Les modèles linéaires généralisés

Guillaume Constantin De Magny, Martin Desruisseaux, Michel Petit

© P. Opic

Les modèles linéaires généralisés (*GLIM*) sont une généralisation bien connue de modèle de régression linéaire dans les cas où la réponse est une variable discrète ou que le modèle est différent des modèles linéaires standards. Les modèles linéaires généralisés utilisés le plus souvent sont des modèles de régression logistiques pour des données binaires et des modèles log-linéaires pour des données non binaires. Le modèle linéaire généralisé contient deux éléments fondamentaux. Le premier est le type d'erreur qui suit une loi normale de variance constante. Le second est la fonction de lien. Dans le modèle linéaire *sensu stricto*, la liaison sous la forme y = ax + b est directement cherchée. Dans le modèle linéaire généralisé le but est de prédire la fonction de lien sous la forme :

$$\text{Log}\,[p / (1 - p)] = ax + b \qquad \text{soit} \quad p = 1/ (1 + e^{-(ax + b)})$$

Un modèle linéaire est un modèle linéaire généralisé d'erreur normale et de lien identité.

Le fichier de données concerne les pêches dont les valeurs des variables sont complètes. Les analyses sont réalisées avec le programme S-plus 2000 de la société MathSoft.

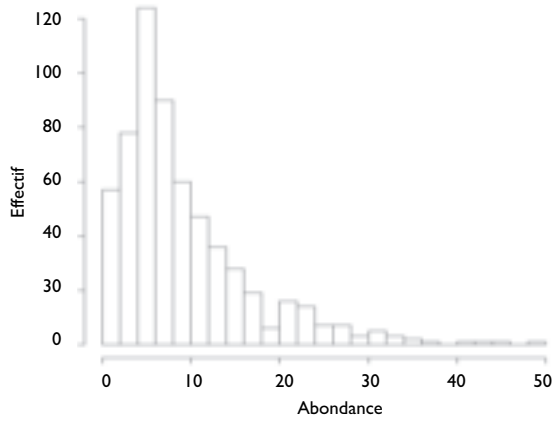
La variable « réponse » du modèle est l'abondance d'espadons « SWO » issue de pêches à la palangre. Les variables explicatrices choisies sont les 29 variables suivantes :

- La distance à la côte (distance_côte).
- La distance minimale séparant la pêche d'une autre le même jour.
- Le nombre d'hameçons (nb_hameçon).

– La bathymétrie au point de pêche.
– La concentration en chlorophylle-*a* déterminée au point de pêche 15, 10 et 5 jours avant, ainsi que le jour même et 5 jours après.
– La température de l'eau en surface au point de pêche 15, 10 et 5 jours avant, ainsi que le jour même et 5 jours après.
– L'anomalie de hauteur d'eau au point de pêche 15, 10 et 5 jours avant, ainsi que le jour même et 5 jours après.
– La composante U du courant géostrophique au point de pêche 15, 10 et 5 jours avant, ainsi que le jour même et 5 jours après.
– La composante V du courant géostrophique au point de pêche 15, 10 et 5 jours avant, ainsi que le jour même et 5 jours après.

Pour construire un modèle, il est nécessaire de connaître la loi de la distribution de la variable de réponse. Un histogramme des abondances d'espadons pêchés permet de la déterminer (fig. 92).

La forme de la distribution des données d'abondance d'espadons pêchés est similaire à une distribution selon la loi de Poisson. Le type de modèle sélectionné est poissonnien et le lien sera logarithmique « log ».



▽ *Fig. 92*
Histogramme de l'effectif des palangres en fonction de l'abondance d'espadons pêchés.

Création du modèle maximal

Les variables seront injectées dans un premier modèle. Aucune interaction ne sera prise en compte dans un premier temps.

Pour l'optimisation des paramètres, le nombre d'itérations utilisées est de 50 000 et la tolérance de convergence de 0,05. Ce critère permet d'éliminer les variables colinéaires.

Lors de l'étape de construction du modèle, le calcul « d'une analyse de variance à un facteur » (*anova*) est réalisé. Une étape supplémentaire est ajoutée ensuite qui fait appel à la fonction « *step* ». Elle a pour but d'éliminer du modèle les facteurs colinéaires. Les facteurs présentés en fin de rapport de la commande *step* sont repris et constituent le nouveau modèle maximal. Ces étapes seront répétées jusqu'à ce que la totalité des facteurs colinéaires soit retirée. Étant donné que le nombre de facteurs est encore élevé, seuls les facteurs les plus significatifs dans *l'anova* qui a précédé seront gardés.

Dans le cadre de l'analyse de l'abondance des espadons pêchés, les facteurs conservés dans le modèle maximal sont :

– La distance à la côte (distance_côte).

– Le nombre d'hameçons (nb_hameçon).

– La concentration en chlorophylle-*a* déterminée au point de pêche 5 jours avant (chloro a5).

– La concentration en chlorophylle-*a* déterminée au point de pêche le jour même (chloro 0).

– La concentration en chlorophylle-*a* déterminée au point de pêche 5 jours après (chloro p5).

– La température de l'eau en surface au point de pêche 5 jours après (SST p5).

– La composante U du courant géostrophique au point de pêche 5 jours avant (U a5).

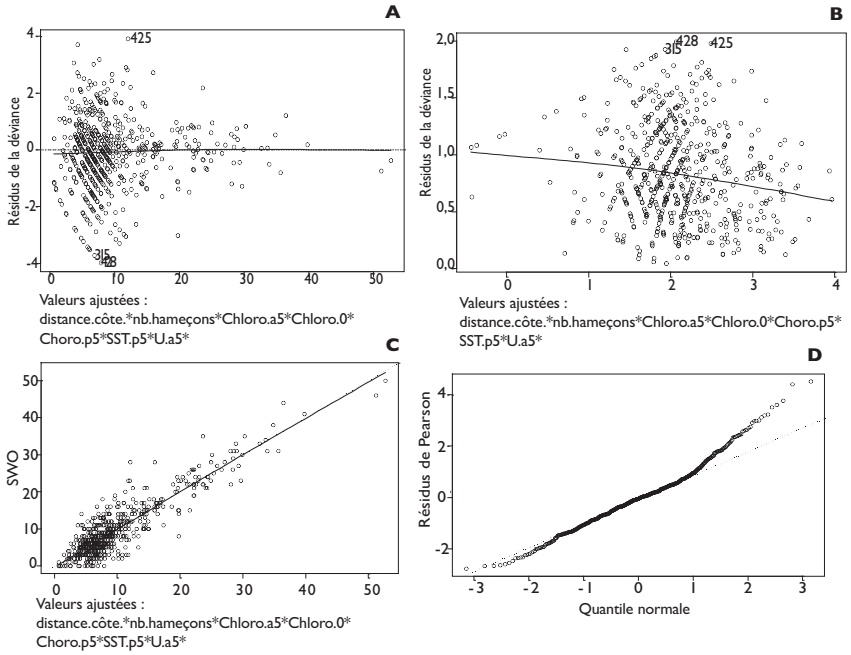
– La composante V du courant géostrophique au point de pêche le jour même (V 0).

Le modèle maximal est créé en ajoutant les facteurs simples et toutes les interactions possibles entre ces facteurs. Il est constitué de 8 variables. Cela fait 255 facteurs de 8 niveaux d'interactions différents résumés dans le tableau 8.

La déviance expliquée par le modèle maximal est de 72,54 %. En d'autres termes, le modèle explique 72,54 % de la variation de l'abondance d'espadons observée.

▽ *Tableau 8*
Récapitulatif de l'information concernant le modèle maximal à 255 termes.

Nombre de facteurs composant les interactions	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Total**
Effectif	8	28	56	70	56	28	8	1	255
Somme de la déviance	1 277,93	214,88	222,78	299,21	186,87	58,68	31,56	0,04	2291,96
Part dans la déviance (en %)	55,76	9,38	9,72	13,06	8,15	2,56	1,38	< 0,01	100
Déviance (en %)	40,45	6,80	7,05	9,47	5,92	1,86	1,00	< 0,01	72,54



▽ *Fig. 93*
Représentation des différents paramètres du modèle.
A : les résidus vs *les valeurs ajustées ;*
B : les valeurs absolues des résidus élevées au carré vs *les valeurs ajustées ;*
C : la réponse vs *les valeurs ajustées,*
D : graphique quantile-quantile normal des résidus (normalité des résidus).

Les différents graphiques de la figure 93 permettent de juger de la qualité du modèle. Le graphique D montre un écart de la distribution des résidus par rapport à la normalité. Il s'agit d'une sous-dispersion de la réponse du modèle pour de faibles valeurs ajustées ainsi qu'une sur-dispersion pour des valeurs ajustées élevées.

Le tableau 8 résume les caractéristiques du modèle.

La première ligne du tableau classe les facteurs par catégories. Il y a dans la colonne 1, les 8 facteurs correspondants aux 8 variables retenues pour le modèle maximal. Dans les colonnes 2 à 8 sont regroupées les interactions d'ordre 2, 3 à 8. La colonne totale indique que le modèle prend en compte 255 termes. La somme de la déviance correspond au cumul des déviances de chacun des termes regroupés dans les colonnes. La part dans la déviance pour chaque classe de facteurs et interactions est calculée de la manière suivante : la somme de la déviance pour chaque classe de facteurs est divisée par la déviance totale (ici égale à 2 291,96). La valeur est donnée en pourcentage. Cela nous renseigne sur la part de chacune des classes de facteurs dans la déviance totale expliquée par le modèle. La dernière ligne du tableau correspond à la déviance expliquée par chacune des classes.

▽ *Tableau 9*
Facteurs simples et interactions de type 2 et 3 hautement significatifs et leurs valeurs de déviances et contribution dans la déviance ainsi que les coefficients déterminés par le modèle.

	Contribution des facteurs dans la déviance (en %)	**Déviance (en %)**	**Coefficients**
Ordonnée à l'origine			- 36,43744
Nb_hameçons	25,36	18,40	- 0,1731539
Distance_côte	22,98	16,67	8,030329
U a5	2,56	1,86	10,10388
Chloro p5	2,22	1,61	- 4259,952
U a5*V	1,70	1,23	- 0,4894207
Distance_côte, U a5,V 0	1,58	1,14	0,05906905
Nb_hameçons, SST p5	1,55	1,13	0,00901742
Nb_hameçons, Chloro a5	1,32	0,96	0,2491969
Nb_hameçons, Chloro p5,V 0	1,13	0,82	- 0,2748209
Chloro 0	0,95	0,69	11501,15
Distance_côte, Chloro p5, U a5	0,81	0,59	2,782335
SST p5	0,80	0,58	38,6342
Distance_côte, U a5	0,80	0,58	- 0,2391531
Chloro a5	0,79	0,57	11551,35
Chloro p5, SST p5	0,77	0,56	199,9692
Distance_côte, Chloro 0	0,60	0,44	93,37972
Nb_hameçons, Chloro a5, U a5	0,55	0,40	- 0,9301072
Nb_hameçons, SST p5, U a5	0,49	0,35	- 0,0007529
Total	**66,97**	**48,58**	

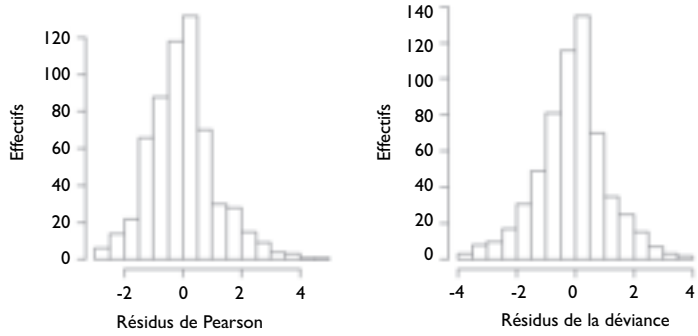
Les facteurs simples sans interactions représentent 55,76 % de la déviance expliquée par ce modèle, 9,376 % par les interactions de type 2 et 9,72 % par les interactions de type 3. La part importante des interactions associant au moins deux facteurs dans cette déviance montre que le modèle est multifactoriel.

Les différents facteurs et interactions de type 2 et 3 hautement significatifs ($Pr < 0,001$) sont résumés dans le tableau 9.

Ainsi, 66,97 % de la déviance sont expliqués par les 18 termes les plus significatifs. En d'autres termes, le modèle explique 48,58 % de la variation de l'abondance d'espadons pêchés avec ces 18 facteurs.

La distribution des résidus est représentée graphiquement afin de vérifier si leur distribution est de forme gaussienne (fig. 94). Les résidus de Pearson sont une version des résidus de travail sur une autre échelle. Leur somme des carrés des écarts est une statistique du Khi2. Les résidus de travail sont déterminés par la soustraction de la valeur ajustée du modèle à la valeur de la réponse.

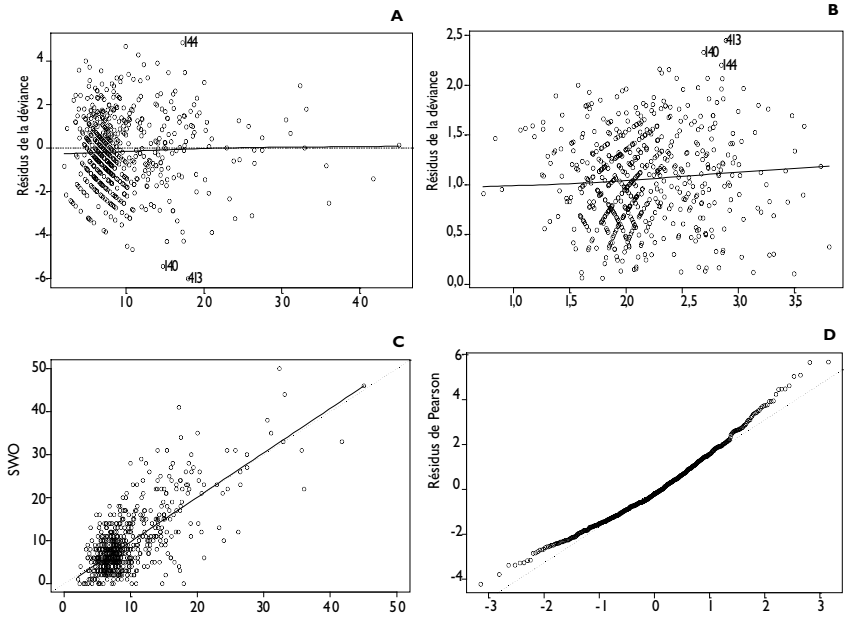
La forme des deux courbes est semblable à une distribution gaussienne. La légère tendance aux deux distributions d'être déséquilibrées vers la partie gauche trouve son origine dans la sur-dispersion du modèle dans le cadre de la prédiction des grandes valeurs.



▽ *Fig. 94*
Histogramme des résidus de Pearson (résidus calculés sur une autre échelle) et de la déviance pour le modèle maximal.

Simplification du modèle maximal :

L'étape suivante consiste à construire un modèle plus simple. Le modèle est constitué des 8 facteurs simples et des 28 interactions à 2 termes uniquement. Il est constitué ainsi de 36 termes. La déviance expliquée par ce modèle est de 47,25 %. La déviance du modèle expliquée par des facteurs significatifs (Pr < 0,001) est de 45,22 % (fig. 95).



▽ *Fig. 95*
Représentation des différents paramètres du modèle simplifié.
A : les résidus vs *les valeurs ajustées ;*
B : les valeurs absolues des résidus élevées au carré vs *les valeurs ajustées ;*
C : la réponse vs *les valeurs ajustées,*
D : graphique quantile-quantile normal des résidus (normalité des résidus).

Ce modèle présente une sur-dispersion des valeurs ajustées faibles et élevées. Le tableau 10 résume les caractéristiques du modèle.

▽ *Tableau 10*
Récapitulatif de l'information concernant le modèle à 36 termes.

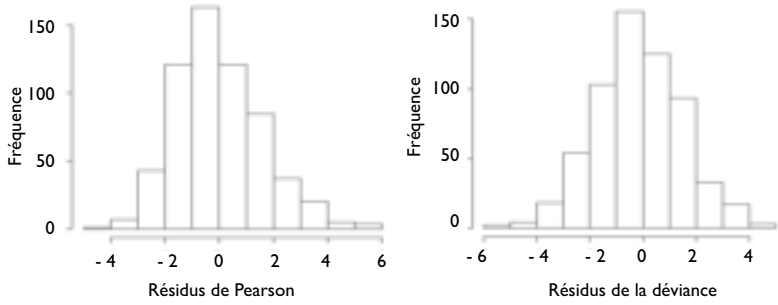
Nombre de facteurs composant les interactions	**1**	**2**	**Total**
Effectif	8	28	36
Somme de la déviance	1 277,93	214,88	1 492,82
Part dans la déviance (en %)	85,61	14,39	100,00
Déviance (en %)	40,45	6,80	47,25

Il apparaît que ce modèle explique 47,25 % de la variation de l'abondance d'espadons pêchés. Les facteurs hautement significatifs de ce modèle sont rassemblés dans le tableau 11.

▽ *Tableau 11*
Facteurs simples et interactions de type 2 hautement significatifs avec leurs valeurs de déviances et contribution dans la déviance ainsi que les coefficients déterminés par le modèle.

	Contribution des facteurs dans la déviance (en %)	**Déviance (en %)**	**Coefficients**
Ordonnée à l'origine			- 2,27
Nb_hameçons	34,88	18,4	0,0044
Distance_côte	31,61	16,67	0,0055
U a5	3,52	1,85	- 0,0039
Chloro p5	3,05	1,6	68,4648
U a5,V 0	2,34	1,23	0,0004
Nb_hameçons, SST p5	2,13	1,12	- 0,0002
Nb_hameçons, Chloro a5	1,82	0,96	0,0081
Chloro 0	1,31	0,69	- 30,882
SST p5	1,10	0,58	0,1645
Distance_côte, U a5	1,10	0,57	0,0001
Chloro a5	1,08	0,57	- 49,6965
Chloro p5, SST p5	1,06	0,55	- 2,1726
Distance_côte, Chloro 0	0,83	0,43	0,0450
Total	**85,83**	**45,22**	

Les facteurs résumés dans le tableau constituent 85,83 % de la déviance expliquée par le modèle, la déviance du modèle étant de 47,25 %. Les histogrammes de la figure 96 représentent les résidus de Pearson ainsi que les résidus de la déviance.



▽ *Fig. 96*
Histogramme des résidus de Pearson et de la déviance pour le modèle.

La normalité de distribution des résidus est vérifiée. La sur-dispersion mentionnée précédemment affecte légèrement la forme des distributions pour les valeurs extrêmes. Dans la démarche d'analyse et de modélisation par la méthode des modèles linéaires généralisés, une double approche par une autre catégorie de ces modèles est intéressante.

Modèle de maximum de vraisemblance

Le même type d'analyse est réalisé en utilisant le modèle de maximum de vraisemblance afin d'avoir une double approche de la modélisation de la variable de réponse.

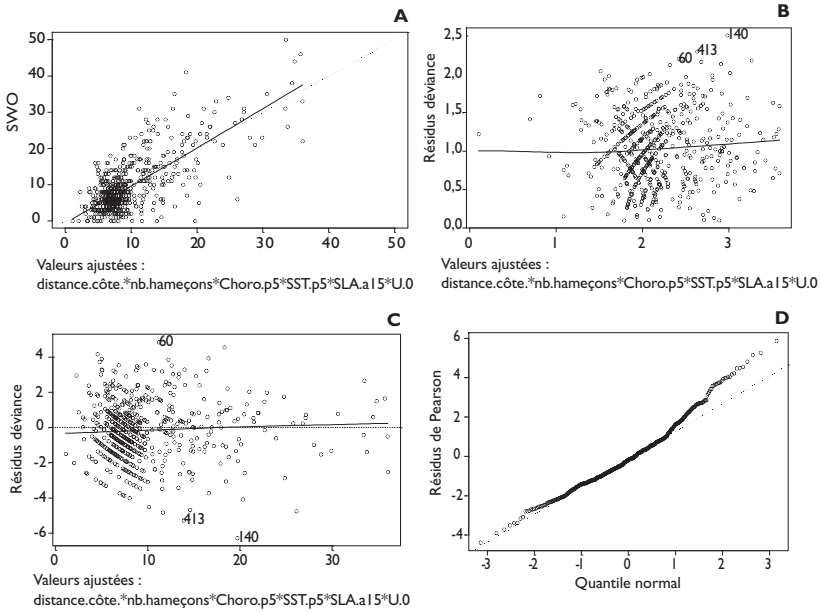
Tous les différents types de modèles, excepté le modèle utilisant la méthode du maximum de vraisemblance, sont associés à une distribution de la famille exponentielle. Pour chacun, ils ont une fonction de variance fixée. Il y a un choix typique de la fonction de lien. La famille des modèles de maximum de vraisemblance ne demande pas de correspondre à une distribution particulière pour la variable de réponse. Ainsi, les modèles dit « quasi » peuvent être définis par différents liens et fonctions de variance.

Les facteurs simples sont déterminés de la même manière que pour les deux précédents modèles. Il s'agit des facteurs suivants :

– La distance à la côte la plus proche.
– Le nombre d'hameçons.
– La concentration en chlorophylle-*a* déterminée au point de pêche 5 jours après.
– La température de l'eau en surface au point de pêche 5 jours après.
– L'anomalie de hauteur d'eau au point de pêche 15 jours avant.
– La composante U du courant géostrophique au point de pêche le jour même.

Le modèle maximal est constitué des 6 variables ci-dessus ainsi que de l'ensemble des interactions possibles entre ces facteurs. Ainsi, 63 termes constituent le modèle. Le type de lien utilisé reste « log », la fonction de la variance est égale à « mu ». Pour l'optimisation des paramètres, le nombre d'itérations utilisées est de 50 000 et la tolérance de convergence de 0,05.

La déviance expliquée par le modèle minimal est 47,36 %. Le paramètre de dispersion donné pour le modèle utilisé montre qu'il est sur-dispersé. Les différents graphiques de la figure 97 le représentent.



▽ *Fig. 97*
Représentation des différents paramètres du modèle.
A : les résidus vs *les valeurs ajustées ;*
B : les valeurs absolues des résidus élevées au carré vs *les valeurs ajustées ;*
C : la réponse vs *les valeurs ajustées,*
D : graphique quantile-quantile normal des résidus (normalité des résidus).

▽ *Tableau 12*
Récapitulatif de l'information concernant le modèle à 63 termes.

Nombre de facteurs composant les interactions	**1**	**2**	**3**	**4**	**5**	**6**	**Total**
Effectif	6	15	20	15	6	1	63
Somme de la déviance	1 277,10	131,92	81,00	38,32	23,76	0,047	1 496,15
Part dans la déviance (en %)	81,62	8,82	5,41	2,56	1,59	0,003	100,00
Déviance (en %)	38,65	4,18	2,56	1,21	0,75	0,001	47,36

De la même manière que dans le précédent modèle, la réponse est sur-dispersée pour les faibles ainsi que les fortes valeurs ajustées. Les tableaux 12 et 13 résument les caractéristiques du modèle.

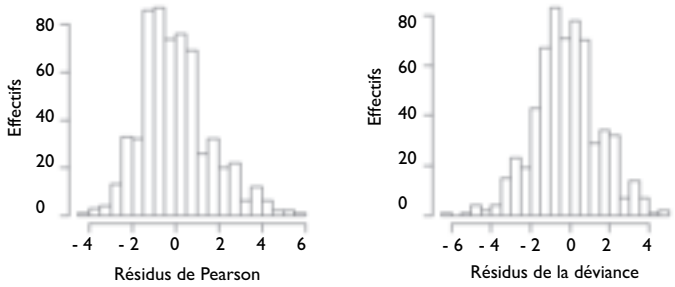
La déviance totale apportée par ce modèle n'explique que 47,36 % de la variation de l'abondance d'espadons pêchés. Les facteurs simples ont une grande part dans la variance expliquée du modèle (81,62 % de la déviance).

▽ *Tableau 13*
Facteurs simples et interactions hautement significatifs avec leurs valeurs de déviances et contribution dans la déviance ainsi que les coefficients déterminés par le modèle.

	Contribution des facteurs dans la déviance (en %)	**Déviance (en %)**	**Coefficients**
Ordonnée à l'origine			- 45,68
Nb_hameçons	38,85	18,40	0,038
Distance_côte	35,20	16,67	0,447
Nb_hameçons, SST p5	3,11	1,47	- 0,001
SLA a15	2,72	1,29	- 1,597
Distance_côte, U 0	2,50	1,19	0,030
U 0	2,00	0,95	- 5,079
SST p5	1,71	0,81	1,848
Total	**86,10**	**40,77**	

Plus nettement, les effets de sur-dispersion du modèle sont visibles sur les graphiques de distribution des résidus (fig. 98).

D'autres ajustements, en particulier concernant les fonctions de variance ont été réalisés. Mais les modèles obtenus n'expliquaient pas autant de variance que le modèle présenté.



▽ *Fig. 98*
Histogramme des résidus de Pearson et de la déviance pour le modèle de maximum de vraisemblance.

Discussion

Les différents modèles construits permettent d'expliquer au maximum 72,54 % de la variation des abondances d'espadons pêchés parmi les 607 pêches considérées. Les facteurs simples décrivent 40,45 % de cette variation et les interactions à 2 termes et plus décrivent les 32,09 % restant. Il apparaît une sur-dispersion pour les grandes valeurs ajustées quel que soit le modèle. La description de la variation de l'abondance des pêches d'espadons apparaît difficilement interprétable car elle fait appel à de nombreux mécanismes et interactions. Il est probable que les mécanismes à l'origine de l'abondance d'espadons lors des pêches sont d'ordre stochastique. Il ressort clairement que le nombre d'hameçons disposés sur la palangre est responsable en grande partie de la variation de l'abondance observée lors des pêches. La part des variables environnementales qui semble influencer le plus cette abondance apparaît 10 fois plus faible que le nombre d'hameçons et la distance à la côte la plus proche quel que soit le modèle. Cela peut permettre de faire un tri entre les différentes variables environnementales prises également à une échelle de temps différente.

L'utilisation d'un réseau de neurones dans un modèle prédictif semble hasardeuse compte tenu du caractère multifactoriel du système. La génération d'une variable aléatoire et introduite en tant que variable indépendante explicatrice dans ce type de système apparaîtrait comme prépondérante dans la description de la variation de l'abondance d'espadons observée. Toutefois, il serait intéressant de compléter ces différentes approches par une analyse de tableaux de contingence.