



Formation à R

Modélisation avec les GLM

SÉBASTIEN ROCHETTE, THINKR

ThinkR

Table des matières

1	Préface	1
2	Présentation de l'étude	1
2.1	Contexte	1
2.2	Objectifs	2
2.3	Données	2
2.4	Covariables	3
2.5	Ajuster un modèle de distribution d'espèces	3
2.6	Exploration des données	4
3	Préparation	4
3.1	Structure des dossiers	4
3.2	Débutons avec R	5
4	Exploration des données	5
4.1	Étapes	5
4.2	Liste des différentes étapes	5
5	Modélisation	7
5.1	Étapes	7
5.2	Interpréter les sorties de modèles	8
5.3	Trouver le meilleur modèle	11
5.4	Prédictions du modèle	12

1. Préface

La version d'origine de cette formation a été créée par Olivier Le Pape et Étienne Rivot à Agrocampus Ouest (Rennes, France). Depuis mon doctorat dans leur équipe, je mets à jour constamment cette formation au gré de ma recherche et de l'évolution du logiciel R.

Generated with R and rmarkdown: Roadmap version - Teacher

2. Présentation de l'étude

Le contexte et les objectifs de votre étude définissent le type de modélisation que vous allez mettre en place sur votre jeu de données.

Ici, nous utilisons les modèles linéaires généralisés pour produire une carte de distribution moyenne de la nourricerie de soles communes de la baie de Vilaine.

2.1. Contexte

- Les zones côtières et les estuaires sont des habitats halieutiques essentiels
 - Zones à forte production
 - Nourriceries
 - Zones restreintes avec de fortes densités (Fig. 1)
- Pression anthropique élevée
 - Perte de surface disponibles (Fig. 2a)
 - Qualité des habitats altérée (Fig. 2b)
- Impact sur le renouvellement des populations
 - Jeune stades = Goulet d'étranglement
 - La taille et la qualité des nourriceries côtières influent sur la production de juvéniles

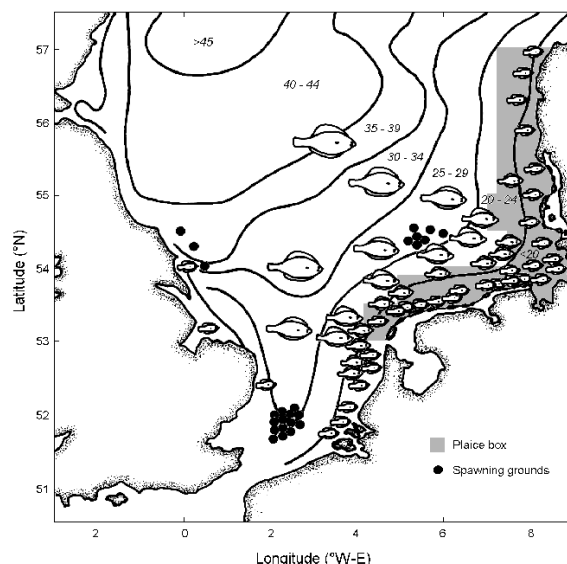


Figure 1 – Plaice box (Rijnsdorp *et al.*)

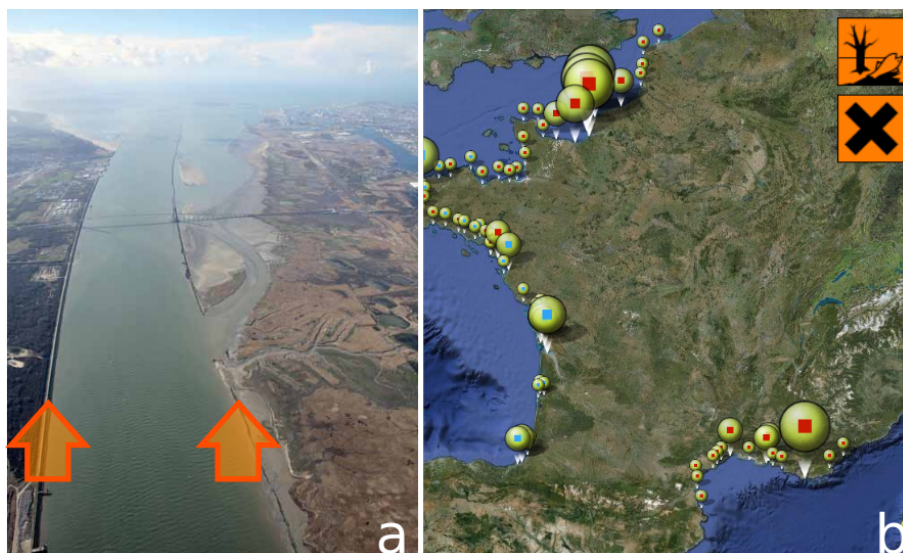


Figure 2 – (a) L'estuaire de la Seine. (b) Niveau de contamination chimique le long des côtes françaises (Ifremer, 2011)

2.2. Objectifs

Déterminer les facteurs ayant une influence sur la distribution des poissons plats (*Solea solea*) en Baie de Vilaine et cartographier la distribution moyenne des densités.

- Cartographier les habitats potentiels nécessite:
 - Connaissance des habitats de juvéniles
 - Campagnes d'échantillonnage dans la zone d'étude
 - Connaissance des covariables environnementales ayant potentiellement de l'influence
 - Cartes exhaustives des covariables environnementales
- Une approche statistique en deux étapes
 - Modèle statistique reliant les densités aux covariables
 - Prédire les habitats potentiels

2.3. Données

Campagne standardisée de chalut à perche dans la baie de Vilaine (Fig. 3)

- 1984 – 2010
- En automne
- Juvéniles de l'année (Âge 0)
 - Nb individus / 1000m²

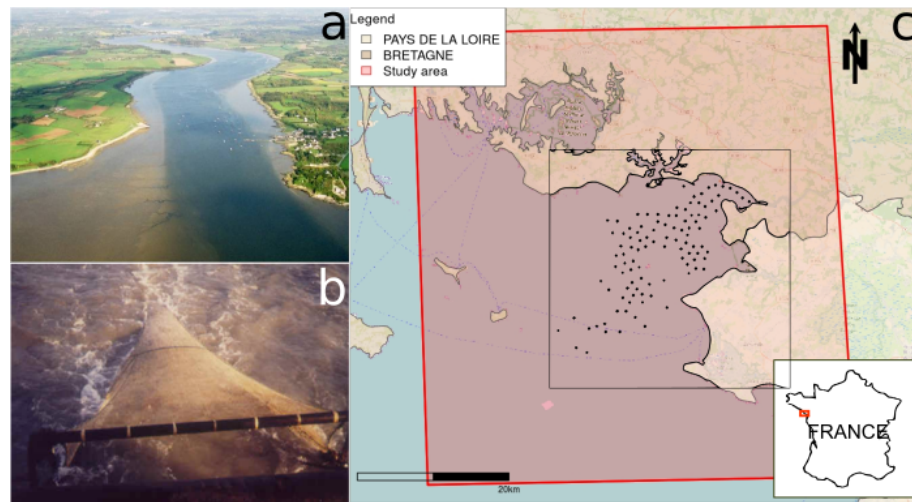


Figure 3 – (a) L'estuaire de la Vilaine. (b) Chalut à perche. (c) Situation des stations d'échantillonnage.

2.4. Covariables

- Bathymétrie (Fig. 4a)
 - MNT à 1000m de résolution
 - Projection Mercator
- Structure sédimentaire (Fig. 4b)
 - Fichier shape de polygones
 - Coordonnées géographiques
- Zones biologiques (Fig. 4c)
 - Combinaison bathymétrie, sédiment, habitat
 - Fichier shape de polygones
 - Coordonnées géographiques

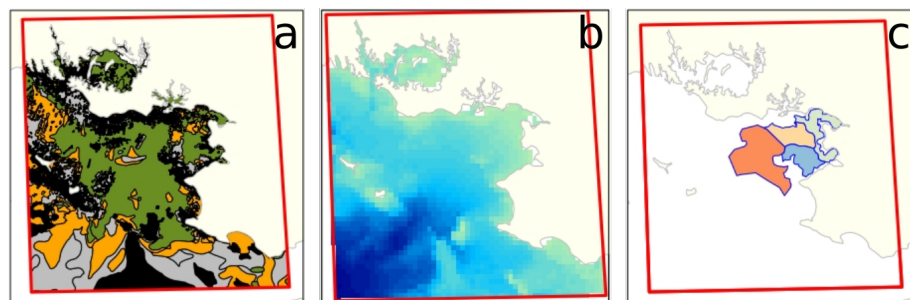


Figure 4 – Covariables en baie de Vilaine. (a) Structure sédimentaire, (b) Bathymétrie et (c) Zones biologiques.

2.5. Ajuster un modèle de distribution d'espèces

- Croiser les données avec les cartes de covariables
 - Utiliser un modèle linéaire
- Utiliser les cartes des covariables pour la prédiction (Fig. 5)
 - Une prédiction pour chaque cellule d'un raster

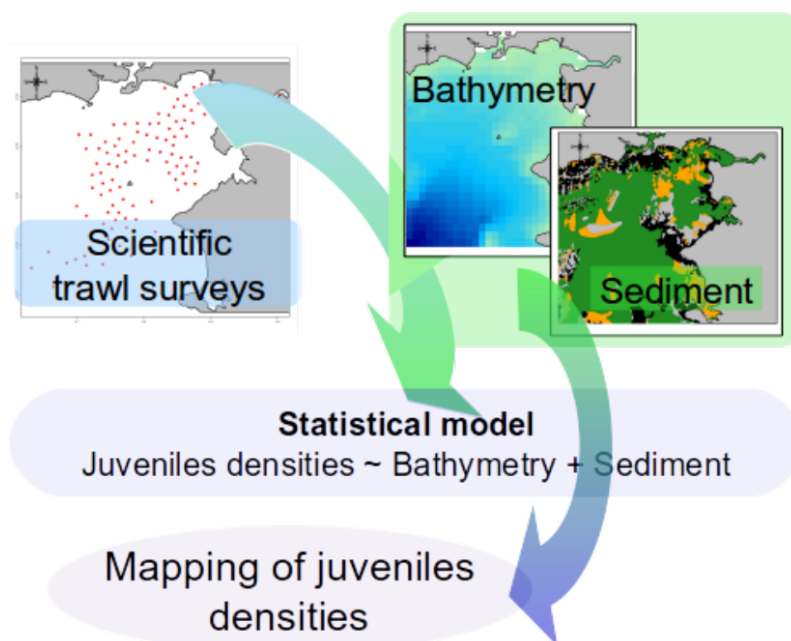


Figure 5 – Procédure pour un modèle de distribution d'espèce

2.6. Exploration des données

Prenez le temps d'explorer vos données avant toutes analyses

- Explorer les données et les covariables
 - Explorer le plan d'échantillonnage
 - Explorer les liens potentiels entre les densités et les covariables
 - Explorer les futurs paramètres de modélisation (interactions, distributions)

Souvenez-vous toujours des objectifs de votre étude !

Question : Que recherchons-nous dans cette exploration ?

3. Préparation

3.1. Structure des dossiers

Il convient de toujours conserver les fichiers originaux : les reprojections entraînent toujours quelques pertes, mieux vaut revenir aux originaux lorsque c'est possible.

L'arborescence de votre dossier de travail est la suivante :

- 01_Original_data
 - DEPARTEMENTS
 - Sedim_GDG_wgs84
 - bathy_GDG_1000_merc (and co)
 - Data_Vilaine_solea.csv
- 02_Outputs
- 03_Figures
- 04_Functions

3.2. Débutons avec R

- Créer un projet Rstudio dans le dossier principal de travail.
- Ouvrez le script R : "Quick_AllDataModel_Teacher.R"
- Lister les différents sous-dossier de travail au début de votre script R

```
# Define working directories -----
WD <- here()
# Folder of original files
origWD <- here("01_Original_data")
# Folder for outputs
saveWD <- here("02_Outputs")
# Folder where to save outputs from R
figWD <- here("03_Figures")
# Folder where complementary functions are stored
funcWD <- here("04_Functions")
```

4. Exploration des données

4.1. Étapes

Souvenez-vous : Définissez ce que vous cherchez, à quelles questions vous souhaiteriez répondre !

- Explorer la répartition du plan d'échantillonnage en fonction des covariables environnementales
- Explorer les données d'observation au regard des covariables environnementales pour détecter de potentielles corrélations
- Explorer les interactions entre les effets des covariables sur les observations
- Explorer les lois de distribution possibles (gaussien, log-normal, ...) des observations en fonctions des combinaisons de covariables

Les scripts qui sont fournis ne sont que des exemples, ils ne sont pas des solutions ! Faites vos propres tests !

4.2. Liste des différentes étapes

- Lire le jeu de données spatialisé (Fig. 6)
- Ajouter une nouvelle covariable : la bathymétrie divisée en classes
 - "< 5 m", "5-10 m", "10-20 m", "20-50 m"
- Explorer la répartition des observations en fonctions des covariables
 - Centrer l'analyse sur l'année, la bathymétrie et le sédiment
 - *Que remarquez-vous ?*
- Explorer les covariables ayant potentiellement des effets sur les densités
 - *Quelles covariables pourraient avoir une influence ?* (Fig. 7)

Les modèles statistiques que nous allons utiliser peuvent se résumer de cette façon :

$$Density = Covar1 + Covar2 + Noise$$

Comme vous le savez, on cherche toujours à savoir si les données sont gaussiennes pour pouvoir procéder à l'analyse statistique. Si elles ne sont pas gaussiennes, nous devons définir le type de distribution pour pouvoir utiliser une transformation de données.

- Explorer la distribution des données
 - *Quelle est la distribution la plus intéressante ?*

La figure 8 montre différents exemples de distributions.

Exemple de l'effet de deux facteurs (Fig. 9)

- *Qu'en pensez-vous ?*
- Explorer les interactions potentielles entre les covariables

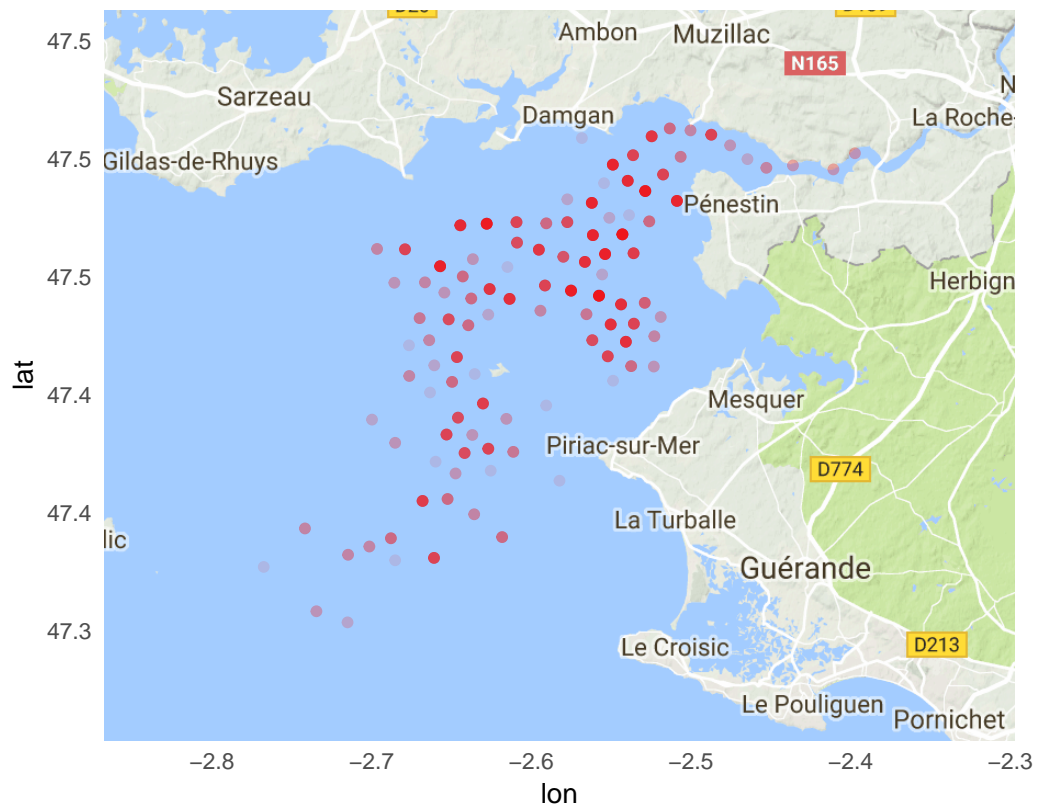


Figure 6 – Répartition des stations d'échantillonnage

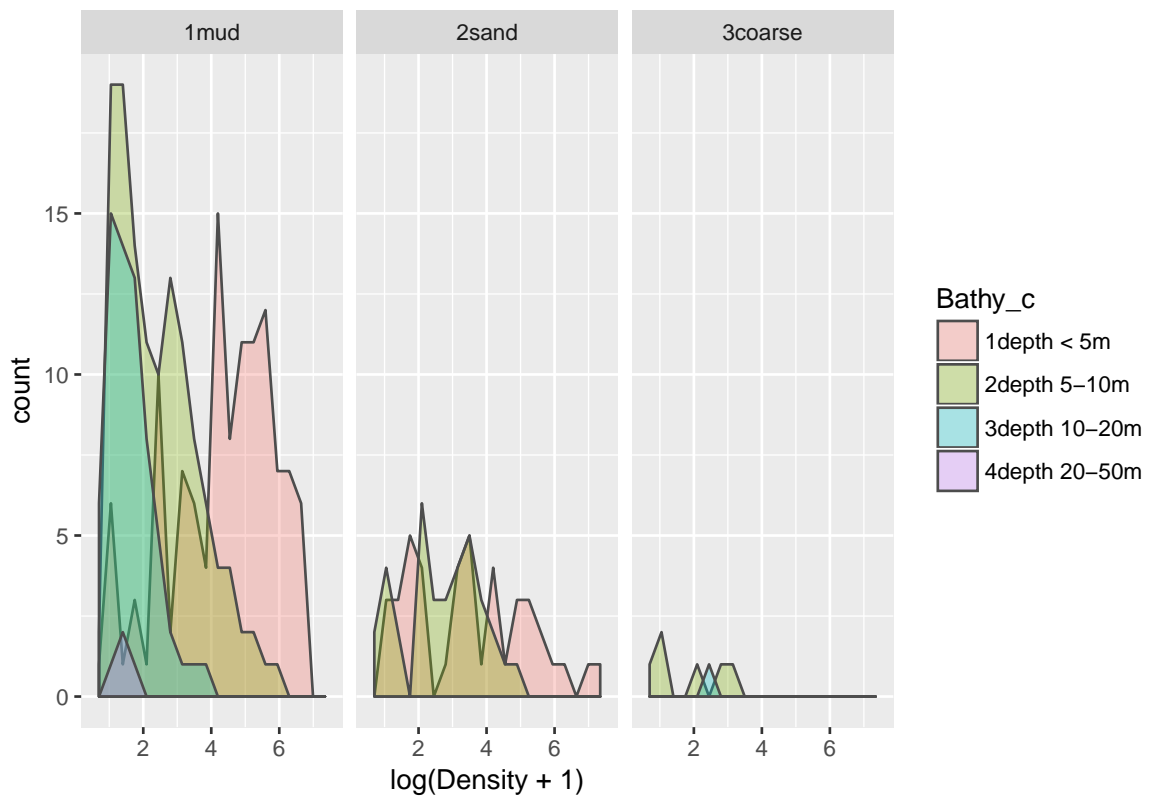


Figure 7 – Densités (log-transformées) en fonction de la bathymétrie et des sédiments

- *Qu'en pensez-vous ?*
- Comme aide à l'interprétation, utiliser l'exemple théorique de la figure 10

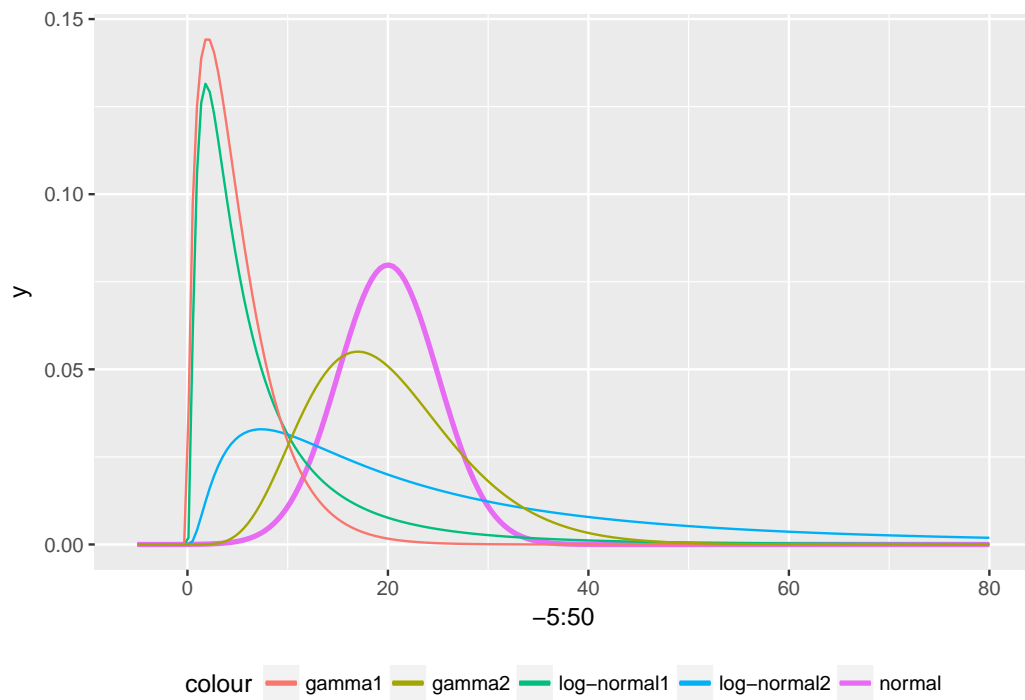


Figure 8 – Différents exemples de distributions

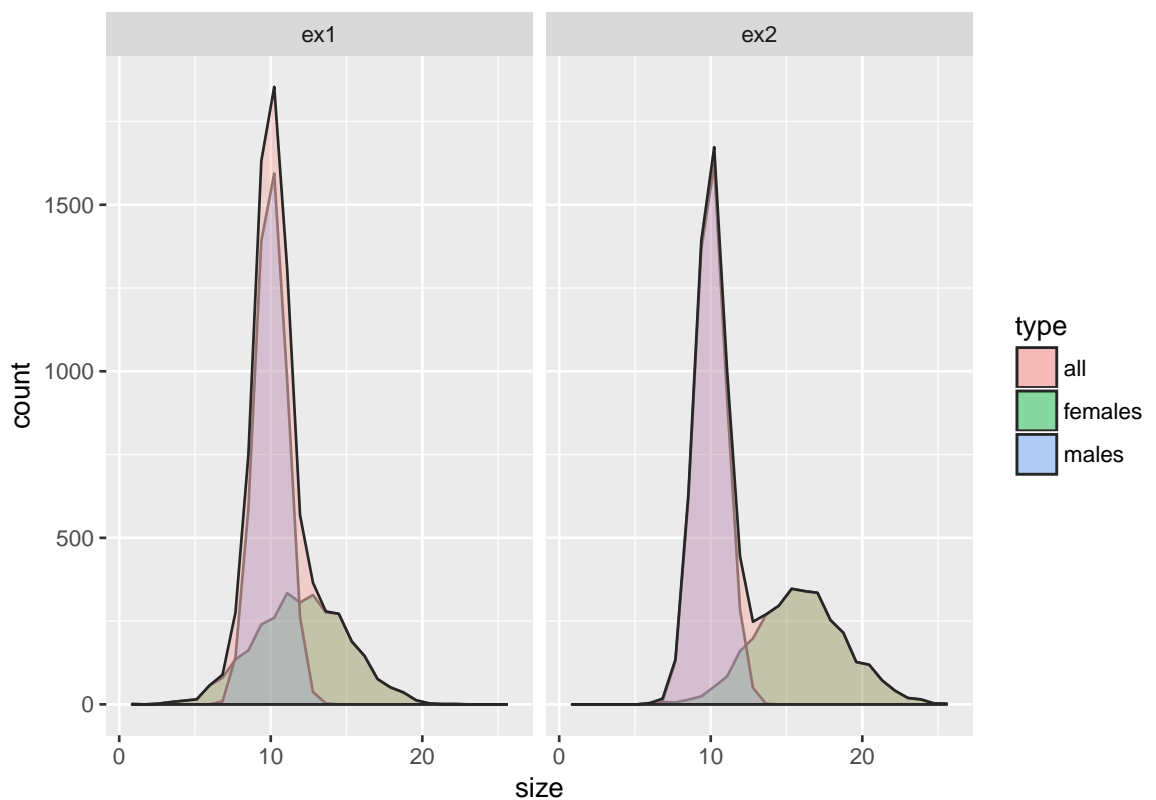


Figure 9 – Différents effets de deux facteurs

5. Modélisation

5.1. Étapes

Souvenez-vous: Définissez ce que vous cherchez, à quelles questions vous souhaiteriez répondre !

- Tester les différentes formes de modèles au regard des combinaisons de covariables et des

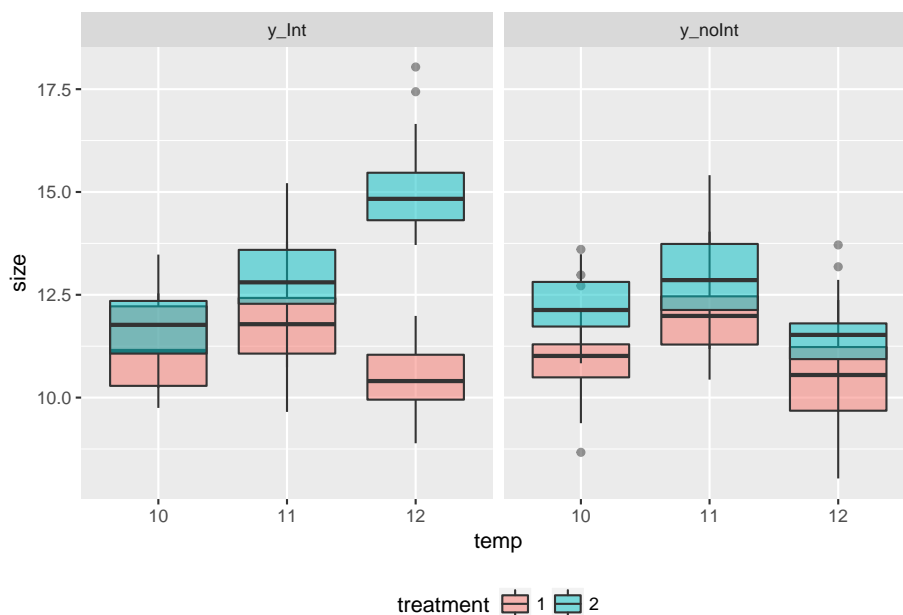


Figure 10 – Interaction entre traitements et temp  rature

formes de distributions des r  sidus

- Comparer les mod  les    l’aide des outils statistiques    disposition (AIC, anova, ...), de la validation crois  e mais aussi de la connaissance du jeu de donn  es et des questions cibl  es
- Analyser les r  sidus des mod  les. Analyser leur distribution et s’assurer que les hypoth  ses de construction sont v  rifi  es.

C’est seulement lorsque les hypoth  ses sur la distribution des r  sidus sont v  rifi  es, que les covariables et les interactions s  lectionn  es peuvent commencer      tre interpr  t  es...

Les scripts qui sont fournis ne sont que des exemples, ils ne sont en aucun cas les meilleures solutions ! Fa  tes vos propres tests !

5.2. Interpr  ter les sorties de mod  les

Lorsque vous ajustez un mod  le lin  aire (lm ou glm), vous pouvez utiliser diff  rents tests statistiques et visuels qui r  pondent    diff  rentes questions. Votre question principale pourrait   tre :

- “Est-ce que mes covariables ont un effet sur mes observations ?”. En r  alit  , ce n’est pas exactement la question    laquelle va r  pondre votre mod  le. Ce serait plut  t “Est-ce que les covariables que j’ai utilis  es expliquent une part de la variabilit   de mes observations ?”

Pour que vous puissiez interpr  ter les diff  rentes sorties de mod  les, dans ce document, nous allons regarder le mod  le suivant :

`lm(Density Bathy + Sedim, data = dataset)`

Ce mod  le n’est pas forc  ment le meilleur mod  le    choisir !

5.2.1 Summary(lm)

Cette fonction montre un tableau de tests de significativit   (Table 1). Ce sont des tests de Student. Ils testent si la valeur estim  e pour un effet est ou non significativement diff  rente de z  ro. Ainsi, si une covariable a un effet non significativement diff  rent de l’effet nul, il est probablement inutile de la conserver dans le mod  le.

- Ici, ce qui est appel   “Intercept” est l’effet de base. Dans une   quation $y = a.x + b$, l’“intercept” serait b . Ici, c’est un peu diff  rent car il y a des covariables au format facteur (“Sedim”). Dans cet exemple, l’“intercept” montre une “p-value” proche de z  ro, ce qui signifie que son effet (`estimate ~ 100`) est significativement diff  rent de z  ro.

Table 1 – Exemple d’une sortie de ‘summary(lm)’

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	173.4	14.27	12.152	0.000
Bathy	14.5	1.71	8.472	0.000
Sedim2sand	-19.5	19.07	-1.020	0.308
Sedim3coarse	-39.4	57.21	-0.689	0.491

- La covariable “Bathy” est continue. Dans une   quation de type $y = a.x + b$ ce serait a . Dans cet exemple, son effet `estimate` ~ 6 , est significativement diff  rent de z  ro (p -value ~ 0).
- La covariable “Sedim” est un facteur    trois niveaux. Dans le “summary”, vous ne pouvez en voir que deux (“2sand”, “3coarse”). En r  alit  , les effets des niveaux de facteur (*estimate*) sont compar  s au premier niveau (“1mud”), pour lequel l’effet est   gal    z  ro. Dans l’  quation $Y = a * Bathy + b[Sedim] + c$, l’estimation de la moyenne Y pour `Sedim = "1mud"` est $Y = a*Bathy + 0 + c$. Dans les r  sultats du tableau, `b["2sand"]` est donc non significativement diff  rent de `b["1mud"] = 0`, et, `b["3coarse"]` avec une p -value = 0.07, n’est pas non plus significativement diff  rent de `b["1mud"] = 0`.

5.2.2 Analyse de r  sidus

Une hypoth  se de construction d’un mod  le lin  aire est l’homosc  dasticit  , aussi appel  e homog  nit   de la variance. Cela signifie que la variance de la r  ponse y est la m  me quelque soit la valeur du pr  dicteur x . Dans un mod  le Gaussien classique ajustant x    y ainsi: $y = a.x + b + \epsilon$, la variable ϵ repr  sente les r  sidus du mod  le. Ils sont suppos  s   tre centr  s sur z  ro et avec une variance Gaussienne, leur distribution suivant ainsi la loi Gaussienne $\epsilon \sim N(0, \sigma)$

Lorsqu’on simule un tel mod  le, par exemple $y = 2.x + 5 + \epsilon$, on observe la figure 11, avec une homog  nit   de la distribution des observations autour de l’ajustement, et une distribution Gaussienne des r  sidus comme le montrent l’histogramme et le “qqplot”.

```
# Example of homogeneous residuals
n <- 1000
epsilon <- rnorm(n, 0, 5)
x <- runif(n, 0, 10)
y <- 2*x + 5 + epsilon

par(mfrow = c(1,3))
plot(x, y, pch = 20)
abline(5, 2, col = "red", lwd = 2)
hist(epsilon, breaks = 20, col = "grey")
qqnorm(epsilon); qqline(epsilon, col = "red")
```

   partir de cet exemple, vous pouvez d  finir les diagnostics graphiques n  cessaires pour v  rifier vos hypoth  ses de construction de mod  le. Lorsqu’on utilise le m  me mod  le que pr  c  demment (Fig.12) :

- **Residuals vs Fitted** - Dans cet exemple, on peut voir que la variabilit   des r  sidus augmente avec les valeurs pr  dites, ce qui va    l’encontre de l’homog  nit   de la variance.
 - **Scale-Location** est en accord avec la figure pr  c  dente car on voit une augmentation de la deviance des r  sidus quand les pr  dictions augmentent.
 - **Normal Q-Q** - Cette figure appel  e “qqplot” montre la divergence entre les quantiles th  oriques d’une loi Gaussienne et les quantiles r  els de la distribution des r  sidus du mod  le. La divergence est importante pour les valeurs   lev  es, ce qui montre une queue de distribution plus longue qu’une loi Normale.
 - **Hist. of residuals** - L’histogramme des r  sidus est en accord avec le qqplot car on voit clairement une distribution qui n’est pas une Gaussienne centr  e sur z  ro. Cette distribution a une longue queue de distribution avec beaucoup plus de valeurs positives que de valeurs n  gatives.
- (ref:RFigResidualsDiagCap) Figures de diagnostic d’un mod  le lin  aire permettant de v  rifier les hypoth  ses de construction.

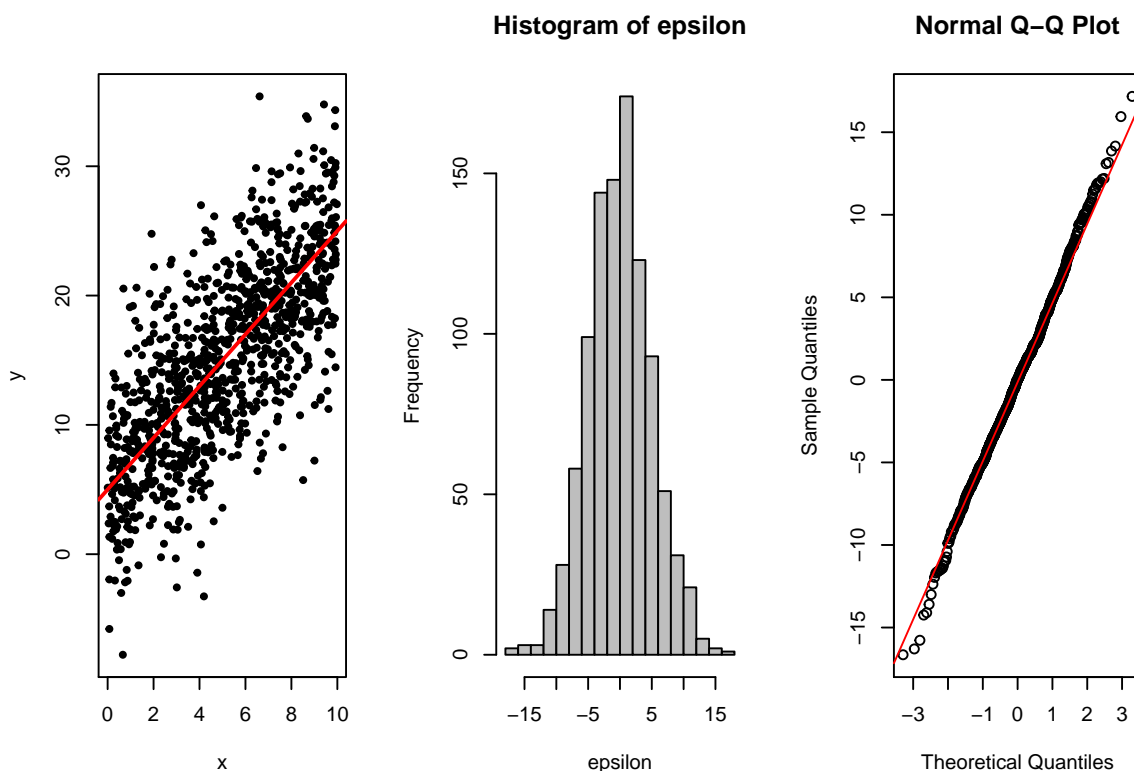


Figure 11 – Simulation d'une relation linéaire entre x et y avec un résidu Gaussien.

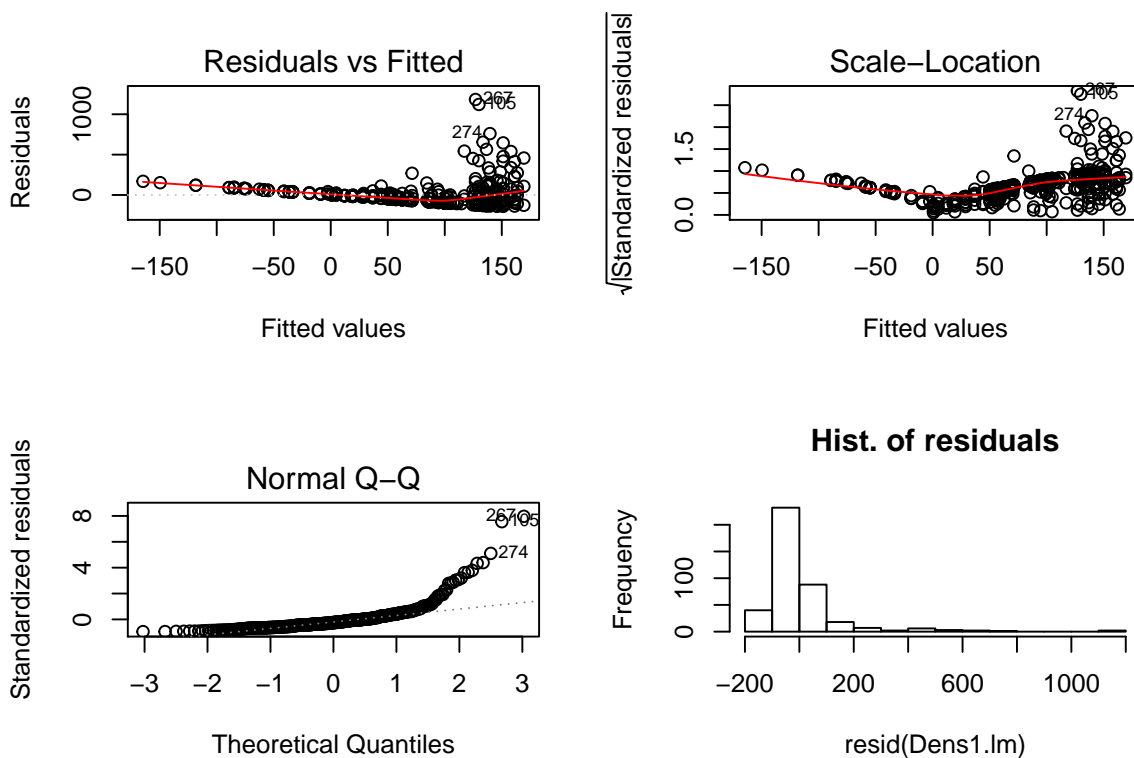


Figure 12 – (ref:RFigResidualsDiagCap)

5.2.3 Analyse de variance

La question à laquelle répond une `anova` (avec un test du Chi-2) est : Est-ce que la covariable ajoutée augmente significativement la vraisemblance du modèle (ou a réduit la déviance résiduelle), comparé au modèle précédent, sans cette covariable ?

Table 2 – Exemple d’une ‘sortie’ de anova(lm)

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Bathy	1	1607692	1607692	71.983	0.000
Sedim	2	32206	16103	0.721	0.487
Residuals	397	8866764	22334	NA	NA

Table 3 – Comparaison de la d  viance expliqu  e par diff  rents mod  les avec un nombre croissant de param  tres

model	Res.Df	RSS	Df	Sum of Sq	Pr(>Chi)	" "	df	AIC
lm1	8	3192	NA	NA	NA		3	92.0
lm2	6	1272	2	1920	0.003		5	86.8
lm3	4	645	2	626	0.144		7	84.1

- NULL est le test pour un mod  le sans covariable, c’est le mod  le qui estime la moyenne : $Density \sim constant$.
- Bathy est le test pour un mod  le uniquement avec la Bathy : $Density \sim Bathy$. La p-value est proche de z  ro, indiquant que le gain de d  viance expliqu  e en ajoutant la Bathy est significativement diff  rent de z  ro.
- Sedim est le test pour un mod  le avec Bathy et Sedim, dans cet ordre : $Density \sim Bathy + Sedim$. La p-value ~ 0.2 indique qu’il y a un risque de 20% que la d  viance expliqu  e en ajoutant la covariable Sedim au mod  le contenant d  j   la Bathy soit nulle.

5.2.4 Crit  res d’Akaike (AIC) et Bayesian (BIC)

L’AIC et le BIC sont des crit  res de qualit   d’ajustement p  nalis  s par le nombre de param  tres estim  s. La description (traduite) de ces fonctions dans R est :

Function g  n  rique calculant “Le Crit  re d’Information” d’Akaike pour un ou plusieurs mod  les ajust  s pour lesquels une “log-vraisemblance” peut   tre obtenue, en utilisant la formule $IC = -2 * \log - likelihood + k * npar$, o   $npar$ represente le nombre de param  tres estim  s, et $k = 2$ pour l’AIC classique, ou $k = \log(n)$ (n   tant le nombre d’observations) pour le BIC ou SBC (Schwarz’s Bayesian criterion).

En effet, plus vous ajoutez de param  tres dans un mod  le, plu svous avez de chances que le mod  le s’ajuste parfaitement aux donn  es (Table 3, Fig. 13). L’AIC diminue avec la d  viance r  siduelle et augmente avec le nombre de param  tres ajust  s. Plus l’AIC est bas, plus parsimonieux est le mod  le.

5.3. Trouver le meilleur mod  le

La fonction `lm` n’est utilis  e que pour des mod  les avec une distribution Gaussienne des r  sidus. Pour tester d’autres types de distributions, il faut utiliser `glm`, avec un param  tre pour la famille de distribution (`family`). Vous pouvez utiliser des distributions qui autorisent une plus grande queue de distribution que la loi Normale. Parmi les familles disponibles, vous pouvez tester `poisson`, `quasipoisson`, `Gamma`, `Log-gaussian`.

- *Quel est le meilleur mod  le au regard des diff  rents crit  res   voqu  s ?*

5.3.1 Exploration des sorties du meilleur mod  le

- *Que pouvez-vous dire sur le diagnostic complet de votre mod  le ?* (Fig. 14)

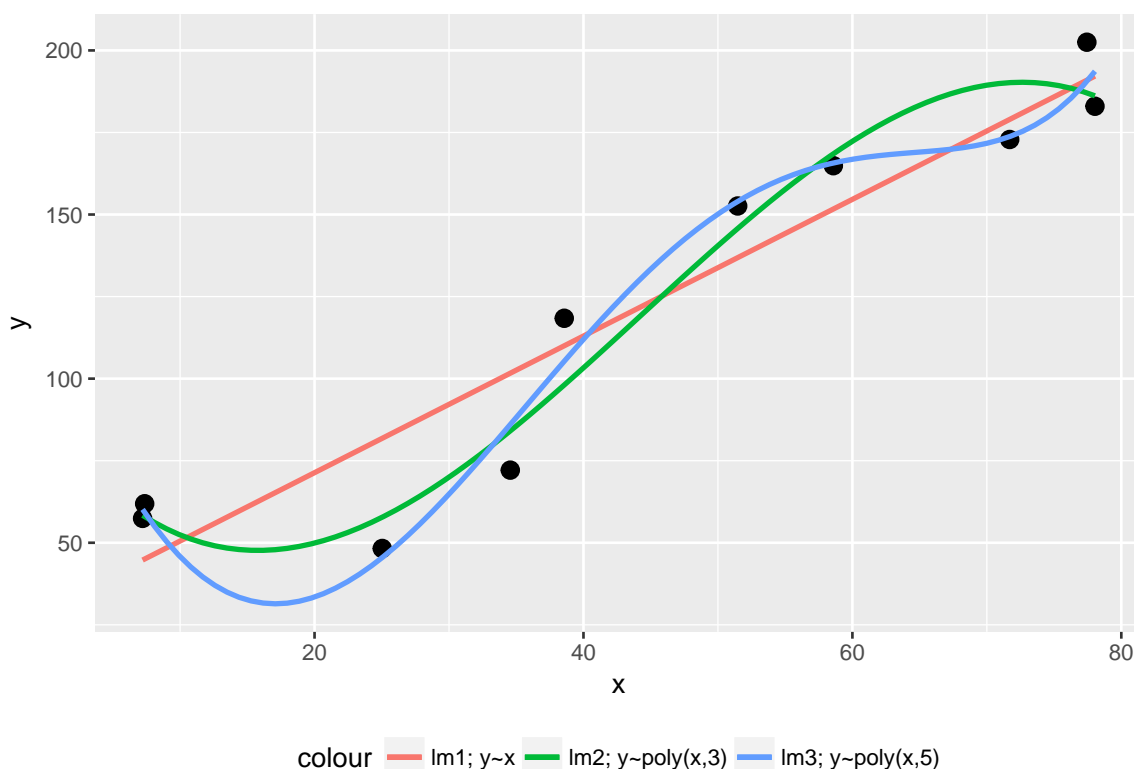


Figure 13 – Diff  rents mod  les ajust  s sur les m  mes donn  es mais avec un nombre de param  tres ajust  s croissant.

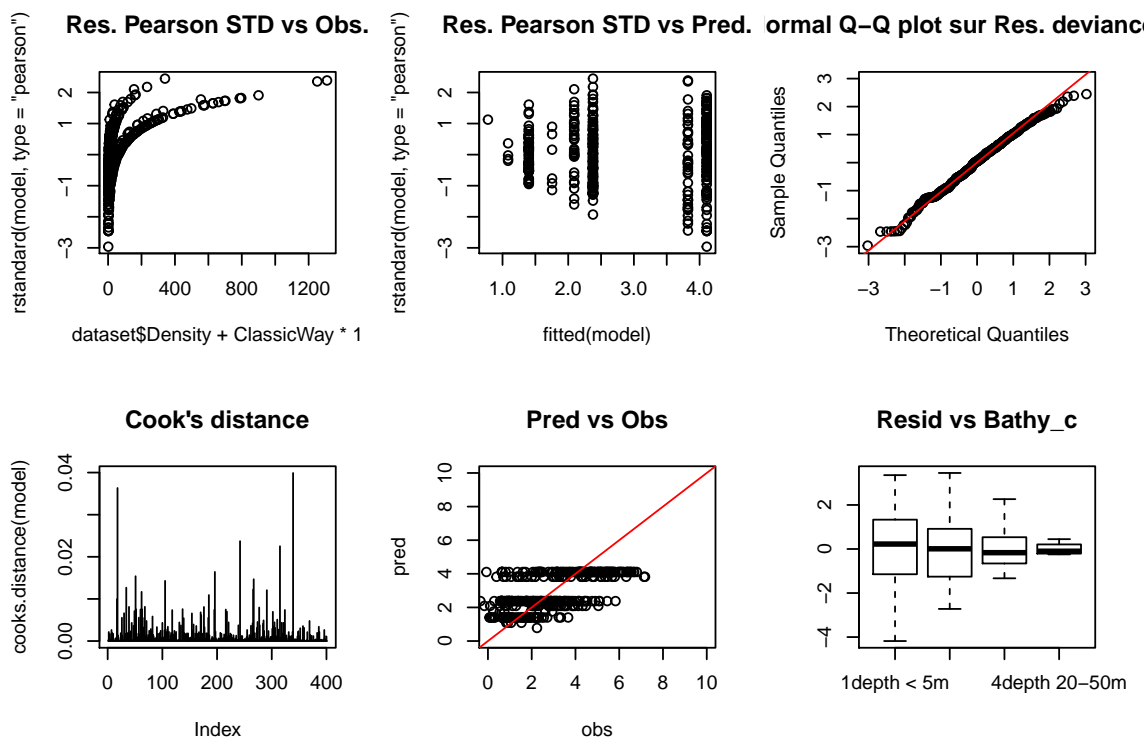


Figure 14 – Diagnostic du meilleur GLM s  lectionn  

5.4. Pr  dictions du mod  le

Lorsque vous   tes satisfaits du mod  le s  lectionn  , vous pouvez faire des pr  dictions

- Utiliser le fichier csv fourni pour voir l'effet des covariables s  lectionn  es (Fig. 15)

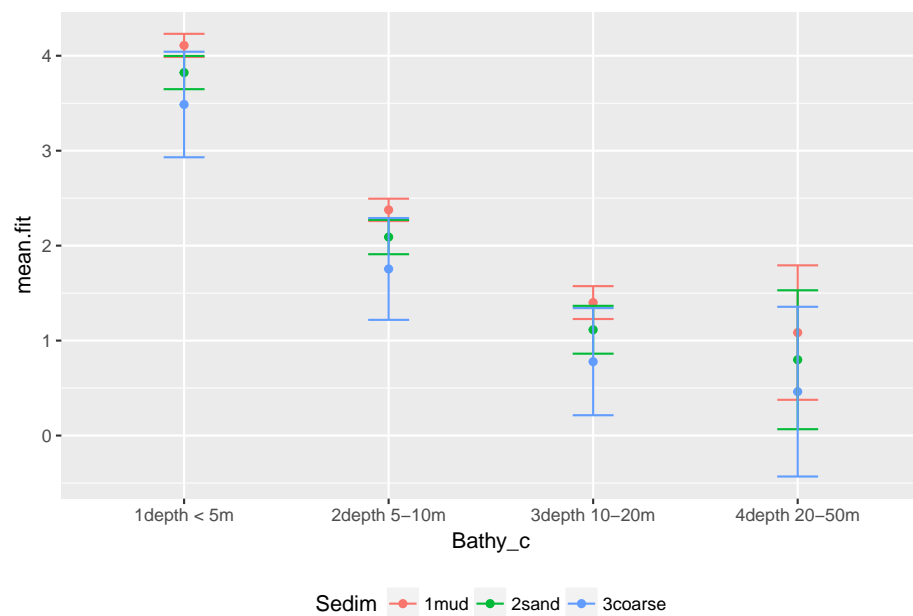


Figure 15 – Prédictions du meilleur GLM sélectionné