UNIVERSITÉ DE REIMS CHAMPAGNE-ARDENNE

UFR SEN
Sciences Exactes et Naturelles

Projet SEP0832
Méthodes d'échantillonnage.

Master 1 Mathématiques et Applications

UFR Sciences Exactes et Naturelles
Parcours Statistique pour l'Évaluation et la Prévision.

PAR
ALIZEE SCHOLLHORN.
GRACE MOLINGO MAMBIKA.
NOUHAILA KHOUNA.
YANNICK GNAGO .

ENSEIGNANT: M. AMOR KEZIOU

.

2021/2022

Table des matières

Introduction

Dans le cadre de ce projet, nous allons étudier l'évolution trimestrielles du nombre d'emplois salariés en 2021 en fonction de l'évolution trimestrielle dans chacun des grands secteurs d'activité à savoir le secteur tertiaire marchand et non marchand, le secteur agricole, le secteur industriel et le secteur de construction. Nous cherchons donc à trouver des réponses aux questions suivantes :

1. Comment l'évolution trimestrielle dans chacun des 5 secteurs d'activité explique l'évolution trimestrielle générale du nombre d'emplois salariés ?
2. Quels sont les secteurs qui exercent une grande influence sur l'évolution trimestrielle durant cette période ?

Afin de répondre à notre problématique, nous allons mettre en place des modèles de régressions linéaires multiples, c'est-à-dire d'établir des relations linéaires entre une variable, dite expliquée, et une ou plusieurs interactions de variables, dites explicatives « l'évolution trimestrielle des 5 secteurs » afin de trouver le meilleur modèle qui décrit au mieux la variable expliquée « l'évolution trimestrielle générale ».

Notre travail se divisera donc en 3 grandes étapes : la construction du modèle de RLM complet et la vérification de ses hypothèses, la sélection des tops modèles et l'estimation de l'erreur de prévision de ces modèles afin de sélectionner le meilleur des ces derniers.

En ce qui concerne la partition des tâches, nous l'avons partagé de la manière suivante :

- Nouhaila Khouna s'est occupée de la construction du modèle RLM complet, de la vérification de ses hypothèses ainsi que la mise en œuvre de la méthode LOOCV pour l'estimation de l'erreur de prévision des modèles.
- Yannick Gnago a trouvé la base de données sur laquelle nous avons travaillé, il s'est occupé de la mise en œuvre de la méthode exhaustive et de la méthode K-fold CV.
- Grace Molingo Mambika a travaillé sur la méthode de recherche pas à pas descendante pour la sélection de modèles ainsi que la mise en place de l'outil bootstrap.
- Alizée Schollhorn a pris en charge les deux méthodes : recherche pas à pas ascendante et algorithme génétique pour la sélection du modèle et la méthode de l'ensemble de validation pour l'estimation de l'erreur de prévision des modèles.

I - Explication de la base de données

Nous utiliserons la base de données relatif aux estimations trimestrielles d'emploi salarié au 3 trimestre 2021 et de l'évolutions par grand secteur d'activité téléchargé sur le site de l'INSEE.

La base de données est sous forme de fichier Excel composée de 3 feuilles (DEP : les données selon les départements de la France métropolitaine, REG : les données selon les régions de la France métropolitaine, précision : les détails sur les années antérieurs).
Nous nous intéressons à la feuille relative aux données selon les départements (DEP).
Dans la base initiale, il y'a 9 variables(colonnes) mais celles qui nous intéresse sont :
– Departement : la liste de tous les départements de la France (100 départements).
– Nbre_demplois_salaries : Nombre d'emplois salariés (en milliers).
– Evol_trimestriel : l'évolution trimestriel d'emplois salariés (en pourcentage)
– agriculture : l'évolution d'emplois salariés dans le secteur agricole (en pourcentage).
– industrie : l'évolution d'emplois salariés dans le secteur industriel (en pourcentage).
– construction : l'évolution d'emplois salariés dans le secteur de construction (en pourcentage).
– tertiaire_marchand : l'évolution d'emplois salariés dans le secteur tertiaire marchand (en pourcentage).
– tertiaire_non_marchand :l'évolution d'emplois salariés dans le secteur tertiaire non marchand (en pourcentage).
Il faut noter qu'un secteur regroupe des entreprises de fabrication, de commerce ou de service qui ont la même activité principale (au regard de la nomenclature d'activité économique considérée).
En France, les principaux secteurs d'activité économique sont : l'agriculture, l'industrie, l'énergie, le Commerce et artisanat et la construction. On les dénomme donc :
–Secteur primaire : l'agriculture et l'agroalimentaire.
–Secteur secondaire : l'industries de transformation des matières premières (produits finis).
–Secteur du BTP (bâtiment et travaux publics) : les activités de conception, de construction et de rénovation de bâtiments (publics et privés, industriels ou non) et d'infrastructures (routes, réseaux, canalisations, etc.).
–Secteur tertiaire marchand : le commerce, transports, activités financières, services rendus aux entreprises, services rendus aux particuliers, hébergement-restauration, immobilier, information-communication.
–secteur tertiaire non marchand : l'administration publique, enseignement, santé humaine, action sociale.
Après la modification de la base pour la rendre plus exploitable, nous avons eu :

	Evol_trimestriel	agriculture	industrie	construction	tertiaire_marchand	tertiaire_non_marchand
1	0.4	1.7	0.4	-1.0	0.6	0.2
2	0.4	1.1	0.2	0.4	0.5	0.2
3	0.8	1.3	0.0	-0.3	1.9	0.1
4	-0.2	-22.5	0.7	0.0	0.9	-0.7
5	-3.0	-42.2	0.3	0.6	-4.8	-0.3
6	0.7	-1.1	-0.1	0.1	1.8	-0.7
7	0.5	5.3	0.2	-0.4	1.2	0.0
8	0.1	0.9	0.2	-0.6	0.3	-0.1
9	-0.3	-2.3	-0.7	-0.2	-0.3	-0.2
10	0.2	-8.0	-0.4	-0.4	0.9	0.5
11	-0.2	-7.1	0.3	0.2	-0.1	0.3
12	0.1	1.7	1.3	-0.4	-0.1	-0.3
13	0.4	-3.7	0.1	-0.1	0.9	-0.2
14	0.3	-1.1	0.1	0.1	0.6	0.1
15	-0.3	1.5	0.1	-1.2	-0.4	-0.1

Showing 1 to 15 of 100 entries, 6 total columns

Figure 1 – La database du projet

II- Construction du modèle de RLM complet

Modèle de RLM complet

Nous cherchons à trouver parmi l'évolution dans les 5 grands secteurs ceux qui expliquent le « mieux » l'évolution trimestrielle d'emplois salariés. En termes statistiques, cela se traduit par la recherche des variables explicatives qui décrivent le « mieux » notre variable cible « évolution trimestrielle ».

On pose :
Y = la variable cible = l'évolution trimestrielle (en%)
X = le vecteur des variables explicatives suivantes :

- Evolution trimestrielle dans l'agriculture (en %)
- Evolution trimestrielle dans l'industrie (en %)
- Evolution trimestrielle dans la construction (en %)
- Evolution trimestrielle dans le tertiaire marchand (en %)
- Evolution trimestrielle dans le tertiaire non marchand (en %)

Dans un premier temps, nous allons étudier le modèle global, celui contenant toutes les variables explicatives :

$$Y = w_0 + w_1.X_1 + w_2.X_2 + \ldots + w_5.X_5 + \varepsilon$$

Où :

- w_i sont les paramètres du modèle RLM
- ε représente les termes d'erreur du modèle non observés

```
Call:
lm(formula = Evol_trimestriel ~ ., data = evol)

Residuals:
     Min       1Q   Median       3Q      Max
-0.23352 -0.04463 -0.00144  0.05117  0.46276

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           -0.011351   0.012754   -0.89   0.3757
agriculture            0.018202   0.001745   10.43  < 2e-16 ***
industrie              0.132707   0.016652    7.97  3.7e-12 ***
construction           0.031454   0.016045    1.96   0.0529 .
tertiaire_marchand     0.454081   0.011284   40.24  < 2e-16 ***
tertiaire_non_marchand 0.341112   0.019480   17.51  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1008 on 94 degrees of freedom
Multiple R-squared:  0.9693,    Adjusted R-squared:  0.9677
F-statistic: 593.3 on 5 and 94 DF,  p-value: < 2.2e-16
```

Figure : Résultats du modèle global de RLM

Vérification des hypothèses :

On remarque que les 4 variables « évolution dans l'agriculture », « l'industrie », « le tertiaire marchand » et « le tertiaire non marchand » sont toutes significatives sauf la variable « construction ». Avant d'aller plus loin dans l'analyse de ce modèle RLM, vérifions d'abord ces hypothèses :

La non-corrélation des erreurs

Afin de tester cette hypothèse, nous allons utiliser du test Durbin-Watson. Sa statistique est comprise entre 0 et 4 et est définie par :
Dans ce cas, nous cherchons à tester l'hypothèse nulle contre l'hypothèse alternative suivantes :

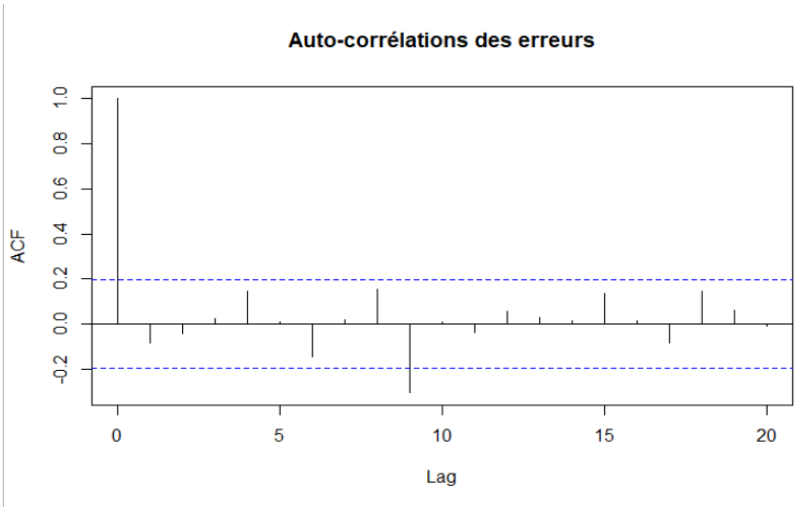
- ➔ H0 : Les erreurs sont non corrélées
- ➔ H1 : Les erreurs sont corrélées

```
        Durbin-Watson test

data:  modele_total
DW = 2.1638, p-value = 0.4196
alternative hypothesis: true autocorrelation is not 0
```

Figure _ Résultats du test Durbin-Watson

La statistique du test de Durbin-Watson étant proche de 2, on en conclut que les erreurs sont non-corrélées. En effet, il existe un deuxième moyen pour vérifier cette hypothèse, celui reposant sur des graphiques. Nous pouvons donc se servir de la fonction « acf » du logiciel R pour tracer la fonction d'autocorrélation des erreurs.



La corrélation des erreurs n'est pas nulle pour les variables qui dépassent l'intervalle de confiance. Cependant, H0 n'est pas significativement différente de zéro. On estime alors que l'hypothèse de non-corrélation des erreurs de notre modèle RLM est bien vérifiée.

Linéarité

On cherche à tester l'hypothèse de linéarité entre la variable réponse « évolution trimestrielle » et les variables explicatives. Pour cela, nous allons appliquer une régression linéaire local des résidus en fonction des valeurs ajustées. On obtient :

$$S_{DW} := \frac{\sum_{i=2}^{n} (\widehat{\varepsilon}_i - \widehat{\varepsilon}_{i-1})^2}{\sum_{i=1}^{n} \widehat{\varepsilon}_i^2}$$

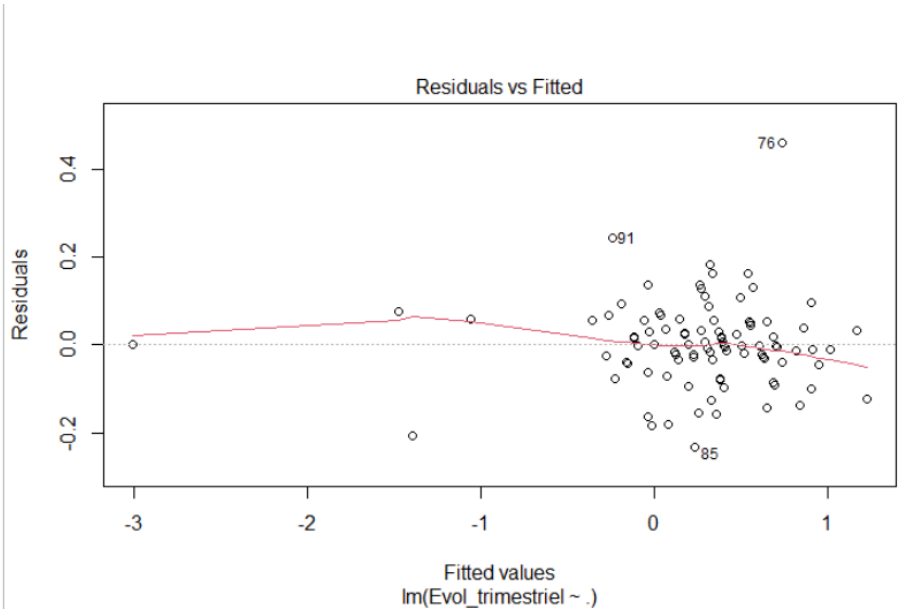


Figure_ RL local des résidus en fonction des valeurs ajustées.

La courbe rouge représente la droite d'ajustement. Elle approche légèrement une courbe horizontale droite, ce qui nous permet de valider la linéarité entre les variables. On estime donc que l'hypothèse de la linéarité de ce modèle RLM est validée.

Homoscédasticité

Pour vérifier cette l'homoscédasticité, nous allons utiliser le test de Breush-Pagan pour évaluer l'hypothèse nulle contre l'hypothèse alternative suivantes :

- ➔ H0 : l'erreur est homoscédastique
- ➔ H1 : l'erreur est hétéroscédastique

```
        Breusch-Pagan test

data:  modele_total
BP = 20.178, df = 5, p-value = 0.001157
```

Figure _ Résultats du test Breush-Pagan

La valeur du p-value est inférieur à 0.05 (proche de 0), l'hypothèse d'homoscédasticité n'est donc pas vérifiée. Ce qui est confirmé avec le graphe suivant :

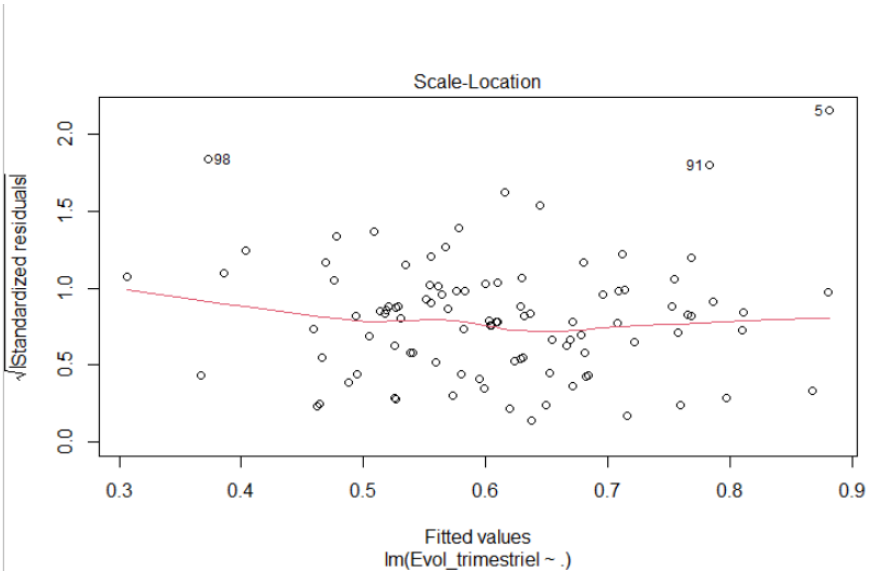


Figure_ RL local des résidus studentisées en fonction des valeurs ajustées.

Pour remédier à ce problème, nous allons appliquer une transformation « $\sqrt{.}$ » sur les valeurs de la variable réponse.

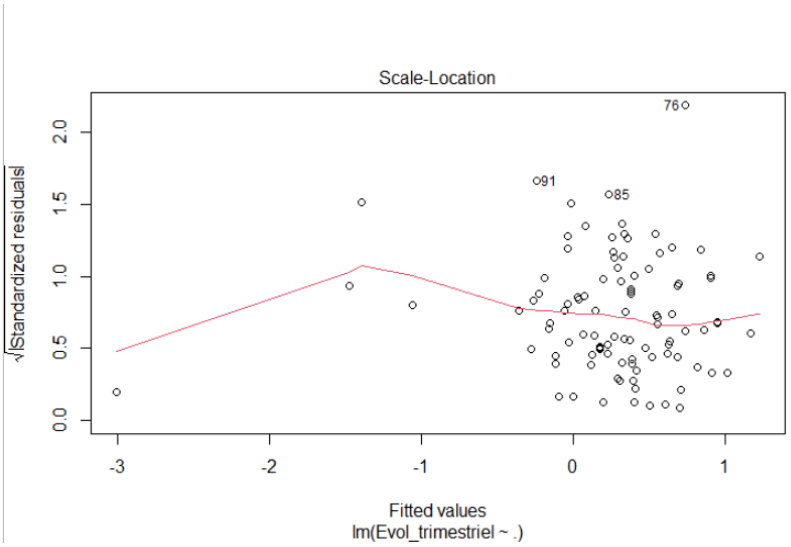


Figure_ RL local des résidus studentisées en fonction des valeurs ajustées.

On remarque que la courbe horizontale est presque parfaite et les points sont réparties des deux côtés de la plage des valeurs prédites de façon homogène. On en conclut que l'hypothèse d'homoscédasticité est désormais vérifiée.

Normalité

On considère l'hypothèse nulle et l'hypothèse alternative suivantes :

H0 : l'erreur est normale

Contre

H1 : l'erreur n'est pas normale

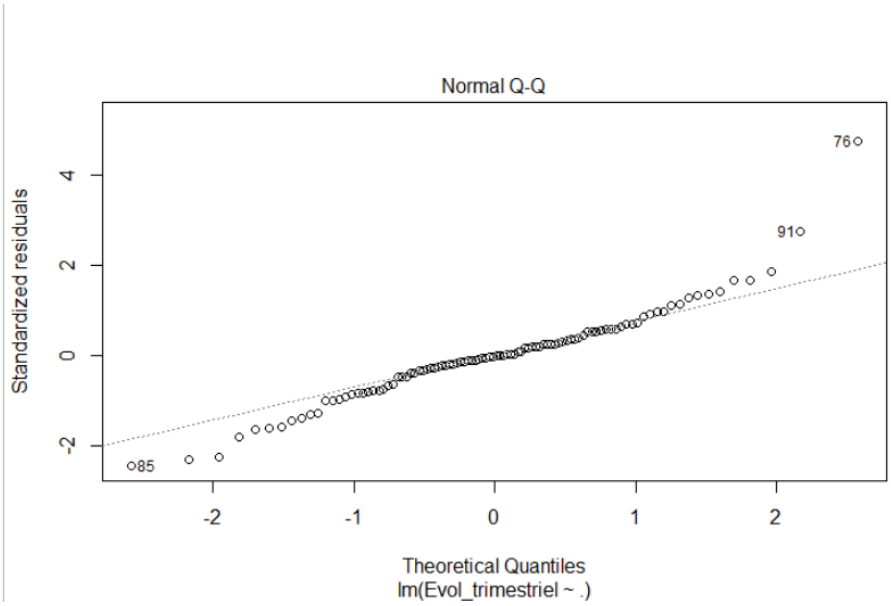


Figure _ représentation Q-Q plot

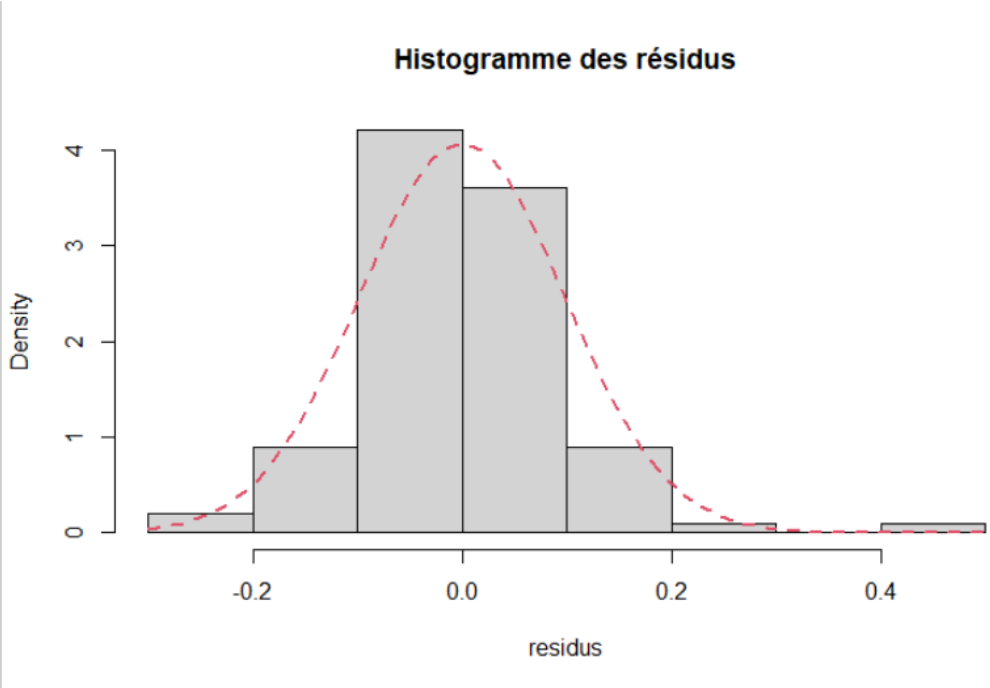


Figure _ Histogramme des erreurs et la densité gaussienne

Les points dans la représentation Q-Q plot se situent presque tous approximativement le long de la droite, ce qui montre que l'erreur suit une loi normale. Cette hypothèse est également confirmée par l'histogramme des résidus ci-dessus.

Classement des variables explicatives

Dans cette partie, nous allons classer les variables explicatives de la plus significative à la moins significatives selon les valeurs croissantes des p-values du test de Fisher. On compte sur les p-value et non pas sur les coefficients des variables car ces dernières ne sont pas distribuées de la même manière.
On cherche à tester l'hypothèse nulle de non-significativité de chaque variable.
Pour i ∈ [0,5] :

H0 : Bi = 0 Contre H1 : Bi ≠ 0

tertiaire_marchand	agriculture	tertiaire_non_marchand	industrie	construction
1.479236e-60	6.643802e-52	2.397102e-31	6.475717e-12	2.863337e-05

Figure _ Résultats du test de Fisher : valeurs croissantes des p-values

On constate que la variable la plus significative est « l'évolution dans le tertiaire marchand », suivie de l'évolution dans « l'agriculture » ensuite dans « le tertiaire marchand » et dans « l'industrie ». Enfin, celle qui explique le moins notre variable réponse est l'évolution dans la construction.

L'influence des outliers sur le modèle RLM :

Dans cette section, nous essayerons de détecter les départements qui contrastent grandement avec les autres départements avec une évolution normale, ainsi que ceux exerçant une grande influence sur l'évolution trimestrielle. Ces départements seront appelés respectivement des outliers et des points leviers extrêmes où un petit changement à leur niveau peut engendrer de grands changements dans l'évolution trimestrielle générale.
On sait que les observations avec des résidus studentisés supérieurs à 3 ou inférieurs à -3 sont considérées des outilers. De plus, un point de levier extrême peut être détecté avec une statistique du levier supérieure à 2*(p+1)/n (ici = 0.12).
Afin de mesurer l'influence d'une observation dans le modèle, nous pouvons se servir de la distance de Cook qui repose sur la condition suivante :

Distance de Cook > 4/(n-p-1) (=0.04) -> observation extrêmement influente

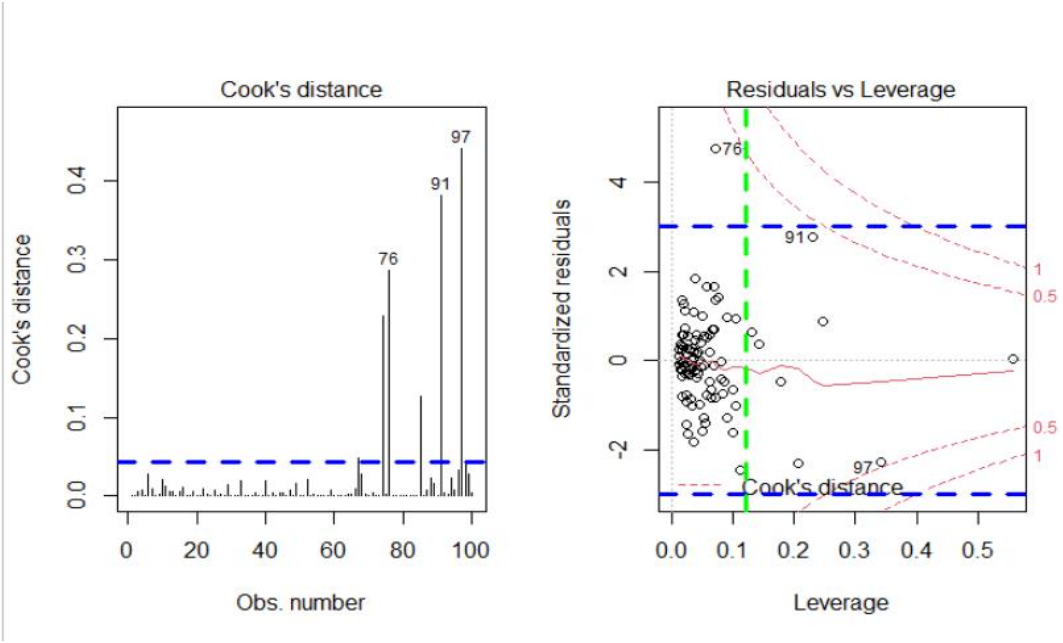


Figure_ Détection des points leviers extrêmes et leurs influences

Ici, on remarque que les départements « Haute-Savoie », « Yonne » et « Val-dOise » ont une influence supérieure à celle du reste des départements. De plus, leur distance de Cook est supérieure à 0.04. On en conclut que ces valeurs extrêmes exercent une grande influence sur l'ajustement du modèle.

Le modèle de régression multiple complet a permis de montrer que l'évolution trimestrielle d'emploi salariés en France est expliqué par toutes les variables exogènes (le secteur tertiaire, le secteur agricole, le secteur industriel) sauf la variable du secteur de construction. A cet effet dans la prochaine partie, nous essayerons de trouver les sous-ensembles des variables explicatives qui conduit au meilleur modèle RLM.

Afin de trouver le meilleur modèle de RLM, nous appliquerons les différents algorithmes de sélection de modeles vu en TP.

III-Sélection du meilleur modèle :

Afin de trouver le meilleur modèle, une multitude de méthodes de sélections de modèle existe. Nous allons donc tester les deux méthodes qui nous semblent être les plus pertinentes, il s'agit de la méthode exhaustive et de la méthode génétique. D'autres méthodes tel que la méthode ascendante et la descente sont disponible en annexe (A et B).

La méthode exhaustive

La recherche exhaustive est une méthode algorithmique qui consiste principalement à construire les $2^P - 1$ modèles possibles (p : les nombre de variables explicatives) et de sélectionner le meilleur modèle selon les critères considérés (BIC,R^2, cp).

1) Le meilleur modèle de RLM avec au maximum une interaction des variables explicatives

Ici, notre but de trouver le meilleur modèle selon les critères sous cette forme : $Y = a \cdot x_1 + b \cdot x_2 + \cdots$,avec Y : l'evolution trimestrielle du nombre d'emploi et x_i: les variables explicatives.

A- Le R² ajusté :

il est défini par:$R_a^2 = 1 - \frac{n-1}{n-k-1}(1 - R^2)$ où n : le nombre de varible observé ,k : le nombre de variables explicatives ($k \leq p$) et R^2: le coefficient de determination multiple.

B- Le Bayésien Information Criterion (BIC):

Il est défini par: $BIC = -2\log L + \log(n)\,(k+2)$ où L : répresente la vraisemblance maximisée.

C- Le cp de Mallow:

Il est défini par: $C_p = \frac{1}{n}(|Y - Y_0 1|^2 + 2(1+k)\sigma^2)$ où Y_0: le vecteur des valeurs adjustées selon le modele utilisant k variables explicatives.

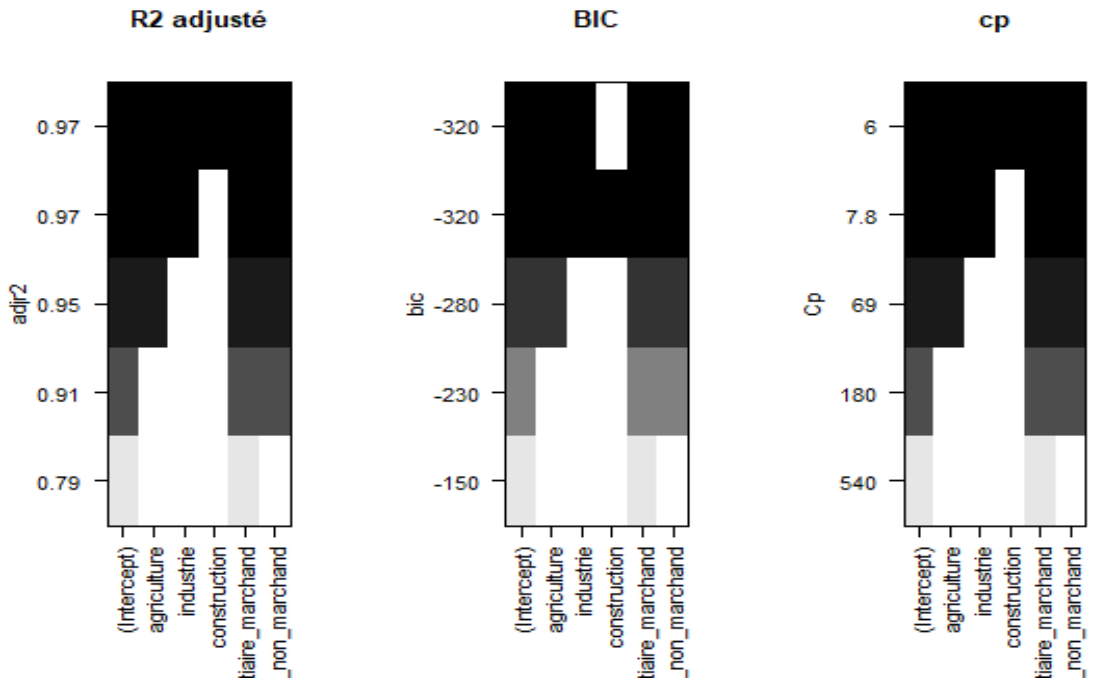


Figure – Les critères de sélection du meilleur modèle avec une interaction de variable.

On obtient 2 modèles différents qui sont sélectionnés selon les 3 critères. Le premier modèle sélectionné suivant le critère du Cp de Mallow et du R^2 ajusté est le modèle qui prend en compte tous les secteurs d'activité. Ensuite, le second modèle sélectionné selon le critère du BIC, donc en choisissant le modèle qui minimise le BIC est celui qui possède toutes les variables explicatives sauf la variable construction. On n'a pas pris pas compte le critère du R^2 pour sélectionner un modèle puisqu'il sélectionne automatiquement le plus grand modèle donc celui qui prend en compte tous les secteurs d'activité.

En utilisant la fonction 'glmulti()' sous R, on trouve que: le meilleur modèle avec une 1 interaction est: $Evol\ trimestriel \sim agriculture + industrie + tertiaire\ marchand + tertiaire\ non\ marchand$

```
Response: Evol_trimestriel
                      Df  Sum Sq Mean Sq  F value    Pr(>F)
agriculture            1 10.1679 10.1679  971.673 < 2.2e-16 ***
industrie              1  0.6266  0.6266   59.883 1.067e-11 ***
tertiaire_marchand     1 16.1749 16.1749 1545.714 < 2.2e-16 ***
tertiaire_non_marchand  1  3.1320  3.1320  299.306 < 2.2e-16 ***
Residuals              95  0.9941  0.0105
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure –les variables exogènes significatives du modèle.

Cela sous-entend que l'évolution trimestriel d'emplois salariés en 2021 dépend fortement de l'évolution d'emplois dans tous les secteurs sauf le secteur de la construction.

2) <u>Le meilleur modèle de RLM avec au maximum deux interaction des variables explicatives</u>

Ici, notre but de trouver le meilleur modèle selon un critère choisit sous cette forme : $Y = a \cdot x_1 + b \cdot x_2 + c \cdot x_1 : x3 + d . x_2 : x_4$ …., avec :Y l'evolution trimestrielle d'emploi salarié et x_i: les variables explicatives.

En tenant compte de notre travail effectué sous R, nous détaillerons les résultats sous le critère de BIC avec la fonction 'glmulti' vu que le temps de compilation est assez long sous le critère AIC (L'Akaike Information Criterion est défini par: $AIC = -2\log(L) + 2(k+2)$).

Le meilleur modèle obtenue avec une 2 interaction sous le critère BIC est:

$$Evol\ trimestriel \sim agriculture + industrie + construction + tertiaire\ marchand + tertiaire\ non\ marchand + construction : agriculture + tertiaire\ marchand : industrie + tertiaire\ non\ marchand : tertiaire\ marchand$$

```
                                        Pr(>F)
agriculture                          < 2.2e-16 ***
industrie                            5.784e-14 ***
construction                         2.962e-06 ***
tertiaire_marchand                   < 2.2e-16 ***
tertiaire_non_marchand               < 2.2e-16 ***
agriculture:construction             0.0974934 .
industrie:tertiaire_marchand         0.0005572 ***
construction:tertiaire_marchand      0.3383514
agriculture:tertiaire_non_marchand   0.0061232 **
tertiaire_marchand:tertiaire_non_marchand 0.0073006 **
Residuals
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure – Les variables exogènes significatives du modèle.

Ce modèle montre que les paires d'interactions des secteurs qui sont significatives pour l'évolution d'emplois salariés sont : le secteur industriel et le tertiaire marchand, le secteur agriculture et le tertiaire non marchand, le secteur tertiaire marchand et le tertiaire non marchand.

L'approche exhaustive permet de trouver les modèles par comparaison selon les critères. Seulement, l'inconvénient est que le temps de calcul devient très important si le nombre de variables est grand. Donc lorsque le nombre de variables est grand, on privilégie souvent les méthodes pas à pas qui consistent à construire les modèles de façon génétique.

La méthode génétique

Une autre méthode de sélection de modèle est l'algorithme génétique. On va donc s'en servir pour trouver de nouveaux modèles candidats pour trouver le meilleur possible, celui qui expliquera au mieux l'évolution trimestrielle.

Comme précédemment, nous allons nous baser sur les critères du BIC et de l'AIC pour sélectionner les modèles.

Selon le critère du BIC, le modèle sélectionné est l'évolution :

Evol_trimestriel ~ agriculture + industrie + tertiaire_marchand + tertiaire_non_marchand + construction:agriculture + tertiaire_marchand:industrie + tertiaire_non_marchand:tertiaire_marchand

Ensuite, selon le critère de l'AIC, le modèle sélectionné est :

Evol_trimestriel ~ agriculture + industrie + construction + tertiaire_marchand + tertiaire_non_marchand + construction:agriculture + tertiaire_marchand:industrie + tertiaire_marchand:construction + tertiaire_non_marchand:agriculture + tertiaire_non_marchand:tertiaire_marchand

Grâce à cette sélection de modèles, nous allons pouvoir estimer les erreurs de chaque modèle ainsi que ceux trouvé pour la méthode exhaustive.

En conclusion, tous les secteurs d'activités semblent avoir un impact dans l'évolution du nombre de salariés dans tous les départements de France. Plus il y a d'entreprises de ce type de secteurs plus il y a des salariés. Ces secteurs favorisent l'évolution du nombre de salariés.

Nous avons trouvé plusieurs meilleurs modèles selon les méthodes et les critères. Pour la suite de l'étude, nous décidons d'en garder que trois (vu les niveaux de significativités des variables exogènes) pour estimer les erreurs de prévision de ces modelés RLM afin de sélectionner celui qui est le plus optimal.

IV- Estimation de l'erreur de prévision des modèles RLM sélectionnés :

Nous estimerons correctement l'erreur théorique de prévision de chaque modèle, afin de choisir le modèle ayant l'erreur estimée la plus faible. Les méthodes de validation croisée permettent d''évaluer efficacement cette erreur. On présente ici trois méthodes différentes (méthode de l'ensemble de validation, méthode K-fold CV) on pourra également retrouver la méthode LOOCV en annexe.

Les modèles sélectionnés sont :

Modèle 1: Evol_trimestriel ~ agriculture+industrie +tertiaire_marchand +tertiaire_non_marchand

Modele 2: Evol_trimestriel ~ agriculture + industrie + tertiaire_marchand + tertiaire_non_marchand + construction:agriculture + tertiaire_marchand:industrie +tertiaire_non_marchand:tertiaire_marchand

Modele 3: Evol_trimestriel ~ agriculture + industrie + construction + tertiaire_marchand +tertiaire_non_marchand + construction:agriculture +terrtiaire_marchand:industrie + tertiaire_marchand:construction +tertiaire_non_marchand:agriculture + tertiaire_non_marchand:tertiaire_marchand

Méthode de l'ensemble de validation

La méthode de l'ensemble de validation est une méthode d'apprentissage et de validation. C'est pourquoi, on a commencé par séparer la base de données en 2 parties, la première contient les 2/3 de la base, ces données seront utilisées pour l'apprentissage et le restant pour tester les prédictions en évaluant la valeur de l'erreur. Cette méthode est privilégiée pour les grandes bases de données mais nous pouvons quand même nous servir.
L'estimation de l'erreur est obtenue en effectuant la régression linéaire de notre variable de l'évolution trimestrielle en fonction des variables des secteurs d'activités puis en prédisant les valeurs sur la base de données de test pour ensuite comparer celles qui sont bien prédites et celles qui ne le sont pas. Le modèle utilisé ici est le meilleur modèle, celui sélectionné dans la partie précédente. Puisque la base de données a été séparé en deux, on peut supposer que l'estimation de l'erreur n'est pas totalement fiable. Donc, pour avoir une meilleure estimation de l'erreur, on a estimé 2000 fois l'erreur en prenant à chaque fois de nouvelles valeurs dans la base de données d'apprentissage, donc la régression linéaire change et donc les valeurs prédites également. On a donc cherché la meilleure estimation de l'erreur pour chacun des 3 modèles qu'on a sélectionné précédemment. Pour avoir une idée visuelle, pour voir comment varie l'estimation de l'erreur en fonction des itérations on a tracer les graphiques pour chaque modèle. (Figure ci-dessous).

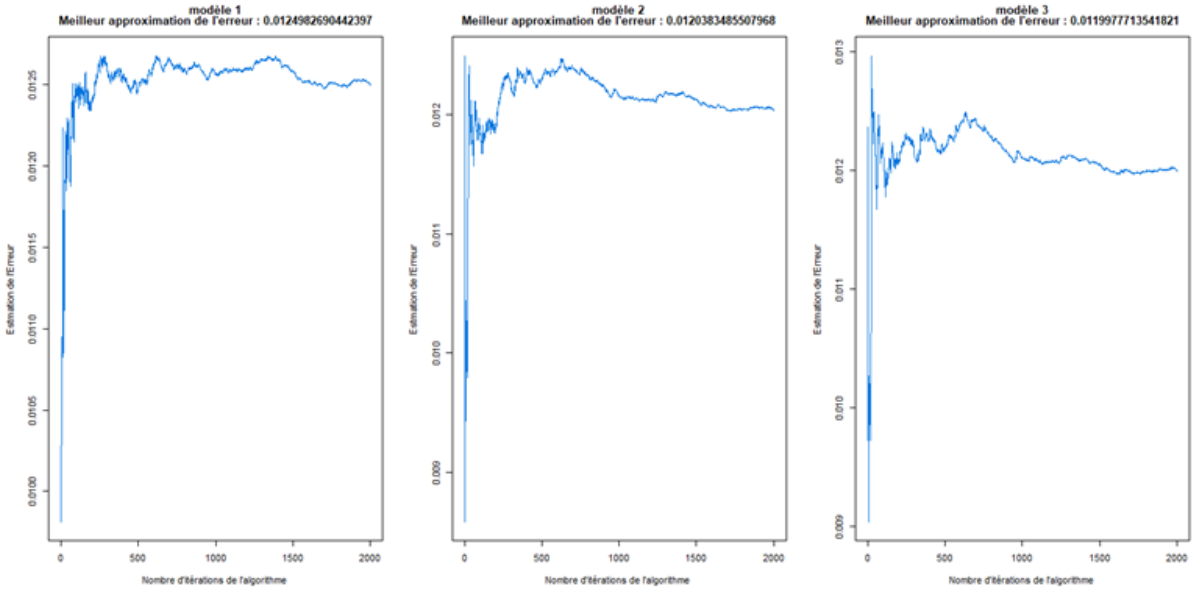


Figure – Estimation des erreurs pour les 3 modèles sélectionnés

On observe donc que plus le nombre d'itérations augmente, plus l'erreur se stabilise et donc approche au mieux l'erreur réelle. Le modèle 3 avec une erreur de 0.0121 possède une plus faible erreur que les deux autres modèles, selon la méthode d'apprentissage/validation.

Maintenant, regardons qu'elles sont les erreurs estimées par la méthode K-fold CV.

Méthode K-fold CV

La méthode K-fold cross validatiion est basée sur la methode précedentes, elle consite à diviser de maniére aléatoire les données en K(ici K=10) groupes et de répeter le calcul de l'erreur K fois en prenant le premier groupe comme ensemble de validation et les K-1 pour ajuster le modéles.

Nous appliquons donc cette méthode au 3 modeles selectionnées dans l'objectif de trouver le meilleur modele pour notre jeu de données.

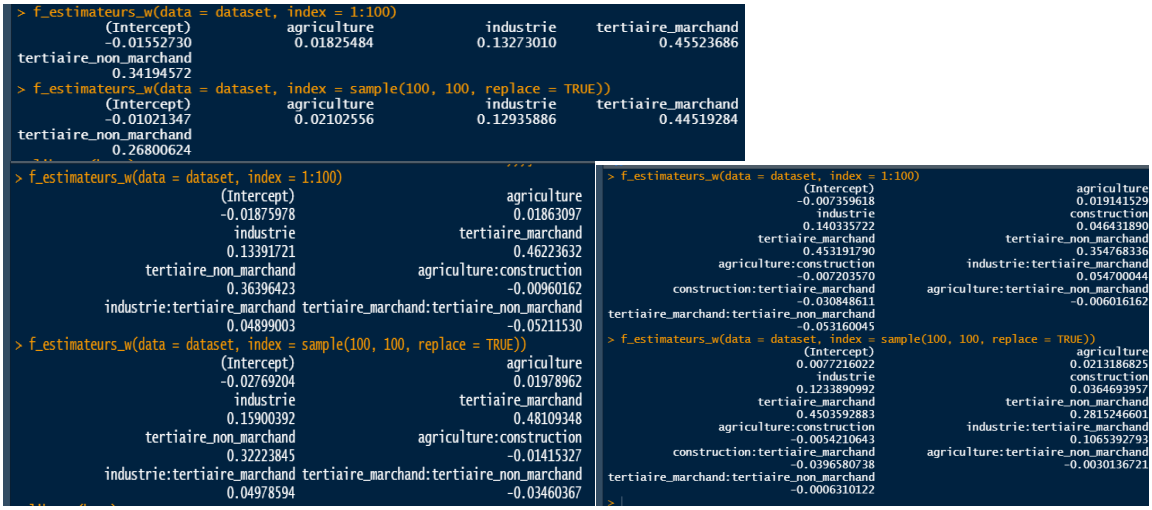
```
[1] "Résultats des estimations par 10-fold CV : "
[2] "Estimation de l'erreur du modele1 =  0.0109598301047113"
[3] "Estimation de l'erreur du modele2 =  0.00980680970537249"
[4] "Estimation de l'erreur du modele3 =  0.0102344468201667"
```

Au regard du résultat, nous pouvons dire que le meilleur modele parmi ces 3 modeles est le deuxieme modele ,i.e, celui dont les composantes expliquent au mieux l'évolution d'emploi trimestrielle salarié.

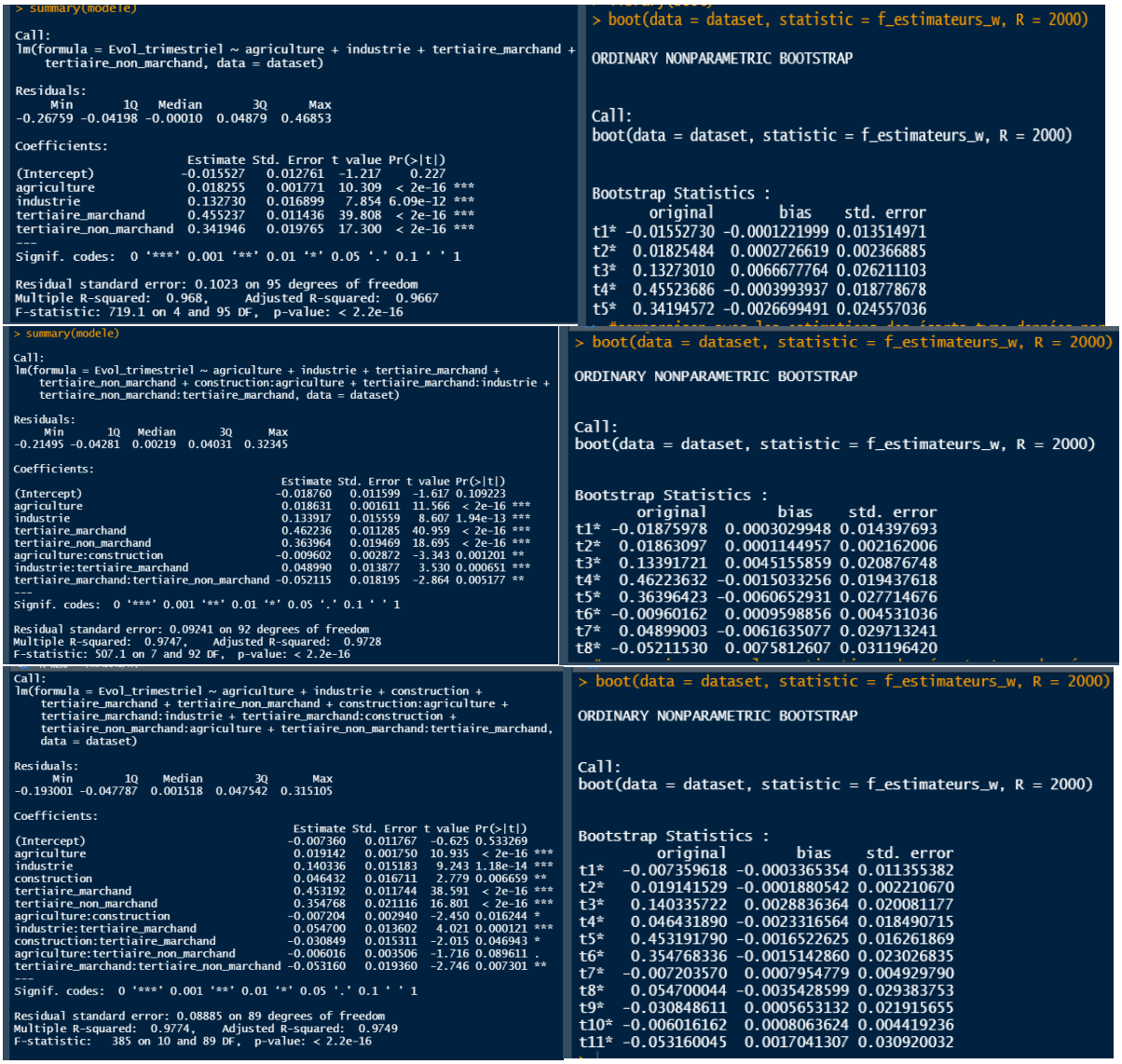
Le bootstrap

Le Bootstrap est un outil performant pour évaluer l'incertitude d'un estimateur ou d'une méthode d'apprentissage. Nous allons l'utiliser pour estimer l'erreur théorique de notre modèle, c'est à dire évaluer l'efficacité de notre estimateur.

Pour ce faire, on s'intéresse aux termes d'erreurs des estimateurs des écart-types. Nous allons tout d'abord construire notre estimateur.



Ensuite, on fait une comparaison des valeurs de notre estimateur et celles de la régression linéaire.



```
> summary(modele)

Call:
lm(formula = Evol_trimestriel ~ agriculture + industrie + tertiaire_marchand +
    tertiaire_non_marchand, data = dataset)

Residuals:
     Min       1Q   Median       3Q      Max
-0.26759 -0.04198 -0.00010  0.04879  0.46853

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       -0.015527   0.012761  -1.217    0.227
agriculture        0.018255   0.001771  10.309  < 2e-16 ***
industrie          0.132730   0.016899   7.854 6.09e-12 ***
tertiaire_marchand 0.455237   0.011436  39.808  < 2e-16 ***
tertiaire_non_marchand 0.341946   0.019765  17.300  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1023 on 95 degrees of freedom
Multiple R-squared:  0.968,    Adjusted R-squared:  0.9667
F-statistic: 719.1 on 4 and 95 DF,  p-value: < 2.2e-16
```

```
> boot(data = dataset, statistic = f_estimateurs_w, R = 2000)

ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = dataset, statistic = f_estimateurs_w, R = 2000)


Bootstrap Statistics :
      original        bias    std. error
t1* -0.01552730 -0.0001221999 0.013514971
t2*  0.01825484  0.0002726619 0.002366885
t3*  0.13273010  0.0066677764 0.026211103
t4*  0.45523686 -0.0003993937 0.018778678
t5*  0.34194572 -0.0026699491 0.024557036
```

```
> summary(modele)

Call:
lm(formula = Evol_trimestriel ~ agriculture + industrie + tertiaire_marchand +
    tertiaire_non_marchand + construction:agriculture + tertiaire_marchand:industrie +
    tertiaire_non_marchand:tertiaire_marchand, data = dataset)

Residuals:
     Min       1Q   Median       3Q      Max
-0.21495 -0.04281  0.00219  0.04031  0.32345

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                              -0.018760   0.011599  -1.617 0.109223
agriculture                               0.018631   0.001611  11.566  < 2e-16 ***
industrie                                 0.133917   0.015559   8.607 1.94e-13 ***
tertiaire_marchand                        0.462236   0.011285  40.959  < 2e-16 ***
tertiaire_non_marchand                    0.363964   0.019469  18.695  < 2e-16 ***
agriculture:construction                 -0.009602   0.002872  -3.343 0.001201 **
industrie:tertiaire_marchand              0.048990   0.013877   3.530 0.000651 ***
tertiaire_marchand:tertiaire_non_marchand -0.052115   0.018195  -2.864 0.005177 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.09241 on 92 degrees of freedom
Multiple R-squared:  0.9747,    Adjusted R-squared:  0.9728
F-statistic: 507.1 on 7 and 92 DF,  p-value: < 2.2e-16
```

```
> boot(data = dataset, statistic = f_estimateurs_w, R = 2000)

ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = dataset, statistic = f_estimateurs_w, R = 2000)


Bootstrap Statistics :
      original        bias    std. error
t1* -0.01875978  0.0003029948 0.014397693
t2*  0.01863097  0.0001144957 0.002162006
t3*  0.13391721  0.0045155859 0.020876748
t4*  0.46223632 -0.0015033256 0.019437618
t5*  0.36396423 -0.0060652931 0.027714676
t6* -0.00960162  0.0009598856 0.004531036
t7*  0.04899003 -0.0061635077 0.029713241
t8* -0.05211530  0.0075812607 0.031196420
```

```
Call:
lm(formula = Evol_trimestriel ~ agriculture + industrie + construction +
    tertiaire_marchand + tertiaire_non_marchand + construction:agriculture +
    tertiaire_marchand:industrie + tertiaire_marchand:construction +
    tertiaire_non_marchand:agriculture + tertiaire_non_marchand:tertiaire_marchand,
    data = dataset)

Residuals:
     Min       1Q   Median       3Q      Max
-0.193001 -0.047787  0.001518  0.047542  0.315105

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                              -0.007360   0.011767  -0.625 0.533269
agriculture                               0.019142   0.001750  10.935  < 2e-16 ***
industrie                                 0.140336   0.015183   9.243 1.18e-14 ***
construction                              0.046432   0.016711   2.779 0.006659 **
tertiaire_marchand                        0.453192   0.011744  38.591  < 2e-16 ***
tertiaire_non_marchand                    0.354768   0.021116  16.801  < 2e-16 ***
agriculture:construction                 -0.007204   0.002940  -2.450 0.016244 *
industrie:tertiaire_marchand              0.054700   0.013602   4.021 0.000121 ***
construction:tertiaire_marchand          -0.030849   0.015311  -2.015 0.046943 *
agriculture:tertiaire_non_marchand       -0.006016   0.003506  -1.716 0.089611 .
tertiaire_marchand:tertiaire_non_marchand -0.053160   0.019360  -2.746 0.007301 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08885 on 89 degrees of freedom
Multiple R-squared:  0.9774,    Adjusted R-squared:  0.9749
F-statistic:   385 on 10 and 89 DF,  p-value: < 2.2e-16
```

```
> boot(data = dataset, statistic = f_estimateurs_w, R = 2000)

ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = dataset, statistic = f_estimateurs_w, R = 2000)


Bootstrap Statistics :
       original        bias    std. error
t1*  -0.007359618 -0.0003365354 0.011355382
t2*   0.019141529 -0.0001880542 0.002210670
t3*   0.140335722  0.0028836364 0.020081177
t4*   0.046431890 -0.0023316564 0.018490715
t5*   0.453191790 -0.0016522625 0.016261869
t6*   0.354768336 -0.0015142860 0.023026835
t7*  -0.007203570  0.0007954779 0.004929790
t8*   0.054700044 -0.0035428599 0.029383753
t9*  -0.030848611  0.0005653132 0.021915655
t10* -0.006016162  0.0008063624 0.004419236
t11* -0.053160045  0.0017041307 0.030920032
```

On trouve de légère différence au niveau de certaines valeurs notamment, l'erreur données par notre estimateur est plus élevée que celle donné par la fonction lm. Aussi, pour chacun des modèles l'estimateur de la variance est élevé et celui du biais est faible : c'est le résultat lorsqu'on est grande dimension.

Par le résultat de la régression, on choisit le deuxième modèle car elle comprend plus de variables significatives. Ce modèle semble être le meilleur et il est évident qu'il soit le meilleur car il a été sélectionné par les deux méthodes exhaustive (étude en petite dimension) et génétique (étude en grande dimension). De plus, il a un biais faible : le but étant souvent de diminuer le biais.

Il est aussi important de comprendre qu'assembler ces secteurs d'activités dans un département permet d'augmenter le nombre d'emplois salaries. Le secteur non marchand a besoin du secteur marchand pour une meilleur efficacité du travail, les employés du secteur marchand sont une aide pour ceux du secteur non marchand. Aussi, la plupart des employés du secteur industriel et du secteur tertiaire marchand sont moins qualifies, d'où l'association des deux secteurs. Associer aussi le secteur de la construction et de l'agriculture peut créer beaucoup d'emplois car on a besoin de beaucoup de personnes dans ces secteurs, qui réalisent beaucoup de tâches et dont le salaire est bas.

Conclusion

Suite de la mise en œuvre de plusieurs méthodes de sélections de modèles et des estimations des erreurs de prévisions, puis du bootstrap, on en a conclu que le secteur de l'agriculture, de l'industrie, du tertiaire marchand et du tertiaire non marchand explique au mieux l'évolution trimestrielle d'emplois salarié dans tous les départements de la France métropolitaine. On y ajoute également des interactions entre les secteurs tel que la construction associée au secteur de l'agriculture, le secteur tertiaire marchand associé au secteur industriel et pour finir, le secteur tertiaire non marchand associé au secteur tertiaire marchand.

Ces interactions deux à deux des secteurs d'activité n'ont pas été assemblées au hasard. Elles représentent les associations des secteurs qui ont un niveau de significativité important, i.e. elles auront une grande influence dans l'évolution trimestrielle d'emplois salariés. En effet, le secteur tertiaire (marchand et non marchand) est le plus représenté avec le plus d'emplois. Ensuite, Il y'a le secteur de l'industrie largement représente dans les secteurs d'activité. En effet, il présente un fort taux d'emplois salariales vu qu'il est considéré comme fournisseur du secteur tertiaire.

Annexe

Méthode ascendante

1) Méthode pas à pas ascendante : Version 1

La méthode forward consiste à faire des sélections de variables. On part donc d'un modèle trivial, c'est-à-dire un modèle avec uniquement l'intercept comme variable explicative. Puis on ajoutera petit à petit des variables afin d'obtenir le meilleur modèle. Posons p le nombre de variables explicatives candidates pour être sélectionné dans le modèle. La valeur p est enfaite la taille de la matrice de design moins un. En effet, la 1ère colonne de la matrice de design est l'intercept, constituée seulement de 1, et les autres colonnes sont les variables explicatives candidates pour le modèle. La matrice de design permet d'effectuer entre autres l'analyse et la gestion des systèmes complexe, ici elle va donc transformer les variables qualitatives en variables quantitatives. Puisque dans notre base de données nous n'avons pas de variables quantitatives donc cela ne change rien à notre nombre de variables explicatives candidates du départ donc on a p=5.
Il existe deux versions de cette méthode. La première version de la méthode forward consiste donc à d'abord construire le modèle trivial $M0$, puis construire p-k modèles, donc 5-k modèles en ajoutant à chaque fois une variable dans Mk. Ensuite, parmi ces modèles, on choisira le modèle qui optimise un critère considéré (AIC, BIC, R^2, R^2 ajusté ou le Cp de Mallow).

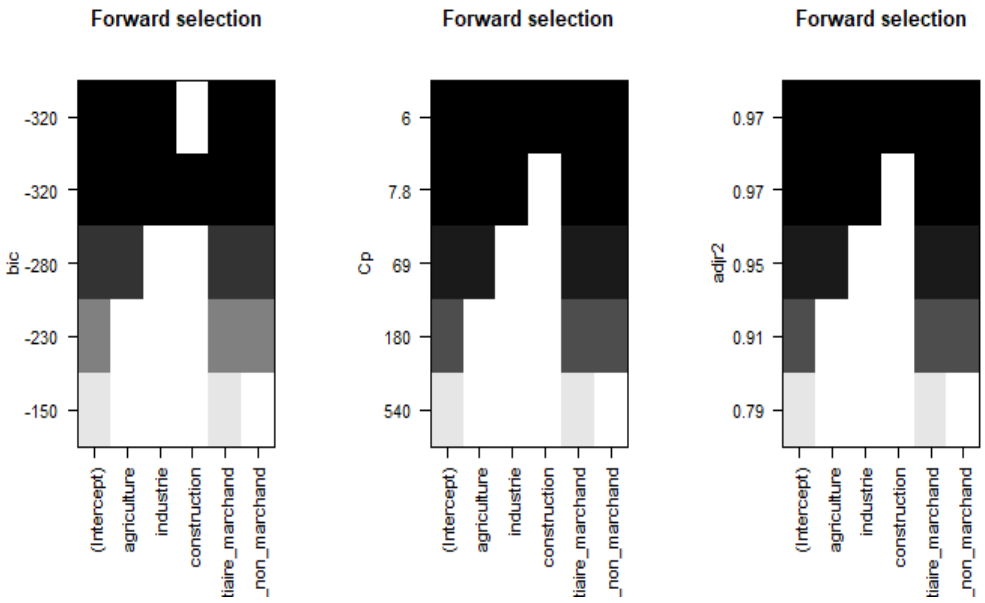


Figure – critères de sélection BIC, CP de Mallow et R^2

On obtient 2 modèles différents qui sont sélectionnés selon les 3 critères. Le premier modèle sélectionné selon le critère du BIC, donc en choisissant le modèle qui minimise le BIC est celui qui possède toutes les variables explicatives sauf la variable construction. Ensuite, le second modèle sélectionné suivant le critère du Cp de Mallow et du R^2 ajusté est le modèle qui prend en compte tous les secteurs d'activité. On ne prendra pas en compte le critère du R^2 pour sélectionner un modèle puisqu'il sélectionne automatiquement le plus grand modèle donc celui qui prend en compte tous les secteurs d'activité.

2) Méthode pas à pas ascendante : Version 2

La seconde version de la méthode forward commence encore une fois avec uniquement l'intercept comme variable explicative. Puis à chaque étape, on ajoute la variable qui diminue le plus le critère choisi. L'algorithme s'arrête lorsqu'il n'y a plus de variables explicatives à ajouter ou alors lorsque l'ajout de variable ne diminue plus critère.

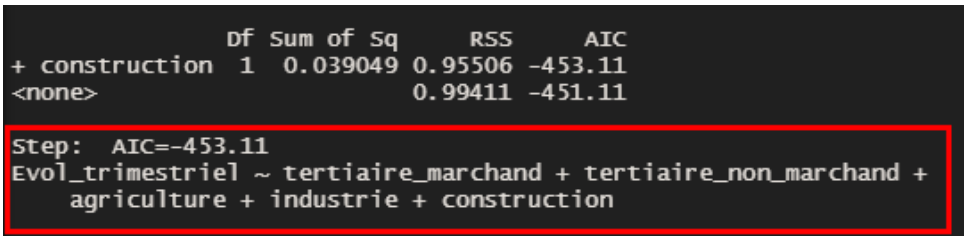


Figure – critères de l'AIC

L'AIC minimal (AIC = -453.11) est obtenu avec le modèle expliqué par tous les secteurs d'activité. Donc le modèle sélectionné par le critère de l'AIC est le modèle complet, celui qui contient toutes les variables explicatives.
Maintenant, passons au critère du BIC, également un critère qu'on cherche à minimiser.

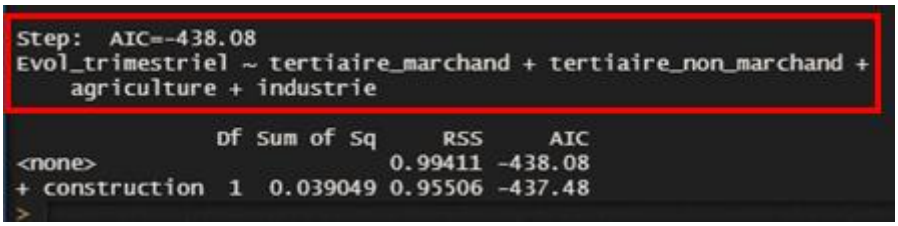


Figure – critères du BIC

Comme dans la version 1 de la méthode forward, le BIC sélectionne le modèle comprenant tous les secteurs d'activités sauf le secteur de la construction. Ici le BIC à une valeur de –438.08 et on voit bien que si le modèle ajouterait le secteur de la construction, le BIC passerait de –438.08 à -437.48 or, ce qu'on veut c'est minimiser le BIC.
Pour finir, passons au critère de Fisher.

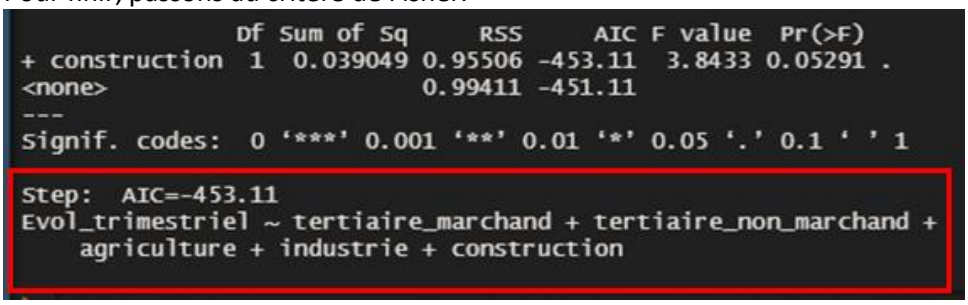
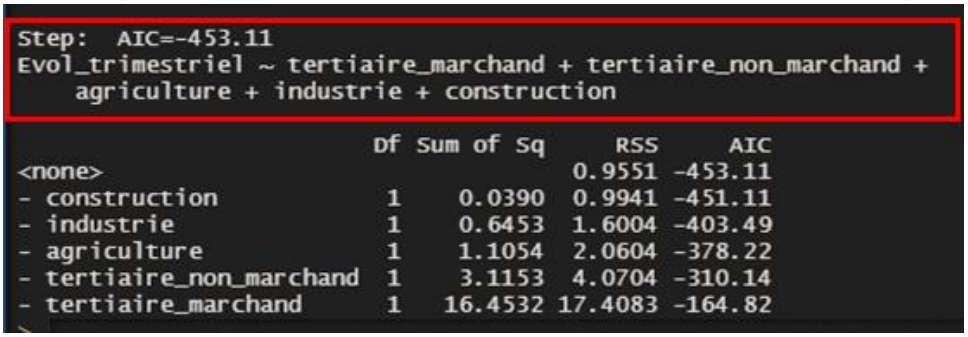


Figure – critères de Fisher

Le test de Fisher conclu que le meilleur modèle selon son critère est également celui qui possède toutes les variables explicatives.

3) Méthode ascendante bidirectionnelle

Pour finir, on va effectuer la méthode ascendante bidirectionnelle. Elle consiste toujours à commencer avec un modèle avec comme unique variable explicative l'intercept, mais cette fois ci, lors de chaque ajout de variables dans le modèle, chaque variable déjà incluse seront remise en causes. Ainsi, si une variable redevient plus significative compte tenu de celle qui vient d'être intégrée alors elle sera exclue du modèle.
Ici encore, on va déterminer le meilleur modèle en fonction des critères AIC, BIC et Fisher.



```
Step:  AIC=-453.11
Evol_trimestriel ~ tertiaire_marchand + tertiaire_non_marchand +
    agriculture + industrie + construction

                        Df Sum of Sq     RSS     AIC
<none>                                0.9551 -453.11
- construction           1    0.0390  0.9941 -451.11
- industrie              1    0.6453  1.6004 -403.49
- agriculture            1    1.1054  2.0604 -378.22
- tertiaire_non_marchand 1    3.1153  4.0704 -310.14
- tertiaire_marchand     1   16.4532 17.4083 -164.82
```

Figure – critères de l'AIC

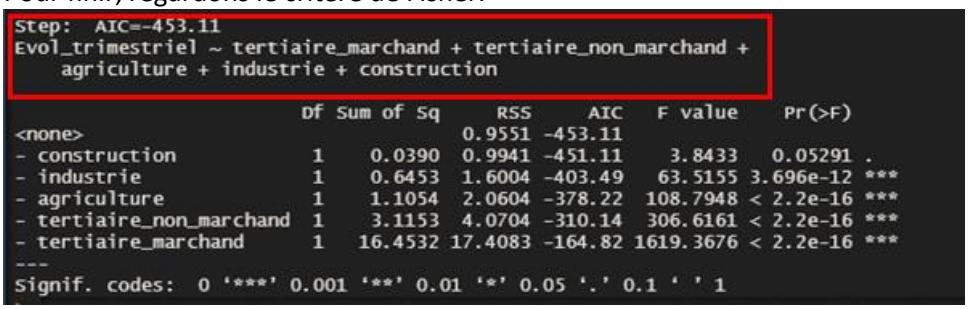
Avec le critère de sélection de modèle AIC, une fois encore on sélectionne le modèle qui possède toutes les 5 variables explicatives. Voyons qu'en est il pour le critère du BIC.

```
Step:  AIC=-438.08
Evol_trimestriel ~ tertiaire_marchand + tertiaire_non_marchand +
    agriculture + industrie

                        Df Sum of Sq     RSS     AIC
<none>                                0.9941 -438.08
+ construction           1    0.0390  0.9551 -437.48
- industrie              1    0.6456  1.6397 -392.65
- agriculture            1    1.1121  2.1062 -367.61
- tertiaire_non_marchand 1    3.1320  4.1262 -300.36
- tertiaire_marchand     1   16.5823 17.5764 -155.44
```

Figure – critères du BIC

Encore une fois, selon le critère de sélection du BIC, on sélectionne le modèle qui possède toutes les variables explicatives sauf la variable du secteur de la construction.
Pour finir, regardons le critère de Fisher.



```
Step:  AIC=-453.11
Evol_trimestriel ~ tertiaire_marchand + tertiaire_non_marchand +
    agriculture + industrie + construction

                        Df Sum of Sq     RSS     AIC   F value    Pr(>F)
<none>                                0.9551 -453.11
- construction           1    0.0390  0.9941 -451.11    3.8433   0.05291 .
- industrie              1    0.6453  1.6004 -403.49   63.5155 3.696e-12 ***
- agriculture            1    1.1054  2.0604 -378.22  108.7948 < 2.2e-16 ***
- tertiaire_non_marchand 1    3.1153  4.0704 -310.14  306.6161 < 2.2e-16 ***
- tertiaire_marchand     1   16.4532 17.4083 -164.82 1619.3676 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure – critères de Fisher

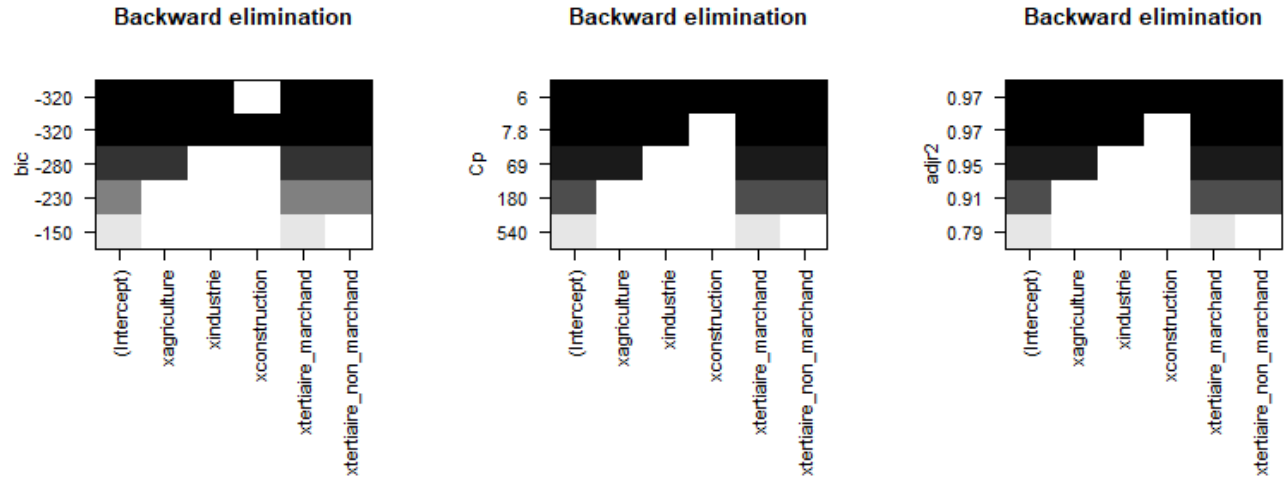
Pour le critère de Fisher aussi, le meilleur modèle est le modèle complet.
Donc pour conclure, on a eu le même résultat quelque soit la méthode de forward utilisé. Le meilleur modèle pour tous les critères sauf celui du BIC est le modèle complet. Et, pour ce dernier, le modèle sélectionné est celui qui contient tous les secteurs d'activité sauf celui de la construction.
Transition : Maintenant que nous avons vu la méthode exhaustive et la méthode ascendante forward, regardons ce qu'il en est du meilleur modèle selon la méthode descendante backward.

Méthode descendante

Dans cette étude de sélection de modèle dans l'algorithme pas à pas, on se servira des deux versions de la méthode descendante ainsi que la méthode bidirectionnelle.

1) Méthode descendante : Version 1

Encore appelé backward selection, la méthode descendante consiste à choisir le modèle qui optimise au mieux un critère donné dans l'algorithme pas à pas d'élimination des variables. Cet algorithme consiste à construire p (nombre de variables explicatives) modèle, puis supprime une variable dans chacun des k modèles consistants construits pour des valeurs de $k = p, \ldots, 1$ et choisit celui qui optimise au mieux un
critère donné.
Dans cette première version de la méthode, on fait la sélection de modèle par les critères BIC, Cp et R2 ajusté.



Selon ces trois critères de sélection, on obtient deux modelés différents. Une par le BIC et l'autre par le Cp de Mallow et le R2 ajusté.
Ainsi, le meilleur modèle de prévision selon le critère BIC est le modèle constitué de tous les secteurs d'activités sauf le secteur de la construction. Et, le meilleur modèle de prévision selon les critères Cp et adjr2 est le modèle constitué de tous les secteurs d'activités.

2) Méthode descendante : Version 2

Encore appelée bockward elimination, la méthode descendante version 2 consiste à éliminer pas à pas dans le modèle complet les variables qui diminuent les critères.
Pour cette deuxième version de la méthode on fait la sélection de modèle par les critères AIC, BIC et Fisher en tenant compte de la valeur de l'AIC.

```
> #selon AIC (k = 2)
> res.select.AIC.bac <- step(modele.complet, data = dataset, direction = "backward", k = 2)
Start:  AIC=-453.11
Evol_trimestriel ~ x

       Df Sum of Sq     RSS     AIC
<none>               0.9551 -453.11
- x     5     30.14 31.0956 -114.81
> #selon BIC (k = log(n))
> res.select.BIC.bac <- step(modele.complet, data = dataset, direction = "backward", k = log(n))
Start:  AIC=-437.48
Evol_trimestriel ~ x

       Df Sum of Sq     RSS     AIC
<none>               0.9551 -437.48
- x     5     30.14 31.0956 -112.21
> #selon le critère de Fisher
> res.select.F.bac <- step(modele.complet, data = dataset, direction = "backward", test = "F")
Start:  AIC=-453.11
Evol_trimestriel ~ x

       Df Sum of Sq     RSS     AIC F value    Pr(>F)
<none>               0.9551 -453.11
- x     5     30.14 31.0956 -114.81   593.3 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Selon le test par tous les critères de choix, le meilleur modelé de prévision est le modelé qui comprend tous les secteurs d'activités. C'est le modèle avec le maximum de secteurs qui minimise au mieux la valeur de l'AIC.

3) Méthode descendante bidirectionnelle

Encore appelé bidirectionnel élimination, la méthode descendante bidirectionnelle remet en cause les étapes de la méthode descendante compte tenu des variables exclues. Elle permet de réintégrer des variables qui redeviennent significatives.

Pour la sélection du meilleur modèle, on utilise les critères AIC, BIC et Fisher.

```
> # Methode descendante bidirectionnelle
> res.select.AIC.bac.both <- step(modele.complet, data = dataset,
+                                 direction = "both", k = 2)
Start:  AIC=-453.11
Evol_trimestriel ~ x

       Df Sum of Sq     RSS     AIC
<none>               0.9551 -453.11
- x     5     30.14 31.0956 -114.81
> res.select.BIC.bac.both <- step(modele.complet, data = dataset,
+                                 direction = "both", k = log(n))
Start:  AIC=-437.48
Evol_trimestriel ~ x

       Df Sum of Sq     RSS     AIC
<none>               0.9551 -437.48
- x     5     30.14 31.0956 -112.21
> res.select.F.bac.both <- step(modele.complet, data = dataset,
+                               direction = "both", test = "F")
Start:  AIC=-453.11
Evol_trimestriel ~ x

       Df Sum of Sq     RSS     AIC F value    Pr(>F)
<none>               0.9551 -453.11
- x     5     30.14 31.0956 -114.81   593.3 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
>
```

Il est évident que le résultat de ce test soit le même que celui de la deuxième version de la méthode descendante car ce dernier sélectionne déjà tous les secteurs. L'ensemble des secteurs d’activités sont déjà significatif dans l’évolution du nombre de salariés dans les départements.

Méthode LOOCV

La méthode leave-one-out Cross-Validation est l'une des méthodes de re-échantillonnage qui va nous servir à estimer l'erreur théorique de notre modèle RLM. Elle permet également de sélectionner le meilleur modèle tout en évitant le problème de sur-ajustement. Le meilleur modèle est donc celui qui possède la plus faible erreur estimée.

Cette méthode consiste à diviser le jeu de données en deux parties. La première partie contient l'observation (X1, Y1) et est utilisé pour la validation pendant que le reste des observations servent pour l'apprentissage du modèle. On répète cette règle n fois pour (X2, Y2), (X3, Y3) et ainsi de suite jusqu'à l'obtention de n estimations de l'erreur théorique. L'estimation donnée par la méthode LOOCV est :

$$\hat{\varepsilon} = \frac{1}{n} \cdot \sum_{i=1}^{n} e_i$$

En appliquant cette méthode sur les 3 modèles sélectionnés précédemment, on obtient :

```
"Résultats des estimations par LOOCV : "
"Estimation de l'erreur du modele1 =  0.012009200147598"
"Estimation de l'erreur du modele2 =  0.0102173628232573"
"Estimation de l'erreur du modele3 =  0.00968889520949844"
```

On remarque que l'erreur de prévision la plus faible est celle du modèle 3, donc c'est le modèle 3 qui explique/prédit le mieux notre variable cible.