



# Formation à R

Modélisation avec les GLM

SÉBASTIEN ROCHETTE, THINKR

ThinkR

## Table des matières

<b>1</b>	<b>Préface</b>	<b>1</b>
<b>2</b>	<b>Présentation de l'étude</b>	<b>1</b>
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
<b>3</b>	<b>Préparation</b>	<b>4</b>
3.1	Structure des dossiers . . . . .	4
3.2	Débutons avec R . . . . .	5
<b>4</b>	<b>Exploration des données</b>	<b>5</b>
4.1	Étapes . . . . .	5
4.2	Liste des différentes étapes . . . . .	5
<b>5</b>	<b>Modélisation</b>	<b>6</b>
5.1	Étapes . . . . .	6
5.2	Interpréter les sorties de modèles . . . . .	7
5.3	Trouver le meilleur modèle . . . . .	8
5.4	Prédictions du modèle . . . . .	9

## 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 - Students*

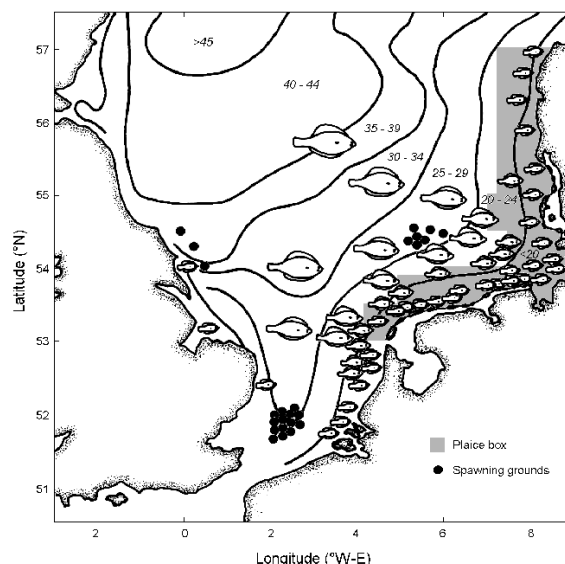
## 2. Présentation de l'étude

*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

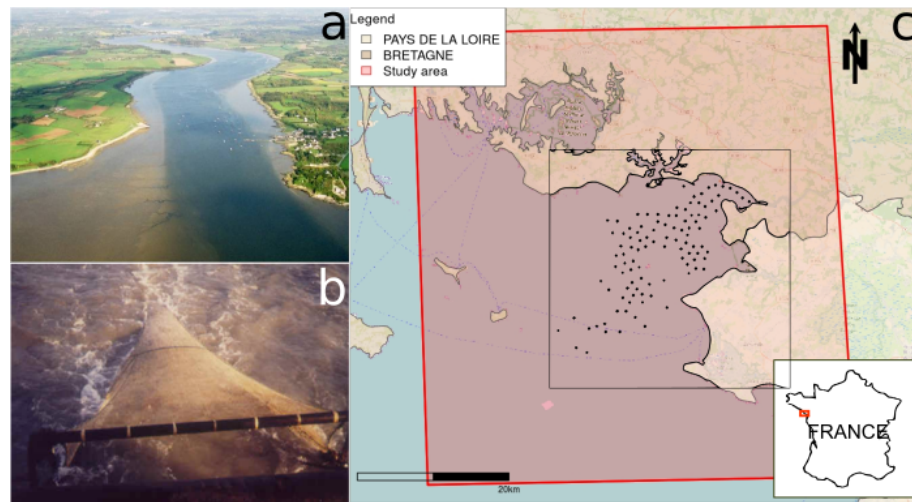
Déterminer les facteurs ayant une influence sur la distribution des poissons plats (*Solea solea*) en Baie de Vilaine et cartographier la distribution moyenne des densités.

- 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)

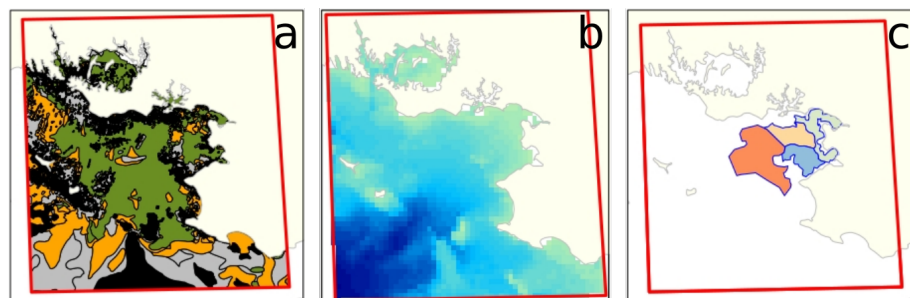
- 1984 – 2010
- En automne
- Juvéniles de l'année (Âge 0)
  - Nb individus / 1000m<sup>2</sup>



**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

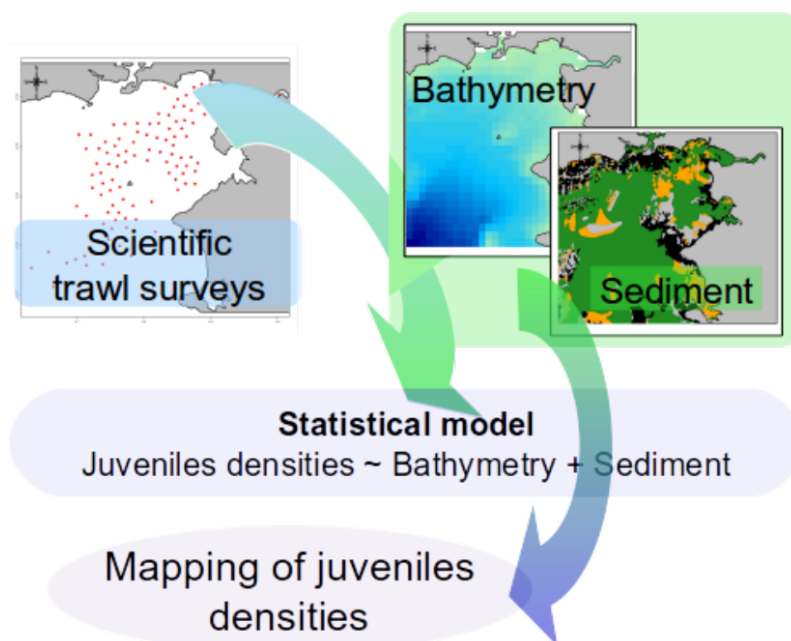


Figure 5 – Proc  dure pour un mod  le de distribution d'esp  ce

## 2.6. Exploration des donn  es

Prenez le temps d'explorer vos donn  es avant toutes analyses

- Explorer les donn  es et les covariables
  - Explorer le plan d'  chantillonnage
  - Explorer les liens potentiels entre les densit  s et les covariables
  - Explorer les futurs param  tres de mod  lisation (interactions, distributions)

*Souvenez-vous toujours des objectifs de votre   tude !*









*Question : Que recherchons-nous dans cette exploration ?*

## 3. Pr  paration

### 3.1. Structure des dossiers

*Il convient de toujours conserver les fichiers originaux : les reprojections entra  nent 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 : "Classic\_AllDataModel\_Student.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 ?*

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

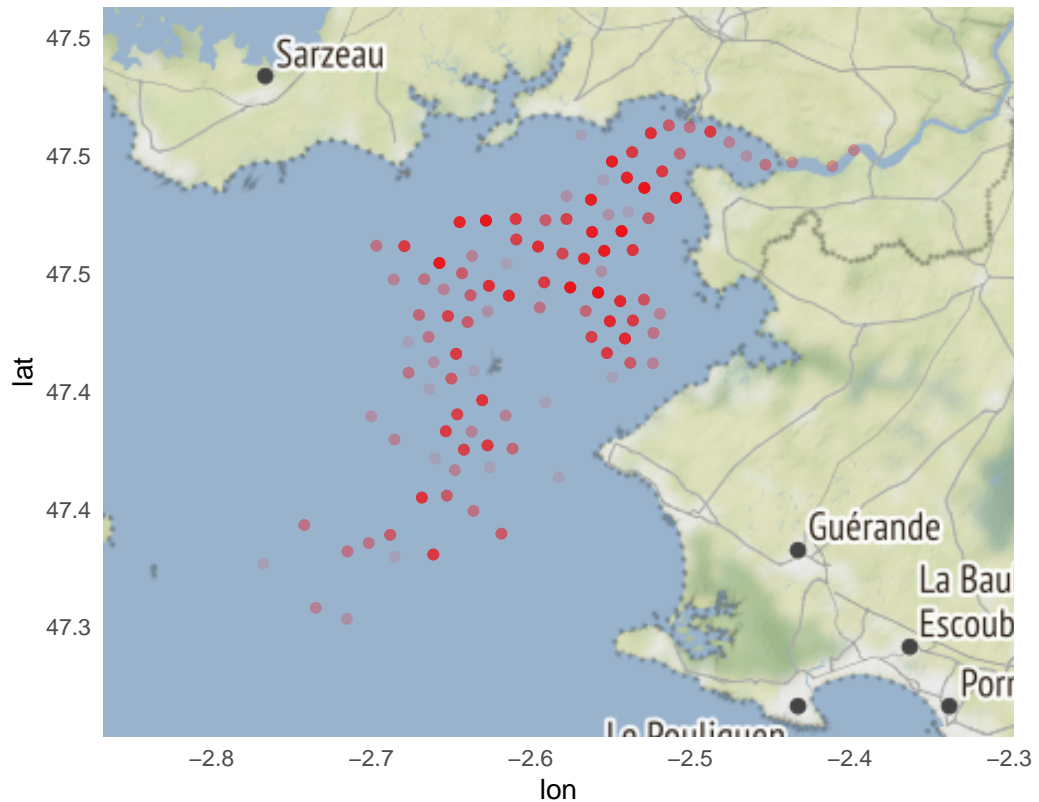


Figure 6 – Répartition des stations d'échantillonnage

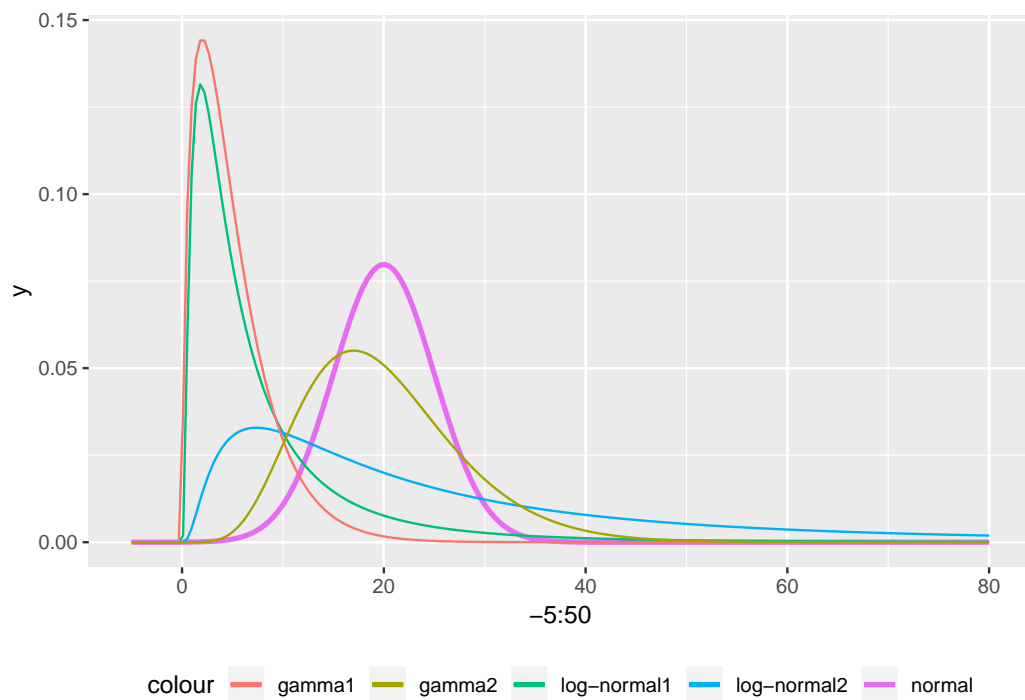


Figure 7 – (ref:RFigDistribCap)

## 5. Modélisation

### 5.1. Étapes

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

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

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	99.944	8.127	12.298	0.000
Bathy	5.921	0.639	9.260	0.000
Sedim2sand	-0.157	12.938	-0.012	0.990
Sedim3coarse	-41.819	23.241	-1.799	0.072

- Tester les diff  rentes formes de mod  les au regard des combinaisons de covariables et des 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.

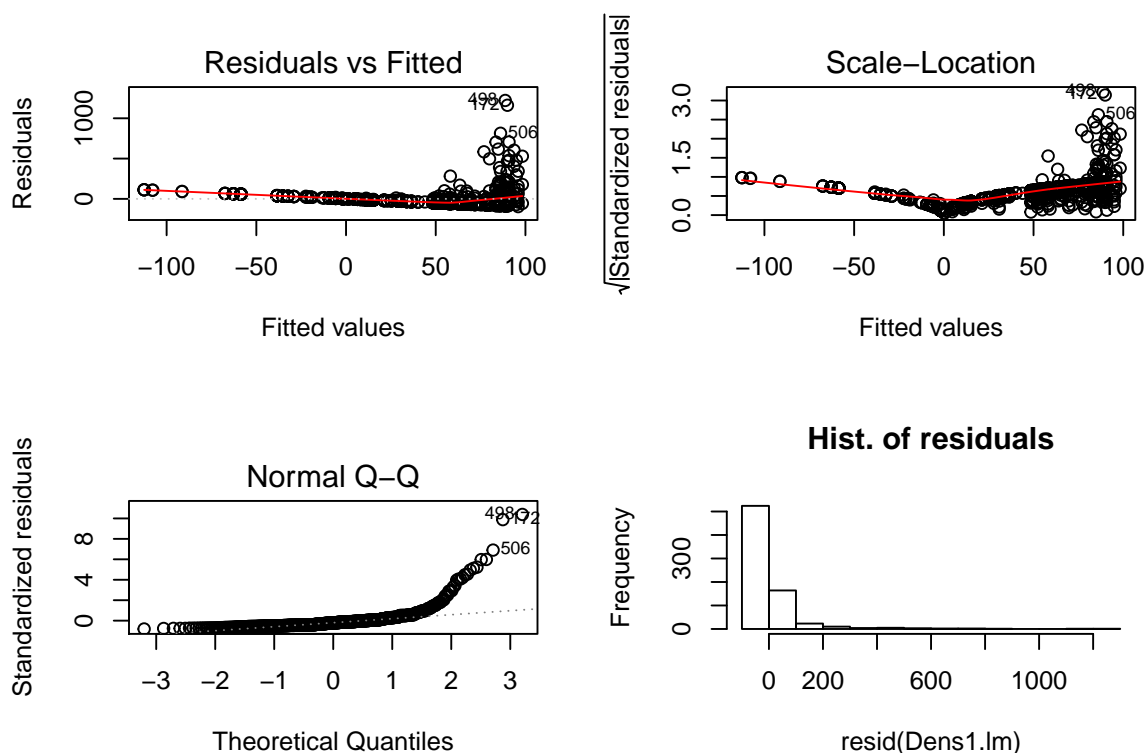
### 5.2.2 Analyse de r  sidus

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  $\epsilon$  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  $\epsilon \sim N(0, \sigma)$

### 5.2.3 Analyse de variance

La question    laquelle r  pond une **anova** (avec un test du Chi-2) est : Est-ce que la covariable ajout  e augmente significativement la vraisemblance du mod  le (ou a r  duit la d  viance r  siduelle), compar   au mod  le pr  c  dent, sans cette covariable ?





**Figure 8** – Figures de diagnostic d’un mod  le lin  aire permettant de v  rifier les hypoth  ses de construction.

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

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Bathy	1	1251584	1251584	89.66	0.000
Sedim	2	45447	22724	1.63	0.197
Residuals	736	10274168	13959	NA	NA

#### 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` repr  sente 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).

### 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 ?*

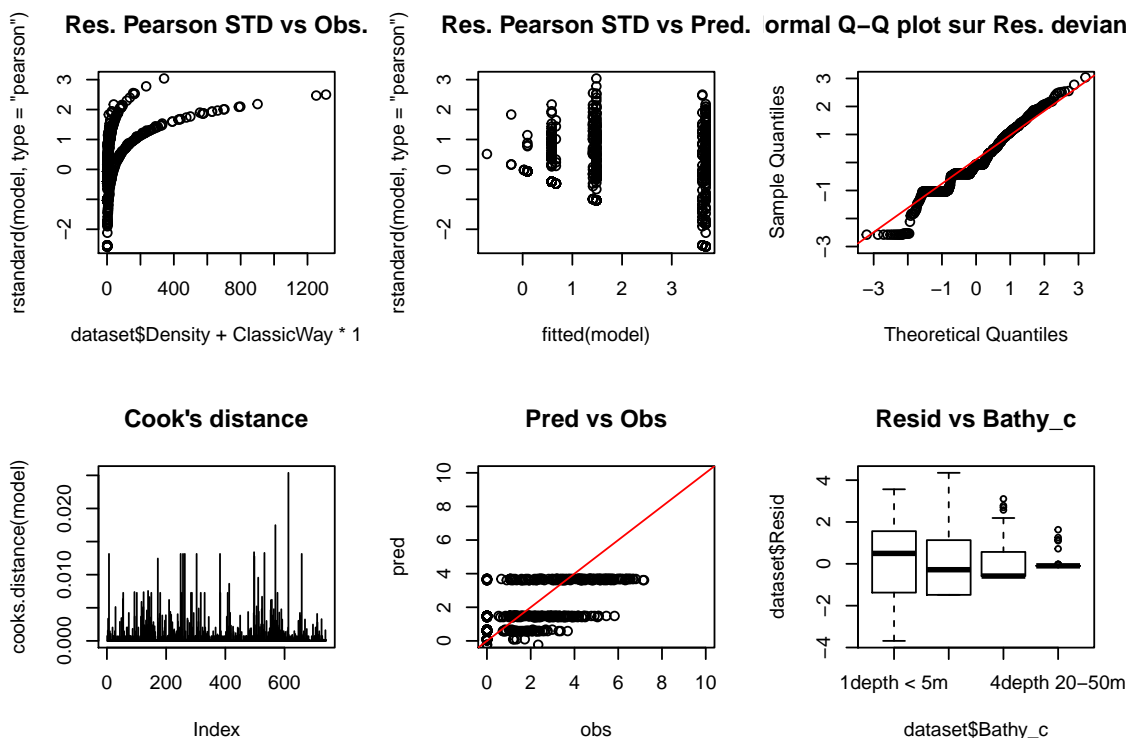


Figure 9 – (ref:RFigGLMResultsCap)

## 5.4. Prédictions du modèle

Lorsque vous êtes satisfaits du modèle sélectionné, vous pouvez faire des prédictions

- Utiliser le fichier csv fourni pour voir l'effet des covariables sélectionnées (Fig. 10)

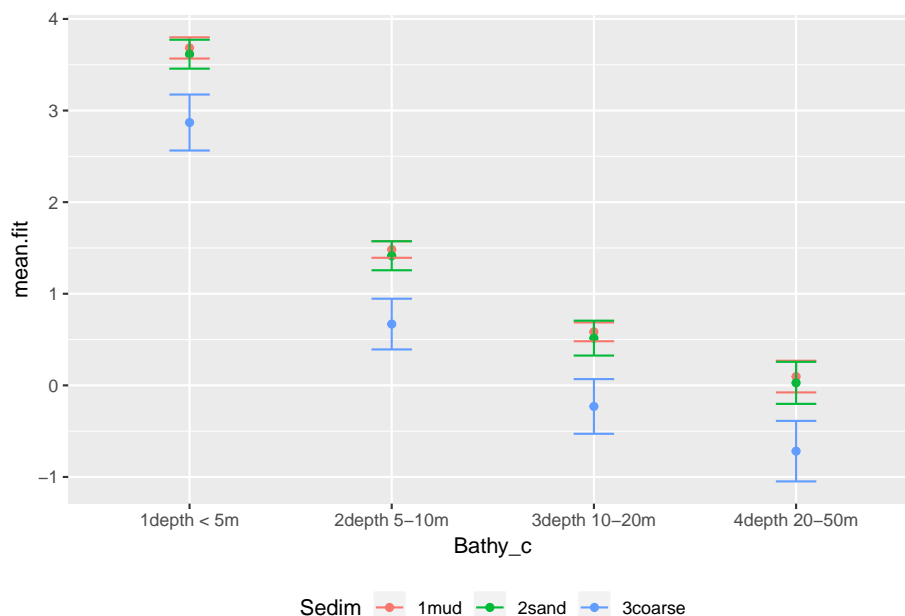


Figure 10 – Prédictions du meilleur GLM sélectionné