MORTALITÉ D'UN COLÉOPTÈRE DE LA FARINE AU CS2

concentration de CS2 (mg/l) cc	répétition 1 y	n	$p = y/n$	répétition 2 y	n	$p = y/n$
49.06	2	29	0.069	4	30	0.133
52.99	7	30	0.233	6	30	0.200
56.91	9	28	0.321	9	34	0.265
60.84	14	27	0.519	14	29	0.483
64.76	23	30	0.767	29	33	0.879
68.69	29	31	0.935	24	28	0.857
72.61	29	30	0.967	32	32	1.000
76.54	29	29	1.000	31	31	1.000

Cet exemple est l'exemple 4.5 "Mortality of confused flour beetles" présenté par Collett.

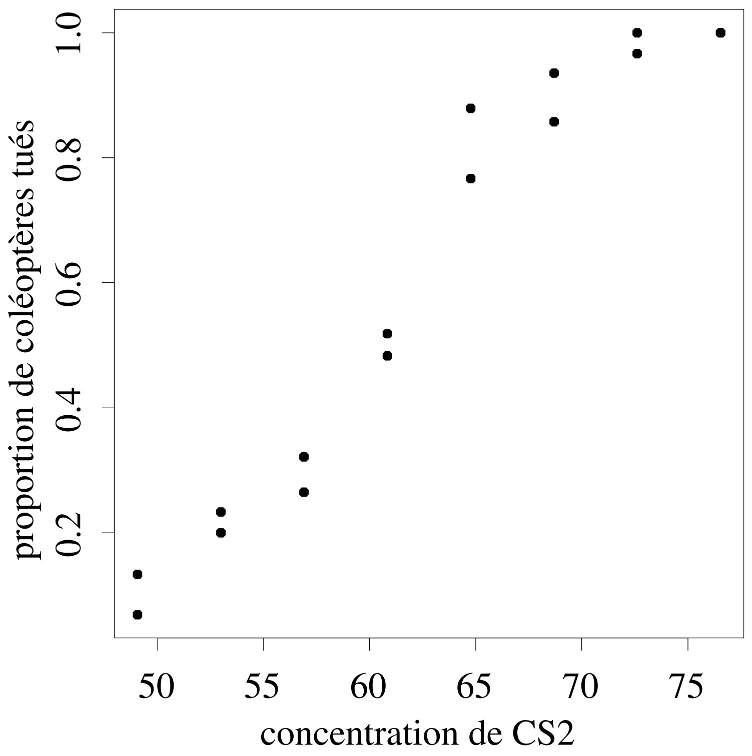
Protocole expérimental :

- des cages renfermant une trentaine de coléoptères sont placées dans des flacons contenant différents volumes de disulphide de carbone liquide.
- au bout de 5 heures, on mesure la concentration de CS2 gazeux dans le flacon et on compte le nombre de coléoptères morts.
- la totalité de l'expérimentation a été répétée deux fois.

On reporte dans le tableau, le nombre y de coléoptères morts, et le nombre n de coléoptères exposés aux différentes concentrations de disulphide de carbone gazeux. y est une variable réponse de type binomial. On peut remarquer que les n_i sont grands et pas tous égaux. cc est une variable explicative continue (régresseur).

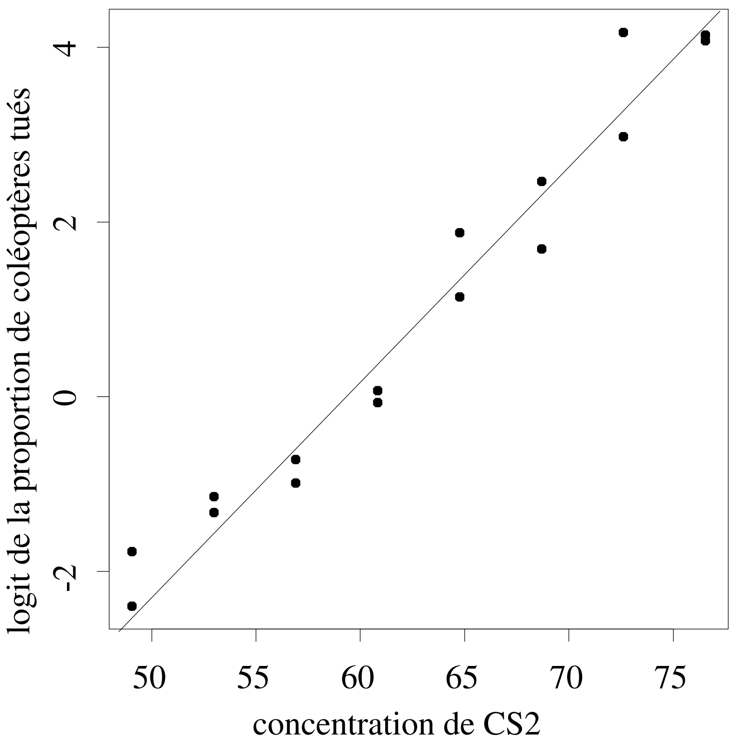
Remarque : le fichier script.chap4.Splus contient l'enchaînement des commandes Splus utilisées pour réaliser les graphiques de ce chapitre.

ÉCHELLE D'ORIGINE



Le graphique des données montre qu'une droite ne résume pas bien l'allure des données : l'allure est plutôt sigmoïdale.

ÉCHELLE LOGIT



On représente les données dans l'échelle du lien pour s'assurer de la linéarité de la liaison : le faible nombre de points ne permet pas de conclure à une structure dans les résidus. L'ajustement dans l'échelle logit apparaît correct.

Remarque : les 3 points ayant les plus fortes ordonnées correspondent à des proportions égales à 1, qu'il n'a été possible de représenter dans l'échelle logit qu'après une transformation logistique empirique (celle-ci étant appliquée à tous les points). Cette transformation logistique empirique est définie par : $\log\left(\frac{y_i+0.5}{n_i-y_i+0.5}\right)$.

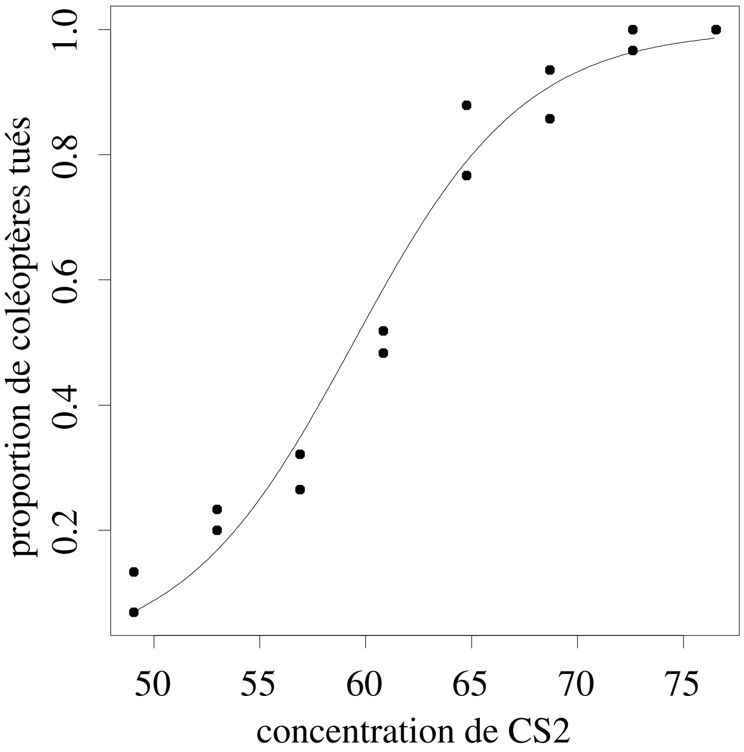
L'estimation des paramètres donne :
$\mathrm{logit}(\widehat{p}) = \widehat{\eta} = -14.808 + 0.249cc$

La déviance du modèle nul vaut 289.14, pour 15 ddl.

Les n_i sont de l'ordre de 30, les données sont groupées, on peut donc utiliser la déviance comme critère d'adéquation :
la déviance résiduelle vaut 12.51 pour 14 ddl. Elle est à comparer à $\chi^2(0.95, 14) = 23.68$. Le modèle n'est pas significativement différent du modèle saturé (pvalue=0.57). L'hypothèse d'adéquation du modèle est acceptée.

Remarque : pour réaliser rapidement, et grossièrement le test d'adéquation, sans se référer à une table du χ^2, on peut simplement comparer la déviance et le nombre de degrés de liberté (l'espérance d'un χ^2 à m ddl vaut m). Ici, la déviance est du même ordre de grandeur que le nombre de ddl, ce qui indique que le modèle est adéquat.

RETOUR À L'ÉCHELLE D'ORIGINE



$\text{logit}(\widehat{p}) = \widehat{\eta} = 14.808 + 0.249cc$

$p = \frac{\exp \eta}{1+\exp \eta}$

En revenant dans l'échelle d'origine par la réciproque de la fonction de lien logistique $\frac{\exp \eta}{1+\exp \eta}$, on constate que la courbe estimée s'ajuste correctement aux données.

Cependant, pour mieux appréhender la qualité du modèle, il faut passer à l'étude des résidus.

LES RÉSIDUS

Résidus *bruts* $y_i - \widehat{y_i}$

Résidus de *Pearson*

$$R_{P_i} = \frac{y_i - \widehat{y_i}}{\text{écart-type}(y_i)} = \frac{y_i - n_i\widehat{p_i}}{\sqrt{n_i\widehat{p_i}(1-\widehat{p_i})}}$$

Résidus de *Pearson standardisés*

$$r_{P_i} = \frac{y_i - \widehat{y_i}}{\text{écart-type}(y_i - \widehat{y_i})} = \frac{y_i - n_i\widehat{p_i}}{\sqrt{n_i\widehat{p_i}(1-\widehat{p_i})(1-h_i)}} = \frac{R_{P_i}}{\sqrt{1-h_i}}$$

Résidus de *Déviance*

$$R_{D_i} = signe(y_i - \widehat{y_i})\sqrt{2y_i \log\frac{y_i}{\widehat{y_i}} + 2(n_i - y_i)\log\frac{n_i - y_i}{n_i - \widehat{y_i}}}$$

Résidus de *Déviance standardisés*

$$r_{D_i} = \frac{R_{D_i}}{\sqrt{1-h_i}}$$

$$\chi^2_{\text{Pearson}} = \sum_i R^2_{P_i} \qquad D = \sum R^2_{D_i}$$

Les résidus bruts sont difficiles à comparer car les valeurs des n_i et des p_i étant variables, ils ne sont pas estimés avec la même précision.

En les divisant par l'écart-type de y_i, on obtient des résidus de Pearson, qui présentent l'avantage d'être plus comparables que les résidus bruts.

On peut également standardiser les résidus bruts par leur écart-type. On obtient alors les résidus de Pearson standardisés, dont la variance est égale à 1.

écart-type$(y_i - \widehat{y_i}) = \sqrt{\widehat{v_i}(1-h_i)}$ où $\widehat{v_i} = n_i\widehat{p_i}(1-\widehat{p_i})$ et h_i est le i[e] élément de la diagonale de la matrice H (les h_i ou "leverage" ou "fonctions d'influence" permettent de repérer les observations qui sont éloignées des autres du fait des valeurs prises par les variables explicatives).

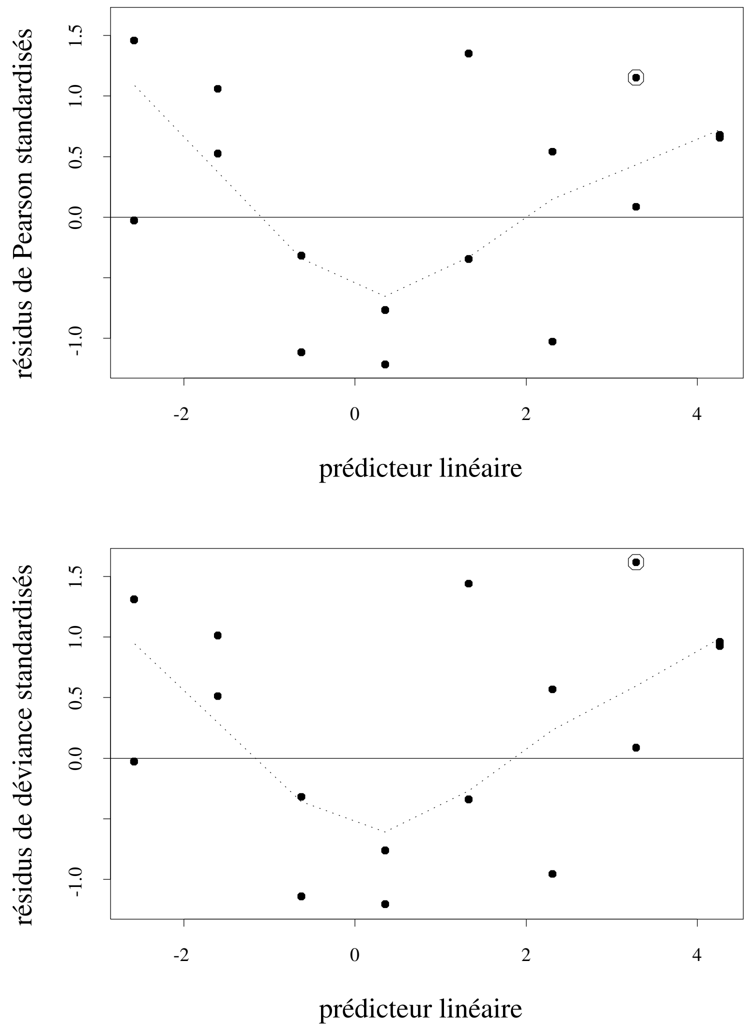
Un autre type de résidu, le résidu de déviance, correspond à la contribution de la i[e] observation à la déviance D.

$D = 2\sum_i \left\{ y_i \log\frac{y_i}{\widehat{y_i}} + (n_i - y_i)\log\frac{n_i - y_i}{n_i - \widehat{y_i}} \right\}$

Les résidus de Pearson présentent l'avantage d'être cohérents avec l'algorithme utilisé. Les résidus de déviance présentent l'avantage d'être cohérents avec la théorie du GLM.

Remarque : on peut noter que la distribution exacte de ces différents types de résidus n'est pas connue.

REPRÉSENTATION DES RÉSIDUS



Revenons à l'exemple CS2 du coléoptère de la farine.

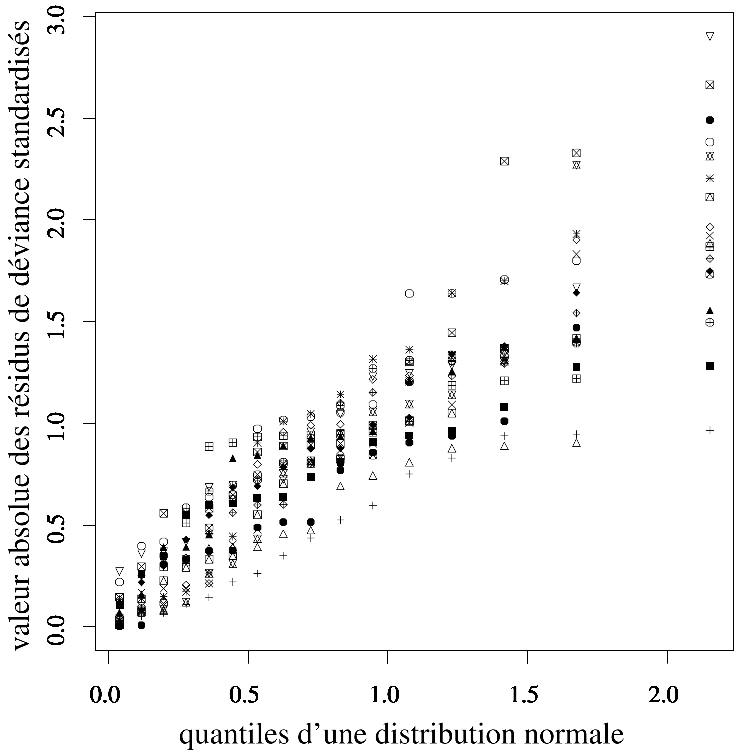
Il apparaît une structure dans les résidus, qui laisse penser qu'on peut trouver un meilleur modèle. Par contre, il n'apparaît pas de point extérieur (point à résidu anormalement élevé).

On constate que la tendance mise en évidence par lissage est la même pour les résidus de Pearson et pour les résidus de déviance. Cependant le résidu le plus élevé ne correspond pas à la même observation dans les deux cas (l'ordre des résidus n'est donc pas le même). Collett préconise l'emploi des résidus de déviance, qui ont l'avantage d'être cohérents avec la théorie du GLM. Dans la majorité des cas, les résidus de Pearson et de déviance sont proches.

Remarque formateur : les résidus de Pearson sont conhérents avec l'algorithme utilisé (algorithme IRLS non présenté aux stagiaires). L'emploi des résidus de Pearson est justifié par le fait que l'algorithme IRLS est équivalent à la maximisation de la vraisemblance.

ADÉQUATION DU PRÉDICTEUR LINÉAIRE

half-normal plot pour
19 échantillons simulés



Si le modèle est correct, les résidus de déviance standardisés sont asymptotiquement distribués selon une loi $\mathcal{N}(0,1)$.

Un "half-normal plot" permet de voir si les résidus sont approximativement distribués selon une loi normale (il permet également de repérer d'éventuels points extérieurs).

Cependant il est possible, lorsqu'on est loin de l'asymptotique, que les résidus ne soient pas approximativement distribués selon une loi normale, et que le modèle soit tout de même correct.
C'est pour cela qu'il est intéressant de construire des enveloppes simulées, qui donnent un moyen de juger de l'adéquation du modèle, sans faire d'hypothèse sur la distribution des résidus.

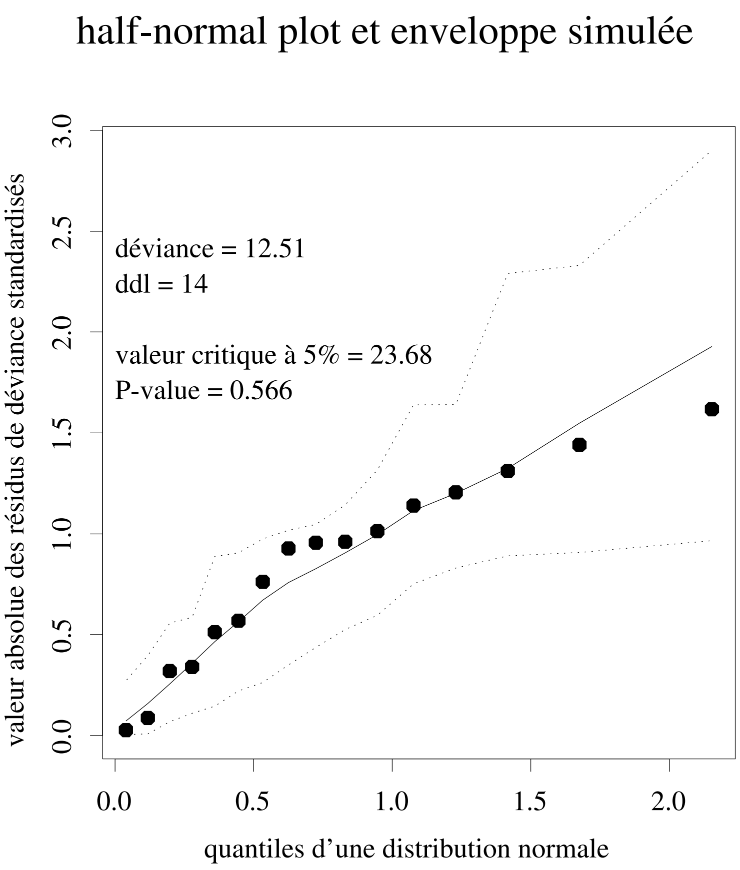
Pour construire l'enveloppe, on simule, pour chacune des valeurs de la variable réponse, 19 nouvelles valeurs, par tirage aléatoire selon une binomiale de paramètres n_i et $\widehat{p_i}$.

On ajuste alors le modèle sur chacun des 19 jeux de données ainsi obtenus, la(les) variable(s) explicative(s) gardant les valeurs d'origine.

Pour chacun des ajustements, on obtient des r_{D_i} que l'on peut ordonner en valeur absolue et représenter dans un "half-normal plot" : $|r_D|_{(1)} < |r_D|_{(2)} < \ldots < |r_D|_{(n)}$

Les valeurs minimale, moyenne et maximale des 19 $|r_D|_{(i)}$, pour $i = 1, 2, \ldots, n$ servent alors à construire l'enveloppe et la ligne moyenne simulées.

ADÉQUATION DU PRÉDICTEUR LINÉAIRE

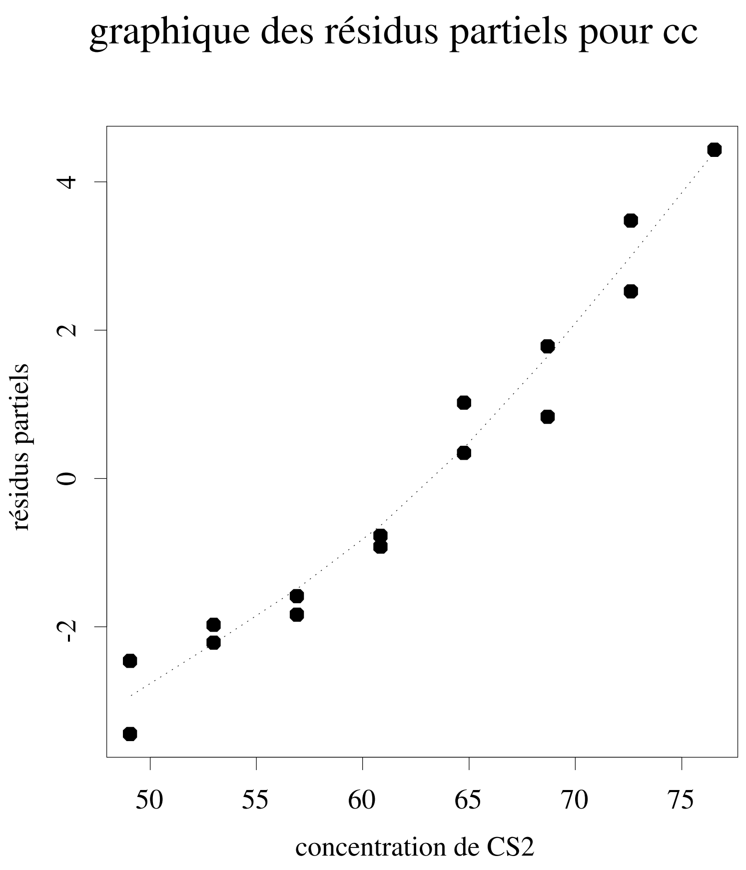


Si le modèle est correct, il y a une chance sur 20 que le résidu le plus élevé se situe en dehors de l'enveloppe, ce qui permet de repérer facilement les points extérieurs (ce sont ceux qui se situent au dessus de l'enveloppe supérieure). Ici, le "half-normal plot" confirme qu'il n'y a pas de point extérieur.

Une déviation systématique des points par rapport à la ligne moyenne simulée indique que le modèle n'est pas approprié (par contre, les points oscillent forcément de façon non aléatoire autour de la ligne moyenne, parce que les résidus ordonnés ne sont bien évidemment pas indépendants les uns des autres).

Bien que les résidus de déviance soient tous contenus dans l'enveloppe, il apparaît une très légère structure par rapport à la ligne moyenne des valeurs simulées, ce qui montre que l'on peut peut-être améliorer le modèle.

ADÉQUATION DU PRÉDICTEUR LINÉAIRE



Est-ce que l'effet de chaque variable est bien pris en compte ?

Pour le savoir, on calcule les résidus partiels qui doivent être liés linéairement à la variable explicative : une tendance non linéaire indique qu'une transformation de la variable est nécessaire.

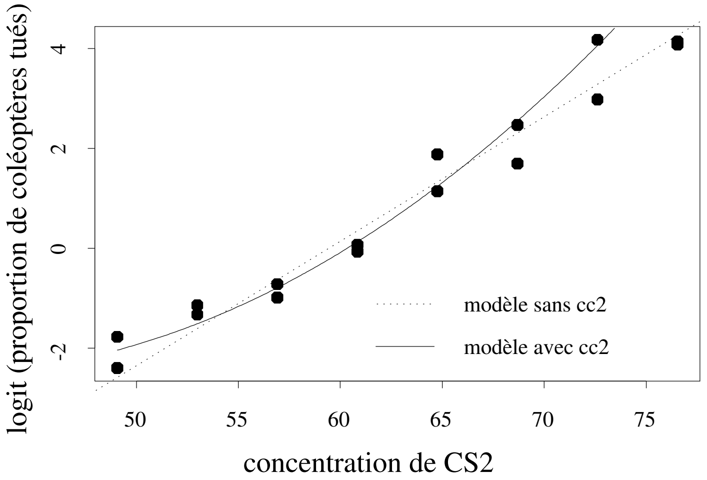
Pour un modèle avec plusieurs variables explicatives, les résidus partiels pour la variable explicative x_j sont donnés par :

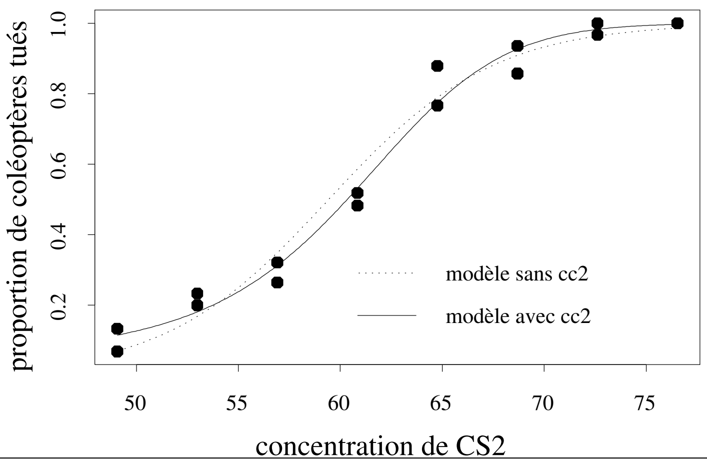
$$\frac{y_i - n_i\widehat{p}_i}{n_i\widehat{p}_i(1-\widehat{p}_i)} + \widehat{\beta}_j x_{ji}$$

Dans le cas de notre modèle avec lien logit, il apparaît une tendance non linéaire dans le graphique des résidus partiels de la concentration, ce qui indique qu'il faudrait introduire une transformation de cette variable explicative.

La forme parabolique de la courbe suggère une transformation quadratique.

INTRODUCTION D'UN TERME QUADRATIQUE





L'estimation des paramètres du modèle quadratique donne :

$$\mathrm{logit}(\widehat{p}) = \widehat{\eta} = 7.968 - 0.517cc + 0.0064cc^2$$

La déviance résiduelle vaut 7.93 pour 13 ddl. Elle est à comparer à $\chi^2(0.95, 13) = 22.36$. Le modèle n'est pas significativement différent du modèle saturé.

Par contre, la valeur très élevée de la pvalue (0.85) pourrait indiquer un modèle surparamétré (on le voit également à travers le rapport "déviance/ddl" qui est très inférieur à 1).

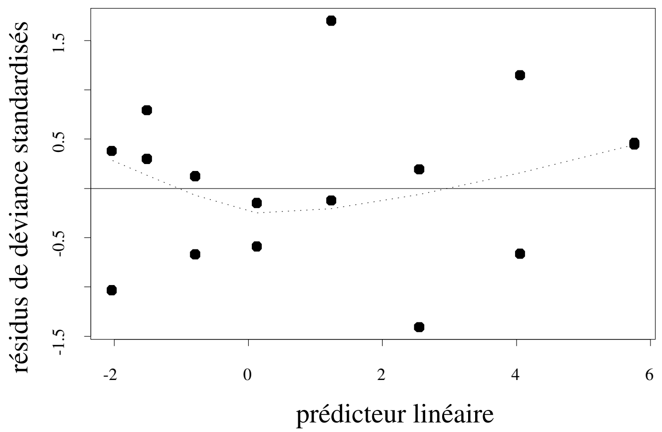
Les modèles étant emboîtés, on peut faire une analyse de déviance qui montre que le modèle quadratique a une déviance significativement plus faible que le modèle précédent :

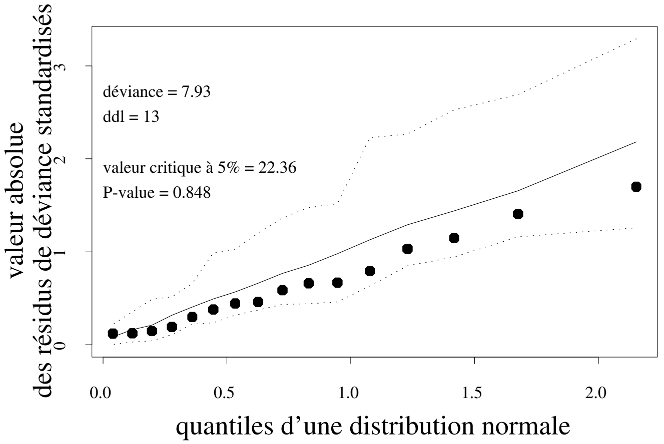
modèle	ddl résiduels	déviance résiduelle	déviance	Pr(Chi)
cc	14	12.51		
$cc + cc^2$	13	7.93	4.58	0.032

Lorsqu'on revient à l'échelle d'origine, on constate un meilleur ajustement de la courbe estimée aux données.

ADÉQUATION DU MODÈLE

Modèle quadratique





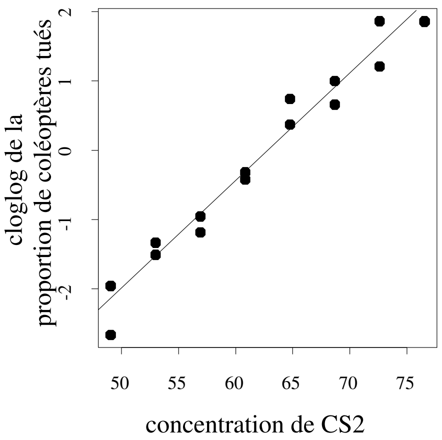
Avec le modèle quadratique, il n'apparaît plus de tendance dans les résidus.

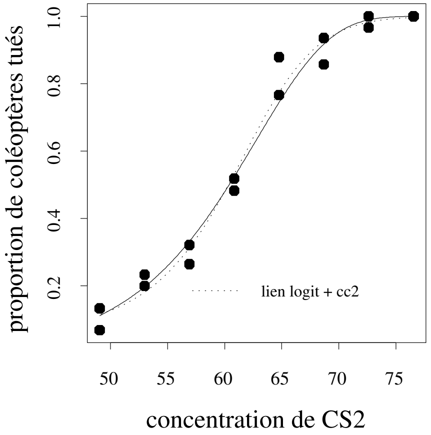
Les résidus sont tous contenus dans l'enveloppe simulée. Cependant, ils se situent nettement en dessous de la ligne moyenne des valeurs simulées, ce qui pourrait à nouveau indiquer un modèle sur-paramétré.

Le terme quadratique améliore l'ajustement (avec le lien logit). Cependant, l'ajustement peut parfois être amélioré par le choix d'un lien plus approprié. Tentons un ajustement avec le lien complémentaire log-log.

CHOIX DU LIEN CLOGLOG

Modèle de degré 1





L'ajustement dans l'échelle cloglog apparaît effectivement bon.

Cependant, on ne peut pas faire d'analyse de déviance car les modèles avec lien logit et lien clog-log ne sont pas emboîtés.

L'estimation des paramètres donne :
$\log(-\log(1-\widehat{p})) = -9.76 + 0.155cc$

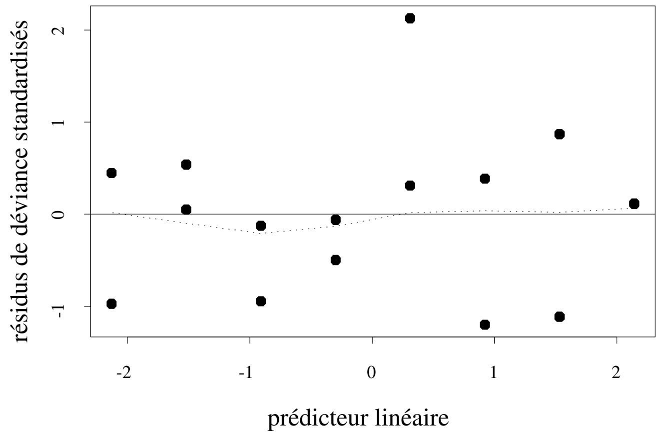
La déviance résiduelle est de 8.67 pour 14ddl. Elle est à comparer à $\chi^2(0.95, 14) = 23.68$. Le modèle n'est pas significativement différent du modèle saturé (P-value $= 0.85$).

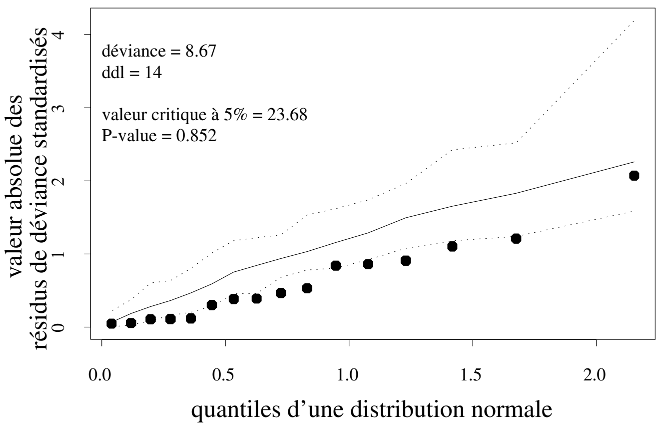
Le rapport "déviance/ddl" est très inférieur à 1, comme dans le cas du modèle quadratique avec lien logit.

Lorsqu'on revient dans l'échelle d'origine par la réciproque de la fonction de lien complémentaire log-log, on constate un très bon ajustement de la courbe aux données.

VALIDATION DU MODÈLE

Modèle de degré 1, lien cloglog





On n'observe plus aucune tendance dans les résidus.

Comme dans le cas du modèle quadratique, le "half-normal plot" pourrait indiquer une sur-paramétrisation du modèle.

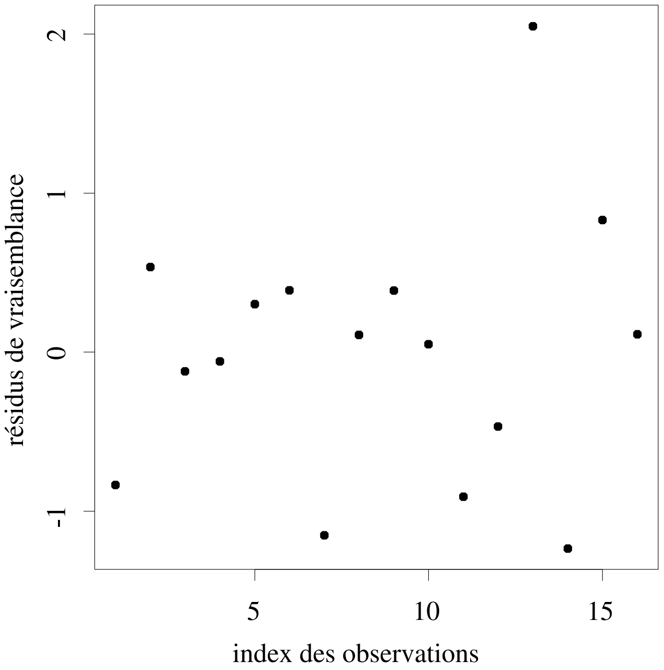
Cependant, on est dans le cas d'un modèle minimal (une seule variable explicative), qu'on ne peut pas essayer de réduire.

Entre le modèle quadratique dans l'échelle logit et le modèle faisant intervenir simplement la concentration dans l'échelle cloglog, on préfèrera le second, qui nécessite l'estimation de moins de paramètres.

Pour valider complètement le modèle, il faut essayer d'identifier d'éventuels points extérieurs (points à résidus élevés), et d'éventuels points influents. Ces derniers correspondent à des observations dont la suppression provoque des changements importants dans l'ajustement du modèle.

Exercice 6.

IDENTIFICATION DES POINTS EXTÉRIEURS



Test de discordance

i^e obs. discordante si $r_{L_i} > \phi^{-1}(1 - \frac{\alpha}{2n})$

valeur critique au seuil 5% : 2.96

Nous avons déjà vu des méthodes graphiques pour repérer d'éventuels points extérieurs (graphique des résidus ou "half-normal plot"). Cependant, il est intéressant de pouvoir tester l'hypothèse "un point est un point extérieur", ce qui, dans le cas de données groupées, est possible grâce à un test de discordance basé sur les résidus de vraisemblance.

Les résidus de vraisemblance sont obtenus par combinaison des résidus de Pearson standardisés et des résidus de déviance standardisés :

$$r_{L_i} = signe(y_i - \widehat{y}_i)\sqrt{h_i r_{P_i}^2 + (1 - h_i) r_{D_i}^2}$$

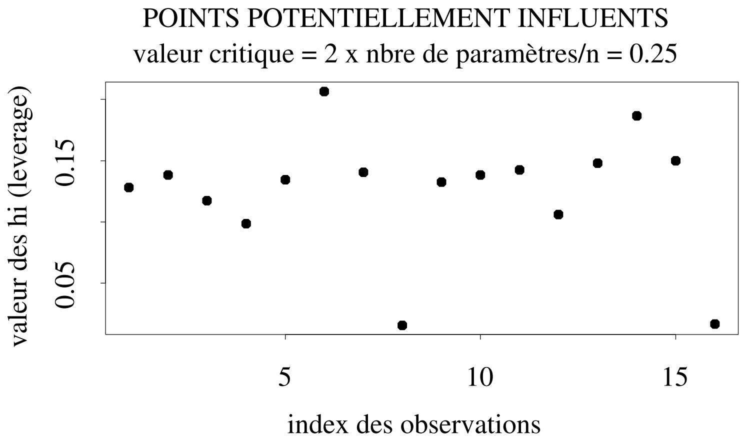
En faisant l'hypothèse que les résidus de vraisemblance sont approximativement distribués selon une loi normale (hypothèse valable uniquement dans le cas de données groupées), on peut déterminer les points extérieurs comme étant ceux dont les résidus sont supérieurs à :

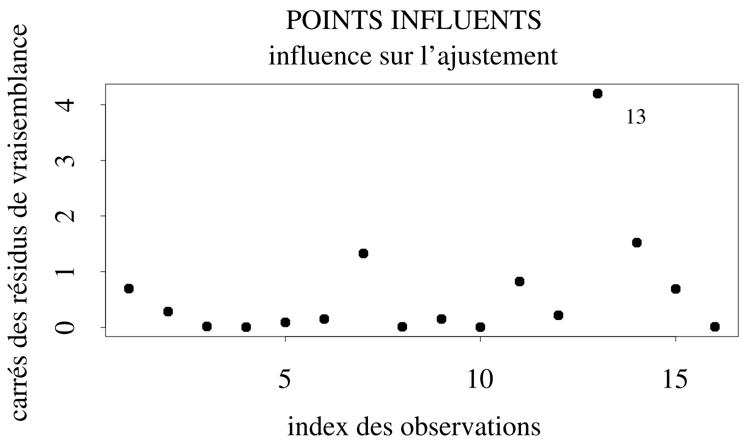
$$\phi^{-1}(1 - \frac{\alpha}{2n})$$

où ϕ correspond à la loi normale standard, α est le risque choisi, et n le nombre d'observations.

Pour notre exemple, on confirme par ce test de discordance qu'il n'y a pas de point extérieur, aucun résidu n'est supèrieur en valeur absolue à la valeur critique 2.96.

IDENTIFICATION DES POINTS INFLUENTS





Les h_i élevés correspondent à des observations qui sont assez éloignées des autres en ce qui concerne les valeurs prises par les variables explicatives. Ces observations, même quand elles ne correspondent pas à des points extérieurs (leur résidu peut être faible), peuvent avoir une forte influence sur les estimations des paramètres du modèle.

Il est donc important de les détecter.

$\frac{2k}{n}$ (où k est le nombre de paramètres du modèle), qui correspond à deux fois la valeur moyenne des éléments diagonaux de la matrice H, constitue une valeur seuil au-delà de laquelle on peut considérer une valeur de h_i comme étant anormalement élevée.

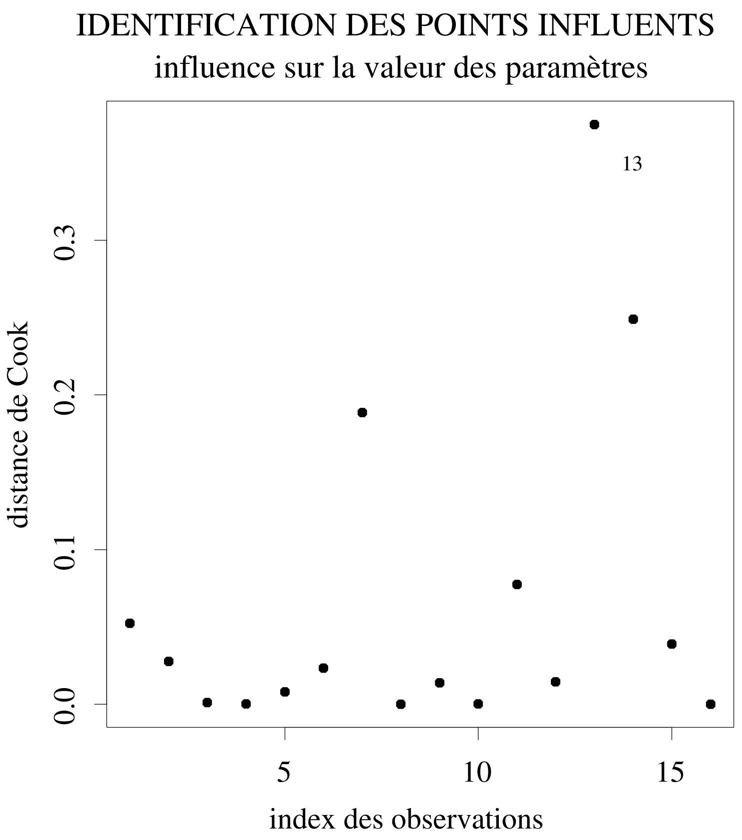
Dans notre exemple il n'y a pas de point à hi anormalement élevé.

Le résidu de vraisemblance étant une moyenne pondérée du résidu de déviance et du résidu de Pearson, on utilise le carré du résidu de vraisemblance pour mesurer l'influence d'une observation sur l'ajustement. Cette mesure est intermédiaire entre le carré du résidu de déviance et le carré du résidu de Pearson. D'après Collett, c'est une approximation du changement de déviance, quand on omet la i^e observation. Remarquons qu'on peut calculer le changement de déviance exact en omettant l'observation et en réajustant le modèle.

Un graphique représentant les $r_{L_i}^2$ permet donc de visualiser l'impact des différentes observations sur l'ajustement du modèle.

Dans notre exemple, la 13e observation apparaît comme étant plus influente que les autres.

IDENTIFICATION DES POINTS INFLUENTS



On peut également regarder l'influence d'une observation sur l'estimation des paramètres du modèle.

La distance de Cook (appelée statistique D par Collett) permet de comparer le vecteur θ (paramètres estimés sur le jeu de données complet) et le vecteur $\theta_{(i)}$ (paramètres estimés sur le jeu de données sans l'observation i).

$$\text{Distance de Cook} = \frac{R_{P_i}^2 h_i}{k(1-h_i)^2} = \frac{r_{P_i}^2 h_i}{k(1-h_i)}$$

(k est le nombre de paramètres du modèle).

La 13^{e} observation apparaît comme ayant le plus fort effet sur l'estimation des paramètres (cependant cet effet semble relativement modéré, et proche de l'effet des autres observations).

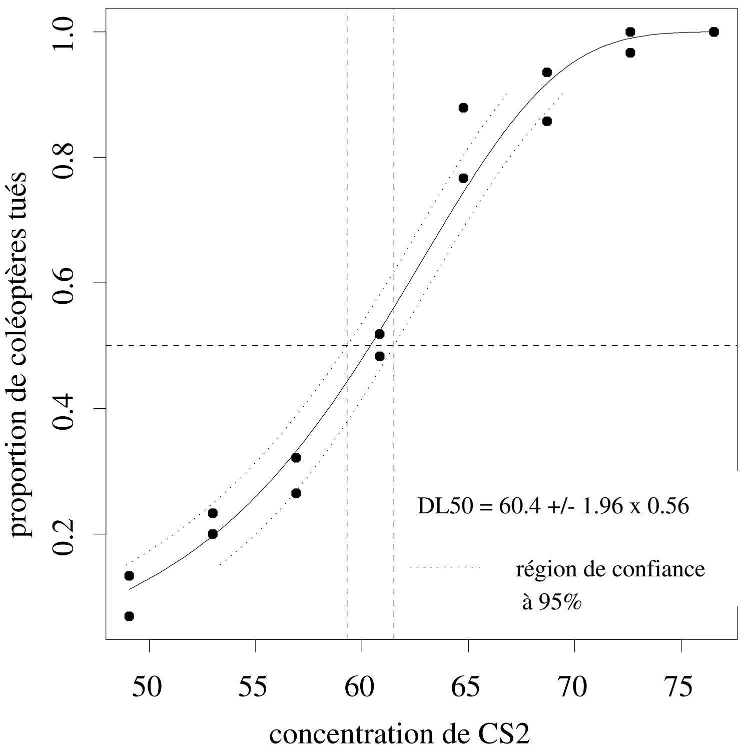
Si on n'a pas de raison particulière d'éliminer cette 13^{e} observation (erreur d'enregistrement ou autre...), il est quand même important de faire un ajustement du modèle sans ce point influent.

On obtient : $\log(-\log(1-\widehat{p})) = -9.67 + 0.153cc$

avec une déviance résiduelle égale à 4.5 pour 13ddl.

La suppression de la 13^{e} observation joue donc effectivement très peu sur l'estimation des paramètres, mais conduit à une réduction de déviance assez importante.

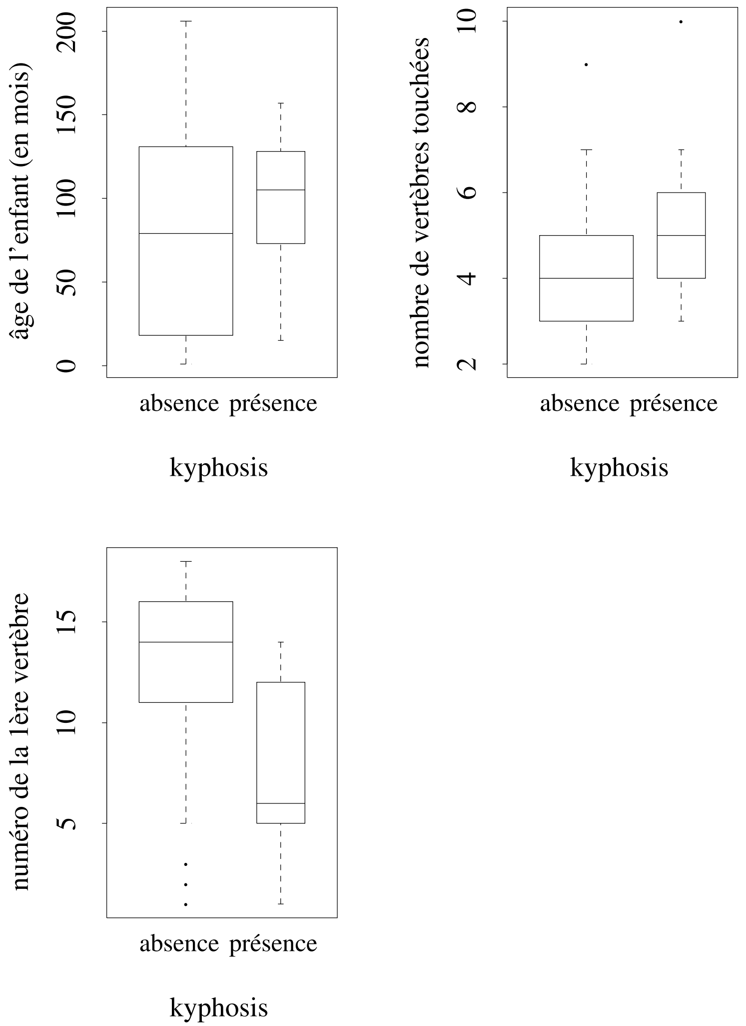
Modèle ajusté, DL_{50}



Une fois le meilleur modèle établi, on peut l'utiliser par exemple pour calculer une dose léthale 50, avec son intervalle de confiance.

On peut également tracer les courbes de confiance pour l'ensemble du domaine d'ajustement.

DONNÉES KYPHOSIS



On prend maintenant un nouvel exemple pour étudier le cas où on a plusieurs variables explicatives.

Un kyphosis est une déformation de la colonne vertébrale pouvant survenir à la suite d'une opération chirurgicale.

Pour 81 enfants ayant subi cette opération, on note :

- la présence ou l'absence d'un kyphosis (variable réponse binaire, qui suit donc une loi de Bernoulli).
- l'âge de l'enfant au moment de l'opération (en mois).
- le nombre de vertèbres touchées par l'opération.
- le numéro de la première vertèbre touchée par l'opération.

On voudrait savoir si ces 3 dernières variables peuvent expliquer l'apparition d'un kyphosis.

Les boxplots montrent un décalage pour les distributions des 3 variables explicatives, entre le sous-échantillon des enfants présentant un kyphosis (effectif 17) et celui des enfants pour lesquels il n'y a pas eu de déformation post-opératoire (effectif 64).

Ce décalage est cependant plus important pour la variable "numéro" que pour les 2 autres.

MODÈLE ADDITIF, LIEN LOGIT

ESTIMATION DES PARAMÈTRES :

	valeur	écart-type	valeur du T
constante	-2.037	1.449	-1.406
numéro	-0.207	0.068	-3.051
nombre	0.411	0.225	1.827
âge	0.011	0.006	1.696

DÉVIANCE DU MODELE NUL : 83.23 pour 80 ddl
DÉVIANCE RÉSIDUELLE : 61.38 pour 77 ddl

TABLE DE DÉVIANCE :

	ddl	déviance	dl résiduels	déviance résiduelle	Pr(Chi)
modèle nul			80	83.23	
numéro	1	15.16	79	68.07	0.0001
nombre	1	3.53	78	64.54	0.0601
âge	1	3.16	77	61.38	0.0756

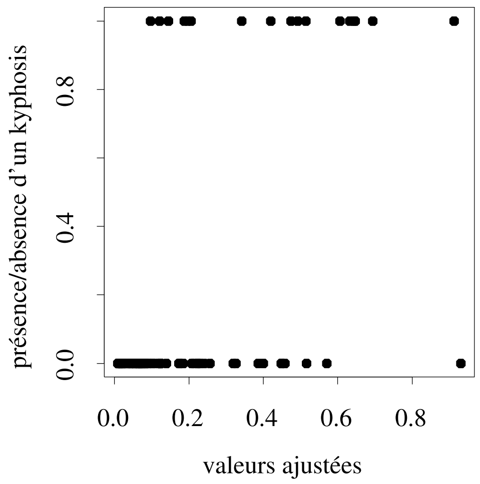
On commence par un GLM avec l'ensemble des variables explicatives et un lien logit.

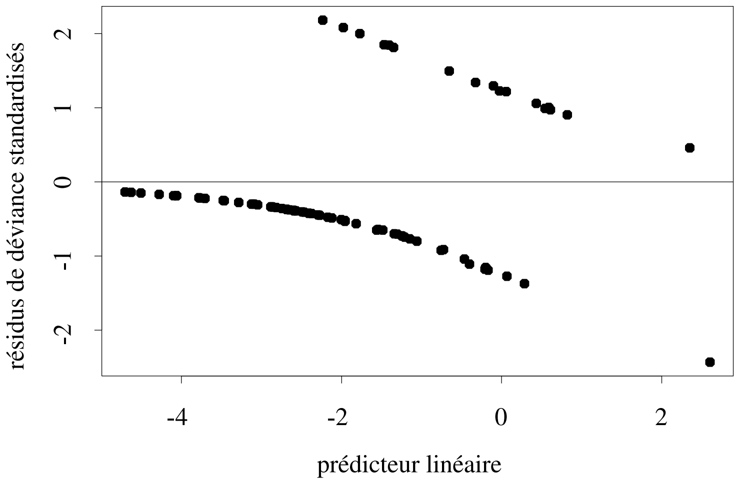
L'estimation des paramètres donne :
$\text{logit}(\widehat{p}) = -2.037 - 0.207\text{numéro} + 0.411\text{nombre} + 0.011\text{âge}$

Ici, on ne peut pas se servir de la déviance pour tester la validité du modèle : dans le cas de données binaires (i.e. non groupées) la déviance résiduelle ne suit pas une loi de χ^2.

C'est la variable "numéro" qui a l'effet le plus important sur la déviance (même lorsqu'elle est introduite en dernier dans le modèle).

VALIDATION DE MODÈLE



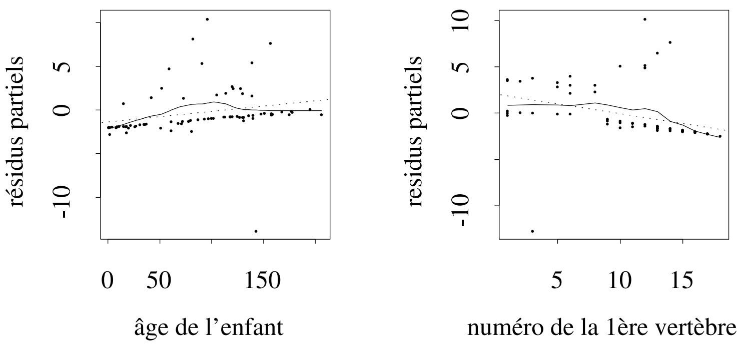


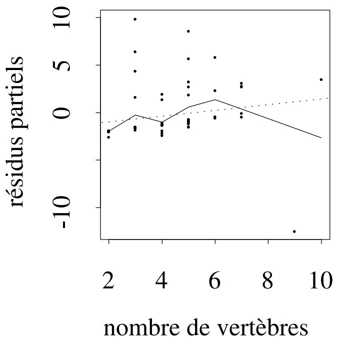
Dans le cas de données binaires, les graphiques "valeurs ajustées - valeurs observées" et "prédicteur linéaire - résidus de déviance standardisés" sont moins aisés à interpréter, car ils présentent des structures particulières dues à la nature binaire des données.

Cependant, on peut voir que ce sont surtout les cas où il y a présence d'un kyphosis qui sont mal prédits par le modèle.

Les résidus négatifs concernent les cas d'absence de kyphosis, et les résidus positifs ceux où il y a présence d'un kyphosis. Si l'ajustement était parfait, tous les résidus seraient nuls, et le graphique du haut se résumerait à un point en haut à droite et un point en haut à gauche.

VALIDATION DU MODÈLE





On représente les résidus partiels pour la variable âge en fonction de l'âge de l'enfant opéré, les résidus partiels pour la variable numero en fonction du numéro de la première vertèbre touchée, les résidus partiels pour la variable nombre en fonction du nombre de vertèbres touchées.

Les lissages superposés sont réalisés à l'aide de la fonction lowess, avec iter=0, f=1/2, c'est à dire sans itération, et une largeur de fenêtre telle, que la fenêtre contient la moitié des observations.

Les graphiques des résidus partiels présentent tous une légère tendance non linéaire, qui indique qu'une transformation des variables explicatives pourrait améliorer le modèle.

MODÈLE QUADRATIQUE

lien logit

TABLE DE DÉVIANCE :

	ddl	déviance	ddl résiduels	déviance résiduelle	Pr(Chi)
modèle nul			80	83.23	
poly(numéro,2)	2	21.53	78	61.70	0.00002
poly(nombre,2)	2	2.19	76	59.51	0.33363
poly(âge,2)	2	11.67	74	47.84	0.00292

SOUS-MODÈLE

TABLE DE DÉVIANCE :

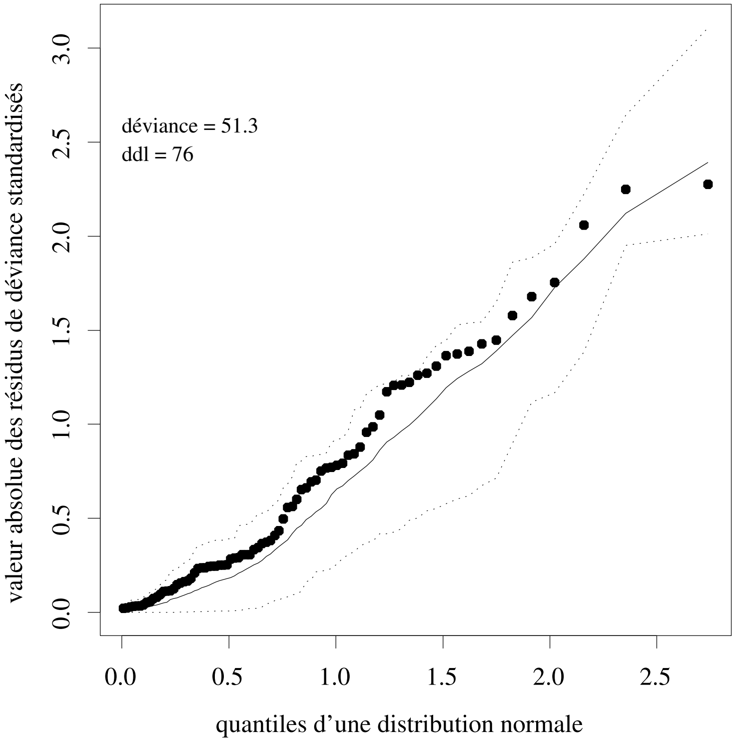
	ddl	déviance	ddl résiduels	déviance résiduelle	Pr(Chi)
modèle nul			80	83.23	
poly(numéro,2)	2	21.53	78	61.70	0.00002
poly(âge,2)	2	10.40	76	51.30	0.00551

Il semble qu'on peut se passer de la variable "nombre".

Le modèle retenant les variables "numéro" et "âge" avec des effets quadratiques semble intéressant.

VALIDATION DE MODÈLE

Modèle quadratique en muméro et en âge

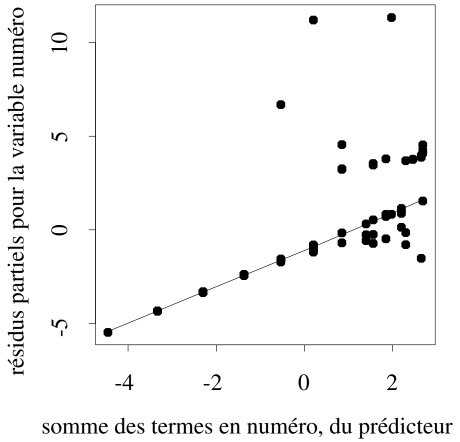


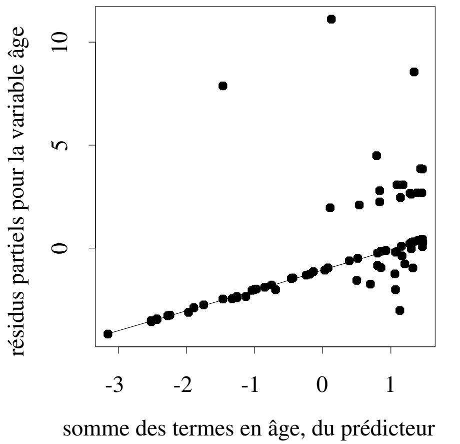
Aucun point ne s'éloigne franchement de l'enveloppe supérieure.

Cependant, les points sont presque systématiquement au-dessus de la ligne moyenne, ce qui laisse penser à une sous-paramétrisation.

VALIDATION DE MODÈLE

Modèle quadratique en muméro et en âge



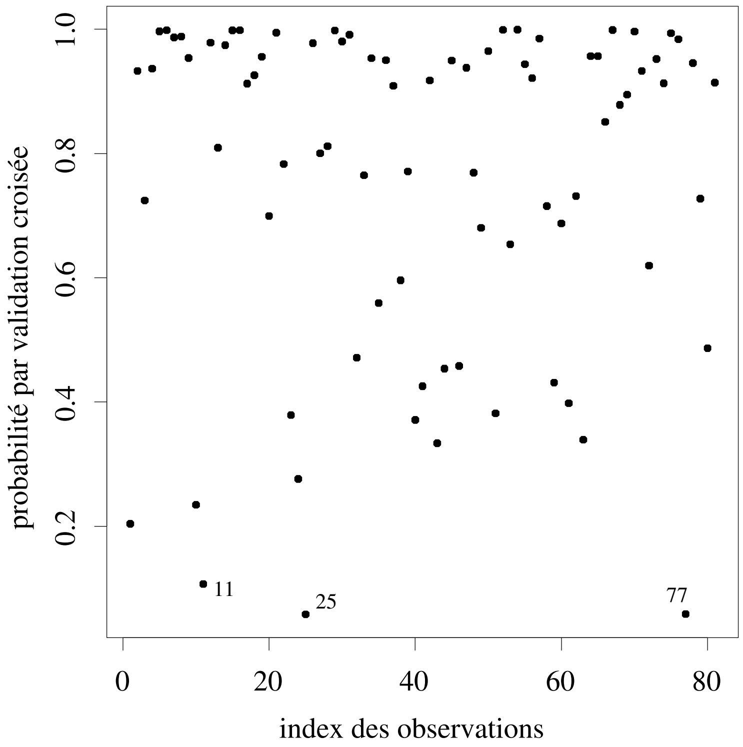


On représente en ordonnée les résidus partiels, pour chacune des variables du modèle. Dans la première représentation, on porte en abscisse la somme des termes en numéro, c'est à dire
$\widehat{\theta_{11}}\, poly(numero, 2)[, 1] + \widehat{\theta_{12}}\, poly(numero, 2)[, 2]$
où $poly(numero, 2)[, 1]$ est le terme linéaire du polynôme de degré 2 en numéro, et $poly(numero, 2)[, 2]$ est le terme quadratique du polynôme de degré 2 en numéro, et où $\widehat{\theta_{11}}$ et $\widehat{\theta_{12}}$ sont les coefficients estimés correspondants. Dans la deuxième représentation, on porte en abscisse $\widehat{\theta_{21}}\, poly(age, 2)[, 1] + \widehat{\theta_{22}}\, poly(age, 2)[, 2]$.

Le lissage superposé est obtenu à l'aide de la fonction lowess, avec les paramètres par défaut, c'est à dire avec 3 itérations, et f=1/3, c'est à dire une largeur de fenêtre déterminée de façon à contenir le tiers des observations.

Les graphiques des résidus partiels montrent une bonne efficacité des transformations de variables réalisées.

IDENTIFICATION DES POINTS EXTÉRIEURS



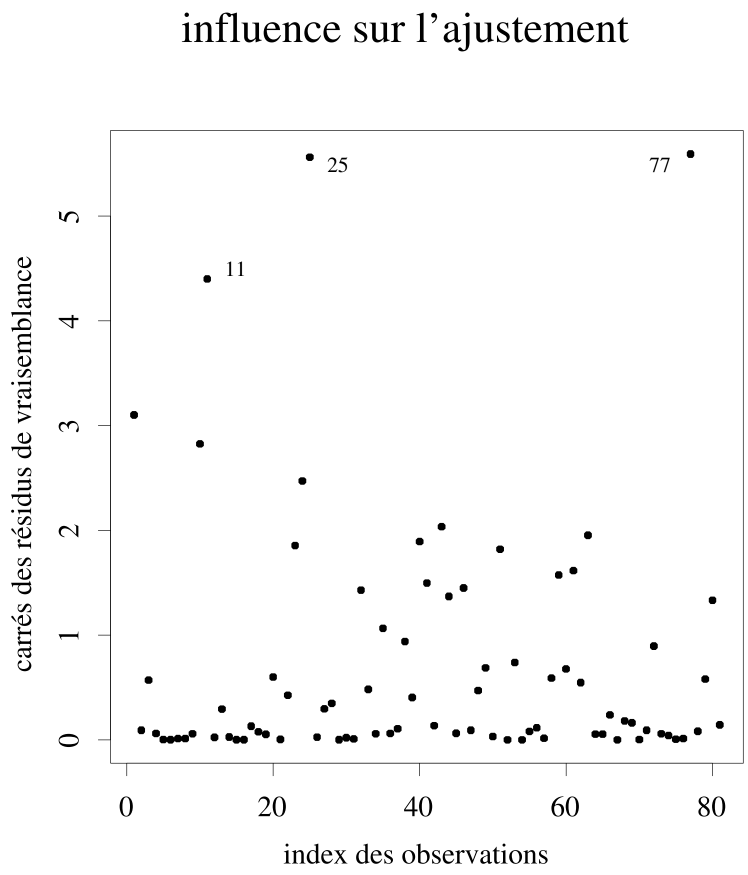
Le test de discordance utilisé dans le cas des données groupées n'est pas valide dans le cas des données binaires, car les résidus de vraisemblance ne sont alors pas distribués selon une loi normale.

Pour mettre en évidence des points extérieurs dans le cas de données binaires, on utilise les probabilités $P(y_i|Y_{(i)})$, c'est-à dire la probabilité de la i^{e} observation sachant les (n-1) autres observations.

Une valeur approximative de ces probabilités de validation croisée est donnée par $\sqrt{1-h_i}\exp -\frac{r_{L_i}^2}{2}$.

Les observations 11, 25 et 77 correspondent à des observations à probabilité faible (mais dont la probabilité est néanmoins non négligeable).

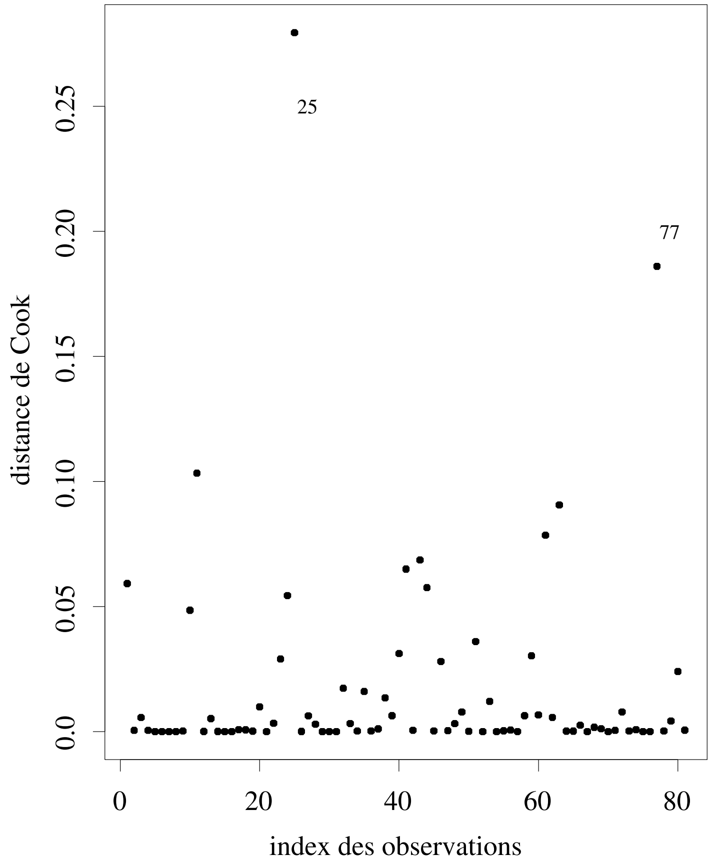
IDENTIFICATION DES POINTS INFLUENTS



Les résidus de vraisemblance les plus grands sont ceux des observations 11, 25, 77. Ces observations sont les plus influentes. Elles n'ont cependant pas une influence anormalement élevée, par rapport aux autres observations.

IDENTIFICATION DES POINTS INFLUENTS

influence sur les paramètres



La distance de Cook fait ressortir les observations 25 et 77 comme étant influentes sur l'estimation des paramètres du modèle.

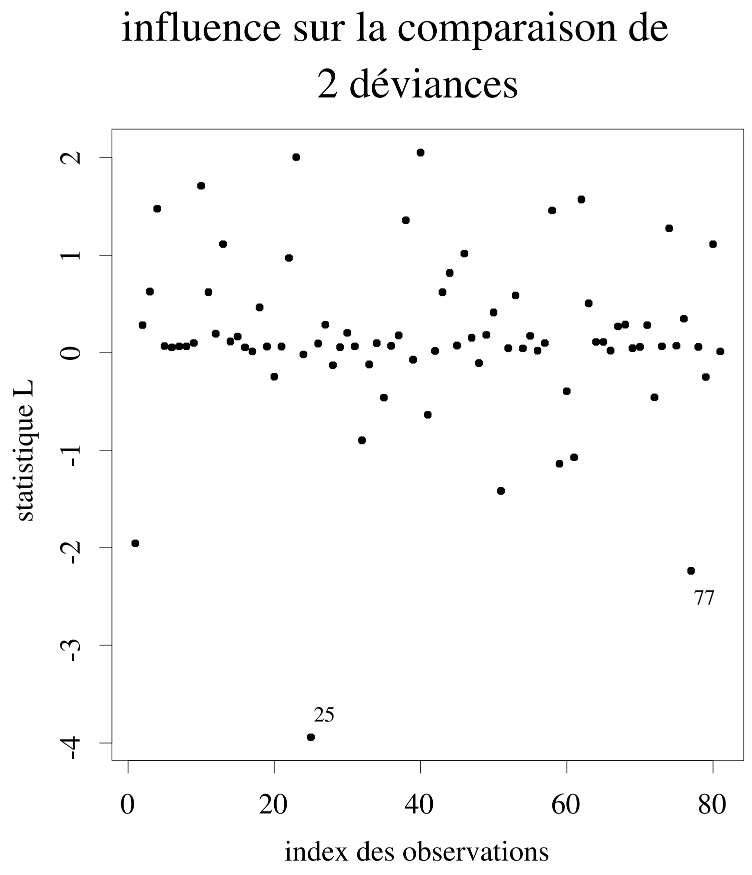
Si on ajuste le modèle après suppression des observations 11, 25 et 77, on obtient :

$\text{logit}(\widehat{p}) = -5.03 - 22.91$ numéro $- 15.93$ numéro$^2 + 10.40$ âge $-$ 25.26 âge^2

La déviance résiduelle vaut 34.1 pour 73 ddl.

L'effet des observations 11, 25 et 77 apparaît effectivement important, à la fois pour la qualité de l'ajustement (avec l'ensemble des observations, la déviance résiduelle était de 51.3) et pour la valeur des coefficients.

IDENTIFICATION DES POINTS INFLUENTS



Il reste à vérifier que la meilleure adéquation du modèle avec effets quadratiques ne soit pas seulement due à une observation particulière.

Cela revient à mesurer l'influence des différentes observations sur la comparaison des déviances des modèles avec et sans effets quadratiques (ce qui est possible parce que ces modèles sont emboîtés).

L'effet de la i^{e} observation sur la déviance est donné approximativement par $r^2_{L_i}$ et donc l'effet sur la différence de déviance entre le modèle 1 et le modèle 2 (1 emboîté dans 2) s'obtient approximativement par :

$$L_i = r^2_{L_i}(\text{modèle 1}) - r^2_{L_i}(\text{modèle 2})$$

où modèle 1 est le modèle sans effet quadratique, et modèle 2 est le modèle avec effet quadratique.

Il apparaît que les observations 25 et 77 ont une influence sur le changement de déviance entre le modèle sans effet quadratique et le modèle avec effets quadratiques :

Avec les observations 25 et 77, la déviance résiduelle passe de 65.29 à 51.29 quand on ajoute le terme quadratique. L'ajout du terme quadratique est donc, *a priori* justifié. Sans les observations 25 et 77, la déviance résiduelle passe de 60.41 à 38.57 quand on ajoute le terme quadratique. L'ajout du terme quadratique est donc parfaitement justifié.

Exercice 7
Exercice 8

ADÉQUATION DU MODÈLE

outil de validation	**type de données**	**usage**
déviance résiduelle	groupées	permet de voir si le modèle est significativement différent du modèle saturé.
graphique des résidus de déviance standardisés en fonction du prédicteur linéaire	groupées	possibilité de meilleur modèle s'il apparaît une structure dans le graphique, mise en évidence de points extérieurs.
"half-normal plot" des résidus de déviance standardisés et enveloppe simulée	groupées non-groupées	possibilité de meilleur modèle si déviation systématique par rapport à la ligne moyenne, mise en évidence de points extérieurs.
graphique des résidus partiels en fonction des variables explicatives	groupées non-groupées	une tendance non linéaire dans le graphique suggère une transformation de la variable explicative et le type de transformation (quadratique, cubique...).
changement de fonction de lien	groupées non groupées	si on n'a pas d'idée sur la distribution des seuils de tolérance, on peut tester différentes fonctions de lien, et retenir le modèle le plus adéquat au regard de la parcimonie, de la déviance, des résidus...

Récapitulation des différents outils de validation qui ont été présentés, et qui permettent de vérifier l'adéquation du modèle. Certains ne s'appliquent qu'à l'un ou l'autre des 2 types de données (groupées ou non-groupées).

On peut mentionner l'existence du "added variable plot" (pour mettre en évidence une variable explicative qu'il faudrait rajouter au modèle), et du "constructed variable plot" (qui renseigne sur la transformation qu'il faudrait réaliser sur une variable explicative). Ces 2 outils n'ont pas été présentés dans le cours car leur justification théorique n'est pas intuitive. Cependant, leur mise en oeuvre ne pose pas de problème, et on peut les utiliser aussi bien pour des données groupées que non-groupées. (Le added variable plot sera utilisé chap. V p. 22).

On pourra aussi se reporter au tableau de Collett résumant les méthodes de validation (Collett, p.178).

ÉTUDE D'OBSERVATIONS PARTICULIÈRES

outil de validation	type de données	usage
graphique indexé des probabilités estimées par validation croisée	non groupées	identification de points extérieurs.
graphique indexé des "leverage"	groupées non groupées	mise en évidence de points potentiellement influents.
graphique indexé des carrés des résidus de vraisemblance test de discordance	groupées non groupées groupées	identification des observations ayant une influence sur la qualité de l'ajustement, identification de points extérieurs.
graphique indexé des distances de Cook	groupées non groupées	identification des observations ayant une influence sur l'estimation des paramètres.
graphique indexé des statistiques L	groupées non groupées	identification des observations ayant une influence sur le changement de déviance entre 2 modèles emboîtés.

Récapitulation des outils permettant de repérer les observations ayant une influence particulière sur la déviance résiduelle et/ou l'estimation des paramètres du modèle.