Etudiantes : Muriel ZOUZZOU

Candice DEMAUGE

SAÉ 4.EMS.01 - EXPLIQUER OU PREDIRE UNE VARIABLE QUANTITATIVE A PARTIR DE PLUSIEURS FACTEURS

Professeur : Mr Bertrand PAGER

I. INTRODUCTION

Le jeu de données utilisé dans cette étude provient de la fusion de deux fichiers contenant des informations sur les véhicules particuliers (VP) rechargeables, les infrastructures de recharge (bornes), et certaines caractéristiques géographiques des territoires.

L'objectif de cette analyse est de comprendre les facteurs qui influencent le **nombre de bornes de recharge installées** dans les communes. En raison de la forte hétérogénéité de cette variable, une **transformation logarithmique** a été appliquée pour obtenir la variable cible log_nbre_pdc (logarithme du nombre de bornes).

Les premières variables explicatives disponibles concernaient principalement les infrastructures de recharge et les véhicules (par exemple, le taux de VP rechargeables ou la puissance nominale installée). Dans un second temps, des **variables socio-démographiques et territoriales** ont été intégrées à l'analyse (revenu_med_disp, Pop_15_64_ans, Superficie) afin d'enrichir le modèle. Cela permet d'examiner si des **facteurs externes**, comme la démographie ou les caractéristiques économiques des territoires, peuvent contribuer à mieux expliquer la présence ou non d'infrastructures de recharge.

```r
getwd()
setwd("C:/Users/murie/Downloads")
# Chargement des données principales
jeu_de_données <- read.csv("fusion_voitures_bornes_densite_residences.csv", header = TRUE, sep = ";")
jeu_de_données$puissance_nominale <- as.numeric(gsub(",", ".", jeu_de_données$puissance_nominale))
# Chargement du fichier secondaire
jeu_de_données2 <- read.csv2("Fichier_final.csv", encoding = "latin1")
jeu_de_données2$Superficie <- as.numeric(gsub(",", ".", jeu_de_données2$Superficie))
jeu_de_données2$Pop_15_64_ans <- as.numeric(gsub(",", ".", jeu_de_données2$Pop_15_64_ans))

# Fusion des bases
vars_utiles <- jeu_de_données2[, c("CODGEO", "Superficie", "Pop_15_64_ans")]
jeu_fusionne <- merge(jeu_de_données, vars_utiles, by = "CODGEO", all.x = TRUE)
```

Les données socio-démographiques proviennent d'une seconde base fusionnée à l'aide de la variable CODGEO :

II. EXPLORATION DES DONNEES

La base de données finale contient 89 observations et 15 variables. Les variables les plus pertinentes retenues pour l'analyse sont :

- nbre_pdc : nombre de bornes de recharge
- log_nbre_pdc : logarithme du nombre de bornes (variable cible)
- taux_vp_rechargeables : part des véhicules particuliers rechargeables

- revenu_med_disp : revenu médian disponible
- puissance_nominale : puissance totale des bornes
- Pop_15_64_ans : population en âge de travailler
- Superficie : surface de la commune

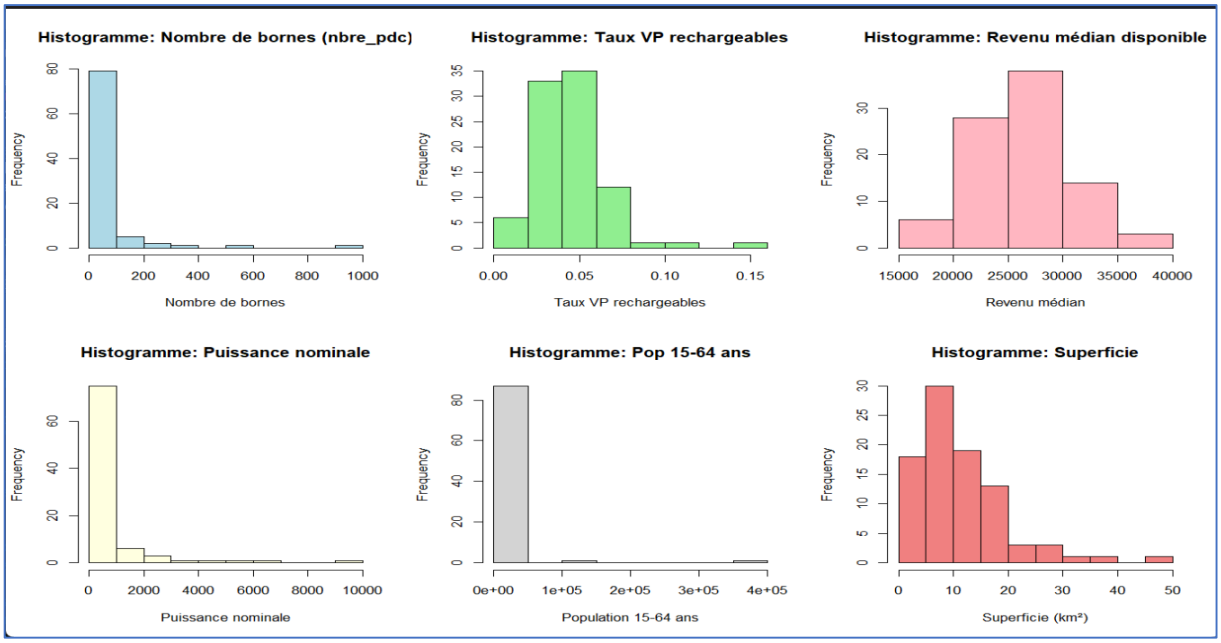
1. STATISTIQUES DESCRIPTIVES

```
#### ANALYSE EXPLORATOIRE ####

# Aperçu des données
str(jeu_fusionne)  # Structure des données
summary(jeu_fusionne[, c("nbre_pdc", "taux_vp_rechargeables", "revenu_med_disp", "puissance_nominale", "Pop_15_64_ans", "Superficie")])
```

Les principales tendances observées sont :

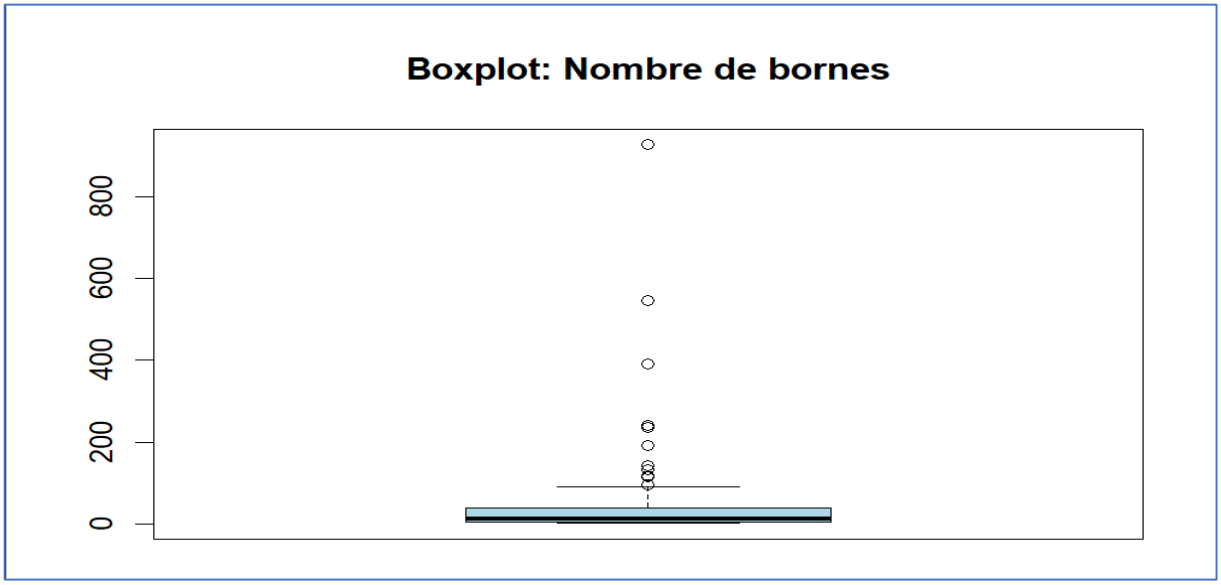
- Le nombre de bornes varie fortement, de 2 à 928 bornes par commune.
- La puissance nominale moyenne est élevée (environ 659 kW), mais la médiane est bien plus basse (132 kW), indiquant une distribution asymétrique à droite.
- Le taux de VP rechargeables atteint en moyenne 4,47 %, avec un maximum de 15,7 %.
- Les variables sociodémographiques présentent aussi des distributions asymétriques, notamment la population et la superficie.

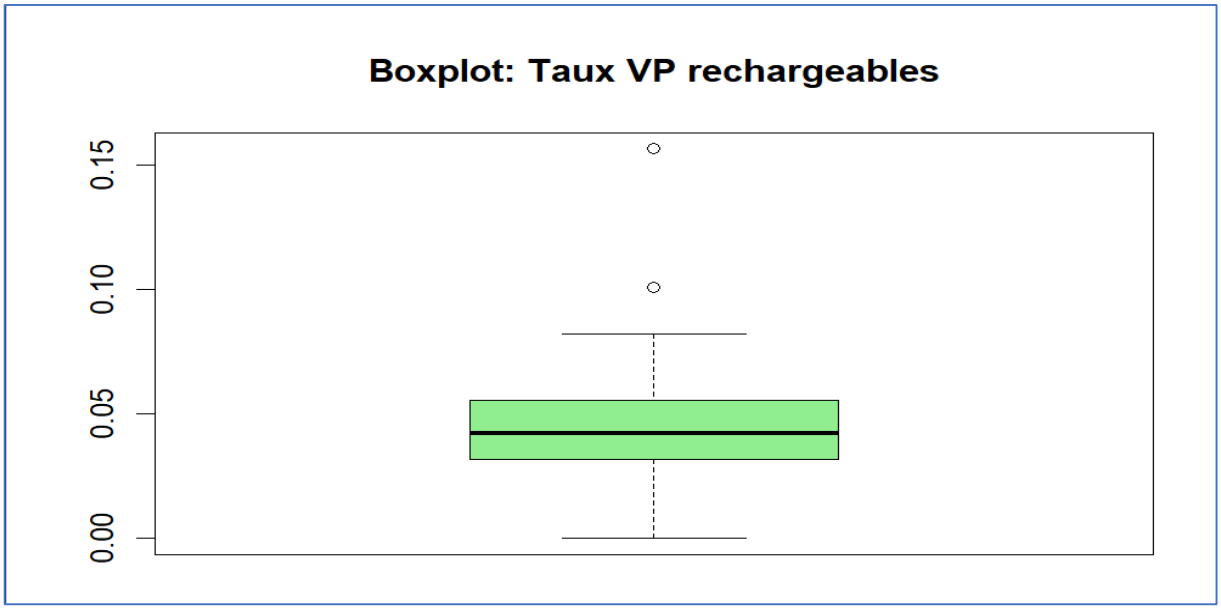
1. Visualisation : histogrammes



Les histogrammes des principales variables montrent des distributions souvent asymétriques, marquées par des valeurs extrêmes. Le nombre de bornes et la puissance installée, par exemple,

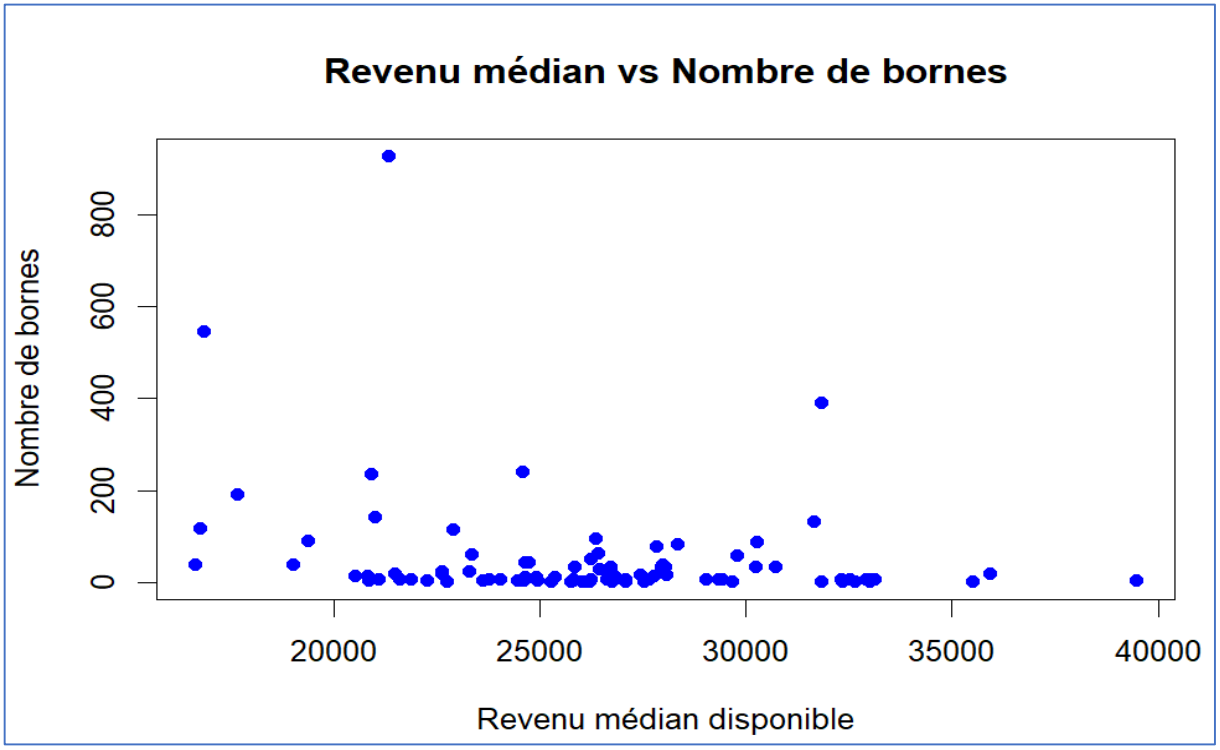
varient fortement d'une commune à l'autre, justifiant l'usage de transformations comme le logarithme pour stabiliser les données. Le taux de véhicules rechargeables, le revenu médian, la population active et la superficie présentent également une forte hétérogénéité. Ces différences structurelles entre communes soulignent l'intérêt de prendre en compte ces variables dans la modélisation, tout en restant attentif aux effets potentiels des outliers.

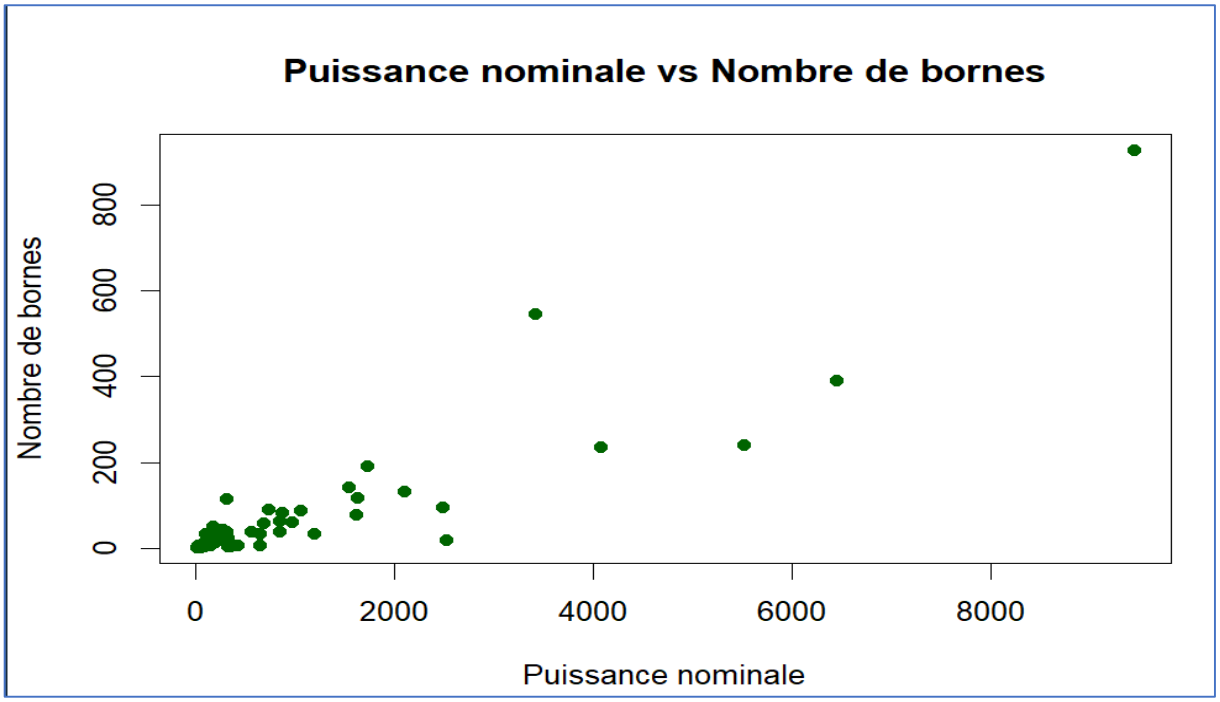




Le premier box plot montre que le nombre de bornes de recharge varie fortement entre les observations, avec une majorité concentrée sous les 50 bornes, mais aussi quelques valeurs très élevées, ce qui indique de fortes disparités. Le deuxième montre que le taux de véhicules

particuliers rechargeables est globalement bas, la plupart des données étant situées entre 3 % et 7 %, mais là aussi, quelques zones dépassent 15 %, ce qui reflète des cas particuliers. En résumé, ces deux graphiques révèlent une forte hétérogénéité dans la répartition des infrastructures et des véhicules électriques.





Les deux nuages de points montrent des liens faibles mais intéressants entre certaines variables économiques et le nombre de bornes de recharge. D'une part, le revenu médian disponible ne semble pas avoir de corrélation claire avec le nombre de bornes : les points sont dispersés, ce qui suggère que la richesse d'un territoire n'explique pas à elle seule le niveau d'équipement en infrastructures. D'autre part, la puissance nominale paraît davantage liée au nombre de bornes, avec une tendance plus visible, bien que l'ensemble reste très variable. En somme, la dynamique des infrastructures de recharge semble plus influencée par des facteurs techniques que purement économiques.

III. MODELISATION

L'objectif est de modéliser le nombre de bornes de recharge en utilisant le logarithme du nombre de bornes (log_nbre_pdc) comme variable dépendante. Cette transformation logarithmique permet de stabiliser la variance des données et de linéariser la relation avec les variables explicatives quantitatives, ce qui améliore la qualité du modèle de régression.

Le modèle initial complet est formulé ainsi :

```
# Modèle initial complet
mod_log <- lm(log_nbre_pdc ~ taux_vp_rechargeables + revenu_med_disp + puissance_nominale + Pop_15_64_ans + Superficie, data = jdd)
View(jdd)
```

La modélisation consiste à estimer une relation linéaire entre le logarithme du nombre de bornes de recharge (log_nbre_pdc) et plusieurs variables explicatives : le taux de véhicules rechargeables, le revenu médian disponible, la puissance nominale, la population active (15-64 ans) et la superficie. En transformant la variable dépendante par un logarithme, on cherche à stabiliser la variance et à mieux respecter les hypothèses du modèle linéaire. Ce modèle initial complet permet d'évaluer l'impact simultané de ces différents facteurs sur le nombre de bornes de recharge, servant de point de départ pour une analyse plus approfondie ou une sélection de variables.

Un modèle de sélection automatique par procédure pas à pas (stepwise) basée sur l'AIC a ensuite été utilisé pour retenir les variables les plus pertinentes :

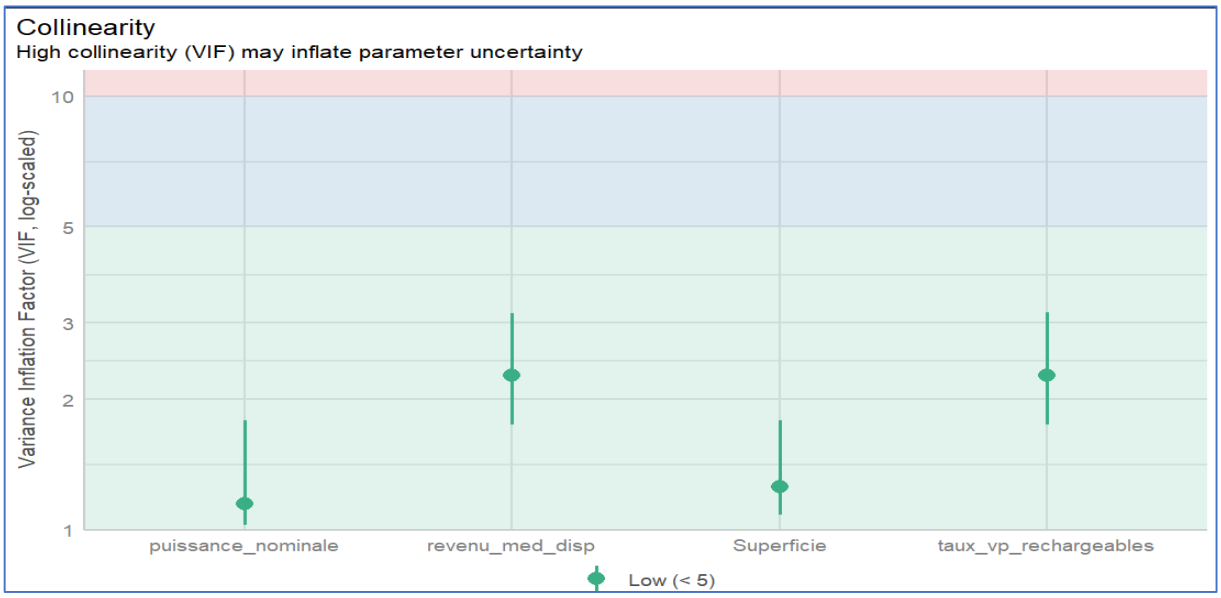
```
# Sélection par step AIC
regBest <- step(mod_log, direction = "both")

# Résumé final du modèle choisi
summary(regBest)
```

Le modèle final, obtenu par sélection stepwise selon le critère AIC, explique environ 61 % de la variance du logarithme du nombre de points de charge (R^2 ajusté = 0,609). Il inclut quatre variables : taux de véhicules rechargeables, revenu médian disponible, puissance nominale et

superficie. Le modèle est globalement significatif (F = 35.3, p < 2.2e-16). Trois variables sont significatives : le taux de véhicules rechargeables (effet positif, p = 0.006), le revenu médian disponible (effet négatif, p = 0.00017) et la puissance nominale (effet positif, p < 0.001). La superficie, bien que conservée, n'est pas significative (p = 0.12). Ces résultats indiquent que plus le taux de véhicules rechargeables et la puissance nominale sont élevés, plus le nombre de points de charge est important. Le revenu médian disponible a un effet inverse, suggérant des dynamiques socio-économiques complexes. La superficie a un impact faible et non significatif.

IV. VERIFICATION DES HYPOTHESES DU MODELE LINEAIRE

2. Colinéarité

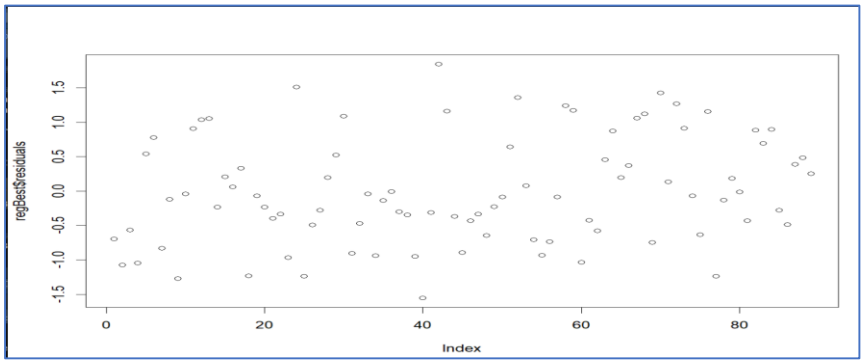


L'analyse des facteurs de inflation de la variance (VIF) montre que toutes les variables explicatives du modèle présentent des valeurs inférieures à 5, ce qui indique qu'il n'y a pas de problème majeur de multicolinéarité entre les prédicteurs. Plus précisément, le taux de véhicules rechargeables et le revenu médian disponible ont des VIF autour de 2.27, ce qui est modéré mais acceptable. La puissance nominale et la superficie ont des VIF proches de 1, ce qui confirme une faible corrélation avec les autres variables. La faible colinéarité garantit la stabilité des estimations des coefficients et renforce la confiance dans l'interprétation des effets individuels de chaque variable sur le logarithme du nombre de points de charge.

Indépendance des résidus

```r
plot(regBest$residuals)
mean(regBest$residuals)
dw <- dwtest(regBest)
print(dw)
```

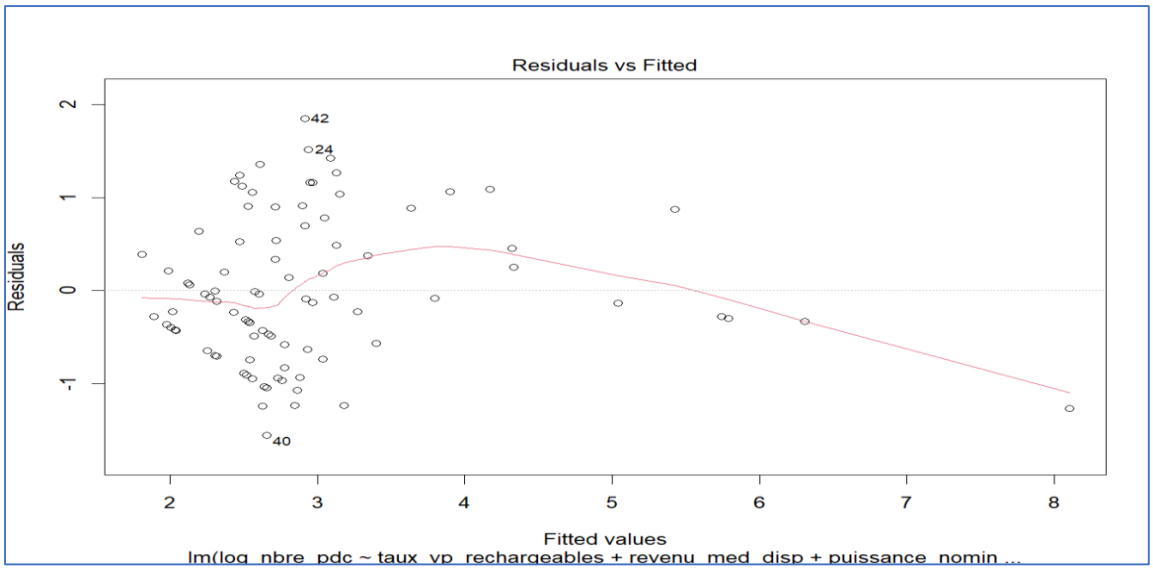
```
> print(dw)

        Durbin-Watson test

data:  regBest
DW = 1.6405, p-value = 0.04325
alternative hypothesis: true autocorrelation is greater than 0
```



Le test de Durbin-Watson donne une valeur de 1.64 avec un p-value de 0.043, ce qui indique la présence d'une autocorrélation positive faible mais significative dans les résidus du modèle. Ce n'est donc pas un « bon signe », mais ce n'est pas non plus catastrophique, c'est un avertissement qu'il faudrait approfondir. En résumé, le modèle fonctionne globalement bien, avec des résidus proches de zéro en moyenne, mais il pourrait être amélioré en tenant compte de cette autocorrélation afin de garantir la fiabilité des tests statistiques et des intervalles de confiance.

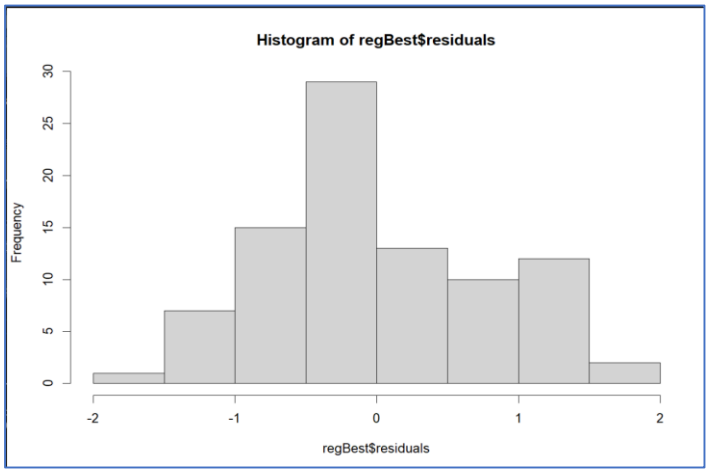
3. Homoscédasticité

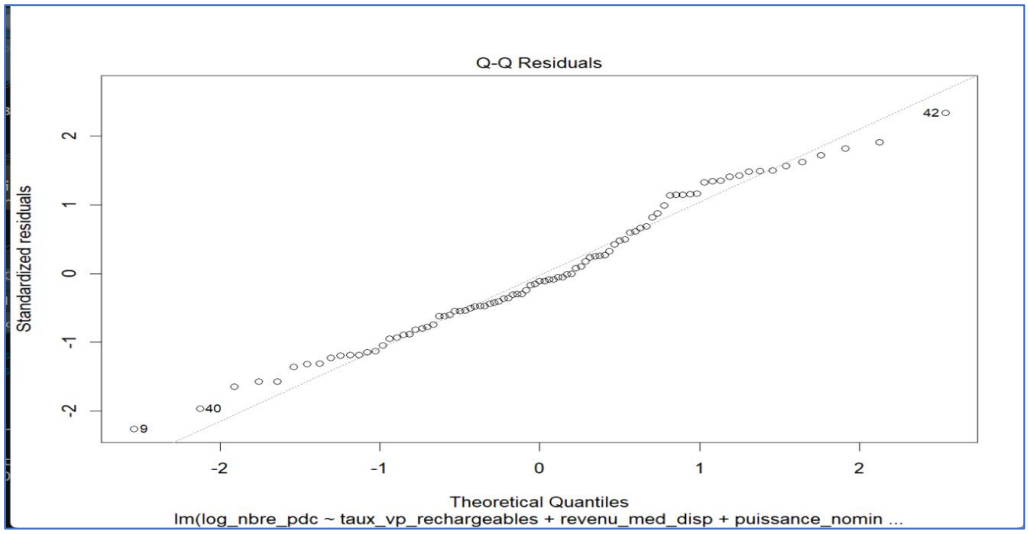
```
> print(bp)

        studentized Breusch-Pagan test

data:  regBest
BP = 3.0465, df = 4, p-value = 0.5501
```



Le test de Breusch-Pagan, avec une statistique de 3.0465 et un p-value de 0.55, ne permet pas de rejeter l'hypothèse nulle d'homoscédasticité des résidus. Autrement dit, il n'y a pas de preuve significative que la variance des erreurs soit inégale selon les valeurs des variables explicatives. Cela suggère que l'hypothèse d'homoscédasticité est respectée, ce qui est un bon signe pour la validité des résultats du modèle et la fiabilité des intervalles de confiance et tests statistiques associés.

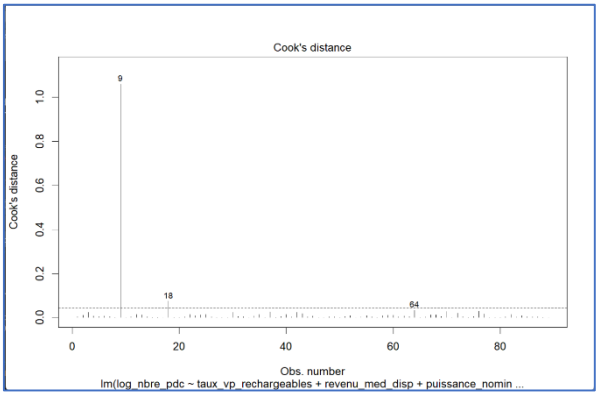
4. Normalité des résidus

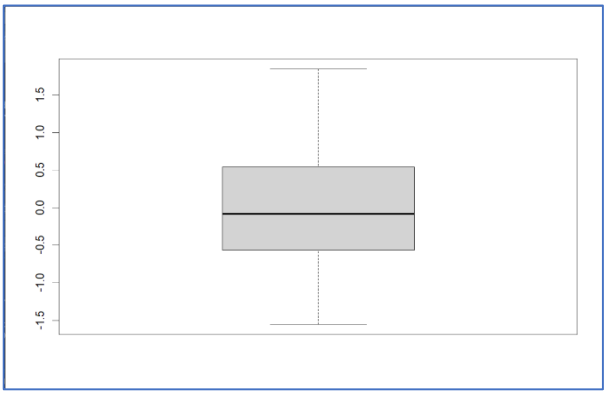




Le test de Shapiro-Wilk donne un W de 0.9726 avec un p-value de 0.0567, ce qui est légèrement supérieur au seuil classique de 0,05. Cela signifie que l'on ne rejette pas l'hypothèse de normalité des résidus au niveau de confiance de 95 %. En d'autres termes, les résidus du modèle suivent globalement une distribution normale, ce qui est une bonne indication pour la validité des tests statistiques dans la régression linéaire. Le graphique histogramme et le plot des résidus confirment visuellement cette distribution normale approximative.

5. Détection des points influents

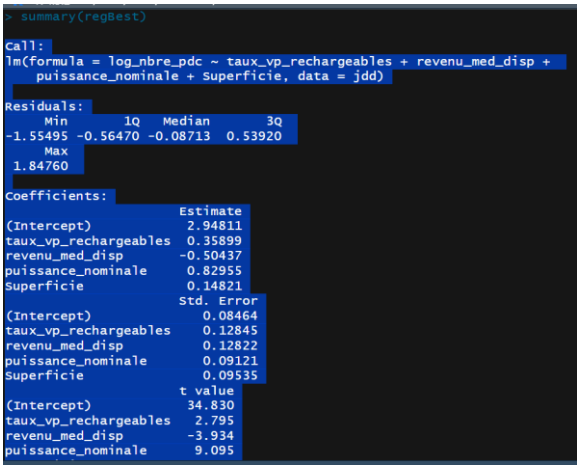




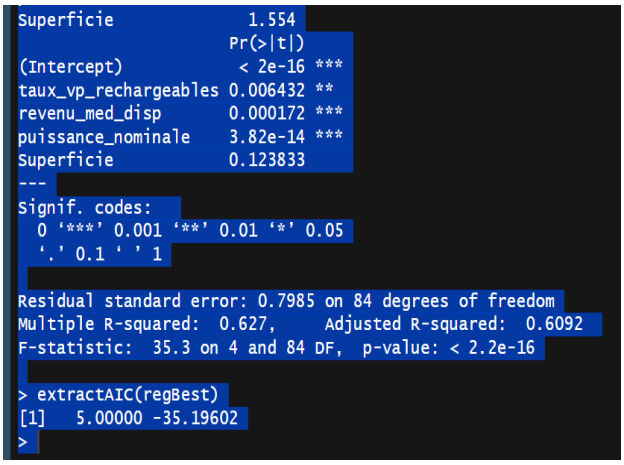
L'analyse des distances de Cook montre que deux observations, aux indices 9 et 18, dépassent le seuil critique fixé à 0,045 environ (calculé comme 4/nombre d'observations). Ces points sont donc considérés comme des observations influentes qui peuvent avoir un impact disproportionné sur la qualité et la stabilité du modèle. Le graphique des distances de Cook (plot 4) confirme visuellement la présence de ces points influents, et le boxplot des résidus permet d'identifier d'éventuelles valeurs extrêmes dans la distribution des erreurs. Il est important de vérifier ces observations (9 et 18) pour comprendre si elles résultent d'erreurs de

saisie, de mesures aberrantes ou de cas particuliers justifiés. Leur prise en compte ou exclusion doit être réfléchie, car elles peuvent affecter la robustesse du modèle. En résumé, la présence de ces points influents constitue un signal d'alerte nécessitant une investigation complémentaire.

V. INTERPRETATION DU MODELE FINAL



```
> summary(regBest)

Call:
lm(formula = log_nbre_pdc ~ taux_vp_rechargeables + revenu_med_disp +
    puissance_nominale + Superficie, data = jdd)

Residuals:
     Min       1Q   Median       3Q
-1.55495 -0.56470 -0.08713  0.53920
     Max
 1.84760

Coefficients:
                   Estimate
(Intercept)         2.94811
taux_vp_rechargeables  0.35899
revenu_med_disp      -0.50437
puissance_nominale    0.82955
Superficie            0.14821
                   Std. Error
(Intercept)           0.08464
taux_vp_rechargeables    0.12845
revenu_med_disp          0.12822
puissance_nominale       0.09121
Superficie               0.09535
                   t value
(Intercept)         34.830
taux_vp_rechargeables   2.795
revenu_med_disp        -3.934
puissance_nominale      9.095
```



```
Superficie              1.554
                       Pr(>|t|)
(Intercept)             < 2e-16 ***
taux_vp_rechargeables 0.006432 **
revenu_med_disp       0.000172 ***
puissance_nominale    3.82e-14 ***
Superficie            0.123833
---
Signif. codes:
  0 '***' 0.001 '**' 0.01 '*' 0.05
  '.' 0.1 ' ' 1

Residual standard error: 0.7985 on 84 degrees of freedom
Multiple R-squared:  0.627,     Adjusted R-squared:  0.6092
F-statistic:  35.3 on 4 and 84 DF,  p-value: < 2.2e-16

> extractAIC(regBest)
[1]   5.00000 -35.19602
>
```

Le modèle linéaire multiple explique la variable log_nbre_pdc à partir de quatre prédicteurs : taux_vp_rechargeables, revenu_med_disp, puissance_nominale et Superficie. Le coefficient d'interception est estimé à environ 2,95, ce qui correspond à la valeur prédite de la variable dépendante lorsque toutes les variables explicatives sont nulles.

Parmi les variables, taux_vp_rechargeables, revenu_med_disp et puissance_nominale sont statistiquement significatives ($p < 0,01$), indiquant un effet réel sur le logarithme du nombre de points de charge. Le coefficient positif pour taux_vp_rechargeables (0,36) et puissance_nominale (0,83) signifie qu'une augmentation de ces variables est associée à une augmentation du nombre de points de charge (sur l'échelle logarithmique).

En revanche, revenu_med_disp a un effet négatif (-0,50), suggérant que des revenus médians disponibles plus élevés sont liés à une diminution du nombre de points de charge. La variable Superficie n'est pas statistiquement significative ($p \approx 0,12$), ce qui signifie que son effet n'est pas clairement établi dans ce modèle. Le modèle explique environ 62,7% de la variance totale de la variable cible (R^2 ajusté = 0,61), ce qui est relativement bon, et la significativité globale du modèle est confirmée par un F-statistique très élevé (35,3) avec un p-value < 2.2e-16.

Enfin, l'AIC du modèle est de -35,2, indiquant une bonne qualité d'ajustement en prenant en compte la complexité du modèle (nombre de paramètres = 5).

En résumé, ce modèle est robuste, avec des variables explicatives pertinentes, même si une variable (Superficie) pourrait être revue ou approfondie.

VI. EXEMPLE DE PREDICTION

```
> valeurs_nouvelles <- data.frame(
+   taux_vp_rechargeables = 0.15,
+   revenu_med_disp = 25740,
+   puissance_nominale = 90,
+   Pop_15_64_ans = 100000,
+   Superficie = 50
+ )
> valeurs_nouvelles_std <- as.data.frame(t((t(valeurs_nouvelles) - moyennes) / ecarts_type))
>
> pred_log <- predict(regBest, newdata = valeurs_nouvelles_std)
> nbre_pdc_pred <- exp(pred_log) - 1
> print(paste("Prédiction nombre de bornes :", round(nbre_pdc_pred,1)))
[1] "Prédiction nombre de bornes : 178.9"
>
```

En appliquant le modèle de régression aux nouvelles données standardisées (taux de véhicules rechargeables à 0,15, revenu médian disponible à 25 740, puissance nominale à 90, population de 100 000 personnes âgées de 15 à 64 ans, et superficie de 50), la prédiction du logarithme du nombre de bornes de recharge est obtenue. Après transformation inverse, cela correspond à une estimation d'environ **179 bornes**. Cette prédiction montre que, dans ces conditions socio-économiques et techniques spécifiques, le modèle anticipe un parc important de bornes, cohérent avec les influences positives de la puissance nominale et du taux de véhicules rechargeables sur le nombre de bornes. Ce résultat chiffré offre une projection tangible pour la planification et le développement des infrastructures de recharge.

VII. CONCLUSION ET PERSPECTIVES

En conclusion, le modèle de régression multiple développé explique de manière significative la variation du nombre de bornes de recharge électrique en fonction du taux de véhicules rechargeables, du revenu médian disponible, de la puissance nominale et de la superficie. Avec un R^2 ajusté de 60,9 %, le modèle présente une bonne capacité explicative, bien que des tests aient révélé une légère autocorrélation des résidus, suggérant une marge d'amélioration. La prédiction effectuée sur un exemple standardisé estime à environ 179 le nombre de bornes à prévoir dans un contexte donné, ce qui témoigne de la pertinence pratique du modèle pour des usages de planification stratégique. Ce résultat permet ainsi d'anticiper de manière éclairée les besoins en infrastructures, en tenant compte de facteurs à la fois socio-économiques et techniques.

Au-delà des résultats quantitatifs, cette SAÉ a permis de mobiliser un ensemble cohérent de compétences professionnelles : interroger les données, les analyser, modéliser avec rigueur et valoriser les résultats dans un cadre appliqué. La structuration d'un jeu de données exploitable, l'interprétation critique des sorties statistiques, et la formulation d'une réponse claire au défi proposé ont constitué une expérience formatrice, concrète et ancrée dans les enjeux actuels de transition énergétique. Ce projet

nous a ainsi permis de relier théorie et application tout en développant une posture réflexive, essentielle dans un contexte de données complexes.

Une analyse spatiale plus poussée, ou l'ajout de variables comme la densité de population ou l'accessibilité aux axes routiers, pourrait constituer une piste de prolongement pertinente pour affiner davantage la modélisation et mieux capturer les dynamiques territoriales liées au déploiement des infrastructures de recharge.