Projet d'étude de Statistiques

Maxime Baba, Alexandre Demarquet, Félix de Brandois, Tristan Gay

2024-01-11

Contents

1 Introduction **2**

2 Analyse descriptive des données **2**

2.1 Analyse unidiemensionnelle 2

2.2 Analyse multidimensionnelle 3

3 Classification des EPCI **7**

3.1 Clustering 7

3.2 Analyse discriminante linéaire 7

4 EMS **8**

4.1 Modèle linéaire 8

4.2 Modèle linéaire généralisé 10

5 Conclusion **10**

List of Figures

1 Boxplot des variables nox_kg,co_kg,so2_kg 2

2 Histogramme de la variable co_kg en brute, scale et scale(log()) 2

3 Corrélation entre les variables 3

4 ACP des variables quantitatives 5

1 Introduction

Le but de ce projet est d'étudier différents polluants mesurés par de nombreux EPCI d'Occitanie.
Nous disposons du jeu de données suivant : `Data-projetmodIA-2324.csv`.

Dans la suite de ce rapport, on utilise les notations suivantes :

- a
- b
- c

2 Analyse descriptive des données

On commence par interpréter les éléments jeu de données.

2.1 Analyse unidiemensionnelle

On s'intéresse dans un premier temps aux variables quantitatives du jeu de données (et en particulier aux émissions de polluants).
La figure 1 présente une visualisation de quelques variables quantitatives brutes.

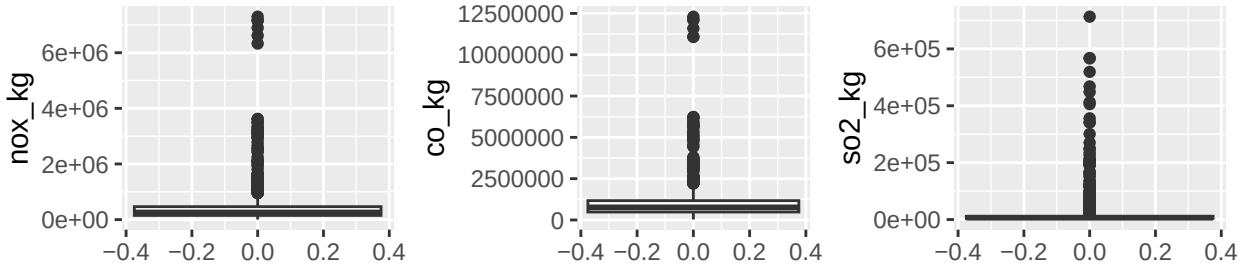


Figure 1: Boxplot des variables nox_kg,co_kg,so2_kg

On observe que ... (pas scale....)

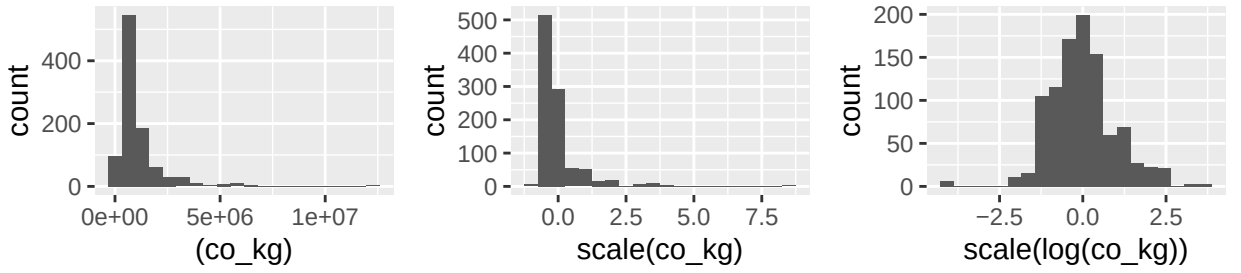


Figure 2: Histogramme de la variable co_kg en brute, scale et scale(log())

On effectue une transformation des données car d'après les boxplots de la figure 1 on remarque une variance énorme de certaines données comme co_kg. En examinant l'histogramme des données quantitatives, on observe une distribution fortement asymétrique. On peut donc appliquer une log-transformation pour normaliser la distribution des données.
Certaines variables ont pour unité la tonne et d'autre le kg on peut donc scale les données. On peut visualiser l'interet de ces transformations grâce à la figure 2 avec la varibale co_kg.
Par la suite, on manipule les variables quantitatives transformées `scale(log())`.

On étudie ensuite la corrélation entre les variables quantitatives.

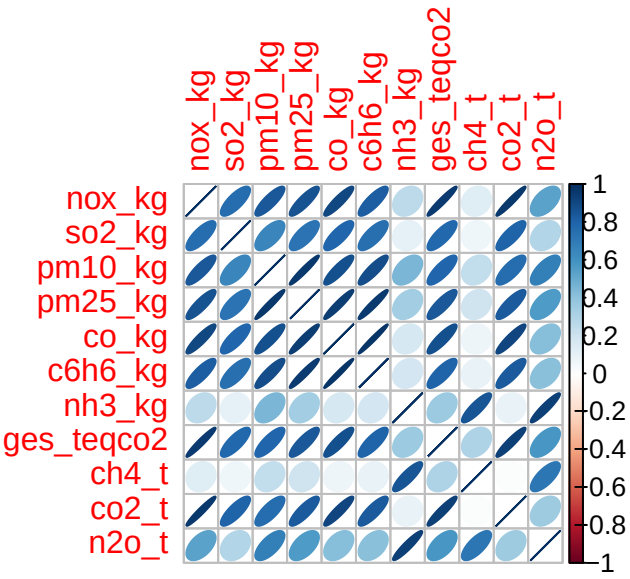


Figure 3: Corrélation entre les variables

L'analyse de la figure 3 nous permet d'identifier rapidement les relations significatives entre nos variables. Les ellipses fortement allongées suggèrent une corrélation plus forte, tandis que les ellipses plus circulaires indiquent une corrélation plus faible.

2.2 Analyse multidimensionnelle

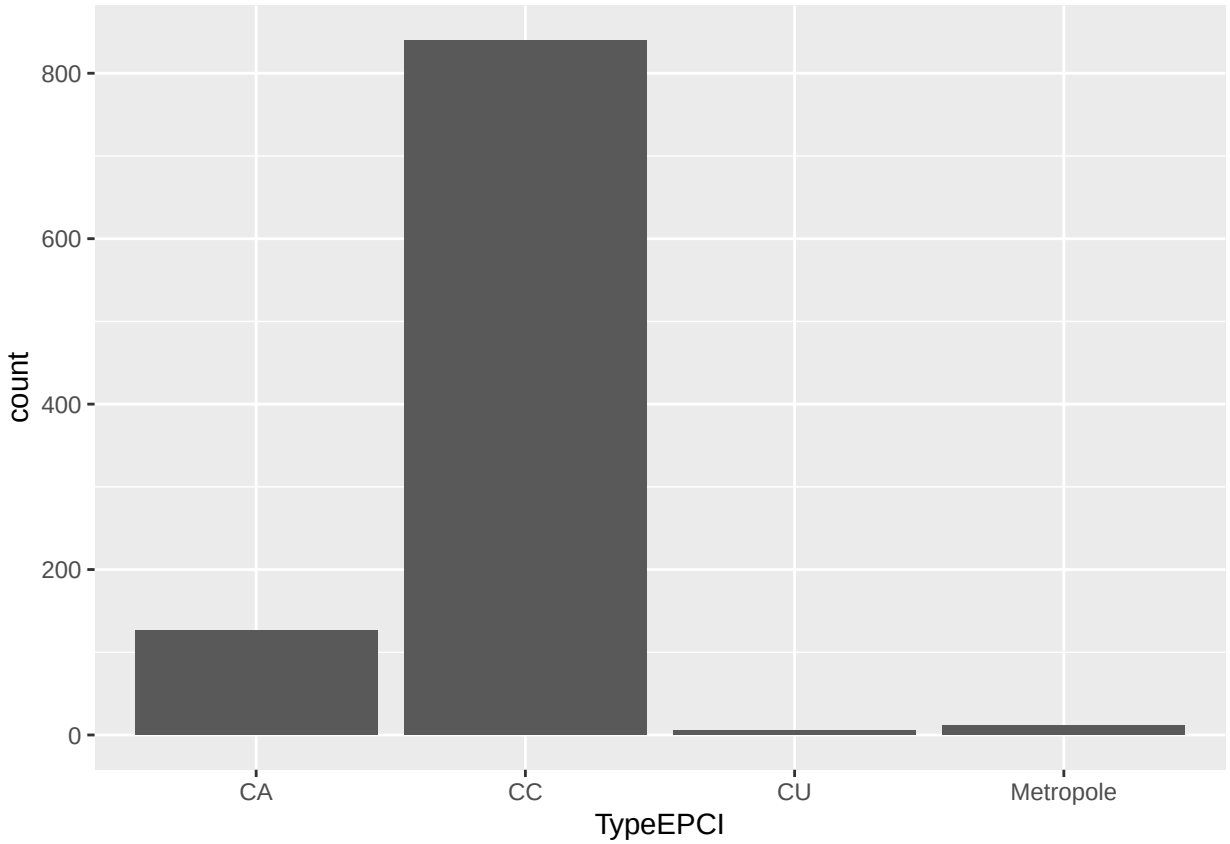
Dans le jeu de données nous avons aussi des variables qualitatives comme le code epci, le lib_epci ou des infos sur les départements.

```
data_quali=data[,c("code_epci","lib_epci","annee_inv","TypeEPCI","nomdepart")]
table(data_quali[,c("nomdepart")])
```

```
##
##                          Ardèche,Gard                                 Ariège
##                                     6                                     48
##                                 Aude                 Aude,Haute-Garonne,Tarn
##                                    48                                      6
##            Aude,Pyrénées-Orientales                                Aveyron
##                                     6                                    102
##                       Aveyron,Lot                          Aveyron,Lozère
```

```
##                                12                                 6
##                              Gard                       Gard,Hérault
##                                78                                 6
##                       Gard,Lozère                      Gard,Vaucluse
##                                 6                                 6
##                              Gers                 Gers,Haute-Garonne
##                                84                                 6
##                       Gers,Landes Gers,Lot-et-Garonne,Tarn-et-Garonne
##                                 6                                 6
##                     Haute-Garonne                Haute-Garonne,Tarn
##                                96                                 6
##                  Hautes-Pyrénées Hautes-Pyrénées,Pyrénées-Atlantiques
##                                48                                12
##                           Hérault                      Hérault,Tarn
##                                90                                 6
##                               Lot                            Lozère
##                                48                                54
##                 Pyrénées-Orientales                             Tarn
##                                66                                72
##                     Tarn-et-Garonne              Tarn,Tarn-et-Garonne
##                                48                                 6
```

```
ggplot(data=data_quali)+geom_bar(aes(x = TypeEPCI))
```



2.2.1 Analyse en Compomantes Principales (ACP) des variables quantitatives

On visualise les individus à partir des émissions de polluants.

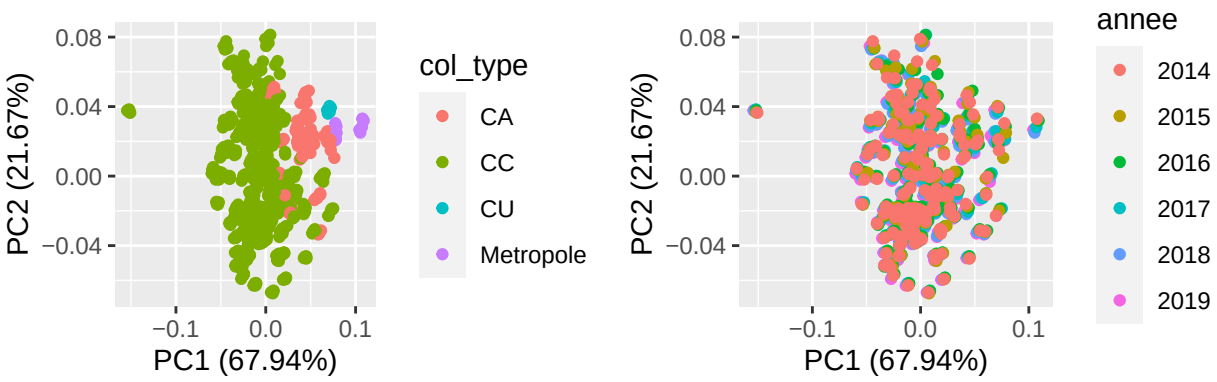


Figure 4: ACP des variables quantitatives

On observe sur la figure 4 que ...

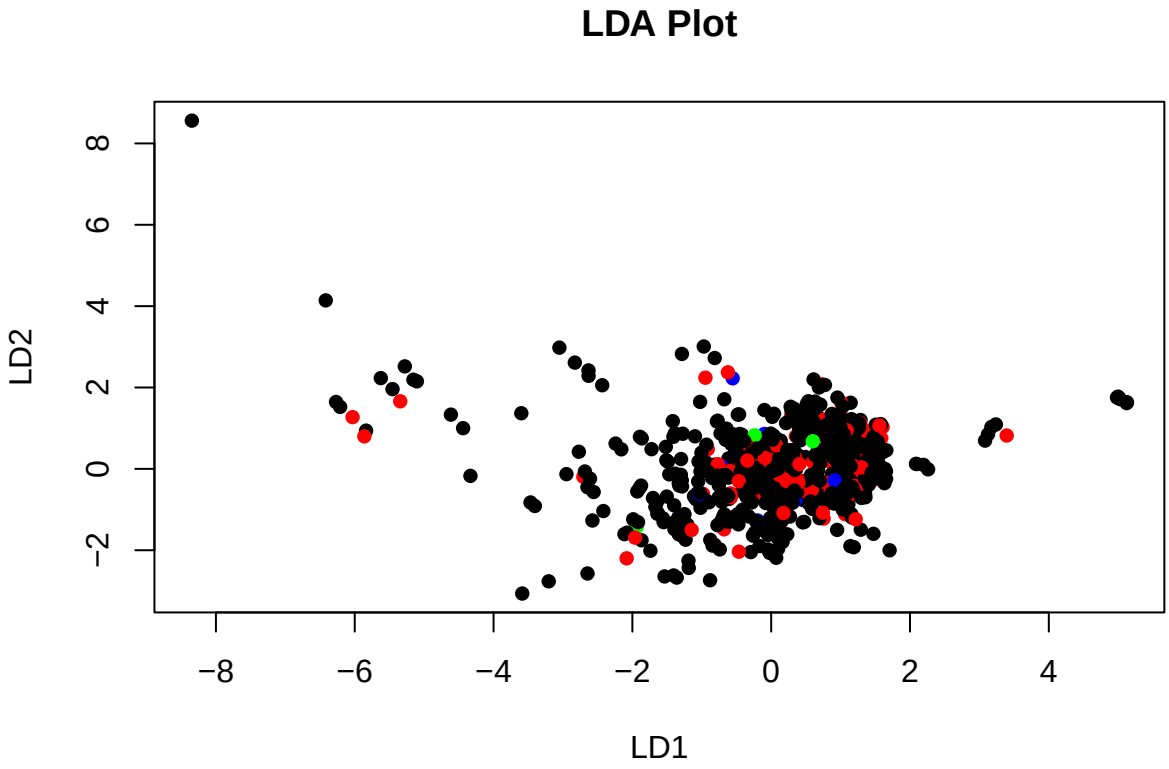
2.2.2 Réduction de dimension (MCA)

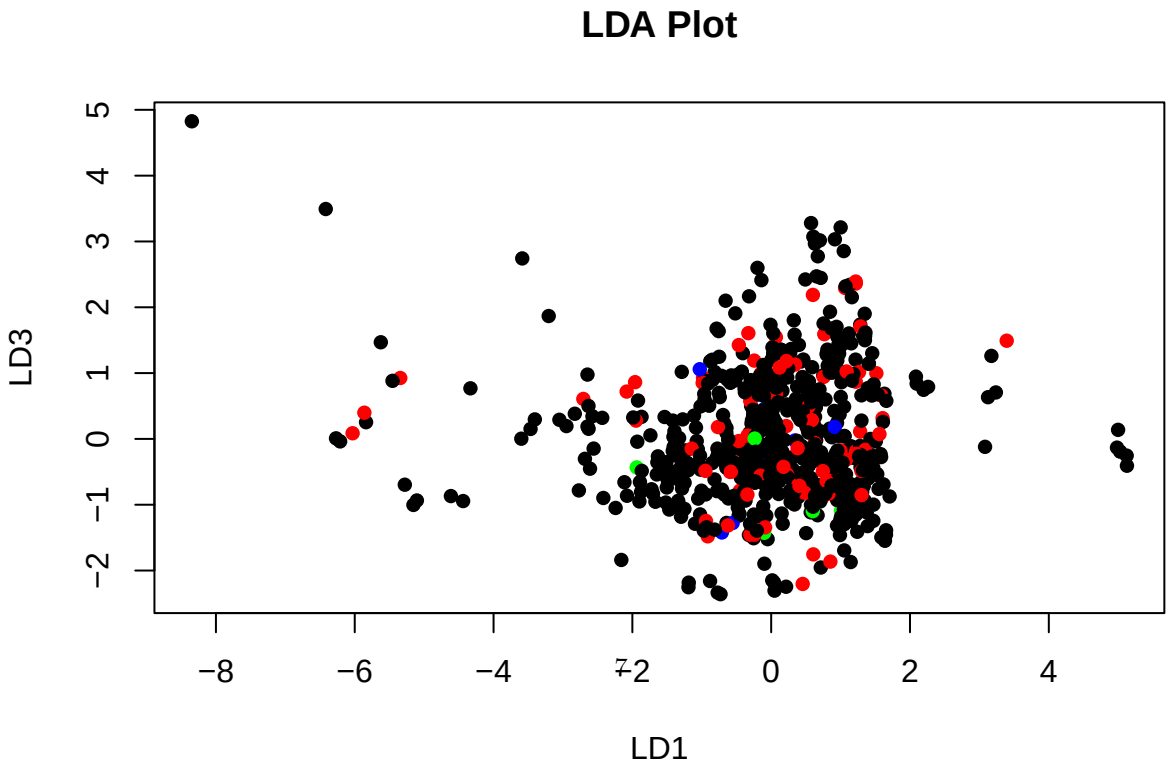
SUPERRRRRR

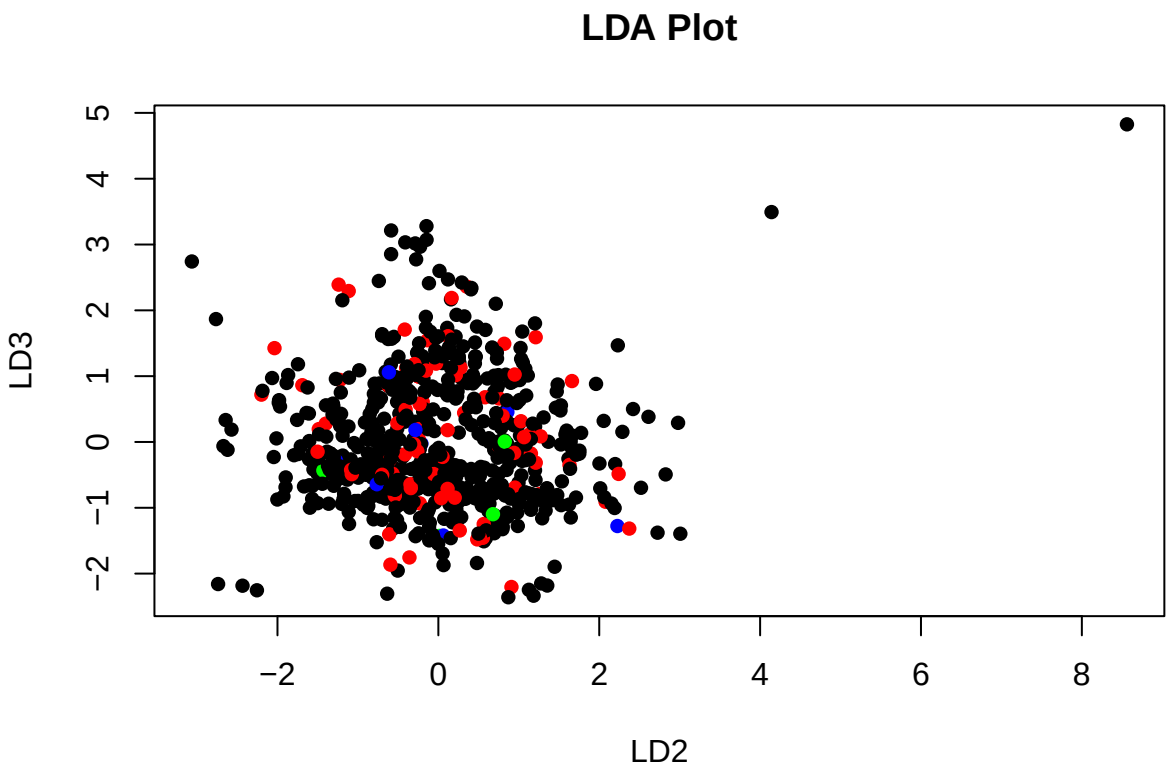
3 Classification des EPCI

3.1 Clustering

3.2 Analyse discriminante linéaire








```
##            prediction
## vrai_valeur  CA  CC  CU Metropole
##   CA         34  10   1         1
##   CC          5 241   0         0
##   CU          0   0   1         0
##   Metropole   1   0   0         2
```

4 EMS

4.1 Modèle linéaire

4.1.1 Modèle d'ANOVA

On explique le gaz à effet de serre en fonction des variables Type et années.

On utilise un modèle d'ANOVA à deux facteurs avec interaction :

$$Y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ij}$$

EXPLIQUER LA SIGNIFICATION DES TERMES DU MODELE

##

```
## Call:
## lm(formula = ges_teqco2 ~ TypeEPCI * annee_inv, data = dlog)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.2367 -0.4233 -0.0383  0.3863  2.8469
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -18.397236  80.407295  -0.229    0.819
## TypeEPCICC                    4.064479  86.227217   0.047    0.962
## TypeEPCICU                    7.818578 377.143642   0.021    0.983
## TypeEPCIMetropole           -19.949436 272.674402  -0.073    0.942
## annee_inv                     0.009761   0.039875   0.245    0.807
## TypeEPCICC:annee_inv         -0.002779   0.042761  -0.065    0.948
## TypeEPCICU:annee_inv         -0.003354   0.187029  -0.018    0.986
## TypeEPCIMetropole:annee_inv   0.010806   0.135222   0.080    0.936
##
## Residual standard error: 0.7644 on 976 degrees of freedom
## Multiple R-squared:  0.4198, Adjusted R-squared:  0.4157
## F-statistic: 100.9 on 7 and 976 DF,  p-value: < 2.2e-16
```

-> Commentaire sur la valeur de R^2 obtenue.

On essaie de simplifier le modèle en enlevant les interactions avec un test de sous-modèle :

$$\begin{aligned}\mathcal{H}_0 : &\quad Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij} \\ \mathcal{H}_1 : &\quad Y_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ij}\end{aligned}$$

On obtient une p-value de $1 > 0.05$.
On ne rejette pas l'hypothèse de nullité des interactions.
On garde donc le modèle suivant :

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}$$

On essaie de simplifier le modèle en enlevant les variables non significatives (on fait 2 tests de sous-modèle) :

$$\begin{aligned}\mathcal{H}_0 : &\quad Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \\ \mathcal{H}_1 : &\quad Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}\end{aligned}$$

et

$$\begin{aligned}\mathcal{H}_0 : &\quad Y_{ij} = \mu + \beta_j + \epsilon_{ij} \\ \mathcal{H}_1 : &\quad Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}\end{aligned}$$

Pour le modèle dépendant uniquement du type d'EPCI, on obtient une p-value de $0.599 > 0.05$.
On peut donc enlever l'année dans le modèle :

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

On essaie à nouveau de simplifier le modèle en enlevant les variables non significatives.

On obtient cette fois une p-value de $0 < 0.05$.
On ne peut donc pas enlever le type d'EPCI dans le modèle.

On vérifie finalement la cohérence du modèle retenu :

On obtient une p-value de $0.99 > 0.05$ donc le modèle est cohérent. On garde donc le modèle :

4.1.2 Régression linéaire

4.1.3 ANCOVA

4.2 Modèle linéaire généralisé

5 Conclusion