 + 100 \theta_1 + 40 \theta_2 + 40 \theta_3 + \varepsilon_2 \\ Y_3 = 100 &= \theta_0 + 180 \theta_1 + 100 \theta_2 + 100 \theta_3 + \varepsilon_3 \\ Y_4 = 60 &= \theta_0 + 100 \theta_1 + 80 \theta_2 + 20 \theta_3 + \varepsilon_4 \\ Y_5 = 70 &= \theta_0 + 150 \theta_1 + 70 \theta_2 + 120 \theta_3 + \varepsilon_5 \end{aligned}$$

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

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

Test de qualité d'un modèle

$$M_1 : Y_n = \theta_0 + \left(\sum_{p=1}^{P-1} Z_{pn} \cdot \theta_p \right) + \varepsilon_n \quad \begin{matrix} P \\ \text{paramètres} \end{matrix}$$

$$\begin{matrix} H_0 : \theta_1 = \dots = \theta_p = 0 \\ H_1 : \exists p; \theta_p \neq 0 \end{matrix}$$

$$M_0 : Y_n = \theta_0 + \varepsilon_n \quad \begin{matrix} 1 \\ \text{paramètre} \end{matrix}$$

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 :

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

Estimation des paramètres

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

Rappel : On quantifie l'utilité du modèle M_1 par rapport au modèle M_0 en considérant la différence entre M_0 et M_1 c'est-à-dire ce qu'apporte M_1 par rapport au modèle le plus simple M_0

Remarque :

Généralement pour la régression la matrice X est inversible.

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 = M1 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

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

Sum of squares = SCE : somme des carrés des écarts

Mean square = CM : carré moyen des écarts

C Total pour corrected Total, en référence au modèle le plus simple à savoir $M_0 : \mu_n = \theta_0$

SCE_{M_0} liée à la variance totale c'est à dire ce qui n'est pas expliqué par le

$$\text{modèle } M_0; SCE_{M_0} = \sum_{n=1}^N (Y_n - \hat{\theta}_0)^2 = 2680$$

SCE_{M_1} ce qui n'est pas expliqué par le modèle M_1 ;

$$SCE_{M_1} = \sum_{n=1}^N \left(Y_n - \sum_{p=1}^4 Z_{p,n} \cdot \hat{\theta}_p \right)^2 = 15$$

Dans cet exemple on teste l'hypothèse nulle

H_0 = aucun régresseur n'a d'influence sur le rendement

Le test F montre que cette hypothèse est acceptable : le modèle testé n'est pas explicatif malgré un R^2 très élevé : inutile d'aller plus loin. F non significatif équivalent à l'intervalle de confiance du R^2 contient 0

On peut remarquer que pour 5 observations (N=5) le modèle a (3+1) paramètres ; il ne reste plus qu'un seul degré de liberté pour l'erreur : le critère $P \leq N-1$ est limite.

Remarque :

Le R^2 tend toujours vers 1 quand on augmente le nombre de régresseurs, alors que le R^2 ajusté se stabilise à une certaine valeur quand un jeu convenable de régresseurs est introduit.

Étude 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 ?

31 pieds de moutarde sont soumis à 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.

Le nombre de racines tubérisées est exprimé en log

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

$$M_1 : \mu = \theta_0 + \theta_1 LT + \theta_2 HF + \theta_3 PR + \theta_4 PA + \theta_5 FE \quad P=6$$

$$\downarrow \begin{array}{l} H_0 : \theta_1 = \theta_2 = \theta_3 = \theta_4 = \theta_5 = 0 \\ H_1 : \exists p: \theta_p \neq 0 \end{array}$$

$$M_0 : \mu = \theta_0 \quad P=1$$

Sortie SAS

Régression linéaire multiple

Exemple : Moutarde et sécheresse

Model : MODEL1 = M₁ 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			

Par le test F on teste l'hypothèse qu'il existe au moins un $\theta_p \neq 0$ contre le modèle le plus simple : modèle M₀ : $\mu = \theta_0$

- F est très élevé donc on rejette l'hypothèse H₀ ; il y a donc au moins une liaison significative entre RC et un des 5 régresseurs
- R² = 0.91 : la proportion de la variabilité expliquée par le modèle M₁ par rapport au modèle M₀ c'est à dire le gain qu'apporte M₁ par rapport à M₀

Sortie SAS : estimation des paramètres

Model : MODEL1		Nbre tubérisations			
dependent variable RC	DF	Parameter Estimate	Standard Error	T for H0 : parameter=0	Prob > T
1	2	3	4	5	6
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)

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

$$\downarrow$$

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

$$H_0 : \theta_5 = 0$$

$$H_1 : \theta_5 \neq 0$$

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

Le test T sur chacun des paramètres est strictement équivalent au test F de chaque sous-modèle avec un θ_1 nul contre le modèle général M_1 . Le F associé est égal au carré du T calculé en colonne 5 du tableau.

On conclut ici que FE n'apporte pas d'amélioration significative au modèle M_1 par rapport au modèle sans FE

Exercice : Qu'en est-il des autres régresseurs ?

Intervalle de confiance des θ_p

$$\widehat{\theta}_p \sim \mathcal{N}(\theta_p, \sigma^2 (X'X)_{pp}^{-1})$$

On montre que :

$$\frac{\widehat{\theta}_p - \theta_p}{\sqrt{\widehat{\text{Var}}(\widehat{\theta}_p)}} \sim t(N - P)$$

Parameter Estimates
Standard Error

(colonne 4) de la sortie SAS

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

$$\widehat{\theta}_p - t_{1-\alpha/2, N-P} \times \sqrt{\widehat{\text{Var}}(\widehat{\theta}_p)} < \theta_p < \widehat{\theta}_p + t_{1-\alpha/2, N-P} \times \sqrt{\widehat{\text{Var}}(\widehat{\theta}_p)}$$

Exemple :

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

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

$(X'X)_{pp}^{-1}$: élément diagonal (ligne p, colonne p) de la matrice $(X'X)^{-1}$

$$\text{Var}(\widehat{\theta}_p) = \sigma^2 (X'X)_{pp}^{-1}$$

σ^2 , variance résiduelle est inconnue. On l'estime par le carré moyen résiduel $\widehat{\sigma}^2$ (= carré moyen de l'erreur dans les tableaux SAS)

On utilise comme estimateur de la variance de $\widehat{\theta}_p$: $\widehat{\text{Var}}(\widehat{\theta}_p) = \widehat{\sigma}^2 (X'X)_{pp}^{-1}$

$\frac{\widehat{\theta}_p - \theta_p}{\sqrt{\widehat{\text{Var}}(\widehat{\theta}_p)}}$ suit une loi de student à N-P degrés de liberté

Quand N-P grand l'intervalle de confiance à 95% est approximativement

$$\widehat{\theta}_p \pm 2 \sqrt{\widehat{\text{Var}}(\widehat{\theta}_p)}$$

Remarque :

Quand N-P est petit (beaucoup de paramètres par rapport au nombre d'observations). On perd beaucoup de précision sur l'estimation des paramètres et l'on obtient des intervalles de confiance très grands.

Exercice :

— Faire construire les intervalles de confiance pour les autres régresseurs

— Faire remarquer qu'il y a correspondance entre la valeur 0 dans l'intervalle de confiance et le fait que le test est non significatif

Région de confiance simultanée des θ cas de 2 régresseurs

Notation

$$v_1^2 = \sum_n (Z_{1n} - Z_{1.})^2$$

$$v_2^2 = \sum_n (Z_{2n} - Z_{2.})^2$$

$$\rho = \frac{\sum_n (Z_{1n} - Z_{1.})(Z_{2n} - Z_{2.})}{\sqrt{v_1^2} \sqrt{v_2^2}}$$

alors

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

La région de confiance simultanée (lorsque N-P est grand)

de (θ_1, θ_2) est la région telle que

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

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

— Le modèle :

$$Y_n = \theta_0 + \theta_1 Z_{1n} + \theta_2 Z_{2n} + \varepsilon_n$$

est équivalent au modèle

$$Y_n = \theta'_0 + \theta_1(Z_{1n} - Z_{1.}) + \theta_2(Z_{2n} - Z_{2.}) + \varepsilon$$

où $\theta'_0 = \theta_0 + \theta_1 Z_{1.} + \theta_2 Z_{2.}$

Dans ce modèle "recentré pour les régresseurs"

$$(X'X) = \begin{pmatrix} N & 0 & 0 \\ 0 & v_1^2 & \rho v_1 v_2 \\ 0 & \rho v_1 v_2 & v_2^2 \end{pmatrix}$$

Et l'on trouve que le bloc de $(X'X)^{-1}$ correspondant à (θ_1, θ_2) est

$$A = \begin{pmatrix} \frac{1}{v_1^2(1-\rho^2)} & \frac{-\rho}{v_1 v_2(1-\rho^2)} \\ \frac{-\rho}{v_1 v_2(1-\rho^2)} & \frac{1}{v_2^2(1-\rho^2)} \end{pmatrix}$$

— Si $\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix}, \sigma^2 A \right)$

alors $\begin{pmatrix} \hat{\theta}_1 - \theta_1, \hat{\theta}_2 - \theta_2 \end{pmatrix} \frac{A^{-1}}{\sigma^2} \begin{pmatrix} \hat{\theta}_1 - \theta_1 \\ \hat{\theta}_2 - \theta_2 \end{pmatrix} \sim \chi^2$ à 2 degrés de liberté ; d'où

l'intervalle de confiance simultanée

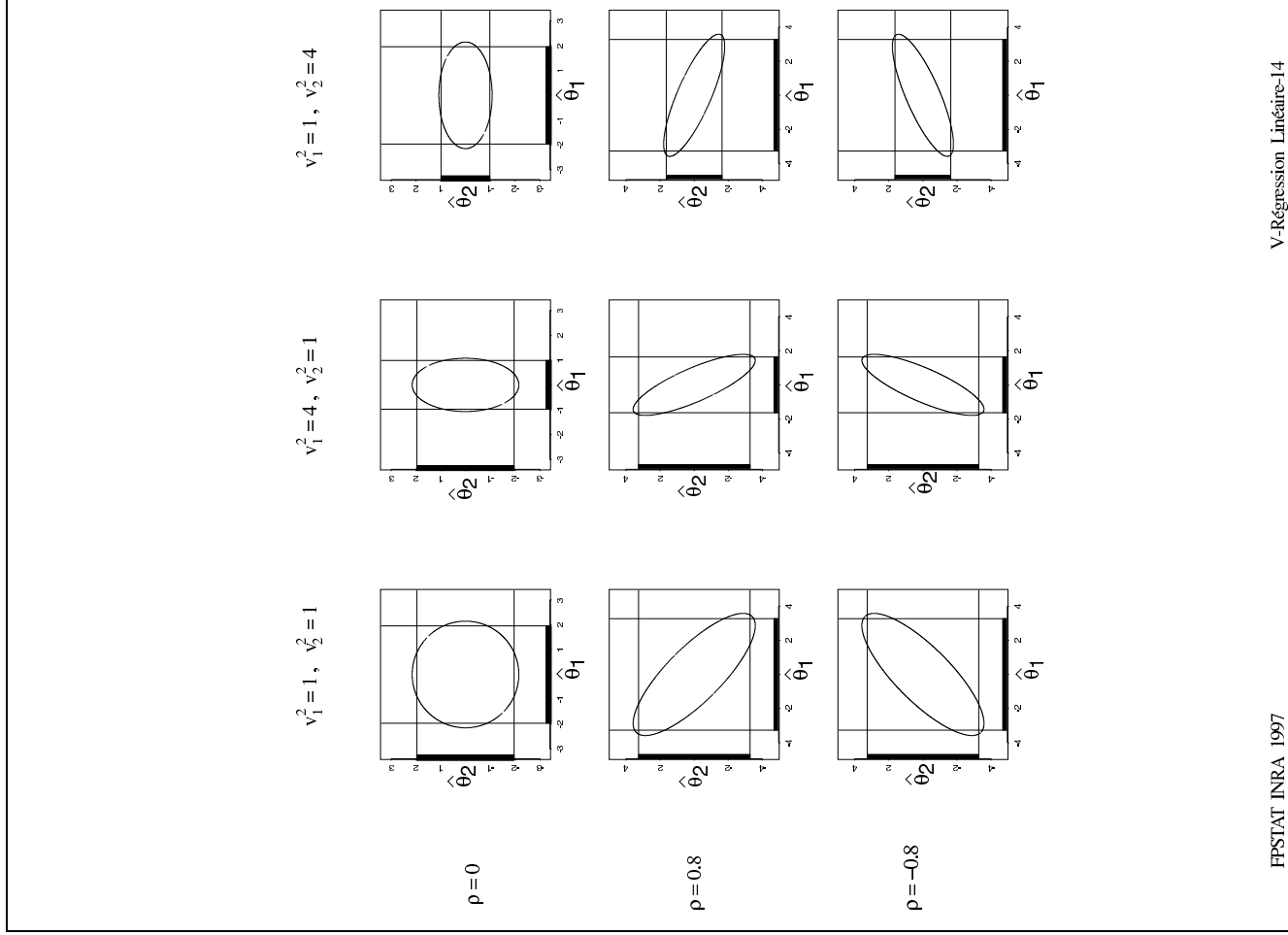
— La figure représente pour $\hat{\sigma}^2 = 1$ les régions de confiance simultanée à 90% pour différentes valeurs de v_1^2 , v_2^2 et ρ (cercle ou ellipses)

— Les intervalles de confiance à 5% de chacun des deux paramètres sont marqués par des boîtes noires

— Le rectangle en pointillé représente une région de confiance simultanée à 90% si l'on suppose la corrélation entre $\hat{\theta}_1$ et $\hat{\theta}_2$ égale à zéro, c'est ce que l'on fait lorsque l'on interprète chacun des paramètres indépendamment des autres

— Lorsque $\rho = 0$, on ne fait pas d'erreur (le rectangle "colle" bien à l'ellipse) mais lorsque ρ est grand cela ne va pas du tout.

Il ne faut pas alors interpréter les paramètres indépendamment les uns des autres



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})

La colinéarité vraie est le fait qu'un des régresseurs est une combinaison linéaire des autres régresseurs et de \mathbb{I}

La colinéarité vraie est rare, de plus il existe un remède évident qui consiste à supprimer un régresseur.

Cependant dans les problèmes réels, on se retrouve souvent face à une quasi colinéarité.

Conséquences de la quasi colinéarité

- tend à augmenter la variance des \hat{Y}
en particulier pour les jeux de données n'appartenant pas à l'échantillon
- tend à augmenter la variance des $\hat{\theta}$
 - ⇒ diminue la puissance des tests sur les paramètres
 - ⇒ risque d'estimation incorrecte des paramètres (ordre de grandeur ou même signe incorrect)
 - ⇒ risque de mauvaise interprétation des paramètres

La variance des prédictions par le modèle augmente

⇒ le modèle perd de son aptitude de prédiction

Les algorithmes de calcul deviennent mauvais

La quasi colinéarité est le fait qu'il y a des valeurs propres quasi nulles dans $X'X$. Ce qui implique que $X'X$ est proche de la non inversibilité.

Comment détecter la quasi colinéarité ?

Regarder la matrice de corrélation des régresseurs

⇒ ne révèle que les corrélations 2 à 2

Utiliser le VIF (Variance Inflation Factor) :

$$M_1 : \mu_n = \theta_0 + \sum_{p=1}^{p-1} Z_{p_n} \cdot \theta_p$$

$$\begin{aligned} &\downarrow \\ H_0 : \theta_j &= 0 \\ H_1 : \theta_j &\neq 0 \end{aligned}$$

$$M_{\theta_j} : \mu_n = \theta_0 + \sum_{p=1}^{j-1} Z_{p_n} \cdot \theta_p + \sum_{p=j+1}^p Z_{p_n} \cdot \theta_p$$

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

$$\text{et } VIF_j = \frac{1}{1 - R_j^2}$$

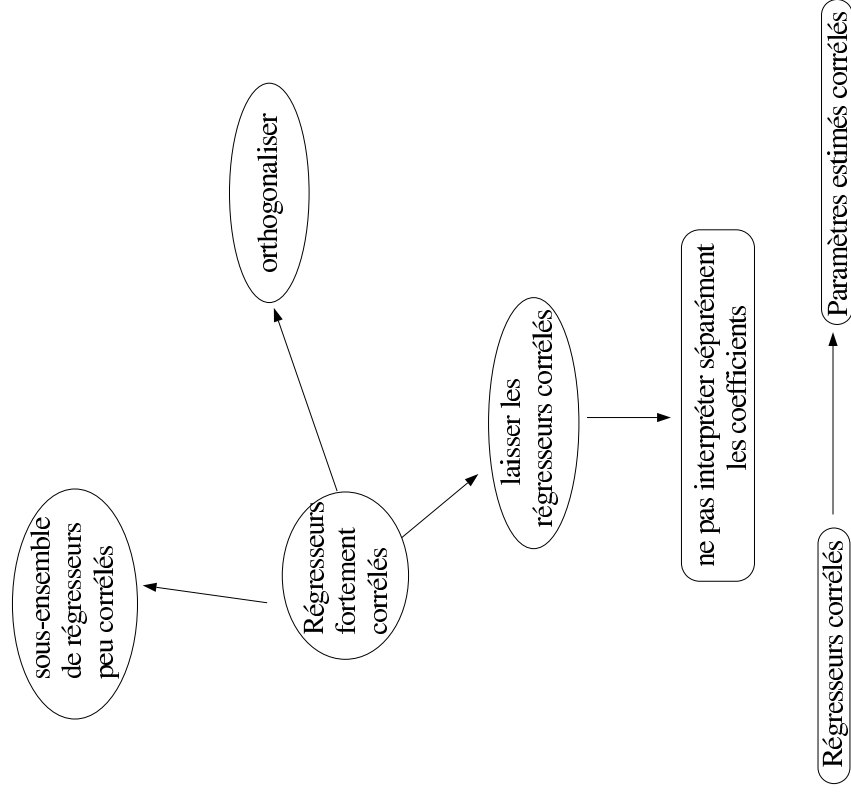
Heuristique : Si $VIF_j > 10$, se méfier

$1 - R_j^2$: tolérance

VIF : inverse de la tolérance

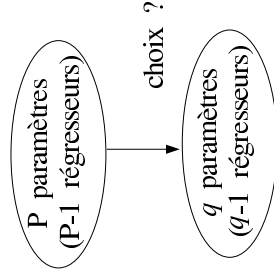
VIF exprime pour chaque régresseur son degré d'indépendance par rapport aux autres.

Comment remédier à la colinéarité



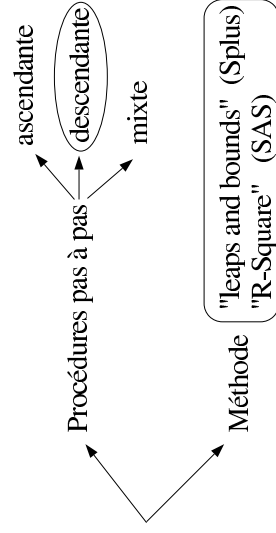
Voir les techniques de l'ACP (module 7-8 : "Analyse multidimensionnelle et classification")

Sélection de Régresseurs



choix délicat

techniques = approches descriptives



- Quand il est trop coûteux de récolter des données sur un grand nombre de variables

- Quand un problème de quasi colinéarité apparaît (surparamétrage du modèle)
⇒ on cherche un modèle avec moins de paramètres, en perdant de la quantité de manière raisonnable

Il est souvent difficile de faire un choix pertinent

⇒ techniques de sélection du (ou des) modèle optimal, à partir de jeu de données.

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

Sélection ascendante :

- On teste l’hypothèse nulle : l’ajout d’un régresseur au modèle n’améliore pas la qualité du modèle
- A chaque étape, on ajoute au modèle le régresseur ayant le F le plus élevé
- On s’arrête quand l’ajout d’un régresseur n’ajoute rien à la qualité du modèle (test F n.s.)

Sélection descendante :

- On part du modèle complet
- Hypothèse nulle : le retrait d’un régresseur ne détériore pas la qualité du modèle
- On enlève à chaque étape le régresseur ayant le F le plus faible
- On s’arrête quand tous les régresseurs restants ont un F significatif

Sélection “stepwise” :

- Comme la sélection ascendante mais en se réservant la possibilité d’enlever des régresseurs ajoutés auparavant

Régression descendante
"backward selection"

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)

On enlève donc à chaque étape, la variable explicative qui apporte le moins d'information sur la variable expliquée, donc le régresseur pour laquelle le retrait est le moins pénalisant par rapport à l'information apportée par le modèle général

On choisit souvent $\alpha = 5\%$ pour décider d'arrêter la procédure mais par défaut SAS utilise $\alpha = 10\%$

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 0 : première étape : modèle général

Par un test F de sous-modèle, on teste la suppression de chacun des régresseurs.

Le sous-modèle correspondant à la suppression de la variable FE n'est pas significativement différent du modèle général M_1 (au risque de 1ère espèce $\alpha = 46.57\%$)

$C(p)$ = statistique de Mallows \rightarrow étudié plus loin

Pour les régresseurs, les sommes des carrés à utiliser sont celles de type II, c'est à dire l'explication apportée par le régresseur lorsqu'il est introduit en dernier dans le modèle. Il n'y a pas lieu de définir un type III, car tous les régresseurs sont des effets de même ordre.

Step 1 : seconde étape : modèle sans la variable FE

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

On teste ici le retrait de chacun des 4 régresseurs restant par rapport au modèle plus général sans FE

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

Modèle sans FE

$$Y_n = \theta_1 + \theta_2LT + \theta_3HF + \theta_4PR + \theta_5PA + \varepsilon_n$$

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

Modèle réduit sans FE, sans HF

$$Y_n = \theta_1 + \theta_2LT + \theta_4PR + \theta_5PA + \varepsilon_n$$

Le sous-modèle sans la variable HF n'est pas significativement différent (au risque α de 11.77%) du modèle sans la variable FE.

On ne peut rejeter l'hypothèse nulle du coefficient de régression associé à la variable HF

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

DF	Sum of Squares	Mean Square	F value	Prob>F
3	64.41092144	21.47030715	79.87	0.0001
27	7.25778824	0.26880697		
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
2	64.00987470	32.00493735	117.01	0.0001
28	7.65883498	0.27352982		
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

On continue la procédure backward.

Step 2 : Troisième étape: modèle général sans FE et HF

Le sous-modèle sans la variable LT n'est pas significativement différent du modèle (sans FE et HF) au risque α de 23.25%.

On pourra donc encore supprimer cette variable.

Step 3 : Quatrième étape: modèle général sans FE, HF et LT

Les deux sous-modèles correspondant à la suppression des variables PR et PA sont significativement différents du modèle sans FE, HF et LT au risque $\alpha = 10\%$. On ne peut supprimer ces deux variables.

On remarquera que ces différentes sélections emboîtées augmentent le risque global d'erreur car en effet on ne sélectionne pas directement PR et PA. Ces techniques ne sont donc pas optimales car on ne maîtrise pas vraiment l'erreur réalisé.

On écrira au tableau le sous-modèle finalement sélectionné par la procédure "backward".

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)

La statistique de Mallows (1976) est l'un des critères donné par les programmes de sélection des régresseurs sous SAS (noté $C(p)$ dans SAS). C'est un critère très populaire parmi d'autres.

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

L'option "Sélection-Rsquare" de proc REG dans SAS permet de classer tous les modèles possibles dans l'ordre :

colonne 1 : nombre de paramètres

colonne 2 : R^2

colonne 3 : C(p) de Mallow

Dans un groupe de modèles ayant le même nombre de paramètres, le meilleur modèle est celui qui a R^2 le plus élevé

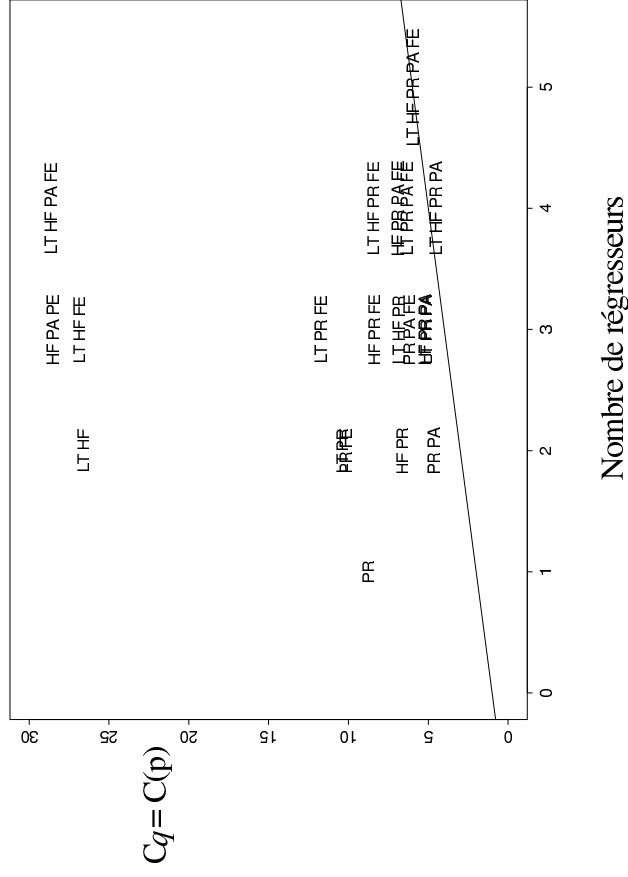
Mais comment choisir le meilleur modèle quel que soit le nombre de paramètres ?

Heuristique :

Tracer le graphe $q \rightarrow C_q$ avec la droite de référence $q \rightarrow q$

En allant vers les sous-modèles à q décroissants, s'arrêter quand le meilleur modèle pour q donné, tout en restant proche de la droite $f(q) = q$, commence à s'écartier du 2ième meilleur sous-modèle à q paramètres

Exemple : Montarde et sécheresse



Les modèles manquants ont un $C(p) > 30$

- Le meilleur modèle à 4 régresseurs est sous la droite : il est probablement sur-paramétré
- Le meilleur modèle à 3 régresseurs est le plus proche de la droite mais un autre modèle à 3 régresseurs est quasi à la même place
- Le meilleur modèle à 2 régresseurs est un petit peu plus loin de la droite mais le 2^{ème} meilleur commence à s'éloigner .
 \Rightarrow on retient le meilleur modèle à 2 régresseurs qui sont PR et PA

Dans cet exemple on arrive au même modèle qu'en sélection descendante mais ce n'est pas toujours le cas.

Comment choisir un bon modèle ?

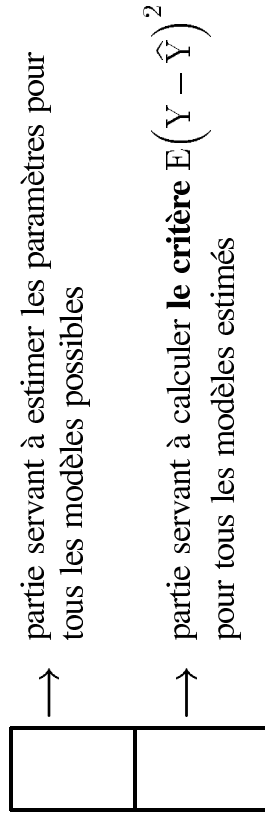
L'heuristique proposée avec le C(p) de Mallows est une bonne technique de choix

MAIS il existe

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

⇒ par exemple :

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



AIC = Akaike's information criterium
BIC = Bayesian information criterium
MSEP = Mean Square Error of Prediction

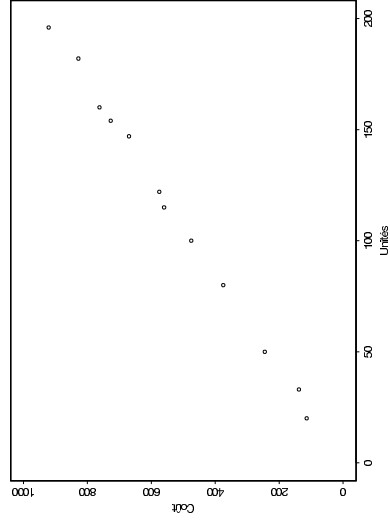
Ordonnée à l'origine nulle

Exemple : Les ateliers

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

Les données

atelier	unités	coûts (\$)
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



La relation semble linéaire entre unité et coût.

On va appliquer un modèle linéaire M_1 : coût = $\theta_0 + \theta_1 \times$ unités + ε

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)

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

$H_0 : \theta_1 = 0$

↓

$H_1 : \theta_1 \neq 0$

$M_0 : \mu_n = \theta_0$

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			

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

Modèle avec Intercept

Le modèle "standard" est testé comme le modèle emboîté le plus simple $\mu_n = \theta_0$

On voit cependant que le paramètre θ_0 (Intercept) n'est pas significativement différent de zéro.

De plus, il est logique que le coût soit nul pour zéro unités produites.

⇒ on a envie de tester le modèle sans intercept $M_1 : \mu_n = \theta_1 Z_{1n}$

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 : Prob> T	Parameter=0
UNITE	1	4.685274	0.03420502	136.976	0.0201

Attention le modèle le plus simple auquel on peut le comparer ne peut être $M_0 : \mu_n = \theta_0$ car ce modèle n'est pas emboîté dans M_1 . SAS teste M_1 par rapport au modèle $M_0 : \mu_n = 0$
 \Rightarrow La ligne U Total pour Uncorrected total est alors $\sum_n Y_n^2$

L'estimation du paramètre θ_1 est valable, ainsi que son intervalle de confiance, sous le **Postulat** (c'est alors un postulat) que **l'ordonnée à l'origine est nulle**

Une Heuristique

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

- Y = volume de bois fort (jusqu'au fin bout)
- variables disponibles : Hauteur et Diamètre à 1.30 m
- N = 50 arbres

Comment choisir les variables pertinentes dans une régression linéaire ?

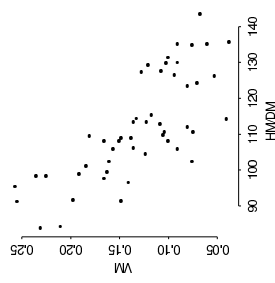
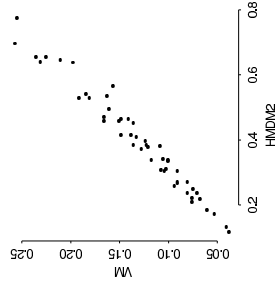
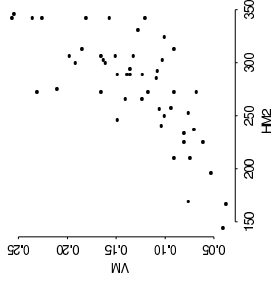
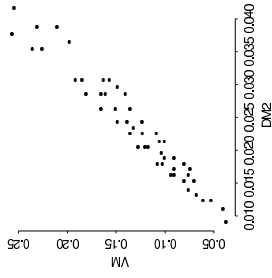
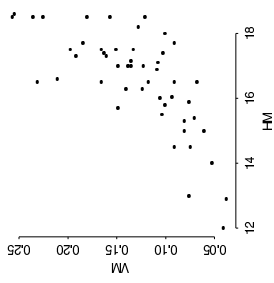
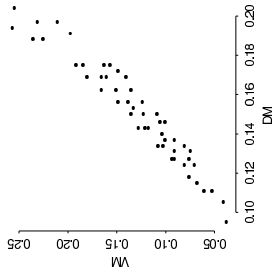
Les variables explicatives sur lesquelles on dispose de données sont elles adaptées au problème ?

En particulier leur relation avec les variables expliquées est-elle linéaire ?

Des changements ou des combinaisons de variables (tout en restant dans le modèle linéaire) sont-ils souhaitables ?
Lesquels ?

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

Première étape



Graphe de la variable à expliquer (**VM**) en fonction des régresseurs (**HM** et **DM**) et de quelques combinaisons de **HM** et **DM**

DM et **HM** montrent une relation avec **VM** qui ne semble pas linéaire mais quadratique

⇒ on essaie HM^2 et DM^2 : la relation s'améliore grandement avec DM^2 , pas avec HM^2

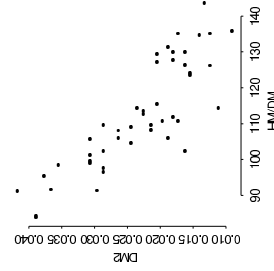
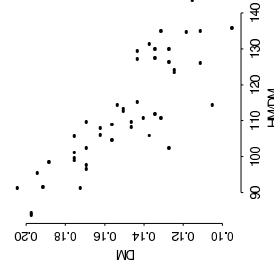
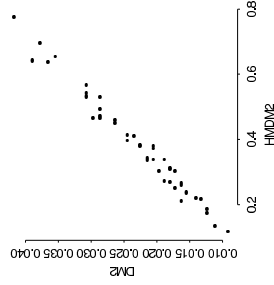
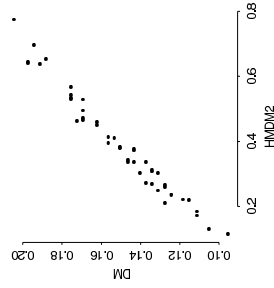
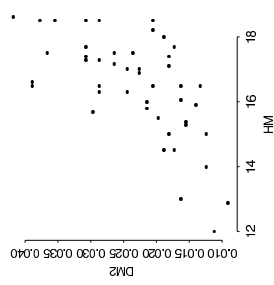
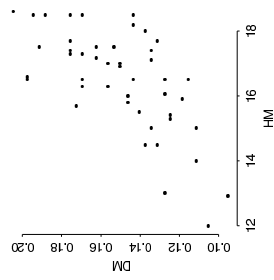
Un volume étant le produit d'une section par une hauteur, corrigé par un coefficient de forme, on essaie $HM \times DM^2$ qui donne apparemment la meilleure relation avec **VM** ; et aussi HM/DM qui montre une relation plus floue avec **VM**.

Deuxième étape

On représente les relations entre les différents régresseurs possibles afin de détecter d'éventuelles colinéarités

— \exists relation forte entre $HM \times DM^2$ et DM ainsi qu'entre $HM \times DM^2$ et DM^2

— il est inutile, voire néfaste, d'introduire ces variables dans le même modèle.



Troisième étape

		$R^2 = 0,9442$		$R^2_{adj} = 0,9430$			
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		
		$R^2 = 0,9574$		$R^2_{adj} = 0,9536$			
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		
		$R^2 = 0,9605$		$R^2_{adj} = 0,9580$			
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		
		$R^2 = 0,9604$		$R^2_{adj} = 0,9587$			
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		
		$R^2 = 0,9557$		$R^2_{adj} = 0,9650$			
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		

On teste les différents modèles, en regroupant les variables en groupes où les colinéarités sont faibles (tous les modèles sont strictement significatifs)

(1) à (4) : groupe de variables **DM2, HM, HM/DM**

— Le modèle **VM=f(DM2)** s'améliore si on ajoute **HM**

— Si on ajoute encore **HM/DM**, le paramètre de **HM** n'est plus significatif

On retient donc le modèle **VM=f(DM2, HM/DM)**

(5) : on teste un modèle différent, avec une seule variable **HMDM2**, selon les indications des graphes précédents.

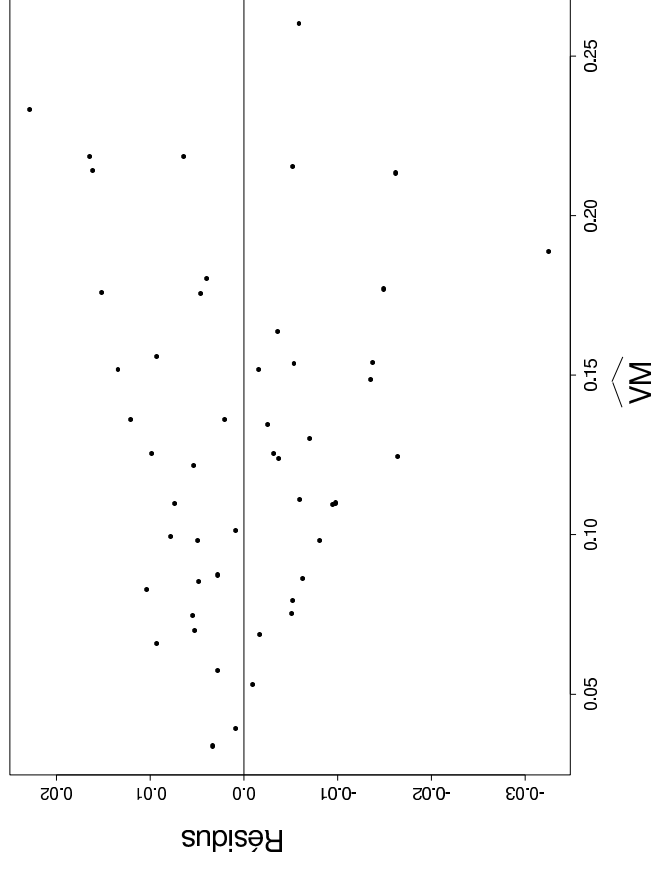
C'est apparemment le meilleur modèle

Remarque :

l'ordonnée à l'origine (intercept) n'est pas significativement différente de 0

Quatrième étape

Modèle : $VM = a + b \text{HMDM2}$



On fait le graphe des résidus ($\hat{\varepsilon}$) du modèle retenu ($VM=f(\text{HMDM2})$) en fonction de la variable prédite VM

- La variabilité semble augmenter quand \widehat{VM} augmente
- Le postulat d'homoscédasticité n'est pas vérifié

Cinquième étape

Model : MODEL1

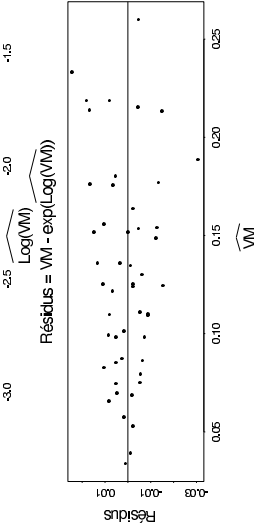
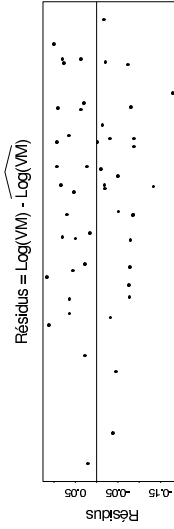
Dependent Variable : LOGVM

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> 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



Le modèle $\text{VM} = f(\text{HMDM2})$ ne respecte pas le postulat d'homoscedasticité. De plus son "intercept" est proche de zéro

⇒ On a envie d'appliquer un modèle multiplicatif du type $\text{VM} = a \text{DM}^b \times \text{HM}^c$ en utilisant la transformation Log pour chaque variable expliquée ou explicative, on a le modèle linéaire suivant :

$$\text{Log}(\text{VM}) = a' + b \text{Log DM} + c \text{Log}(\text{HM})$$

— Les résidus du 1er graphe ne montrent pas de structure. Les postulats sont respectés

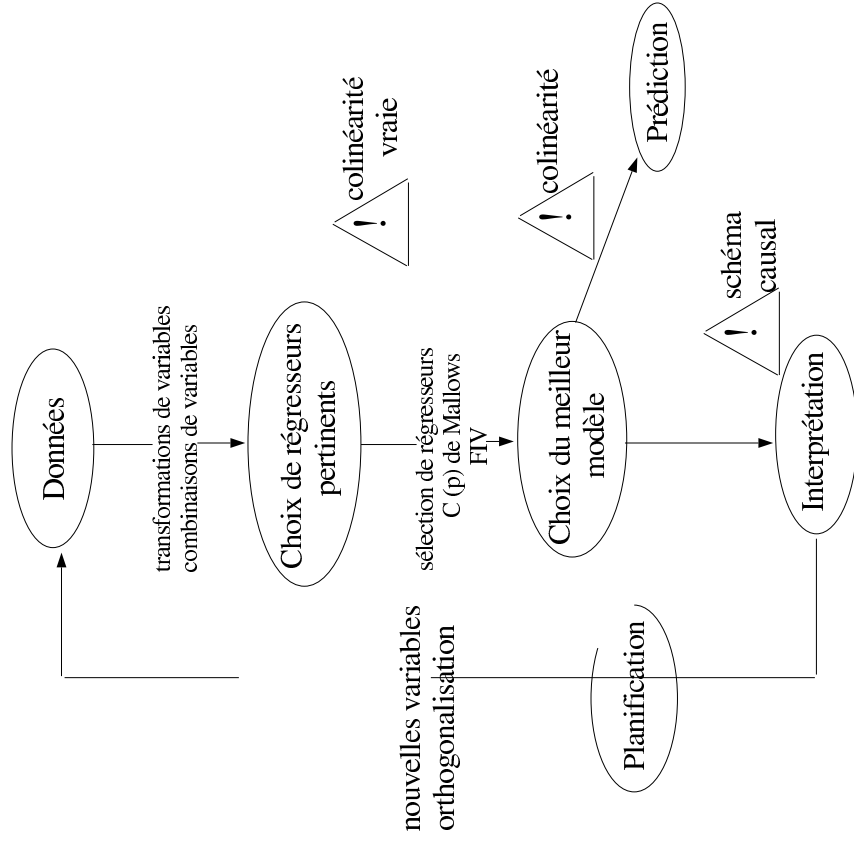
— Les paramètres de Log (DM) et de Log (HM) sont proches respectivement de 2 et de 1 (ils n'en sont pas significativement différents)

Tout est parfait sauf que, quand on revient aux variables initiales, on introduit un biais important car la transformation n'est pas linéaire.

⇒ se méfier en particulier des intervalles de confiance des volumes estimés. Les résidus du 2ème graphe présentent d'ailleurs le même aspect que ceux du modèle $\text{VM} = f(\text{HMDM2})$

Régression linéaire

Conclusion



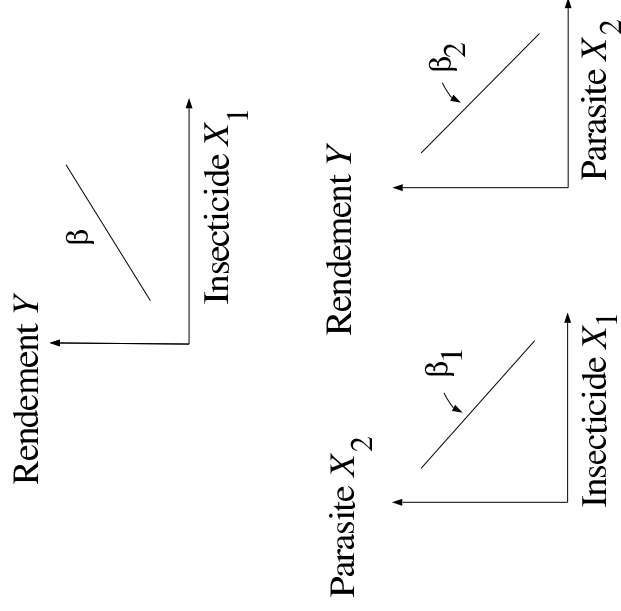
Choix des régresseurs pertinents :

- linéarité de la relation variable expliquée/régresseur
⇒ transformation de variable éventuelle
- homoscedasticité (cas fréquent de la variance augmentant avec la moyenne)
⇒ transformation de variable
- interprétables en relation simple avec la réalité (ou la théorie) physique ou biologique

Attention toujours à la colinéarité

Attention aussi à l'interprétation causale

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



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

Attention à l'interprétation d'un modèle

L'insecticide X_1 semble avoir un effet favorable sur le rendement .

En fait il y a une variable cachée liée aux deux autres : la taille de la population de ravageurs X_2