Plan

STRATÉGIE DE CHOIX DE MODÈLE

- non orthogonalité
- instabilité des paramètres

DEUX EXEMPLES :

- Armillaire
- folle avoine

Le but de ce chapitre est de présenter des exemples réels de stratégie de choix de modèle, et de mettre en évidence les deux difficultés principales que sont la non orthogonalité d'un modèle, et le risque d'instabilité des paramètres. L'exemple Armillaire illustrera principalement la première difficulté, et l'exemple folle avoine la seconde.

Exemple armillaire, objectif

Influence du pH du sol
sur
l'abondance de l'Armillaire

Champignon décomposeur de bois
et
parasite opportuniste

⇒ colonisation du bois mort par l'Armillaire

Un exemple réel de choix de modèle est présenté, l'exemple Armillaire. Une stratégie de choix de modèle est proposée. Les problèmes liés à la non orthogonalité, déjà discutés dans le cadre du module "modèle linéaire" seront discutés.

L'Armillaire est un champignon basidiomycète lignivore qui est un des principaux décomposeurs de bois en forêt.

On s'y intéresse car c'est un parasite opportuniste qui intervient dans le dépérissement de l'Erable à sucre. Il est capable de tuer des arbres préalablement affaiblis par d'autres facteurs.

Le dépérissement de l'Erable à sucre est plus accentué sur les sols les plus pauvres, avec un pH acide.

⇒ Y a-t-il davantage d'Armillaire sur ces sols ?

Étude de la colonisation du bois mort par l'Armillaire (1996)

19 bois morts examinés

échantillonnés les plus proches de 19 points déjà matérialisés sur le terrain.

Variable binaire :
→ colonisé par l'Armillaire
→ non colonisé par l'Armillaire

1				13
	2	8	14	
	3	9	15	
	4	10	16	
	5	11	17	
	6	12	18	
7				19

Échantillonnage sur une parcelle

La variable étudiée est la variable groupée mesurée sur chaque parcelle : "nb bois morts colonisés / nb regardés".

On suppose que cette variable suit une loi binomiale : ceci suppose qu'il y a indépendance entre les différentes pièces de bois mort d'une même parcelle, pour ce qui est de la colonisation par l'Armillaire. Cette supposition n'est pas excessive car on ne note pas de structure spatiale particulière dans la distribution des pièces de bois mort colonisées ou non par l'Armillaire.

Le plan d'échantillonnage choisi, et représenté sur le schéma est un plan systématique.

Exemple Armillaire, dispositif

1985 : dispositif

4 blocs (Bl)
X
3 facteurs :

- Chaulage (Ch)
- Engrillagement (Gr)
- Herbicide (He)

+ leurs combinaisons (8 par bloc)
⇒ **32 parcelles en tout**

Coupe de régénération :
50 % de la surface terrière enlevée.

Ce dispositif s'insère dans une étude plus large sur le dépérissement et la régénération de l'Erable à sucre (en Pennsylvanie). Trois facteurs sont étudiés :

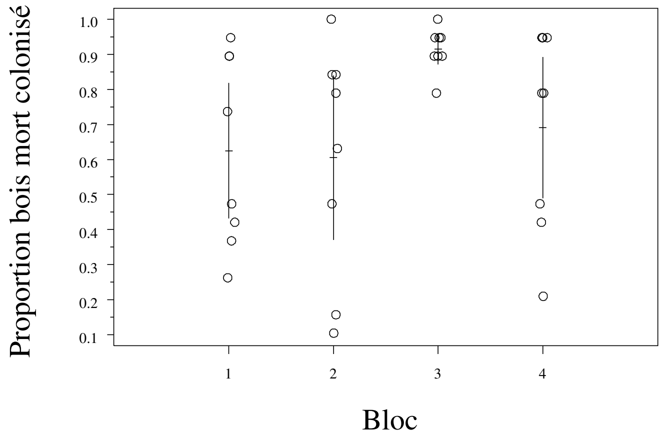
- chaulage : les sols très acides ne sont pas propices à l'Erable à sucre. Le chaulage est destiné à augmenter le pH de ces sols.
- engrillagement : une forte population de chevreuils broute les régénérations et entraîne une modification du sous-bois néfaste à la régénération des érables. L'engrillagement sert à protéger les régénérations.
- herbicide : le traitement herbicide vise à modifier la nature du sous-bois pour le rendre plus propice aux semis d'érables.

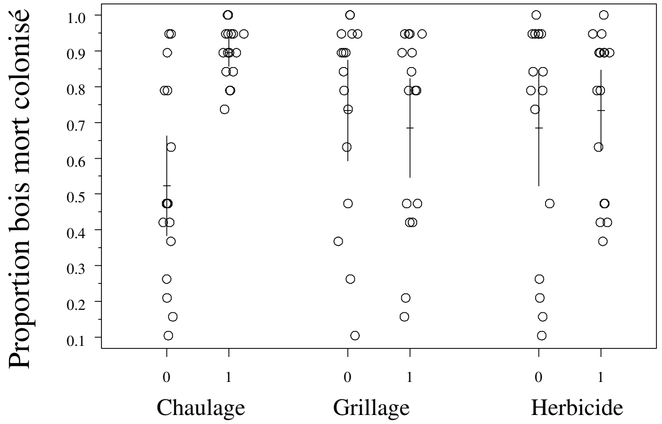
L'effet de ces traitements sur l'abondance de l'Armillaire ne nous intéresse pas directement. Mais il en a été tenu compte car ils entraînent une modification drastique de la végétation.

Le dispositif est orthogonal.

Surface terrière : $\sum$ (surfaces des troncs des arbres).

Impact des facteurs sur la colonisation du bois mort



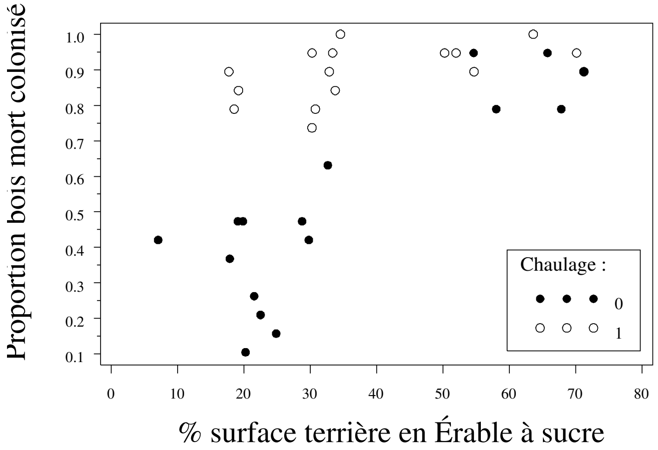


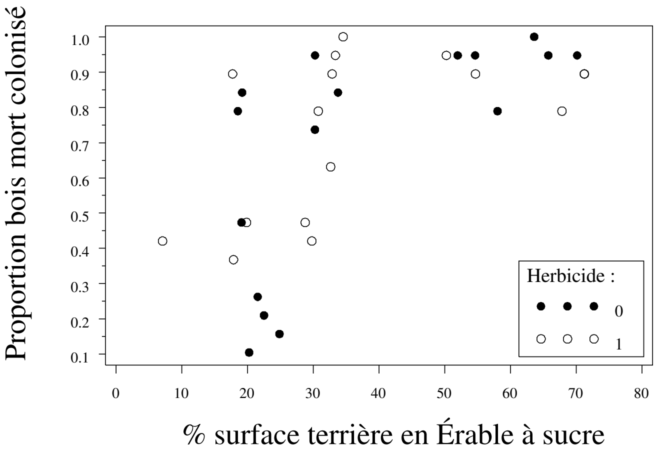
Les observations faites au niveau de chaque parcelle sont représentées, ainsi que les moyennes et leur intervalle de confiance à 95 %.

La colonisation du bois mort par l'Armillaire est globalement plus élevée dans le bloc 3, elle est moins variable également.

A première vue, le chaulage semble avoir un effet sur la colonisation du bois mort par l'Armillaire.

Impact des facteurs sur la colonisation du bois mort





La colonisation par l'Armillaire varie aussi avec le pourcentage de surface terrière en Erable à sucre.

Type et quantité de bois mort présent

Forte influence du type de bois mort

	Origine du bois mort	
	Erable à sucre	Autres essences
Colonisation par l'Armillaire	83 %	55 %

Prise en compte?

- Proportion d'érable à sucre dans le peuplement (PropES).
- Surface terrière d'érable à sucre enlevée en 85 (EScoupe).
- Surface terrière totale enlevée en 85 (Totcoupe).

Comment peut-on tenir compte de la quantité et du type de bois mort présent dans les parcelles?

Lors de la prise de données, on remarque que le bois mort provenant d'érables à sucre est colonisé plus fréquemment que le bois mort provenant d'autres essences. Type de bois mort et quantité sont donc des variables apparemment importantes, dont on décide de tenir compte dans le modèle.

La quantité de bois mort au sol n'a pas été mesurée directement, mais une bonne proportion du bois mort a été généré au moment de l'éclaircie de 1985 et on connaît la surface terrière totale enlevée lors de cette coupe (totcoupe), de même que la surface terrière en érable à sucre enlevée (EScoupe).

Il y a eu aussi du dépérissement dans les parcelles, donc une autre source de bois mort. On peut se dire que la proportion de bois mort provenant d'érables à sucre dépend de la proportion d'érable à sucre dans le peuplement (PropES).

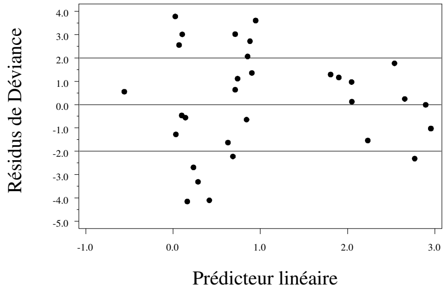
Type et quantité de bois mort présent

Choix du régresseur :

	Déviance résiduelle	Chi2	DDl	Pr>Chi
Modèle nul	229.95			
Totcoupe	216.34	13.61	1	0.0002
Modèle nul	229.95			
EScoupe	174.39	55.56	1	0.0000
Modèle nul	229.95			
PropES	138.25	**91.70**	1	0.0000

Chacun des trois modèles à une variable explicative (Totcoupe, EScoupe, PropES) est testé par rapport au modèle nul (test du rapport de vraisemblance; c'est le lien logit qui est choisi).

Chacune des trois variables permet une réduction significative de la déviance. PropES, la proportion d'érable à sucre dans le peuplement est celle qui apporte la plus forte diminution de déviance. Un examen des données montre que c'est en partie parce que le bois mort est très fréquemment colonisé dans le bloc 3 qui comporte beaucoup d'érable à sucre.



Représentation des résidus pour le modèle utilisant PropES comme régresseur.

⇒ une transformation de PropES ne semble pas s'imposer.

Modèle complet

Class Level Information

Class	Levels	Values
GR	2	0 1
CH	2	0 1
HE	2	0 1
BL	4	1 2 3 4

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	12	16.6245	1.3854
Pearson Chi-Square	12	17.3018	1.4418
Log Likelihood	.	-260.0528	

LR Statistics For Type 3 Analysis

Source	DF	ChiSquare	Pr>Chi
PROPES	1	4.5236	**0.0334**
BL	3	1.1781	0.7583
CH	1	59.4951	**0.0000**
CH*BL	3	5.1135	0.1637
HE	1	0.2064	0.6496
HE*BL	3	13.5797	**0.0035**
GR	1	0.0748	0.7844
GR*BL	3	3.0955	0.3771
GR*HE	1	0.9237	0.3365
CH*HE	1	0.2634	0.6078
GR*CH	1	0.4574	0.4988

Les n_i étant égaux à 19, donc pas trop petits, on peut utiliser la déviance pour valider le modèle. La déviance vaut 16.6 (ddl=12, p=0.16). Le modèle complet est donc accepté.

La proportion d'érable à sucre dans la parcelle, le chaulage et l'interaction herbicide * bloc ont un effet significatif.

Ce sont des statistiques du rapport de vraisemblance de type 3 qui sont reportées : elles correspondent à des statistiques du rapport de vraisemblance de type 1, lorsque le terme testé est placé en dernier dans le modèle (même un effet principal est placé après les termes d'interaction dans lesquels il intervient).

Corrélation entre paramètres

Modèle avec PropES et bloc uniquement :

Estimated Correlation Matrix

	PropES	Bl1	Bl2	Bl3
PropES	1.000			
Bl1	-0.099	1.000		
Bl2	-0.044	0.561	1.000	
Bl3	**-0.756**	0.349	0.240	1.000

⇒ introduction séquentielle des 2 variables :

LR Statistics For Type 1 Analysis

Source	Déviance	DF	ChiSquare	Pr>Chi
Intercept	229.9			
PropES	138.2	1	91.7	**0.0001**
Bl	135.2	3	3.0	**0.39**
Intercept	229.9			
Bl	178.3	3	51.6	**0.0001**
PropEs	135.2	1	43.1	**0.001**

Il y a une forte corrélation entre les variables PropES et Bloc, due à la forte colonisation du bois mort dans le bloc 3, là où il y a beaucoup d'érable à sucre. Cela entraîne une **corrélation élevée entre les paramètres** correspondants.

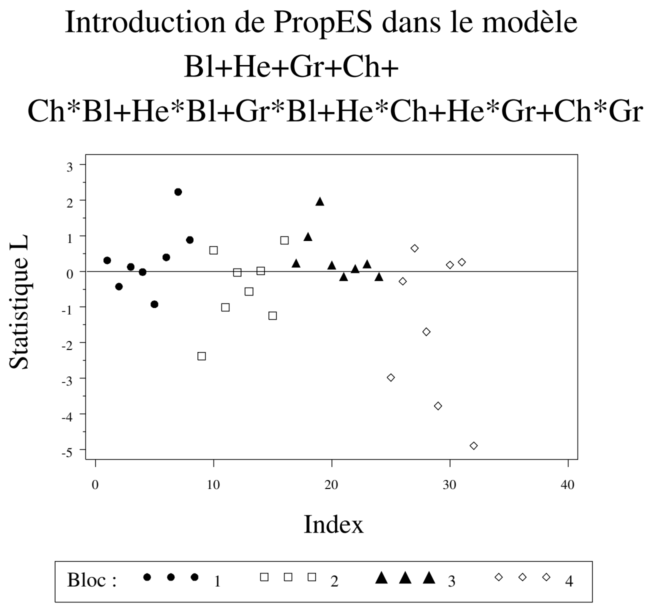
Cela a aussi une autre conséquence : regardons le tableau des diminutions de déviance associées à l'ajout de chacune des variables lorsqu'elles sont introduites successivement. La diminution de déviance entraînée par l'ajout du bloc au modèle nul est significative. Elle ne l'est plus, lorsque le bloc est ajouté au modèle contenant la variable PropES. La diminution de déviance entraînée par l'ajout de la variable PropES au modèle nul est supérieure à la diminution de déviance entraînée par l'ajout de PropES au modèle comprenant le facteur bloc.

On voit sur cet exemple que **l'ordre d'introduction des termes dans le modèle a une grande importance**. L'examen d'une unique table d'analyse de déviance est insuffisant, et risque de conduire à des erreurs d'interprétation, lorsque certaines des corrélations entre les paramètres sont non négligeables. Une solution peut être de tester les termes d'intérêt par des tests de sous-modèle (test du rapport de vraisemblance obtenu en comparant le modèle complet avec le sous-modèle ne contenant pas le terme à tester).

Ce phénomène de **non orthogonalité** n'est pas nouveau : il a déjà été abordé dans le module "modèle linéaire". Le GLM présente cependant une particularité par rapport au modèle linéaire. Avec un modèle linéaire, lorsque le plan est orthogonal, toutes les corrélations entre les paramètres sont nulles, et la table d'analyse de variance est unique, la décomposition des sommes de carrés ne dépend pas de l'ordre d'introduction des termes dans le modèle. Avec un GLM, même si le plan est orthogonal, les corrélations entre les paramètres ne sont pas nulles, et la table d'analyse de déviance n'est pas unique. Plus les corrélations entre les paramètres sont élevées, plus la table d'analyse de déviance varie selon l'ordre d'introduction des termes dans le modèle.

Statistique L

Contribution des observations à la déviance associée à PropES



On représente la contribution de chaque parcelle à la déviance associée au terme PropES ajouté au modèle complet :
Nb colonisés / nb total =
Bl+He+Gr+Ch+Ch*Bl+He*Bl+Gr*Bl+He*Ch+He*Gr+Ch*Gr.

3 parcelles du bloc 4 ont une contribution élevée. Le bloc 4 est le plus hétérogène, pour ce qui est de la quantité d'érable à sucre dans les parcelles. Dans le bloc 3, où la quantité d'érable à sucre est élevée, ou dans les blocs 1 et 2, où elle est uniformément faible, l'introduction de la variable PropES n'améliore pas l'estimation.

A posteriori, on peut dire que les blocs n'ont pas été si bien choisis, en particulier le bloc 4 qui est hétérogène.

Remarque : si on enlève les 3 parcelles du bloc 4, dont la contribution est élevée, PropES n'apporte pas d'explication significative, lorsqu'ajoutée au modèle complet. Par contre, l'ajout de PropES au modèle nul, ou au modèle "BLOC" (cf. transparent précédent) apporte une explication significative.

Rappel : la statistique L est la différence des carrés des résidus de vraisemblance entre deux modèles emboîtés.

Second modèle proposé

Class Level Information

Class	Levels	Values
CH	2	0 1
HE	2	0 1

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	26	30.1926	1.1613
Pearson Chi-Square	26	29.0665	1.1179
Log Likelihood	.	-266.8369	

LR Statistics For Type 3 Analysis

Source	DF	ChiSquare	Pr>Chi
PROPES	1	37.5531	0.0000
HE	1	15.9413	0.0001
HE*PROPES	1	12.1970	0.0005
CH	1	35.8300	0.0000
CH*PROPES	1	6.2152	0.0127

Un examen des données montre que les blocs se différencient essentiellement par la proportion d'érable à sucre. De plus, dans le bloc 4, la proportion d'érable à sucre est variable. Un modèle où le facteur bloc serait remplacé par la variable PropES serait plus intéressant, plus informatif.

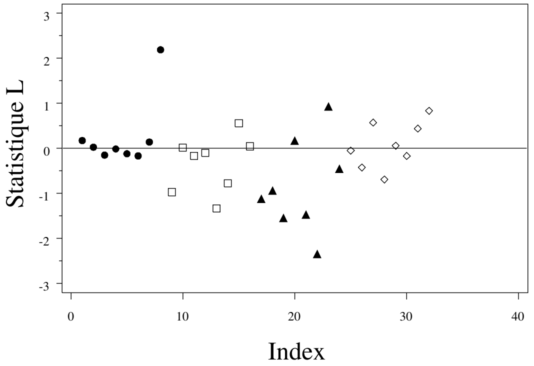
Le modèle PropES+He+He*PropES+Ch+Ch*PropES est donc ajusté. Avec un rapport déviance/ddl de 1.16, il est accepté (déviance = 30.2, ddl=26, p=0.26).

A priori, au vu du rapport déviance/ddl, il n'y a pas de surdispersion.

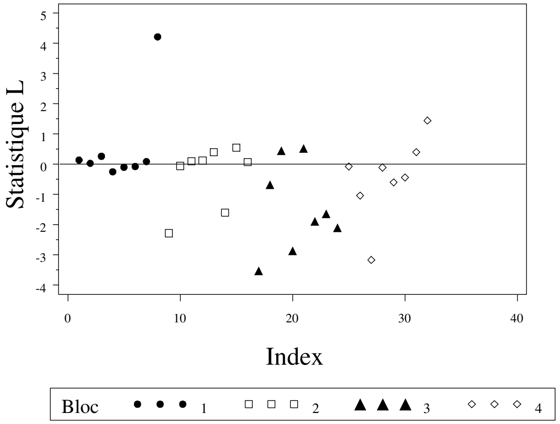
⇒ L'hypothèse d'indépendance entre les bois morts d'une même parcelle pour la colonisation par l'Armillaire est justifiée *a posteriori*.

Validation du modèle

Introduction de PropES*Ch dans le modèle comprenant
PropES, Ch, He, et PropES*He



Introduction de Propes*He dans le modèle comprenant
PropES, Ch, He et PropES*Ch



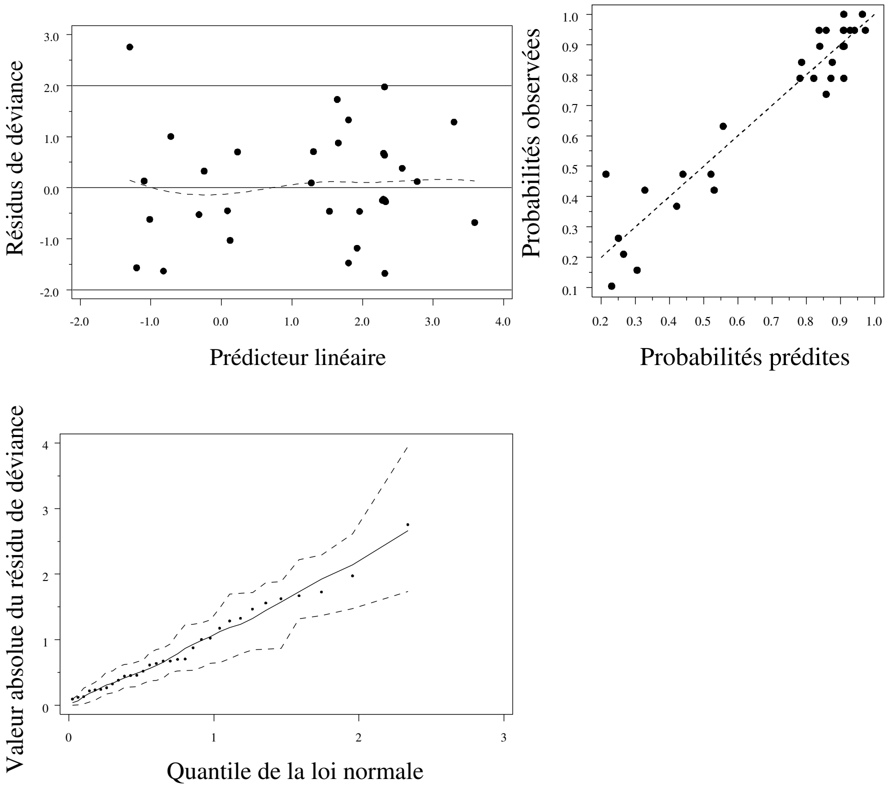
La contribution de chaque parcelle à la diminution de déviance occasionnée par l'ajout de l'interaction PropES*Ch au modèle PropES+Ch+He+PropES*He est représentée, ou par l'ajout de PropES*He au modèle PropES+Ch+PropES*Ch.

La dernière parcelle du bloc 1 se distingue nettement des autres. Le fait de supprimer cette parcelle renforce encore la significativité des interactions PropES*Ch et PropES*He.

Ces graphes confirment donc la réalité des interactions PropES*Ch et PropES*He.

Remarque : ces représentations graphiques constituent un outil (parmi d'autres) de validation de modèle, dans la mesure où elles permettent de s'assurer qu'aucun des termes du modèle ne doit sa significativité à une seule observation (une seule parcelle).

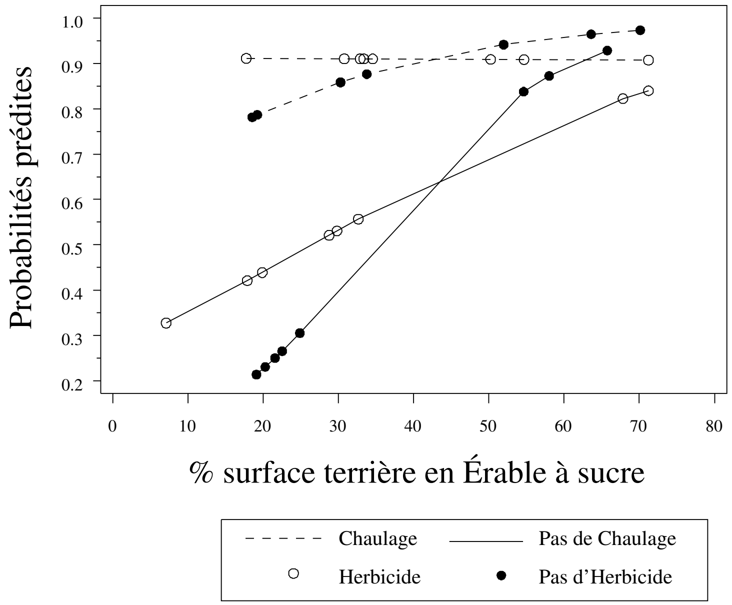
Validation du modèle



Le graphique des résidus de déviance en fonction du prédicteur linéaire et le graphique semi-gaussien des résidus semblent aussi indiquer que le modèle est correct.

Les probabilités de colonisation prédites sont en bonne adéquation avec les probabilités de colonisation observées.

Résultats



La proportion estimée de débris ligneux colonisés est représentée en fonction de la proportion de surface terrière en érable à sucre (PropES). Le chaulage augmente assez fortement la colonisation du bois mort dans les parcelles où il y a peu d'érable à sucre, mais pas là où il y en a beaucoup (bloc 3 et partie du bloc 4). Le traitement herbicide augmente la colonisation du bois mort, mais seulement dans les parcelles où il y a peu d'érable à sucre.

Données

Bloc	Parcelle	Ch	He	Gr	Totcoupe	EScoupe	PropES	nb. colonises	nb. observes
1	11	1	0	0	36.7	12.5	30	14	19
1	12	1	1	0	49.0	16.3	33	17	19
1	13	0	1	0	90.1	20.2	18	7	19
1	14	0	0	0	70.5	17.5	22	5	19
1	15	1	0	1	54.5	32.0	30	18	19
1	16	1	1	1	61.1	11.8	18	17	19
1	17	0	1	1	60.7	29.0	30	8	19
1	18	0	0	1	44.0	5.8	19	9	19
2	21	0	0	0	70.9	12.0	20	2	19
2	22	1	1	0	63.3	17.4	35	19	19
2	23	0	1	0	86.7	29.8	33	12	19
2	24	1	0	0	49.6	18.6	34	16	19
2	25	1	0	1	76.2	5.0	19	16	19
2	26	0	0	1	61.6	11.5	25	3	19
2	27	1	1	1	84.6	24.4	31	15	19
2	28	0	1	1	74.1	15.2	29	9	19
3	31	0	0	1	77.5	39.2	55	18	19
3	32	1	1	1	54.6	37.2	55	17	19
3	33	1	0	1	66.0	53.6	70	18	19
3	34	0	1	1	75.1	51.4	68	15	19
3	35	0	1	0	44.7	32.0	71	17	19
3	36	1	1	0	67.0	44.7	71	17	19
3	37	1	0	0	78.9	49.8	64	19	19
3	38	0	0	0	85.9	65.5	66	18	19
4	41	1	1	1	61.2	4.3	33	18	19
4	42	0	0	1	109.4	17.9	23	4	19
4	43	0	1	1	73.5	0.1	7	8	19
4	44	1	0	1	101.8	15.1	19	15	19
4	45	0	1	0	77.5	18.1	20	9	19
4	46	1	0	0	73.8	36.5	52	18	19
4	47	1	1	0	46.2	39.9	50	18	19
4	48	0	0	0	77.2	45.1	58	15	19

Ceux qui auraient le courage de refaire les calculs à partir des données présentées ici ne trouveront pas exactement les mêmes résultats, car certaines données ont été arrondies.

Matrice des corrélations entre paramètres (Modèle complet)

	PROPES	Bl1	Bl2	Bl3	He0	He0* Bl1	He0* Bl2	He0* Bl3	Gr0	Gr0* Bl1	Gr0* Bl2	Gr0* Bl3
Bl1	-0.5											
Bl2	-0.4	0.6										
Bl3	**-0.8**	0.6	0.6									
He0	-0.3	0.4	0.5	0.6								
He0*Bl1	**0.7**	-0.7	-0.5	-0.6	-0.6							
He0*Bl2	**0.6**	-0.5	-0.6	-0.6	-0.6	0.7						
He0*Bl3	**0.5**	-0.4	-0.4	-0.6	-0.5	0.6	0.6					
Gr0	**-0.7**	0.4	0.3	0.6	0.3	-0.5	-0.4	-0.4				
Gr0*Bl1	**0.6**	-0.5	-0.3	-0.5	-0.2	0.4	0.4	0.3	-0.7			
Gr0*Bl2	**0.7**	-0.4	-0.5	-0.6	-0.3	0.5	0.4	0.4	-0.8	0.7		
Gr0*Bl3	0.4	-0.3	-0.2	-0.5	-0.2	0.3	0.2	0.2	-0.6	0.5	0.5	
Ch0	0.2	0.3	0.4	0.2	0.3	0.0	0.0	0.0	-0.1	0.2	0.2	0.1
Ch0*Bl1	-0.2	-0.5	-0.3	-0.1	-0.0	0.1	-0.0	-0.0	0.2	-0.2	-0.2	-0.1
Ch0*Bl2	-0.1	-0.3	-0.6	-0.1	-0.1	-0.0	-0.0	-0.0	0.2	-0.2	-0.1	-0.1
Ch0*Bl3	-0.4	-0.1	-0.2	-0.1	-0.1	-0.2	-0.1	-0.2	0.3	-0.3	-0.3	-0.2
Ch0*He0	-0.2	0.1	-0.1	-0.0	-0.1	-0.1	-0.1	-0.1	0.2	-0.1	-0.1	-0.1
Ch0*Gr0	0.4	-0.2	-0.1	-0.3	-0.1	0.3	0.3	0.2	-0.7	0.3	0.3	0.2
He0*Gr0	-0.4	0.2	0.1	0.2	-0.2	-0.3	-0.2	-0.2	-0.0	-0.2	-0.2	0.0

	Ch0	Ch0* Bl1	Ch0* Bl2	Ch0* He0	Ch0* Gr0	He0* Gr0
Ch0*Bl1	-0.6					
Ch0*Bl2	-0.6	0.6				
Ch0*Bl3	-0.6	0.5	0.5			
Ch0*He0	-0.5	0.1	0.2	0.3		
Ch0*Gr0	-0.1	-0.1	-0.2	-0.2	-0.1	
He0*Gr0	-0.1	0.1	0.1	0.2	0.1	-0.0

Surparamétrisation et instabilité des paramètres

facteur A	Proportion
1	$\boxed{0/4}$
2	5/6

$$\begin{aligned} logit(p) &= \beta_0, & \text{si A=1} \\ &= \beta_0 + \alpha_2, & \text{si A=2} \end{aligned}$$

	estimation	écart-type
β_0	$\boxed{-12.37}$	$\boxed{146.9}$
α_2	$\boxed{13.97}$	$\boxed{146.9}$

Pourquoi ?

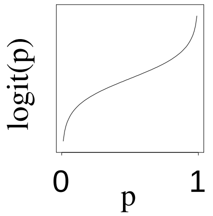
$$\left.\begin{array}{c}\text{jeu de données particulières}\\ \text{surparamétrisation}\end{array}\right\} \Rightarrow \text{instabilité}$$

Avant de passer à un autre exemple réel de choix de modèle, montrons, à partir d'un tout petit exemple fictif, l'origine d'un problème fréquent, celui de l'instabilité des paramètres.

Pourquoi ?

- $\widetilde{p}_1 = 0$, or $\log(\frac{\widehat{p}_1}{1-\widehat{p}_1}) = \widehat{\beta}_0 \Rightarrow \widehat{\beta}_0 = -\infty$

  qd $\widehat{p}_1 \to 0$, $\boxed{\widehat{\beta}_0 \to -\infty}$
- répercussion sur α_2.
  α_2 très grand. Pourquoi ?
  pour compenser la grande valeur négative de β_0
  (car $\beta_0 + \alpha_2$ est fini) .
- les variances des paramètres sont infinies. En effet, var$(\widehat{\beta}_0)$ = var(logit$(\widetilde{p})$). La dérivée de logit$(\widetilde{p})$, lorsque $\widetilde{p} \to 0$, tend vers $+\infty$.

  C'est un problème d'instabilité des paramètres. Ici, le problème vient du fait que toutes les observations sont égales à 0, pour le niveau 1 du facteur A, autrement dit il vient du fait que la proportion observée est nulle. Le même problème est observé dans le cas d'une proportion observée exactement égale à 1.

  Une instabilité des paramètres peut provenir du jeu de données (données creuses), ou d'une surparamétrisation (modèle trop complexe, par rapport aux données).

Surparamétrisation

Quoi ?

- trop grand nombre de paramètres ...

Quand ?

- données creuses
- modèle trop complexe

Diagnostic

- message de non convergence (éventuellement).
- $\left\{\begin{array}{lll} \text{certains} & \widehat{\text{paramètres}} & \text{très grands.} \\ \text{certains} & \widehat{\text{écarts-type}} & \text{très grands.} \end{array}\right.$

Conséquences

- paramètres instables
- estimations peu précises

Qu'est-ce que c'est ?

- trop grand nombre de paramètres ... par rapport au nombre de données (un modèle qui reproduit tout ou partie des données avec autant de paramètres qu'il y a de données est peu utile; "Un modèle qui prend en compte toute la variabilité n'est pas plus utile qu'une carte à l'échelle 1/1" J. ROBINSON).

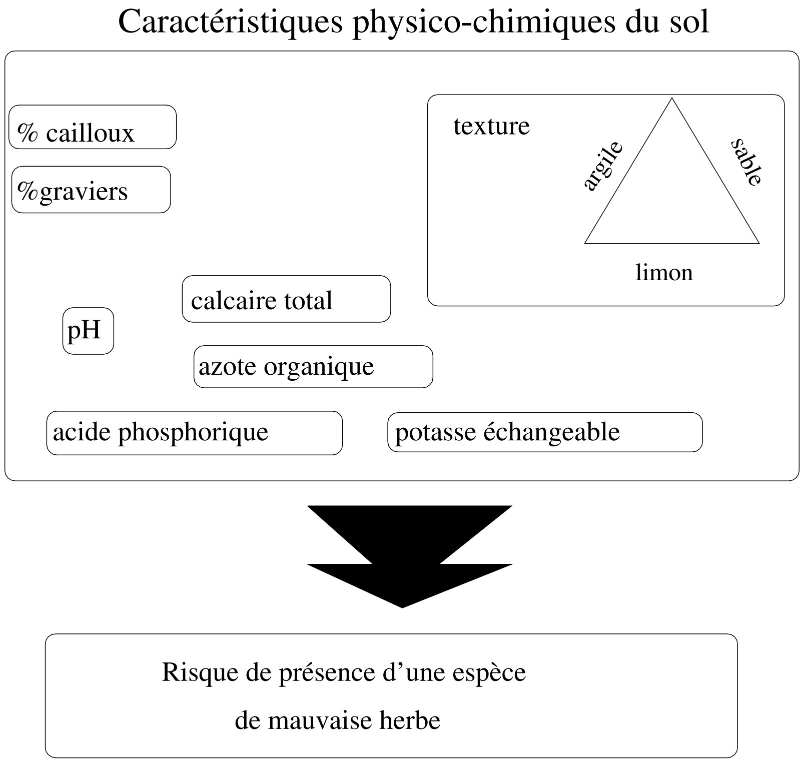
Quand cela peut-il arriver ?

- données creuses (sparses) : si certaines combinaisons des niveaux des facteurs explicatifs (certains traitements) donnent peu d'observations binaires.
- modèle trop complexe : un trop grand nombre de variables explicatives, de termes d'interaction (d'ordre trop élevé) sont considérés dans le modèle.

conséquences :

- paramètres instables : une faible modification des données entraîne une grande modification des paramètres estimés.
- estimations peu précises : grands écarts-type.

Exemple malherbologie, objectif



On présente l'exemple Malherbologie dans le but d'illustrer le problème de l'instabilité des paramètres, par un exemple réel. C'est aussi l'occasion de montrer un autre exemple de stratégie de choix de modèle.

On cherche à savoir si la présence d'une espèce de mauvaises herbes dans une parcelle peut être expliquée et prédite, à l'aide des caractéristiques physico-chimiques du sol. On utilise pour cela la méthode du GLM. Si une telle prédiction était possible, elle pourrait être utile pour décider, en fonction du type de sol, du type de traitement, mécanique ou chimique, et pour réduire les quantités d'herbicide utilisées.

On s'intéresse à l'espèce folle avoine. Les caractéristiques physico-chimiques suceptibles de prédire le risque de présence de l'espèce sont :

- le % de cailloux et de graviers,
- la texture, qui tient compte des % d'argile, de sable et de limon. La texture est déterminée au moyen d'un diagramme en triangle, représentant les pourcentages d'argile, de limon, et de sable. Les textures sont codées de 1 à 17 selon les proportions. Elles peuvent être ordonnées de la plus fine (1 : argile lourde) à la plus grossière (17 : sable), en échangeant l'ordre de quelques unes des textures (4 et 5, 7 et 8 par exemple). cCette modification de l'ordre a été faite, suite à discussion avec le spécialiste.
- le ph, qui n'est pas mesuré s'il y a présence de calcaire, qui est mesuré sinon.
- le calcaire total,
- l'azote organique,
- l'acide phosphorique,
- la potasse échangeable.

Exemple malherbologie, protocole

Côte d'Or 1969 $\longrightarrow$ 1976
245 parcelles

218 espèces inventoriées

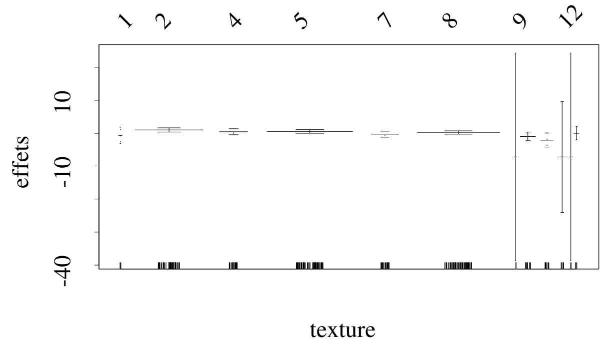
parcelle	$\rightarrow$ 3 années	culture automne + culture printemps $\downarrow$ relevés botaniques notation abondance 0 à 5

0/1 :
0 : absence de l'espèce
1 : présence de l'espèce pour un relevé au moins

- Les données ont été recueillies en Côte d'Or, entre 69 et 76. 245 parcelles agricoles ont été observées. Sur chaque parcelle, on a inventorié les espèces observées. Au total, 218 espèces ont été inventoriées.
- Les semences ont une grande longévité : une dizaine d'années pour les sols régulièrement travaillés. Une espèce peut très bien ne pas apparaître dans la culture, mais être présente dans le sol, et apparaître dans une autre culture. C'est pour cette raison que chaque parcelle est observée 3 années consécutives, de manière à ce qu'il y ait au moins une culture d'automne et une culture de printemps. A chaque observation de la parcelle, l'abondance de folle avoine (et des autres espèces) est notée.
- On s'intéresse au potentiel de la parcelle à héberger l'espèce. Pour chaque parcelle, on résume donc l'ensemble des observations par une unique valeur, 0 ou 1 :
  0 : si la folle avoine n'a jamais été rencontrée,
  1 : si la folle avoine a été rencontrée au moins une fois.

Modèle : texture

```
Coefficients:
                  Value Std. Error
 texture1 -6.931472e-01  1.2247449
 texture2  9.903987e-01  0.3247992
 texture4  4.054651e-01  0.4564355
 texture5  5.465437e-01  0.2678999
 texture7 -3.184537e-01  0.4646602
 texture8  2.076394e-01  0.2640294
 texture9 -7.201716e+00 15.7257175
texture10 -9.808293e-01  0.6770032
texture11 -2.079442e+00  1.0606602
texture12 -7.201716e+00  8.4057496
texture14 -7.201716e+00 15.7257175
texture15  1.110223e-16  1.0000000
```



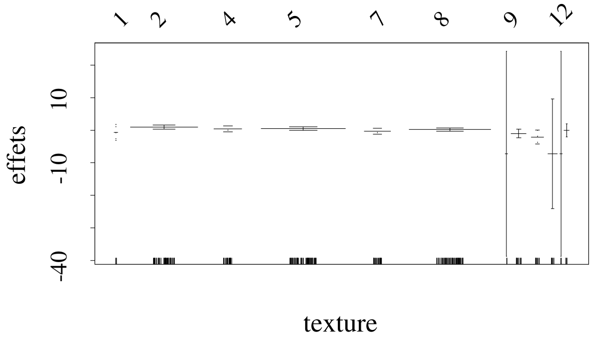
On cherche d'abord à expliquer la probabilité de présence par la texture. La variable texture est particulièrement intéressante pour prédire, car l'information est cartographiée, et donc directement disponible, elle ne nécessite aucune mesure sur le terrain.

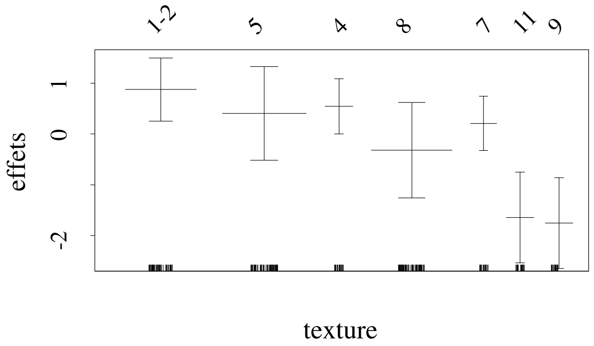
Certaines des textures ont des effectifs très faibles, de l'ordre de 1 ou 2, le modèle obtenu est donc surparamétré.

Les effets représentés sont les paramètres estimés $\widehat{\alpha}$ dans le modèle $\text{logit}(p) = \mu + \alpha_{\text{texture}}$ avec la contrainte $\sum_{\text{texture}} \alpha_{\text{texture}} = 0$.

La largeur des traits horizontaux représentant les effets est proportionnelle à l'effectif. Les “petits traits en bas” (tickmarks) représentent aussi les effectifs. (Ce graphique est obtenu avec la fonction plot.gam de Splus).

Modèle : texture





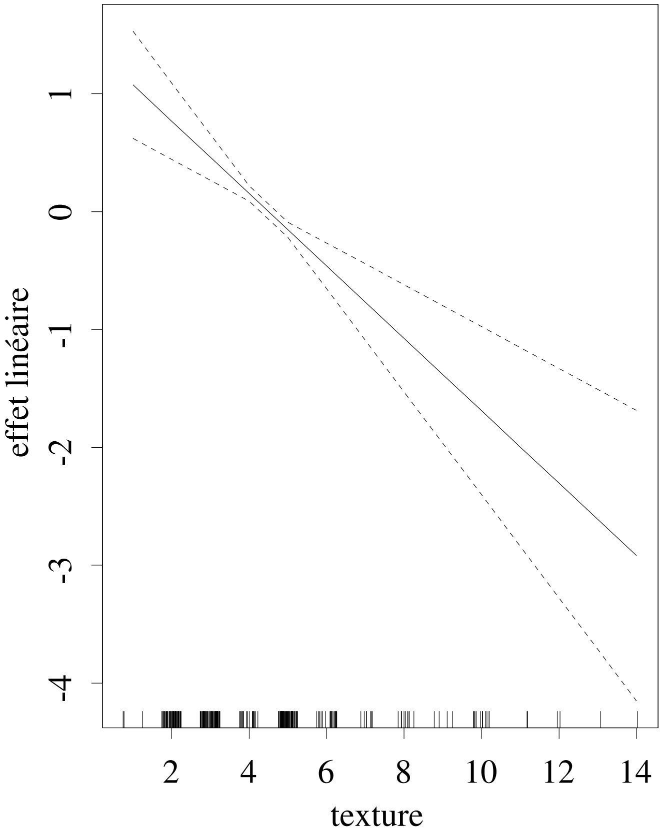
effectifs des textures regroupées						
1-2	5	4	8	7	11-12 15	9-10-13 14-17
51	60	20	58	19	20	17

On regroupe donc les textures proches de façon à obtenir des effectifs plus importants. Les textures 1 et 2 sont regroupées, de même que les textures 11, 12 et 15 d'une part, 9, 10, 13, 14 et 17 d'autre part. 51 parcelles ont une texture de type 1 ou 2, 60 parcelles ont une texture de type 5 ...

Le modèle obtenu n'est plus surparamétré. On représente les effets et intervalles de confiance. Les textures les plus fines sont conviennent mieux à la folle avoine que les textures les plus grossières.

Remarque : les textures 3,6 et 16 n'ont pas été observées parmi les parcelles étudiées.

Modèle : effet linéaire de la texture

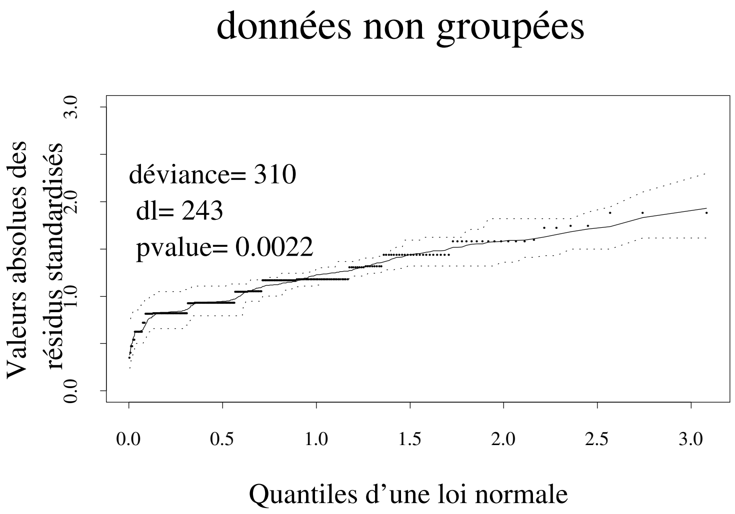


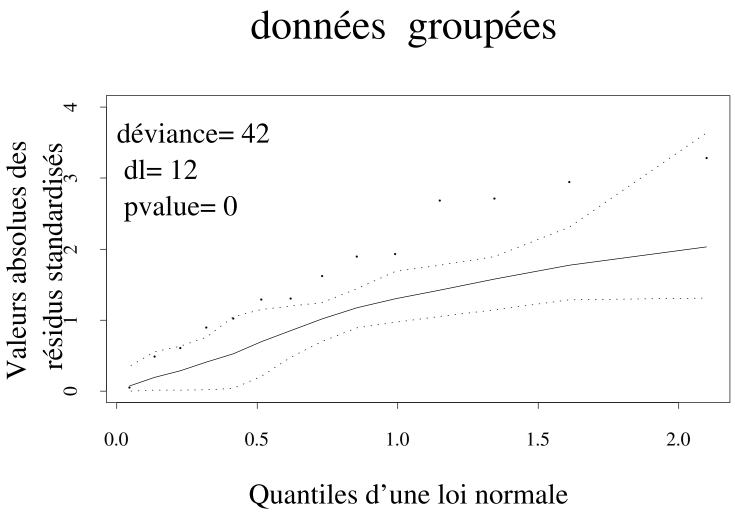
On cherche à diminuer le nombre de paramètres, on remplace le facteur texture par un effet linéaire de la texture (on revient aux textures initiales non groupées). On code les textures de 1 à 17. On suppose qu'il y a autant d'écart entre les niveaux 1 et 2 qu'entre les niveaux 4 et 5, et que l'écart est plus grand entre les niveaux 2 et 6 qu'entre les niveaux 5 et 6.
Le modèle n'est pas statistiquement différent du modèle précédent ($p=0.24$). Il est adopté.

Cette représentation graphique est l'analogue de la précédente, adaptée au cas d'un régresseur, alors que la précédente était adaptée au cas d'un facteur. C'est la pente estimée qui est représentée, ainsi qu'un intervalle de confiance pour la pente (il ne s'agit pas d'un intervalle de confiance autour de la droite).

L'effet et un intervalle de confiance sont représentés graphiquement. Les textures les plus fines sont “préférées” aux textures les plus grossières.

Modèle : effet linéaire de la texture





Les n_i varient de 1 à 60. Plusieurs des n_i valent 1, 2 ou 3. On n'utilise donc pas la déviance pour valider le modèle, mais le graphique semi-normal des résidus de déviance, avec enveloppe simulée.

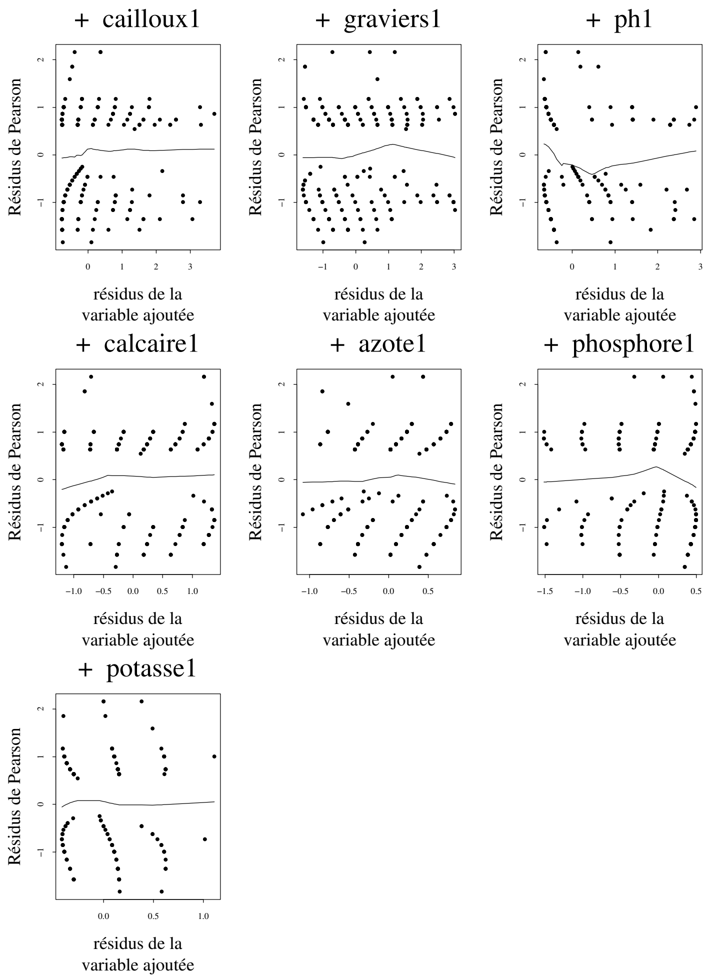
Le premier graphique semi-normal des résidus est obtenu à partir du modèle ajusté sur les données non groupées. Il n'est pas très facile à interpréter, à cause des "paliers" qui proviennent du fait que les données auraient pu être groupées, et qu'elles ne l'ont pas été. On remarque que de nombreux résidus sont très proches du bord de l'enveloppe, plusieurs sont en dehors de l'enveloppe.

On réalise la même représentation, mais à partir de l'ajustement sur les données groupées. Ce graphique révèle très clairement l'inadéquation du modèle.

Conclusion : dans tous les cas où les données peuvent être groupées, mieux vaut les grouper. Non seulement le graphique semi-normal est plus informatif, mais il en est de même pour les graphiques des résidus en fonction du prédicteur.

Question : on se demande s'il est licite d'interpréter un graphique semi-normal avec enveloppe, dans le cas de données non groupées.

Graphiques des variables ajoutées



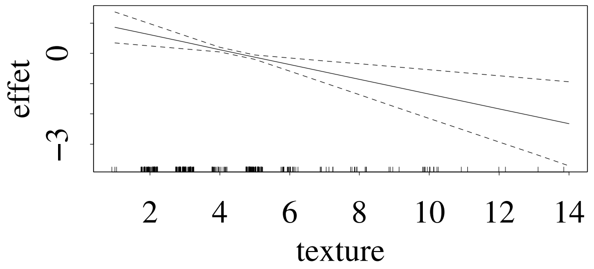
- La variable texture n'explique pas tout. Quelles variables peut-on ajouter, qui apporteraient une information complémentaire ?
- On présente un graphique des variables ajoutées, pour chacune des variables suceptibles d'être ajoutées au modèle. On superpose un lissage.

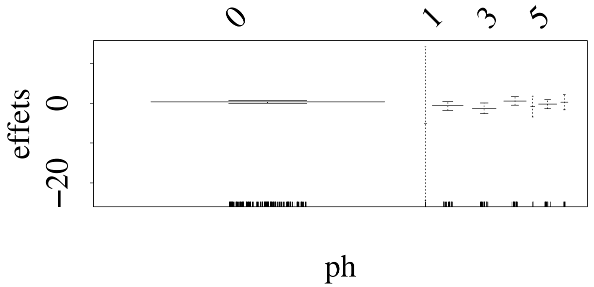
  graphique de la variable ajoutée : added variable plot (Collett) :
  les résidus de Pearson sont représentés en fonction de la variable explicative candidate, après "suppression" de l'effet des variables déjà incluses dans le modèle (texture ici). C'est un résidu que l'on porte en abscisse, c'est le résidu de la régression linéaire multiple expliquant la variable à ajouter par les autres variables explicatives du modèle.

  On recherche une structure des résidus de Pearson, qui pourrait être expliquée par la variable ajoutée. Ici, il semble que la variable pH pourrait apporter une explication complémentaire.

  ⟶ on ajoute donc la variable ph, sous forme de facteur (pH ne peut pas être introduit sous forme de régresseur, car le niveau 0 du facteur pH correspond en fait à une présence de calcaire, il correspond à un pH non mesuré).

Modèle : texture.lin + ph





	pH							
	0	1	2	3	4	5	6	7
n	167	2	22	17	16	3	13	5

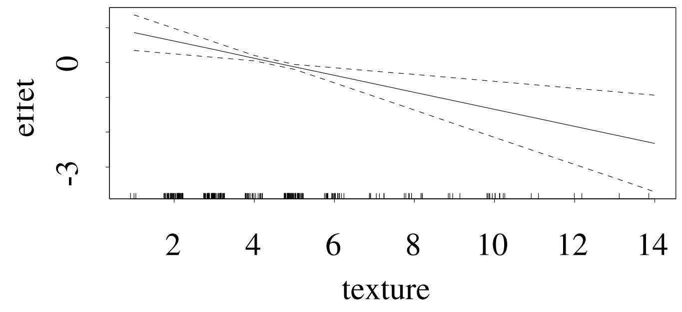
Le modèle obtenu est surparamétré. En effet, la classe 1 de la variable ph n'a été observée que pour 2 parcelles.

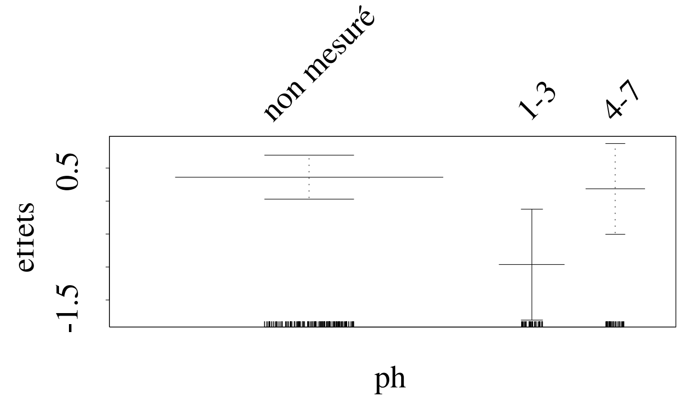
Codage de la variable pH :

0	pH non mesuré	présence de calcaire
1	pH 4 à 5	sol fortement acide
2	pH 5.1 à 6	sol acide
3	pH 6.1 à 6.5	sol faiblement acide
4	pH 6.6 à 6.9	sol très faiblement acide
5	pH 7	sol neutre
6	pH 7.1 à 7.5	sol faiblement alcalin
7	pH 7.6 à 8.5	sol alcalin

La modalité 0 de la variable pH doit être considérée comme une modalité à part, puisqu'elle correspond à un pH non mesuré, pour cause de présence de calcaire. Elle ne peut pas être ordonnée par rapport aux autres modalités.

Modèle texture.lin + ph





On regroupe les pH 1 à 3 d'une part, 4 à 7 d'autre part. Le pH 0 (non mesuré) est toujours considéré comme une classe à part. On réajuste le modèle.

Les sols à texture fine, calcaire, non acides sont favorable à la folle avoine.

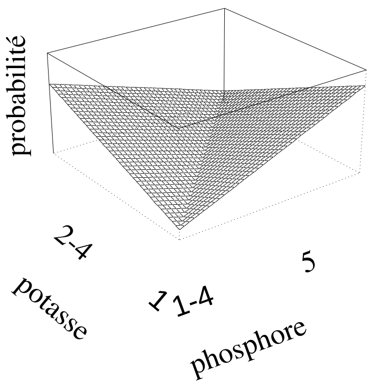
On recherche ensuite d'autres variables à ajouter : on trace de nouveau les graphes des variables ajoutées, on essaye d'inclure dans le modèle des variables suceptibles d'apporter une explication supplémentaire. Aucune des variables n'apporte d'explication significative.

Recherche d'interactions

modèle additif : toutes les var. expl.

+ interactions d'ordre 2

→ phosphore : potasse (p=1 %)



{ à valider avec de nouvelles données
difficile à mesurer

→ retour au modèle additif.

On passe à une phase de recherche d'interaction. Pour cela, on repart du modèle additif avec toutes les variables explicatives potentielles. On envisage toutes les interactions d'ordre 2, par une procédure automatique de sélection de variables, basée sur le critère d'Akaïke (on minimise le critère "Déviance + 2 . nombre de paramètres"). On ajoute les interactions qui diminuent le plus fortement la déviance, avec un nombre de paramètres le plus faible possible.

Seule l'interaction phosphore x potasse est significative. Le modèle obtenu est surparamétré, ce qui nous amène à regrouper des classes de phosphore et de potasse.
Le terme phosphore x potasse est significatif, mais à 1% seulement. Comme l'introduction de ce terme résulte d'une recherche parmi un grand nombre de termes possibles (ça ressemble à une partie de "pêche à la ligne"), il faut être prudent. Il faudrait valider l'hypothèse d'une interaction phosphore x potasse à l'aide d'autres données.

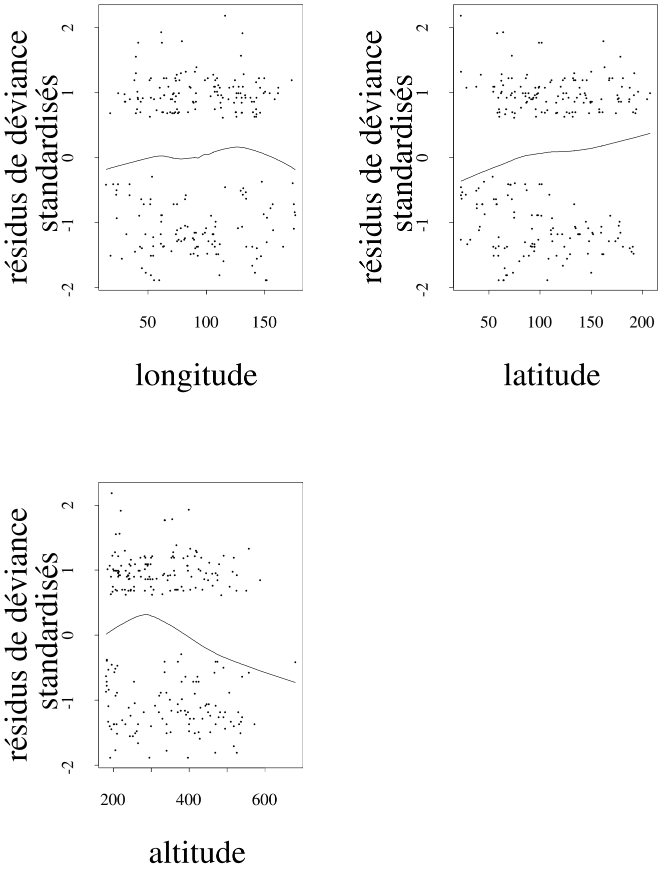
Les effets estimés de l'interaction sont représentés. Si cette interaction existe, elle s'interprète comme : les sols très bien pourvus en phosphore, ou non carencés en potasse sont favorables à la folle avoine, mais pas les sols qui présentent les 2 caractéristiques à la fois.

Ceci n'est pas une conclusion, mais seulement une hypothèse. On a finalement préféré ne pas retenir ce terme dans le modèle,

- à cause des doutes que l'on a, de la faible significativité du terme, au regard de la partie de pêche à la ligne à laquelle on s'est livré.
- parce que l'utilisation de ce modèle nécessiterait des mesures au niveau de la parcelle.

⟶ retour au modèle additif : poly(texture,1) + ph.

Structure des résidus?



- Modèle : poly(texture,1) + ph.
  Il existe une autre information, disponible pour chaque parcelle, non encore utilisée, ce sont les coordonnées x, y, z, latitude, longitude et altitude.

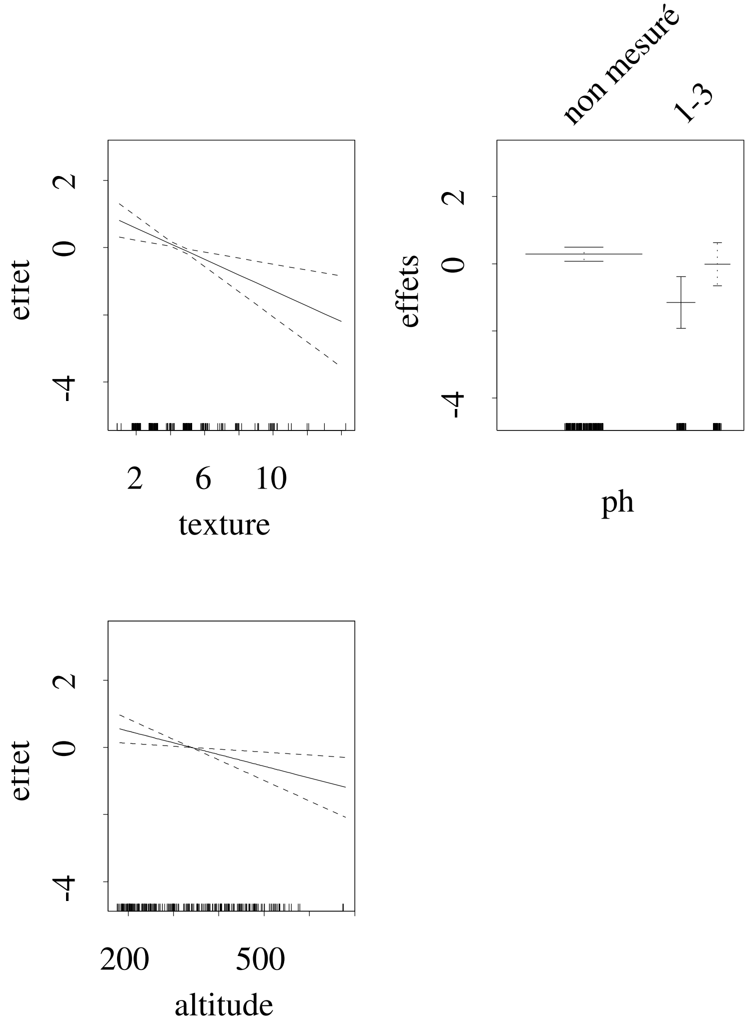
- On recherche une structure des résidus en fonction de chacune de ces variables. On représente les résidus de déviance en fonction de chaque variable, on superpose un lissage. On trouve une structure, en fonction de la latitude, et de l'altitude essentiellement. Il reste encore une structure non expliquée dans les résidus.

- L'introduction de l'altitude dans le modèle est plus intéressante que l'introduction de la latitude. Ou alors, il faudrait utiliser latitude et longitude pour déterminer des régions.

  ⟶ on introduit un effet linéaire de l'altitude.
  On a aussi essayé d'introduire un effet quadratique, qui était significatif, mais en examinant les statistiques L, on s'est aperçu qu'il ne devait sa significativité qu'à 1 ou 2 parcelles. On est donc revenu à un effet linéaire.

Remarquons qu'il aurait été tout à fait judicieux de réaliser un "added variable plot".

Modèle final : texture, ph, altitude

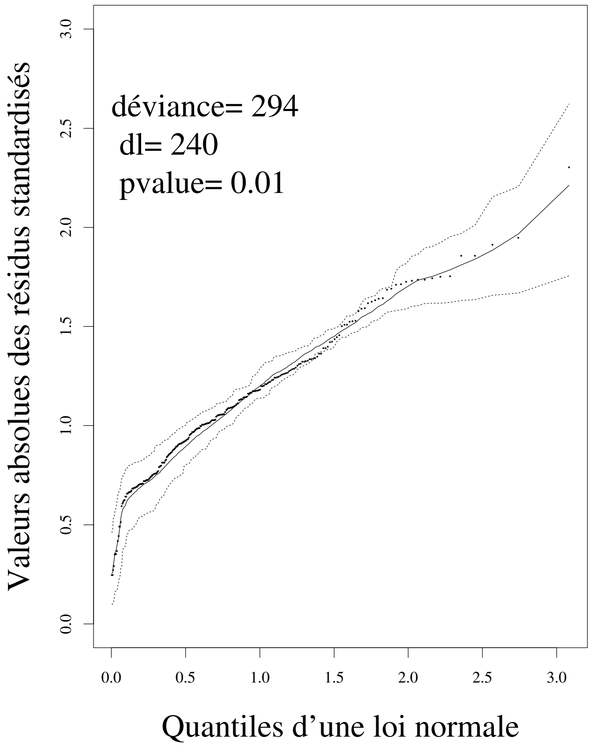


Le modèle final est : poly(texture,1) + pH + altitude

On représente les effets estimés et intervalles de confiance, tous à la même échelle.

La folle avoine apprécie les textures fines, les sols calcaires, ou neutres à basiques, de faible altitude.

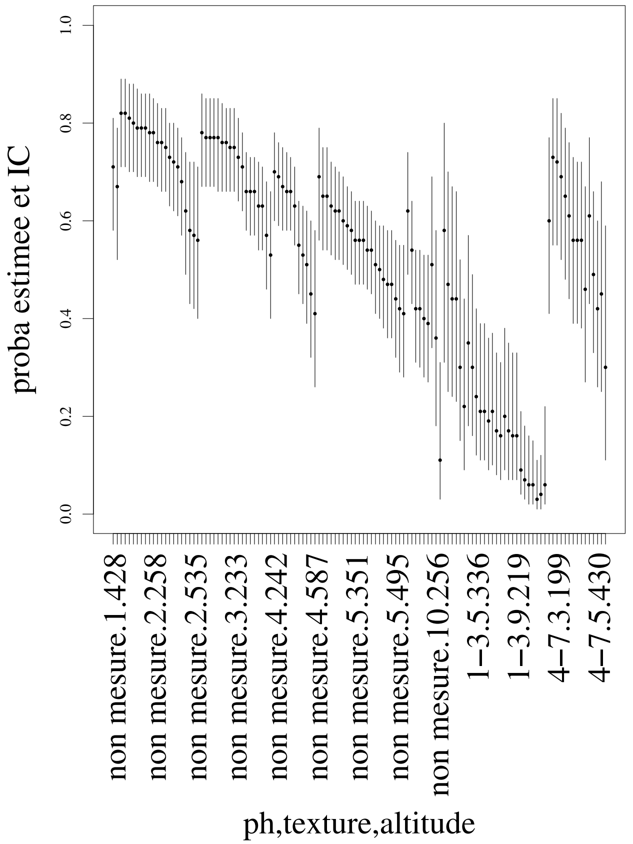
Modèle final : texture, ph, altitude



On valide le modèle obtenu. Cette fois, on a 2 régresseurs dans le modèle, donc tous les n_i valent 1. Les données sont non groupées. La déviance n'est plus utilisable pour valider le modèle (elle est affichée automatiquement sur le graphique, mais cela ne signifie pas qu'il faille l'interpréter, de même que la p-value correspondante. L'affichage se fait dans tous les cas, la responsabilité de son utilisation ou pas incombe à l'utilisateur). Le graphique semi-normal des résidus assorti de son enveloppe prend tout son intérêt. Le modèle semble adéquat.

On vérifie aussi, à l'aide de la distance de Cook et de la statistique L, l'absence de points influents.

Probabilités estimées et IC



Le modèle obtenu peut être représenté par le graphique des effets et intervalles de confiance. Il est également intéressant de présenter les probabilités estimées, assorties d'un intervalle de confiance, pour les différents types de parcelles observées, après les avoir ordonnées en fonction du ph, de la texture, puis de l'altitude. Toutes les parcelles sont représentées, il y a un “tickmark” par parcelle, mais les labels correspondants sont échantillonnés selon un pas constant. La probabilité de présence varie de 0.1 à 0.8 selon les caractéristiques de la parcelle.

Stratégie de sélection de modèle

stratégie idéale : descendante

Données

$\downarrow$

Modèle le plus complet possible

$\downarrow$

Validation

$\downarrow$

Simplification du modèle

$\rightarrow$ interprétation + facile
$\rightarrow$ + grande précision

La stratégie idéale de sélection de modèle est une stratégie descendante. Elle consiste à poser le modèle le plus complet possible, incluant tous les facteurs, toutes les covariables, tous les termes d'interaction jusqu'au degré choisi *a priori*, puis à valider ce modèle, et enfin, à supprimer du modèle les termes non significatifs, tout en maintenant une certaine cohérence, c'est à dire en ne supprimant pas les termes qui interviennent dans des interactions de niveau plus élevé. Cette simplification permet à la fois d'accroître la précision de l'estimation des paramètres, et d'obtenir un modèle plus simple, ne contenant que des termes significatifs, plus facilement interprétable.

Sélection de modèle

Comparaison de modèles emboîtés
Tests du rapport de vraisemblance

sélection

$\swarrow \quad \searrow$

descendante ascendante

mixte

utiliser la connaissance
Réduire le nombre de tests

Comparer des modèles non emboîtés

Minimiser le critère d'Akaïke :
AIC = Dev + 2. nb param

Dans de nombreux cas, cette stratégie descendante ne peut être adoptée, car le modèle le plus complet possible ne peut être estimé (la convergence n'est pas obtenue, ou difficilement), ou bien le modèle complet ne peut être validé (modèle instable), ou encore, dans le cas de données binaires non groupées, les tests de sous-modèle par rapport au modèle saturé ne sont pas valides. De plus, lorsque les n_i sont petits, par exemple tous égaux à 1, on ne peut pas valider le grand modèle par un test, les tests de sous-modèle comparant un sous-modèle au grand modèle ne sont pas valides, dans la mesure où ils reviennent à faire un test d'adéquation.

On peut alors adopter une stratégie ascendante, ou mieux, une stratégie mixte. La stratégie ascendante consiste à partir du modèle nul, et à ajouter à chaque pas le terme le plus significatif. La stratégie mixte permet d'utiliser le mieux possible toute la connaissance du phénomène : la connaissance *a priori* des facteurs et variables influents, des variables et facteurs que l'on souhaite voir intervenir dans le modèle, ou pas, selon qu'ils sont par exemple, faciles ou difficiles à mesurer. On pose un premier modèle en fonction des connaissances que l'on a et des hypothèses que l'on fait.

Ces stratégies, qu'elles soient ascendantes, descendantes ou mixtes, mettent en jeu des tests de sous-modèle. Il est souhaitable de réduire le nombre de tests réalisés. Attention à l'interprétation des pvalues, si elles sont le résultat implicite d'un trop grand nombre de tests (ie : attention, lorsque le choix de modèle s'apparente à une partie de pêche à la ligne, il est nécessaire de valider le modèle obtenu avec un jeu de données indépendant, ou par des méthodes de validation croisée).

Et si l'on souhaite comparer deux modèles qui ne sont pas emboîtés ? On peut choisir celui pour lequel le critère d'Akaïke est le plus faible.

Sélection de modèle

Ne pas interpréter indépendamment les uns des autres des paramètres fortement corrélés

→ polynômes orthogonaux
→ choix de var. expl. peu corrélées
→ régression sur composantes principales

Attention aux relations causales :

$$\begin{tikzcd} \mathbf{X} & & \mathbf{Y} \\ \mathbf{X'} \arrow[u] \arrow[urr] & & \end{tikzcd}$$

X : var. expl. d'intérêt
X' : var. expl.

On a déjà vu les problèmes liés à la non orthogonalité : il ne faut pas interpréter indépendamment les uns des autres des paramètres fortement corrélés (c'est aussi le cas en modèle linéaire). Il faut donc prendre l'habitude d'utiliser les polynômes orthogonaux. Il faut, si possible, choisir parmi les variables explicatives possibles, un sous-ensemble de variables explicatives peu corrélées. On peut utiliser les méthodes de régression sur composantes principales.

Polynômes orthogonaux : c'est le même principe qu'en régression linéaire simple, où le fait de centrer la variable explicative rend nulle la corrélation entre ordonnée à l'origine et pente. On applique ce principe aux termes d'un polynôme. Ainsi, la corrélation entre le coefficient de degré 1 et le coefficient de degré 2 est nulle, et les coefficients peuvent être interprétés indépendamment l'un de l'autre.

Supposons que le schéma causal soit le schéma représenté ci-contre, c'est à dire que la variable explicative X' agit sur Y, alors que X n'agit pas sur Y. Pour une interprétation correcte, il est nécessaire de tester l'effet de X, ajusté par rapport à X', c'est à dire, après avoir tenu compte de l'effet de X'. L'interprétation sera fausse si on considère un modèle où X' n'intervient pas, ou bien, si on base l'interprétation sur la table d'analyse de déviance obtenue avec X' introduit après X.

RÉSUMÉ

difficultés	comment diagnostiquer ?
non orthogonalité	matrice de corrélation des param
instabilité	écart-type des paramètres

Choix de modèle :

par ordre décroissant de préférence,

stratégie descendante
stratégie mixte
stratégie ascendante

On détecte une non orthogonalité à partir de la matrice de corrélation des paramètres, et une instabilité à partir des écarts-type des paramètres. Non orthogonalité et instabilité sont à éviter dans la mesure du possible.

Les stratégies de choix de modèle sont ordonnées par ordre de préférence. On préfère une stratégie descendante lorsque c'est possible. Si ça n'est pas possible, on peut adopter une stratégie mixte, à condition d'avoir une connaissance *a priori* du phénomène modélisé. À défaut, on adopte une stratégie ascendante.