Modélisation et prévision du nombre d'allocataires RSA payés par le Conseil général du Bas-Rhin

Hamza Sadaoui

Modélisation et prévision du nombre d'allocataires RSA payés par le conseil général du BAS-RHIN

Sadaoui.hamza@etu.unistra.fr

Université de STRASBOURG

UFR de Mathématique et d'Informatique

Master 1 Statistique

Le 26 aout 2013

Remerciements

Je souhaiterai tout d'abord remercier Monsieur Emmanuel BASTIAN pour m'avoir accueilli au sein de la direction de l'insertion professionnelle et de l'action sociale.

Je tiens également à remercier ma tutrice de stage, Anne EHRHART pour sa disponibilité, ses conseils et ses soutiens. Un grand merci à Charlotte SEITER, Fabien BIVERT et Carole Lambliotte pour leurs soutiens leurs bons humours.

Table des matières

Introduction :

Il y a plus vingt ans, on pouvait vivre en France sans avoir droit au moindre revenu. La loi du 1er décembre 1988, en créant le RMI (revenu minimum d'insertion), a mis fin à cette situation. Toute personne résidant régulièrement en France a désormais droit à un revenu minimum.

En 1998, dix ans après la création du RMI, a été inventé un système d'intéressement, permettant aux allocataires du RMI de conserver une partie de leurs allocations pendant la première année de retour au travail. Une loi de 2006 a transformé ce mécanisme en prime forfaitaire. Parallèlement, pour inciter au retour au travail, a été créée en 2001 une prime pour l'emploi. Mais il a été montré que la prime pour l'emploi (PPE) n'a pas eu l'effet incitatif attendu : son effet, dilué sur une très large population, est trop faible. Ces différentes réformes, conjuguées avec les différentes aides locales, ont abouti à un système d'une complexité telle qu'il est difficile de prédire les revenus d'une personne qui reprend un travail, augmente son activité ou tout simplement bénéficie d'une augmentation de salaire. Pendant cette période, l'écart entre le revenu minimum et le salaire minimum s'est accru.

Le système n'est plus tenable, une augmentation du RMI accroît le nombre des situations dans lesquelles le retour au travail n'est pas rémunérateur, Les augmentations du SMIC, quant à elles, n'ont pas évité l'émergence de travailleurs pauvres, et une augmentation du coût du travail peut avoir un effet d'éviction du marché du travail de personnes peu qualifiées.

C'est pour répondre à ces problèmes – retour au travail non rémunérateur, pauvreté au travail, soutien au pouvoir d'achat des personnes à faible revenu, complexité du système d'aide – qu'a été conçu **le revenu de solidarité active RSA.**

1 Présentation de l'organisme d'accueil et du sujet d'étude

1.1 Présentation du Conseil Général

1.1.1 Historique et missions

Apparue après la Révolution française de 1789, la création du Conseil Général en France fut déjà à cette période de l'histoire, souhaité pour assurer une égalité parfaite à tous les citoyens du territoire en termes de droits et d'accès aux services. Mais cette institution a évolué au cours du temps, et les missions qui ont été confiés au Conseil Général ont été étoffées.

Le Conseil général est donc une institution indépendante de l'Etat chargée de gérer les affaires d'un département, sa mission vise à améliorer la vie quotidienne de ses habitants. Il intervient pour ce faire dans 5 domaines : l'action sociale, l'équipement routier et portuaire et les transports, l'éducation, la culture et le patrimoine, l'environnement et le tourisme. Il apporte une aide aux communes et enfin il contribue au développement économique et social du territoire.

1.1.2 L'organisation générale

Le Conseil Général du Bas-Rhin (CG67) est la collectivité qui gère les affaires du département du Bas-Rhin. Le département ou a été recensé en 2009, 1 091 018 habitants répartis dans 44 cantons. Le CG67 compte environ 3700 agents répartis dans 5 pôles (cf. organigramme p.3):

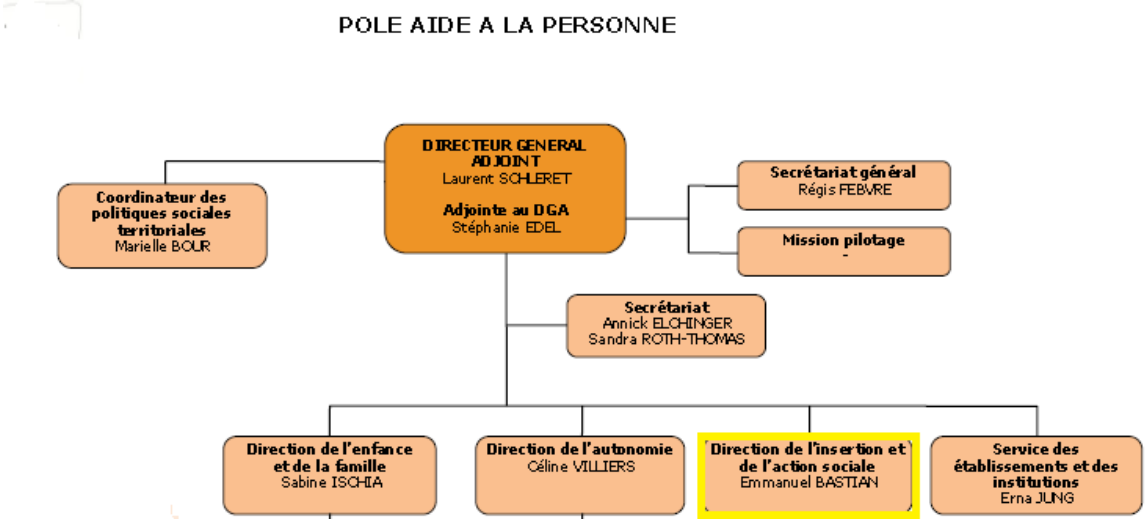
- Le Pôle Fonctionnel (PF)
- Le Pôle Aménagement du Territoire (PAT)
- Le Pôle Aide à la Personne (PAP)
- Le Pôle Epanouissement de la Personne (PEP)
- Le Pôle Développement des Territoires (PDT)

S'y ajoute la direction générale des services et le cabinet du Président du Conseil Général. Actuellement il s'agit de Guy-Dominique KENNEL, élu en 2008, également élu conseiller général du canton de Woerth.

1.1.3 Pole aide à la personne (PAP)

J'ai effectué mon stage au pôle aide à la personne qui s'occupe des missions d'action sociale et de solidarité confiées au Conseil Général par les lois de décentralisation. Il intervient dans quatre grands domaines : l'enfance et la famille ; l'insertion sociale et professionnelle et la lutte contre les exclusions sociales ; la santé publique ; les personnes âgées et les personnes en situation de handicap.

Figure n°1 : Organigramme du pole aide à la personne



1.2 Présentation du sujet d'étude

1.2.1 Méthode de projection au conseil général de Bas Rhin :

La méthode utilisée au conseil général de Bas Rhin, pour faire des projections du nombre d'allocataires RSA est une méthode d'extrapolation.

Cette méthode consiste à prolonger l'évolution passée ; il faut choisir :

- Jusqu'à quelle date on remonte ;
- Quelles sont les observations les plus importantes (pondération des Observations).

1.2.1.1 Problématique :

- De telles méthodes peuvent-elles être efficaces ? ;

- Permettent-elles au conseil général et à la direction de l'insertion professionnelle et de l'action sociale à un moment donné de mieux rationnaliser ses dépenses inscrites dans son cahier de charges ?;

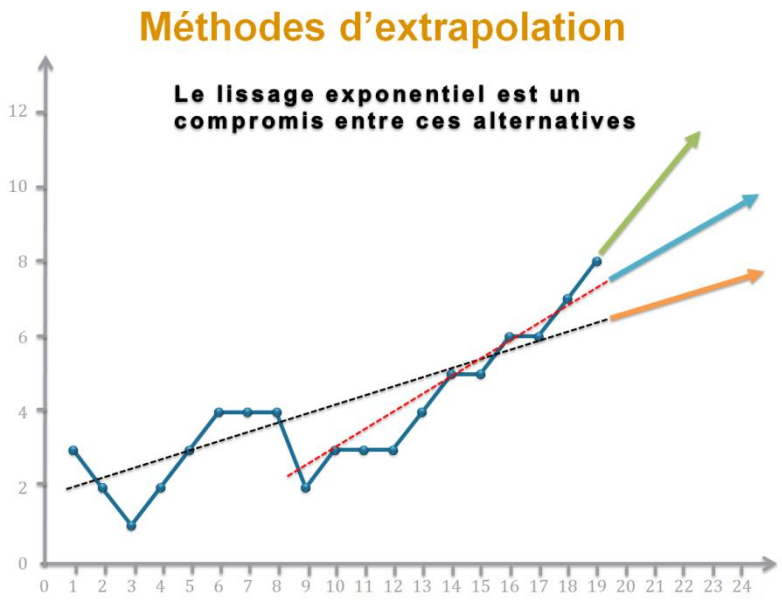
- Sont-elles à mesure de remédier à ce problème de prévision du nombre d'allocataires RSA au Bas Rhin ? ;

Des questions qui nous mènent à poser notre problématique principale ;

- Pouvons-nous trouver d'autres techniques plus fiables et mieux adaptés au contexte ?

Pour essayer de répondre à ces questions, nous allons nous attarder à la modélisation du nombre d'allocataires RSA payés par le conseil général du Bas Rhin et à la prévision à bases d'outils mathématiques qui sont en quelque sorte des maquettes très simplifiées des multiples relations directe ou indirectes entre les allocataire RSA et les demandeurs d'emploi , dans le souci de matérialiser, voir la mise en place d'un système informatique de Prévisions tenant en compte l'évolution des dites variables.

Figure n°2 : graphique de la méthode d'extrapolation



Fondamental
Les méthodes de lissage exponentiel sont un compromis entre ces trois types d'extrapolation puisqu'elles tiennent compte de toutes les observations, mais en diminuant leur importance au fur et à mesure que l'on remonte dans le passé.

1.2.1.2 Principe des méthodes de lissage exponentiel

- Les méthodes de lissage exponentiel sont des méthodes de prévision à court terme ;
- Elles supposent que le phénomène étudié ne dépend que de ses valeurs passées ;
- Ce sont des méthodes d'extrapolation qui donnent un poids prépondérant aux valeurs récentes : les coefficients de pondération décroissent exponentiellement en remontant dans le temps ;

- Chacune des méthodes dépend d'un ou plusieurs paramètres (paramètres de lissage) compris entre et ;
- Le poids de chacune des valeurs passées se calcule à partir de ces paramètres Complémentaire.

Cependant, ces méthodes de lissages exponentielles, on ne peut les appliquer que si on a 48 observations au minimum.
Le dispositif RSA a été mis en application à partir de 2009, l'évolution du nombre d'allocataires RSA est suivie chaque mois, mais nous ne pouvons toujours pas utiliser ces méthodes vu que l'on ne dispose que de 37 observations mensuelles

Cela nous mènent à chercher d'autres méthodes, qui prennent en compte les différents facteurs qui impactent directement ou indirectement sur l'évolution du nombre d'allocataires RSA.

1.3 Le revenu de solidarité Active :

1.3.1 Définition du RSA :

Le RSA, Revenu de Solidarité Active, est une allocation qui se substituera aux allocations minimum existantes comme le RMI ou l'API.

De plus elle remplacera les dispositifs financiers d'aide au retour à l'emploi comme la prime pour l'emploi, le PRE ou la prime forfaitaire de retour à l'emploi

Le RSA c'est pour ceux qui ne travaillent pas un revenu minimum et pour ceux qui travaillent un complément de revenu. C'est donc un instrument "mixte" qui met fin aux cloisonnements entre dispositifs et qui supprime les trous dans le dispositif. Le RSA est à la fois un moyen de garantir que le retour au travail procure des revenus supplémentaires et un puissant instrument de lutte contre la pauvreté.
Par sa double fonction, le RSA restera donc un revenu minimum pour les personnes ne travaillant pas, mais également un complément de revenu pour ceux qui travaillent.

1.3.2 Composantes du revenu de solidarité active

On distingue trois types d'allocataires du Revenu de solidarité Active.

1. Bénéficiaires du RSA Socle

Il s'agit des foyers bénéficiaires du RSA socle seul qui n'ont pas de revenu d'activité, ou bien, dont les membres ayant un emploi sont en période de cumul intégral (1)

2. Bénéficiaires du RSA Socle et Activité

Les bénéficiaires du RSA socle et activité ont de faibles revenus d'activité et l'ensemble de leurs ressources est inférieurs à un montant forfaitaire (intervenant dans le calcul du montant de la prestation et dépendant de la composition du foyer)

(1) *Le cumul intégral consiste à neutraliser l'ensemble des revenus d'activité pour le calcul du rSa, pendant une période de trois mois suivant la reprise d'emploi, dans la limite de quatre mois au cours des douze derniers mois.*

3. Bénéficiaires du RSA Activité Seul

Les foyers bénéficiaires du RSA activité seul qui ont de faibles revenus d'activité et dont l'ensemble des ressources est supérieur au montant forfaitaire.

1.4 Demandeurs d'emploi

1.4.1 Définition des Demandeurs d'emploi

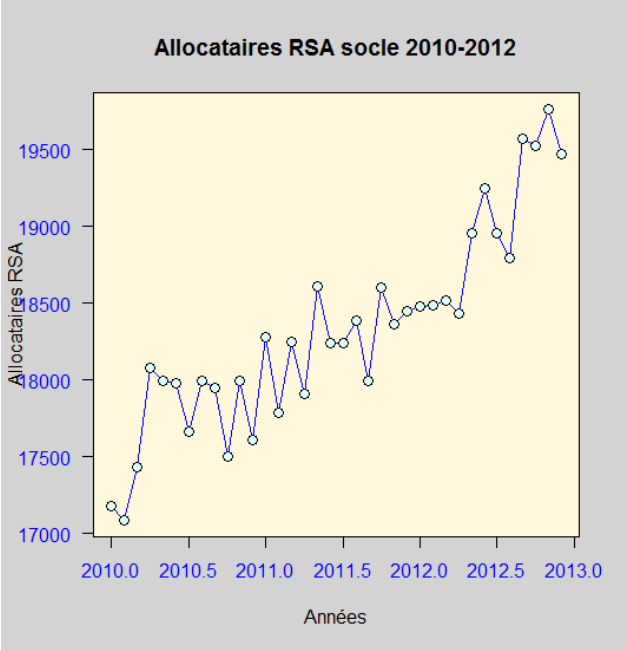
Les demandeurs d'emploi sont les personnes qui s'inscrivent à Pôle Emploi. Ces demandeurs sont enregistrés à Pôle Emploi dans différentes catégories de demandes d'emploi en fonction de leur disponibilité, du type de contrat recherché et de la quotité de temps de travail souhaité.

1.4.2 Catégories d'inscription :

- Catégorie A : Demandeurs d'emploi tenus de faire des actes positifs de recherche d'emploi, sans emploi.
- Catégorie B : Demandeurs d'emploi tenus de faire des actes positifs de recherche d'emploi, ayant exercé une activité réduite courte (i.e. de 78 heures ou moins) au cours du mois.
- Catégorie C : Demandeurs d'emploi tenus de faire des actes positifs de recherche d'emploi, ayant exercé une activité réduite longue (i.e. de plus de 78 heures) au cours du mois.
- Catégorie D : Demandeurs d'emploi non tenus de faire des actes positifs de recherche d'emploi (en raison d'un stage, d'une formation d'une maladie…), y compris les demandeurs d'emploi en convention de reclassement personnalisé (CRP) et en contrat de transition professionnelle (CTP), sans emploi .
- Catégorie E : Demandeurs d'emploi non tenus de faire de actes positifs de recherche d'emploi, en emploi (par exemple : bénéficiaires de contrats aidés).

1.5 Quelques statistiques sur le RSA et les demandeurs d'emploi :

Figure n°3 : évolution du nombre d'allocataires RSA payés par le CG67



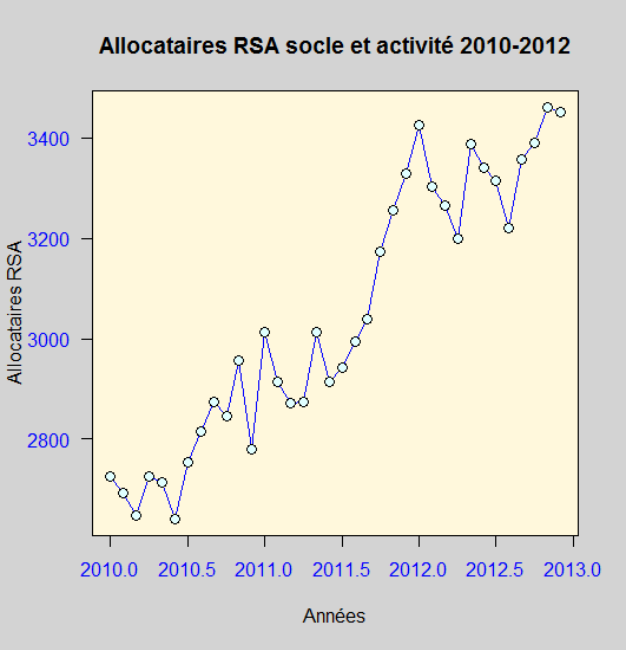
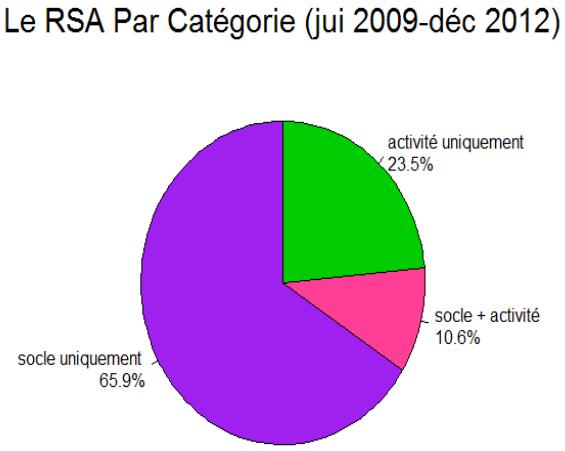
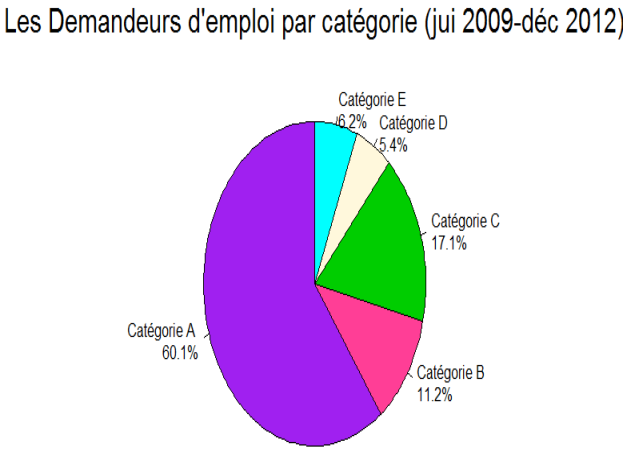


Figure n°4 : parts des allocataires RSA des demandeurs d'emploi





Conclusion :

Dans ce chapitre, nous avons présentés le conseil général du Bas Rhin, ses missions ainsi que ses différents pôles.

Le pole aide à la personne ou j'ai effectué mon stage, plus précisément la direction d'insertion professionnelle et de l'action sociale qui s'occupe de l'attribution de l'allocation RSA socle, et le RSA socle et activité a constaté que le nombre d'allocataires RSA ne cesse d'augmenter d'une année à une autre, c'est pour cette raison que la direction d'insertion professionnelle et de l'action sociale s'intéresse à l'évolution des postulants pour une allocation RSA.

Nous avons présentés également la méthode utilisée par le conseil général du Bas Rhin pour prévoir le nombre d'allocataires RSA.

Nous avons donnés quelques définitions de nos variables d'intérêts, et les différentes variables explicatives qui sont les catégories des demandeurs d'emploi.

Dans le chapitre suivant, nous donnerons quelques concepts théoriques de base des différents méthodes que nous utiliserons par la suite dans la pratique afin d'estimer l'enveloppe budgétaire consacrée par le conseil général du Bas Rhin aux allocataires

2. Quelques Concepts de base de la régression linéaire généralisée

2.1 Définition :

Un modèle linéaire généralisé pour $(x_i, y_i)_{i=1,\ldots..n}$ est une distribution pour la suite (y_i) déterminée par la donnée de
– une famille exponentielle à un paramètre de nuisance $f(,\theta,\varphi)v$
– une fonction r (dont la réciproque est appelée fonction de lien)
– une valeur φ_* et un vecteur de régression B_* avec les propriétés suivantes (les régresseurs x_i sont déterministes) :
– indépendance des y_i
– $y_i \sim f(y,\theta,\varphi)v(dy)$
– $\grave{b}(\theta_i) = r(x_i B^*)$
La dernière relation détermine θ_i en fonction de $x_i B^*$

La fonction $\acute{B}$ est bien inversible, en raison de la stricte convexité de b. Ceci se résume un peu rapidement par les propriétés suivantes :

La loi de y_i est issue de la famille $E(y_i) = r(x_i B^*)$	…. (1,1)

À quoi on doit ajouter la caractérisation de φ_* : $Var(y_i) = \varphi_* V(r(x_i B^*))$

Notons en particulier que, dans le cas d'une seule variable explicative, la fonction de lien fait que la droite de régression devient une courbe de régression et que pour tous ces modèles, à part le modèle gaussien, la variance augmente avec la moyenne (plus $\widehat{y_i}$ est grand, moins les points sont attirés par la courbe de régression).

2.2 Aspects pratiques.

Il y a priori beaucoup de choix à faire pour déterminer le modèle puisqu'il faut choisir la famille et la fonction de lien ; la table 1.2 décrit les fonctions de lien **g** usuelles. Voici quelques indications utiles pour le choix du modèle, elles se résument à dire que l'encadré (1.1) doit avoir un sens et à privilégier le lien canonique (proposé par défaut par les logiciels) :

- Le choix de la famille exponentielle : Dans l'écrasante majorité des cas le choix parmi les cinq familles présentées précédemment est quasiment déterminé par les valeurs prises par y (support de v).
  Si plusieurs choix sont possibles les tracés de résidus normalisés permettront souvent de décider du plus adéquat car les modèles proposent un comportement différent de la variance comme fonction de $u = r(xB)$
- La fonction de lien sera quant à elle guidée par les considérations suivantes

1. Le lien canonique $r = \acute{B}$ est un choix naturel et numériquement avantageux car beaucoup de formules se simplifient considérablement du fait que $\theta_i = x_i B$. Il est très généralement préféré si rien ne s'y oppose.
Ce choix peut introduire des distorsions (p.ex. $E(y_i) = e^{x_i B}$ *au lieu de* $y_i = x_i B$) qui peuvent être corrigées par des changements de variables sur x (p.ex. en passant au logarithme).
2. Interprétation de $E(y) = r(xB^*)$: Si r a un domaine de définition restreint (p.ex.$R_+ \; si \; R(u) = \frac{1}{u}$), il faut que ce domaine soit réaliste pour x_i De plus quand x_i varie, $r(x_i)$ doit prendre des valeurs raisonnables pour $E(y_i)$: rester borné si l'on a choisi une loi binomiale, rester positif si l'on a choisi une loi gamma, etc.

lien	$\eta = g(u)$	$u = r(\eta)$	Loi . Can	D_r	$r(D_r)$
Identité	u	η	Normale	R	R
logarithme	$\log(u)$	e^{η}	Poisson	R	R_+
probit	$\varphi_{-1}(u)$	$\varphi(\eta)$		R	[0 ,1]
puissance	$-u^a$	$(-\eta)^{1/a}$	Gamma	R_-	R_+

Table 1.2 – Les fonctions de lien usuelles. φ Désigne la fonction de répartition de la Gaussienne. Une colonne indique la loi pour laquelle le lien est canonique.

2.3 Modèle poissonnien

Dans notre exemple : On compte sur plusieurs années le nombre d'allocataires RSA par rapport au nombre de demandeurs d'emploi
yi = nombre d'allocataire RSA payé par le conseil général du Bas Rhin.
xi = nombre de demandeurs d'emploi
Le premier régresseur à 2 modalités et le deuxième régresseur en a 5. Le modèle naturel est Poissonnien, ce qui donne avec lien canonique

$$y \sim P(u), \log(u) = xB$$

2.4 Estimation de $\varphi_* \; et \; B^*$:

La consistance de $\hat{B}$ implique (sous certaines hypothèses) que :

$$\hat{\varphi} = \frac{1}{n}\sum_i V(\hat{u_i})^{-1}(y_i - \hat{u_i})^2$$

Est un estimateur consistant de φ_*

Un algorithme d'estimation de B^* : L'algorithme de Newton pour la maximisation de $L(B)$ est : $B_{new} = B - L''(B^{-1})L'(B^{-1})$.
Malheureusement la matrice de dérivée seconde est généralement difficile à calculer. On préfère la remplacer par l'approximation $-\widehat{\Gamma_n}$, d'où l'algorithme ; $B_{new} = B + (\widehat{X^T}\widehat{D}X)^{-1}(\widehat{X^T}D(\hat{y} - \hat{u})$ où tout est calculé avec la valeur courante de$B$.

3. Quelques Concepts de base des séries chronologiques

L'étude des séries chronologiques correspond à l'analyse statistique d'observation régulièrement espacée dans le temps, dans le but de représenter des phénomènes aléatoires qui évoluent dans le temps. Le modèle obtenu sera par la suite utilisé, selon les objectifs recherchés comme la prévision ou le contrôle. Sur ce fait, cette section est consacrée à une présentation des techniques d'analyse des séries chronologiques. on va d'abord donner quelques définitions de base qui permettront par la suite de présenter les différents modèles susceptibles de modéliser une série chronologique (AR, MA, ARMA, etc.)

On va étudier par la suite les caractéristiques statistiques en termes de stationnarité de ces séries.

3.1 Fonction d'auto-corrélation simple et partielle :

A/ Fonction d'auto-corrélation simple (FAC) :

La fonction d'auto corrélation simple d'un processus X_t, tϵ T de moyenne E(X_t), noté ρ_k ou ρ_k est définie par : $\rho(k) = \rho_k = \frac{\gamma(k)}{\gamma(0)}$ quel que soit k ϵT.

Elle mesure la corrélation de la série avec elle –même décalé de k périodes :

- $\boldsymbol{\rho(k)}\epsilon$ [-1,1] et $\rho(0)$ =1
- $\gamma(k)$ désigne la fonction d'auto-covariance telle que :
  $\gamma(k)$ =E [(X_t-m) (X_{t-k}-m)]
- $\boldsymbol{\rho(0)}$désigne la fonction de variance
- les fonctions de $\boldsymbol{\rho(k)}$et $\gamma(k)$ sont symétriques : $\boldsymbol{\rho(k) = \rho(-k)}$ et $\gamma(k) = \gamma(-k)$

Remarque : la représentation graphique de r(k) est appelée : corrélogrammes.

B/ Fonction d'auto-corrélation partielle (FAP) :

On peut définir la fonction d'auto-corrélation partielle (FAP) de retard k comme étant le coefficient de corrélation partielle entre Y_t et Y_{t-k}, c'est-à-dire comme étant la corrélation entre Y_t et Y_{t-1} l'influence des autres variables décalées de k périodes (Y_{t-1}, Y_{t-2}, Y_t ……Y_{t-k+1}) ayant été retirée.

3.2 Les opérateurs linéaires :

A/ L'opérateur de retard :

On aura souvent à considérer une variable en fonction de son passée, il est donc commode de définir un opérateur qui transforme une variable X_t en sa valeur passée. C'est l'opérateur retard désigné par la lettre B et tel que :

$B(X_t)= X_{t-1}$ et $B^K(k)= X_{t-k}$

B / L'opérateur d'avance (forward) :

$F(X_t)=X_{t+1}$; et $F^n(X_t) = X_{t+n}$

C / L'opérateur de différence ordinaire :

L'opérateur de différence ordinaire noté ∇ associé au processus (X_t ,tϵ T) tel que : $\forall\, t\epsilon\, T$; $\nabla X_t = X_t - X_{t-1} = X_t - (BX_t) = (1- B)X_t$, et par construction, nous obtiendrons l'opérateur de la d ème différence noté ∇^d tel que : $\nabla^d x_t = (1 - B)^d X_t$

3.3 Processus aléatoire stationnaire :

3.3.1 Processus stationnaire au sens strict

Définition : Xt, t$\in$ T est strictement ou fortement stationnaire

- VAR ($X_t < +\infty$); t$\in$ T (finie et indépendant de temps)
- $E(X_t) = m$, constante et indépendante de t, t$\in$ T
- $\forall$ (t, h) $\in T^2$, Cov $(X_t, X_{t+k})= E[(X_t - m)(X_{t+h} - m)] = \gamma(h)$ indépendant de t.

Ceci implique que la série ne comporte ni tendance, ni saisonnalité et plus généralement aucun facteur n'évolue avec le temps.

3.3.2 Processus faiblement stationnaire :

Définition : X_t stationnaire (au second ordre)

- $\forall t, E(X_t) = m$ (moyenne constante)
- $\forall t, s, cov(X_t, X_s) = \gamma(|t - s|)$

(Covariance symétrique, invariante par translation) En particulier $\forall$ t, $var(X_t) = \boldsymbol{\gamma}(0)$ (variance constante).

Propriété : de décomposition de Wold :

Un processus stationnaire (X_t) s'écrit "généralement" : $X_t = m + \sum_{j=0}^{+\infty} a_j\, \varepsilon_{t-j}$ avec $\sum_{j=0}^{+\infty} a_j^2 \leq +\infty$, et ε_t bruit blanc.

3.3.3 Le processus bruit blanc

Un processus bruit blanc ε_t t $\in Z$ est une suite de variables aléatoires non corrélées de moyenne nulle et de variance constante σ^2, il est donc caractérisé par la fonction d'auto covariance suivante :

$$E(\varepsilon_t, \varepsilon_{t-1}) = \begin{cases} \sigma^2 & \text{si } h = 0 \quad \forall\, t \in Z \\ \mathbf{0} & \textbf{si } h \neq 0 \end{cases}$$

On parle souvent de bruit blanc gaussien, il s'agit d'un bruit blanc dont la distribution marginale suit une loi normale.

3.3.4 Modèle autorégressif AR(p)

Ce sont les processus (X_t) du type :

$\forall X_t, X_t = \sum_{i=1}^{P} \emptyset i \, X_{t-i} + \varepsilon_t$

Ou $P \geq 1$, ($\emptyset i$) $1 \leq i \leq P$ réel et ε_t est un bruit blanc (gaussien).

Convention d'écriture :

$\varepsilon_t = \emptyset(B) \, Xt$

Ou $\emptyset(B) = 1 - \emptyset_1 B - \emptyset_2 B^2 - \emptyset_3 B^3 - \ldots\ldots \emptyset_p B^p$

$\emptyset_1, \emptyset_2 \ldots\ldots \emptyset_p$ Sont des paramètres réels indépendants de t

ε_t est une suite de variables aléatoires indépendantes et identiquement distribuées.

Et B est l'opérateur retard

3.3.5 Modèle moyenne mobile MA(p) :

Ce sont les processus (Xt) du type :

$\forall Xt, Xt = \varepsilon t - \sum_{i=1}^{q} \theta i \, Xt - i$

Ou q ≥ 1 , (θi) $1 \leq i \leq q$ réel, et ε_t est un bruit blanc (gaussien).

Convention d'´écriture :

$Xt = \theta(B) \, \varepsilon_t$

Ou $\theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \theta_3 B^3 \ldots\ldots\ldots\ldots - \theta_q B^q$, et B est l'opérateur retard : $B^k \, \varepsilon_t = \varepsilon_{t-k}$

ε_t Est une suite de variables aléatoires indépendantes et identiquement distribuées.

Un processus moyen mobile est par définition stationnaire, car il constitue une combinaison linéaire de bruit blanc

Caractérisation de l'ordre d'un MA(q) :

$$\forall h \geq 0, \rho(h) = \frac{cov(Xt, Xt+h)}{\sqrt{VAR(Xt)VAR(Xt+h)}} = \frac{\gamma(h)}{\gamma(0)}$$

Pour un processus MA(q) : (calcul simple)

$$\gamma(h) = \begin{cases} (1+\theta_1^2 + \theta_2^2 + \ldots\ldots\ \theta_q^2)\sigma^2 & \text{si } h = 0 \\ (\theta_h + \theta_{h+1}\,\theta_1 + \ldots\ldots\ldots + \theta_q\,\theta_{q-h})\,\sigma^2 & \text{si } 1 \leq h \leq q \\ 0 & \text{si } h > q \end{cases}$$

3.3.6 Les modèles ARMA (p, q) :

Définition : ces processus constituent une extension naturelle des processus AR et MA, des processus mixtes dans le sens où ils incorporent simultanément des composantes AR et MA ce qui permet d'obtenir une description plus parcimonieuse des données, donc ce sont les processus (Xt) du type :

$\forall t,\ X_t = \theta_0 + \sum_{i=1}^{P} \emptyset i\ X_{t-i} - \sum_{j=1}^{q} \theta j\ \varepsilon_{t-j} + \varepsilon_t$

Ou $p \geq 1, q \geq 1, (\emptyset i)\ 1 \leq i \leq P\ et\ (\theta i)\ 1 \leq i \leq q$ réel, et ε_t est un bruit blanc (gaussien).

Notation symbolique : $\emptyset(B)\ X_t = \theta_0 + \theta(B)\ \varepsilon_t$

Remarque : si $\emptyset$ a ses racines en dehors du cercle unité, alors on peut écrire :

$X_t = \sum_{j=0}^{\infty} a_j\ \varepsilon_{t-j}$ (sous forme AR(+∞))

3.4 Processus aléatoire non stationnaires :

Rappelons qu'un processus est stationnaire au second ordre si ses moments d'ordre 1 et 2 sont indépendants du temps. par opposition, un processus non stationnaire est un processus qui ne satisfait pas l'une de ou l'autre de ces deux conditions.

3.4.1 Le processus TS

Définition : ($x_t, t \in Z$) est un processus TS s'il peut s'écrire sous la forme $x_t = f(t) + z_t$

Où f(t) est une fonction du temps et z_t est un processus stochastique stationnaire

Dans ce cas, le processus X_t s'écrit comme la somme d'une fonction déterministe du temps et d'une composante stochastique stationnaire, éventuellement de type ARMA: Dès lors, il est évident que le processus ne satisfait plus la définition de la stationnarité du second ordre. En effet, on montre immédiatement que **E (x_t)= f(t)+z** où z= **E (z_t);** dépend du temps, ce qui viole la seconde condition de la définition d'un processus stationnaire.

3.4.2 Le processus DS

Définition : Un processus non stationnaire (X_t ; t∈Z) est un processus DS (Differency Stationary) d'ordre d; où d désigne l'ordre d'intégration, si le processus filtré défini par $(1- B)^d X_t$ est stationnaire. On dit aussi que (X_t ; t∈Z) est un processus intégré d'ordre d; noté I(d)

Si $d = 1$ on dit que le processus est au premier ordre. Et Il s'écrit :

$$(1-B)\, x_t = \beta+\varepsilon_t \Leftrightarrow x_t = x_{t-1}+\beta+\varepsilon_t$$

L'introduction de la constante β dans le processus DS permet de définir deux processus différents :

- $\beta = 0$ le processus DS est dit sans dérive. Il s'écrit : $x_t = x_{t-1}+\varepsilon_t$
- $\beta \neq 0$ le processus DS est dit avec dérive. Il s'écrit : $x_t = x_{t-1}+\beta+\varepsilon_t$

3.4.3 L'extension aux processus ARIMA

- si la série étudiée est de type TS, il convient de la stationnariser par régression sur le temps et le résidu d'estimation est alors étudié selon la méthodologie de Box Jenkins, le modèle est toujours dans ce cas un ARMA (p, q)
- si la série étudiée est de type DS , il convient de la stationnariser par passage aux différence selon l'ordre d'intégration I=d, la série est alors étudiée selon la méthodologie de Box Jenkins, qui permet de déterminer les ordre q , p, on note ce type de modèle ARIMA(p,d,q).

A) Définition : processus ARIMA (p, d, q)

Ce sont les processus (Xt) du type :

$\emptyset(B)(1 - B)^d X_t = \theta_0 + \theta(B)\, \varepsilon_t$ et (εt) est un bruit blanc

Où ∅ sans racine unitaire, p, d, q ≥ 0

$\emptyset(B)= 1- \sum_{i=1}^{q} \theta i\, Xt - i \;\; \forall\, i \leq p,\; \theta i \in R \; et \;\; \theta p \in R^*$

$\theta(B) = \sum_{i=0}^{q} \theta_i B^i \quad \forall\, i \leq p,\; \theta i \in R \; et \;\; \theta p \in R^*$

Remarque : processus ARIMA non stationnaires, mais d fixé tel que $(1 - B)^d X_t$ stationnaire.

B) Dérivation des processus :

ARIMA : $\emptyset(B)(1 - B)^d Xt = \theta_0 + \theta(B)\, \varepsilon t$

Intérêt 1 : permet de traiter les chroniques avec tendance

- $X_t = a_t + b + u_t$ alors ;

$(1 - B)X_t = a + (1 - B)u_t$ est stationnaire

- $X_t = a_t^2 + b_t + c + u_t : (1 - B)^2$ X_t est stationnaire

Intérêt 2: permet de stabiliser la variance.

- $X_t = X_{t-1} + u_t$ (marche aléatoire), alors ;

$(1 - B)Xt = u_t$ est stationnaire

- Dans les deux cas : racine unité dans la partie AR.
- tests statistiques envisageables (Dickey-Fuller).

3.5 Modélisation des séries chronologiques unies variées

3.5.1 Identification du modèle ARIMA (p, d, q)

0. Transformation de la chronique (log, exp, √, Box-Cox) pour stabiliser la variance.
1. Identification de d.
Indices de non-stationnarité: ça se voit sur le graphique (tendance, saisonnalité), ACF ne décroit pas assez vite et ou périodique (période τ).

Si X_t déjà stationnaire, alors $Y_t = X_t - X_{t-1}$ à un premier pic de l'ACF proche de −1.

2. Identification de p et q. pour cela il s'agit de l'Utilisation des corrélogrammes.

Il existe cependant des méthodes d'identifications automatiques, basées sur les critères d'informations.

*/Critères d'informations :

Il existe des critères d'information qui mesurent l'écart entre la vraie loi inconnu et celle du modèle proposé ; **critère d'Akaike** 1970 appelés aussi AIC tel que : AIC= $\log \sigma^2 + \frac{2(p+q)}{N}$

3.5.2 Estimation des paramètres du modèle :

Estimation des paramètres θ_j et φ_i : $\forall t$, $Xt = \sum_{i=1}^{p} \emptyset_i X_{t-i} + \varepsilon_t - \sum_{j=1}^{q} \theta_j X_{t-j}$

Hypotheses: (εt) i.i.d. $\sim N(0, \sigma^2)$

Vraisemblance des T observations Xt :

$$L[(Xt)|(\varphi_i),(\theta_j),\sigma] = \frac{1}{(2\pi\sigma^2)^{T/2}\sqrt{DET\Omega}} \exp\left(-\frac{1}{2\sigma^2} X'\Omega^{-1}X\right)$$

où Ω dépend des φ_i et θ_j et $X = (X_t)_{1 \le t \le T}$

(Idée : plus la vraisemblance est grande, plus il est probable que le modèle pour les (X_t) est bon)

→ Estimation des paramètres $(\varphi_i),(\theta_j),\sigma$ par la méthode (numérique) du maximum de vraisemblance

3.5.3 Validation :

Examen de la validité du modèle

1. Significativité des paramètres estimés (Student)

2. Résidus = bruit blanc ?; Résidus = écart entre observation X_t et prévision $\hat{X}_{t-1}$ (1)

→ Vérifier que les résidus (et) sont bien des réalisations d'un bruit blanc (εt).

- vérification graphique : moyenne nulle, variance constante, pas de "structure".
- ACF et PACF ne doivent pas avoir de pics significatifs.
- Portmanteau Test (Box-Pierce) :

Soit ρ(h) l'ACF des résidus. Alors : $(T - K)\sum_{h=1}^{k} \rho(h)^2$ suit un χ^2 à K – p – q

d.d.l., sous hypothèse (εt) bruit blanc gaussien.

***/ Test de Ljung et Box : la définition est dans la partie pratique**

***/ Test de normalité :** Il convient de vérifier la normalité des résidus, ceci est possible par un test de Jarque et Bera ou de Shapiro Wilk.

Sélection de modèle : comment choisir entre différents modèles ARIMA(p,d,q)

Plusieurs possibilités :

- Minimisation des critères habituels : RMSE, MAE.
- Minimisation de critères d'information :
  AIC $= -2\log(L)+2(p+q)$
  ou BIC = SBC $= -2\log(L)+(p+q)\log(T)$.

(dans les deux cas, compromis entre vraisemblance et nombre de paramètres : prévision vs "sur apprentissage".)

- Critères basés sur le pouvoir prédictif...

Remarque : le choix entre déférents modèles se fait selon un critère.

3.5.4 Prévision :

$\Phi(B)(1-B)^d X_t = \Theta(B)\varepsilon_t$

On connait (X_t) jusque la date t = T, on cherche $\hat{X}_{T+h}$.

On écrit X_{T+h} sous la forme :

$$X_{T+h} = \sum_{i=1}^{p+q} \Psi_i X_{T+h-I} + \varepsilon_{T+h} - \sum_{j=1}^{q} \theta_j \varepsilon_{T+h-i} \qquad (*)$$

Soit $\forall\, h \leq 1$, $\hat{X}_{T+h} = E(X_{T+h} / (X_t)_{t \leq T})$ alors,

$\widehat{X}_T(h) = \sum_{i=1}^{p+q} \Psi_i \, X_{T+h-i} - \sum_{j=1}^{q} \theta_j \, \varepsilon_{T+h-I}$ (∗∗)

car $E(\varepsilon_{T+h-j} / (X_t)_{t \leq T}) = \varepsilon_{T+h-j}$ si $j \geq h$ et 0 sinon

$\varepsilon_{T+h-j} \in vect$ $(X_t)_{t \leq T}$ si $j \geq h$ et (ε_t) non corrélés).

→ formule d'actualisation utilisant l'estimation des Ψ_i , θ_j , ε_t

Remarque : avec (∗) et (∗∗), $X_{T+1} - \hat{X}_T(h) = \varepsilon_{T+1}$.

Partie pratique

La démarche suivie

1/ La première étape consiste à trouver un bon modèle qui peut prendre en compte la relation qui existe entre les bénéficiaires RSA et les catégories des demandeurs d'emploi, pour cela on a utilisé les modèles linéaires généralisés vu qu'il s'agit d'une mesure de comptage.

Dans l'état actuel d'accès aux données, on a pris en compte la variable des demandeurs d'emploi (Catégories, A, B, C, D, E), le recours à la séparation de cette variables en sous 5 autres variables explicatives du nombre d'allocataires RSA socle et RSA socle et activité est due à la différence qui existe entre les observations , celles-ci agissent différemment par rapport à notre variable d'intérêt (le nombre de bénéficiaire RSA payée par le conseil général du Bas Rhin) , de plus , il faut ajouter que les demandeurs d'allocation RSA socle représente plus de 60 % de l'ensemble des demandeurs d'allocation RSA , idem , pour les variables explicatives ou les demandeurs d'emploi Catégorie A qui représente plus de 65 % de l'ensemble des demandeurs d'emploi inscris à Pole emploi.

Cette étape consiste à modéliser nos variables d'intérêts (RSA socle et RSA socle + activité), faire des tests statistiques (étude de significativité des estimateurs, de colinéarité entre les variables, vérification des hypothèses sur les résidus, et enfin la sélection du meilleur modèle)

2/
Après avoir étudié les différentes relations qui lient nos variables d'intérêts, aux autres variables explicatives, la deuxième étape consiste à prédire l'évolution de ces différentes variables explicatives retenues dans les modèles finales.
Pour cela on a utilisé deux autres méthodes statistiques liées aux séries chronologiques;

- La méthode de lissage exponentiel

Les méthodes de lissage exponentiel sont un compromis des méthodes d'extrapolations (qui consistent à prolonger l'évolution passée), puisqu'elles tiennent compte de toutes les observations, mais en diminuant leur importance au fur et à mesure que l'on remonte dans le passé.

- La méthodologie de Box et Jenkins

Dans les séries temporelles, cette méthode est appliquée à des processus autorégressif et moyen mobile (modèle ARMA ou ARIMA), dans le but de trouver une meilleur estimation d'une série (CatA, CatC) à partir des valeurs précédentes et ce pour faire des prévisions.

A noter que ;

- Dans la première étape de modélisation, nous avons travaillés sur une période de 39 mois (de décembre 2009 jusqu'à février 2013)

- Dans la deuxième étape de prévision des séries retenues des demandeurs d'emploi, nous avons travaillés sur des données mensuelles (de janvier 1996 jusqu' à février 2013).
- Les projections calculées sont basées sur les données non consolidées du RSA.

4. Régression linéaire généralisé

Sur ce qui suit, les résultats concernent le nombre d'allocataires RSA socle, concernant la modélisation du RSA socle et activité, les résultats et le modèle final sont donnés dans l'annexe 2.

Comme on l'a expliqué précédemment, afin d'étudier l'évolution du nombre d'allocataire RSA, il faut analyser et connaître les différentes variables qui ont un impact sur le RSA, le nombre des demandeurs d'emploi avec les différentes catégories à un impact considérable sur l'évolution du nombre des demandeurs d'allocation RSA.

Pour étudier cette relation, on utilise la régression linéaire généralisée, avec une fonction lien de type « Poisson » puisque il s'agit de mesure de comptage,

Le modèle initial s'écrit de la façon suivante :

$RSA\_S \rightarrow \mathcal{P}(\mu)$, $\quad Log(\mu) = \alpha + \beta_1 * CatA_i + \beta_2 * CatB_i + \beta_3 * CatC_i + \beta_4 * CatD_i + \beta_5 * CatE_i$

Ou encore ;

$$\mathbb{E}(RSA\_S_i) = e^{\alpha + \beta 1 * CatAi + \beta 2 * CatBi + \beta 3 * CatCi + \beta 4 * CatDi + \beta 5 * CatEi}$$

Tel que :

RSA_S : variable expliquée qui représente le nombre des demandeurs d'allocation RSA socle,

$CatA_i$: variable explicative qui représente le nombre des demandeurs d'emploi catégorie A dans l'année i,

CatB : variable explicative qui représente le nombre des demandeurs d'emploi catégorie B dans l'année i,

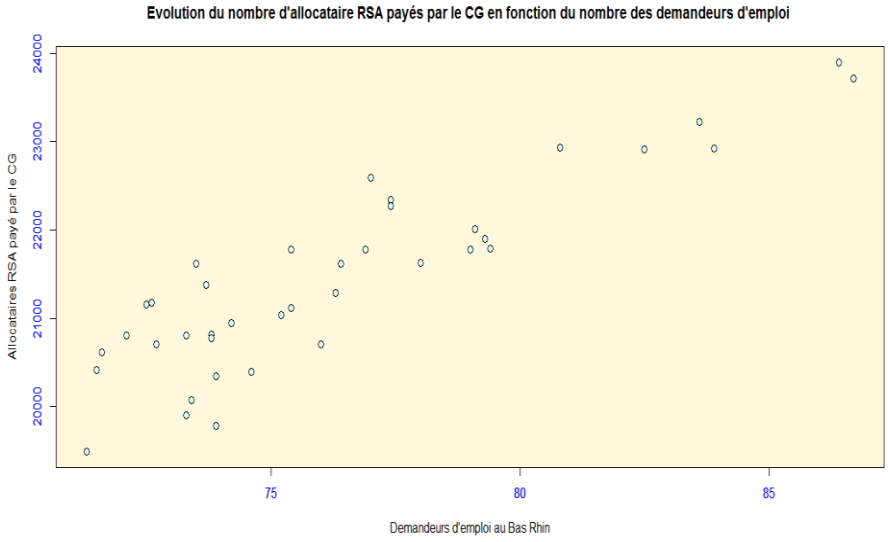
CatC : variable explicative qui représente le nombre des demandeurs d'emploi catégorie C dans l'année i,

CatD : variable explicative qui représente le nombre des demandeurs d'emploi catégorie D dans l'année i,

CatE : variable explicative qui représente le nombre des demandeurs d'emploi catégorie E dans l'année i,

4.1 Relation entre le nombre d'allocataire RSA payé par le conseil général du Bas Rhin et les demandeurs d'emploi :

Figure n°5 : évolution du nombre d'allocataires RSA payer par le CG67 par rapport au nombre des demandeurs d'emploi



- Il existe une relation linéaire entre le nombre d'allocataires RSA payés par le conseil général du Bas Rhin et les demandeurs d'emploi inscrits à pôle emploi.
- Le coefficient de corrélation entre ces deux variables est de 90.75, le nombre des allocataires RSA payés par le CG est fortement corrélé avec le nombre des demandeurs d'emploi.

4.2 Tableau des estimateurs :

	Estimate	Std.	Error	z value	Pr(>|z|)
(Intercept)	9.133039	0.035681	255.967	< 2e-16	***
CatA	0.007045	0.001092	6.451	1.11E-10	***
CatB	0.008644	0.006315	1.369	0.1711	
CatC	0.01337	0.00249	5.37	7.87E-08	***
CatE	0.021563	0.008104	2.661	0.0078	**
CatD	0.00284	0.002322	1.223	0.2213	

La p value des paramètres des variables CatA, CatC,CatE et de la canstante sont inférieurs à 0 donc ils sont significativement différents de 0 .

$$E(RSA\_S) = e^{9.133+0.007*CatAi+0.008*CatBi+0.013*CatCi+0.021*CatDi+0.02*CatEi} \quad \dots (1)$$

4.2 Diagnostics du modèle linéaire généralisé :

4.2.1 Résidus et graphiques des résidus :

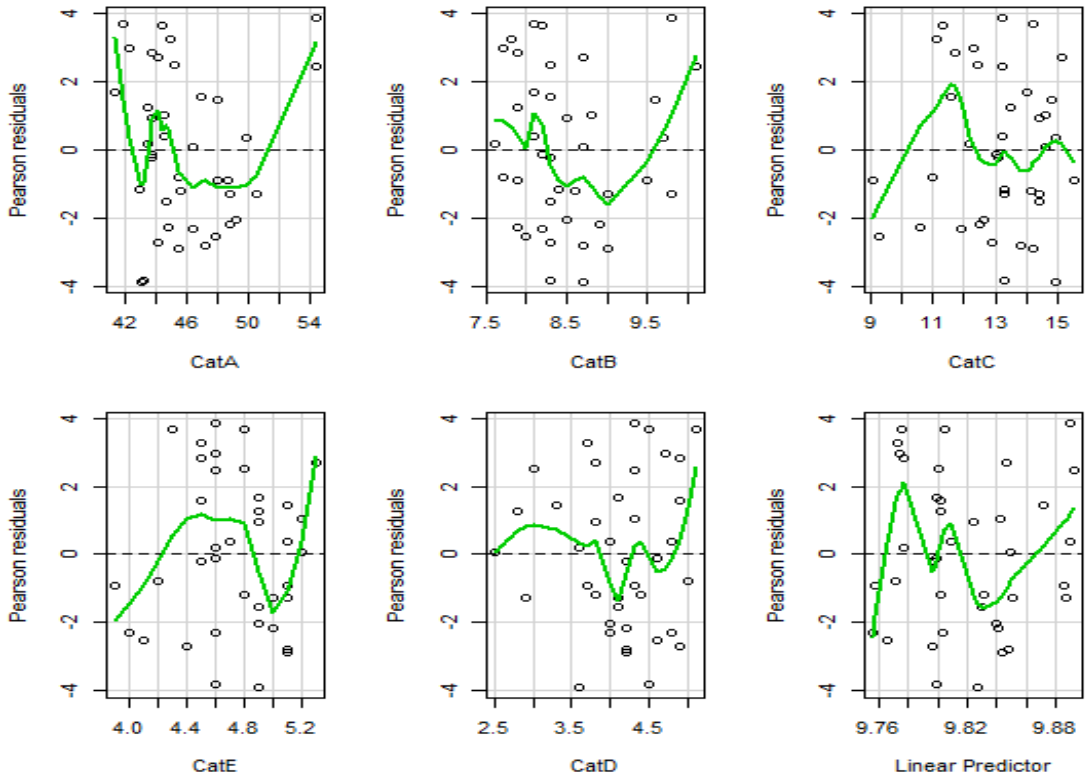
La différence entre les diagnostics du modèle linéaire et ceux du modèle linéaire généralisée réside dans la définition des résidus, dans le modèle linéaire, le résidu est la différence entre « $y - \hat{y}$ » et dans le modèle linéaire généralisé, le résidu est « $\varepsilon = \mathrm{E}\,(\mathrm{y}/\eta) - y$».

Les résidus de Pearson :

Ces résidus sont calculés par la formule suivante :

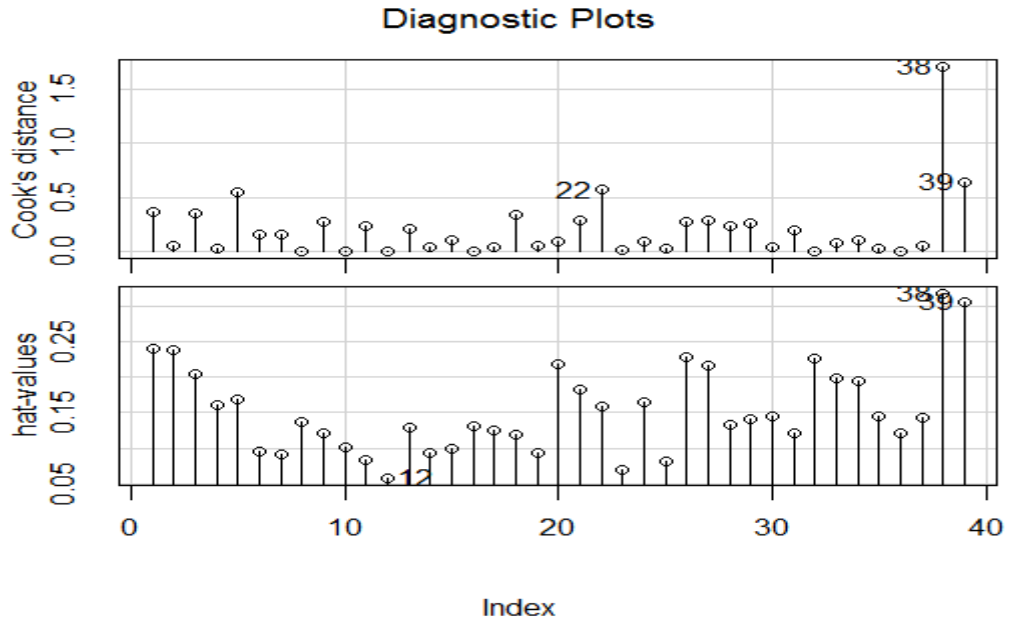
« $\varepsilon_{PSi} = \frac{y_i - \hat{u}}{S\sqrt{(1-h_i)}}$»

Pour calculer les résidus de Person nous devons définir les hat-values hi du glm.



Sur les graphiques ci-dessus, on constate que les résidus par rapport aux variables explicatives sont répartis de manière homogène.

4.2.2 Influence des observations :



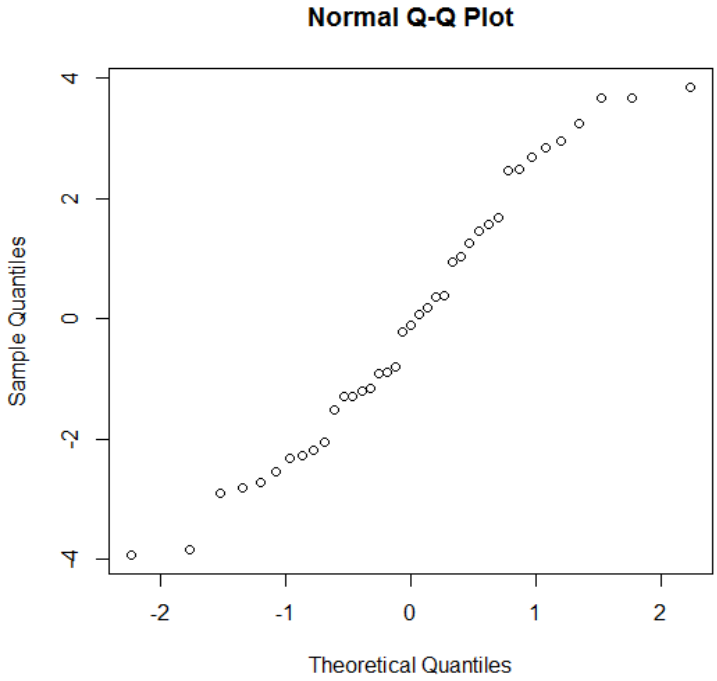
Sur les graphiques ci-dessus, on constate l'existence de trois observations influentes, on supprime ces observations, puis on refait la modélisation par la régression linéaire généralisée, pour comparer les deux modèles ; nous obtiendrons le tableau suivant :

	Est.1	SE.1	Est.2	SE.2
(Intercept)	9.13E+00	3.57E-02	9.26E+00	3.90E-02
CatA	7.04E-03	1.09E-03	2.56E-03	1.23E-03
CatB	8.64E-03	6.32E-03	1.68E-02	6.52E-03
CatC	1.34E-02	2.49E-03	7.55E-03	2.67E-03
CatD	2.84E-03	2.32E-03	9.28E-05	2.37E-03
CatE	2.16E-02	8.10E-03	4.14E-02	9.15E-03

Les estimateurs ont changés de valeurs, la formule obtenue en supprimant les trois valeurs s'écrit :

$$E(RSA\_S) = e^{9.26+2.56*CatAi+1.68*CatBi+2.56*CatCi+9.28*CatDi+9.15*CatE} \quad \ldots (2)$$

4.2.3 La normalité des résidus :



Le tracée QQ plot montre un assez bon ajustement à la loi normale, le test de Schapiro-Wilk peut confirmer cette hypothèse

>shapiro.test(resd1)

Shapiro-Wilk normality test

data: resd1

W = 0.9591, p-value = 0.1669

La p-valeur (p-value =0.1669) étant strictement supérieure à a=5%, le test est significatif , nous décidons de ne pas rejeter H0 et donc d'accepter l'hypothèse de normalité des résidus.

4.2.4 Test de non corrélation des résidus

Le test de Durbin-Watson permet de détecter une autocorrélation de la forme :

$\varepsilon_{i+1} = \rho\varepsilon_i + \eta_i, \quad \eta_i \sim \mathbb{N}(0, \sigma_\eta)$

On calcule le coefficient de Durbin-Watson à partir des résidus $\varepsilon_i = \mathbb{E}(y/\eta) - y$

$DW = \frac{\sum_i (\varepsilon_{i+1} - \varepsilon_i)^2}{\sum_i \varepsilon_i^2}$

On notant : $\rho=\left(\frac{\sum_i(\varepsilon_{i+1}*\varepsilon_i}{\sum_i \varepsilon_i^2}\right)$ si les résidus forment un processus autorégressif d'ordre 1, c'est-à-dire suivent le modèle $\varepsilon_{i+1} = \rho\varepsilon_i + \eta_i$, alors DW vaut à peu près 2(1-ρ).

Ou, DW ≅ $2\left(\frac{\sum_i(\varepsilon_{i+1}*\varepsilon_i}{\sum_i \varepsilon_i^2}\right)$

Liens entre les valeurs ρ et DW:

Si 0<ρ < 1 ⇒ DW compris entre 0 et 2

Si 0< ρ > -1 ⇒DW compris entre 2 et 4

S'il n'y a pas d'auto-corrélation d'ordre 1 ⇔ρ proche de 0, donc DW proche de 2.

```
>dwtest(mod1)
```

```
        Durbin-Watson test
             data:  mod1

DW = 1.5929, p-value = 0.03488
```

la valeurs de Durbin Watson est proche de 2 , on conclut donc que les résidus sont non corrélés.

4.2.5 La colinéarité dans le modèle linéaire:

```
> vif(mod1)
```

CatA	CatB	CatC	CatD	CatE
7.945360	11.629551	10.308702	1.458575	5.397707

La valeur VIF (ou la tolérance, soit l'inverse du VIF (1/VIF)) permet de vérifier la prémisse de multi colinéarité. Nous cherchons à obtenir une valeur VIF près de 1. Si elle est de 10, c'est problématique. Conséquemment, si la tolérance est équivalente à 10, il y a un problème sérieux. Probablement que les corrélations entre 2 variables prédicatrices ou plus sont trop élevées.

En regardant les proportions de la variance, on constate que les 2 variables « CatA », et la « CatB », ont des VIF supérieurs à 10 , ce qui veut dire qu'elles ont un problème de multi colinéarité, il convient donc d'éliminer l'un des deux paramètres.

4.2.6 Tableau de corrélation :

	CatA	CatB	CatC	CatD	CatE	RSA_S
CatA	1					
CatB	0.74491679	1				
CatC	-0.0210940	0.5770209	1			
CatD	-0.0865213	-0.2340353	-0.4000759	1		
CatE	0.05923367	0.5346073	0.87917155	-0.5233555	1	
RSA_S	0.6096264	0.8726043	0.67449031	-0.3133781	0.66555806	1

La variable CatA est fortement corrélée avec la variable CatB, il en est de même pour la variable CatE qui est fortement corrélée avec la variable CatC, il est important de prendre tous ces corrélations en considération dans la procédure de sélection du bon modèle.

4.3 Critère de choix de modèles

L'objet de ces critères de choix est de comparer des modèles entre eux.

Par définition l'AIC (Akaike Informative Criterion) pour un modèle à p paramètres est défini par

$$\mathbf{AIC = -2L + 2}\ p$$

La philosophie est simple : plus la vraisemblance est grande, plus grande est donc la log-vraisemblance L et meilleur est le modèle. Cependant si l'on met le nombre maximum de paramètre (ce qui est le modèle saturé) alors, L sera maximum. Il suffit donc de rajouter des paramètres pour la faire augmenter. Pour obtenir un modèle de taille raisonnable il sera donc bon de la pénaliser par une fonction du nombre de paramètre, ici **2** p.

Un autre critère de choix de modèle le BIC (Bayesian Informative Criterion) pour un modèle à p paramètres estimé sur n observations est défini par

$$\mathbf{BIC = 2}p + p\ \mathbf{log(n)}$$

L'utilisation de ces critères est simple. Pour chaque modèle concurrent le critère de choix de modèle est calculé et le modèle qui présente le plus faible est sélectionné

4.4 Sélection du modèle :

Nous avons utilisés la sélection automatique à l'aide de la fonction « **bestglm** », présente dans la librairie « **bestglm** », le modèle sélectionné par cette méthode est le suivant :

```
> modelF
BIC
BICq equivalent for q in (0.417828197229577, 0.628808555710825)
Best Model:
              Estimate   Std. Error    z value      Pr(>|z|)
(Intercept) 9.146762314 0.0245911706 371.953108  0.000000e+00
CatA        0.008418324 0.0003872785  21.737129 9.144290e-105
CatC        0.016255065 0.0016426998   9.895335  4.361468e-23
CatE        0.015402622 0.0074041012   2.080283  3.749963e-02
```

Les différentes méthodes qu'on a essayé que soit la sélection descendante , la sélection ascendante , et progressive conserve au moins quatre variables explicatives et ceux malgré les colinéarités entre ces variables et la forte corrélation

entre elles , le modèle finale ainsi retenu après l'étude des différents diagnostics sur le modèle linéaire généralisé est le suivant :

	Estimate	Std.Error	z value	Pr(>|z|)	Pr(>|z|)
(Intercept)	9.1749643	0.0205072	447.4	<2e-16	***
CatA	0.0085381	0.0003827	22.31	<2e-16	***
CatC	0.0192592	0.000783	24.6	<2e-16	***

Tous les coefficients des estimateurs sont significatifs, ainsi le modèle s'écrit mathématiquement de la manière suivante :

$$E(RSA\_S) = e^{9.1749643 + 0.0085381 * CatAi + 0.0192592 * CatCi} \quad \dots (3)$$

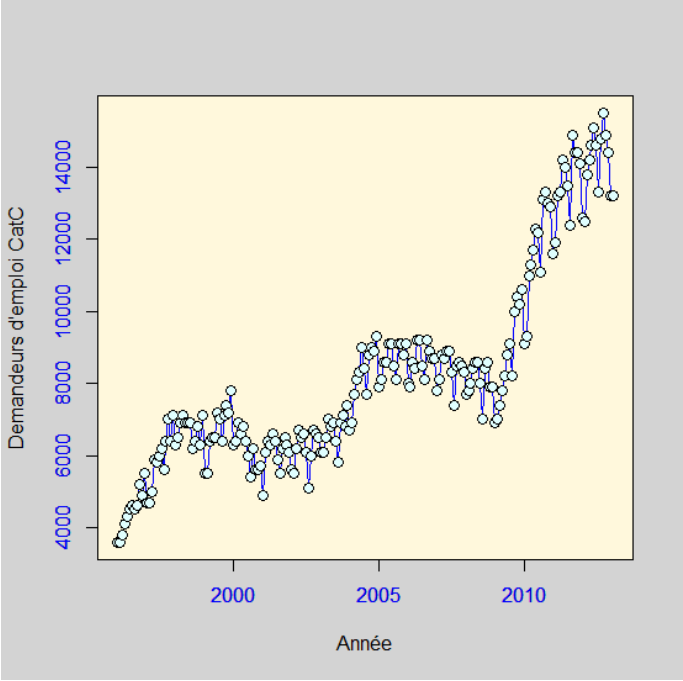
5. Application de la méthodologie de box Jenkins pour la série CatC

La méthodologie de Box & Jenkins vise à formuler un modèle permettant de représenter une chronique avec comme finalité de prévoir des valeurs futures. De ce fait, l'objet de cette méthodologie est de modéliser une série temporelle en fonction de ses valeurs passées et présentes afin de déterminer le processus ARIMA adéquat Cette méthodologie suggère une procédure à trois étapes :

- Identification du modèle
- Estimation du modèle
- Validation du modèle (Test de diagnostique)

Pour tout étude économétrique, à long terme où à court terme, la stationnarité des séries est nécessaire. Pour cela, on va étudier et identifier l‟ordre d‟intégration pour chacune des séries présentées en appliquant les tests DF et ADF.

Evaluation graphique :



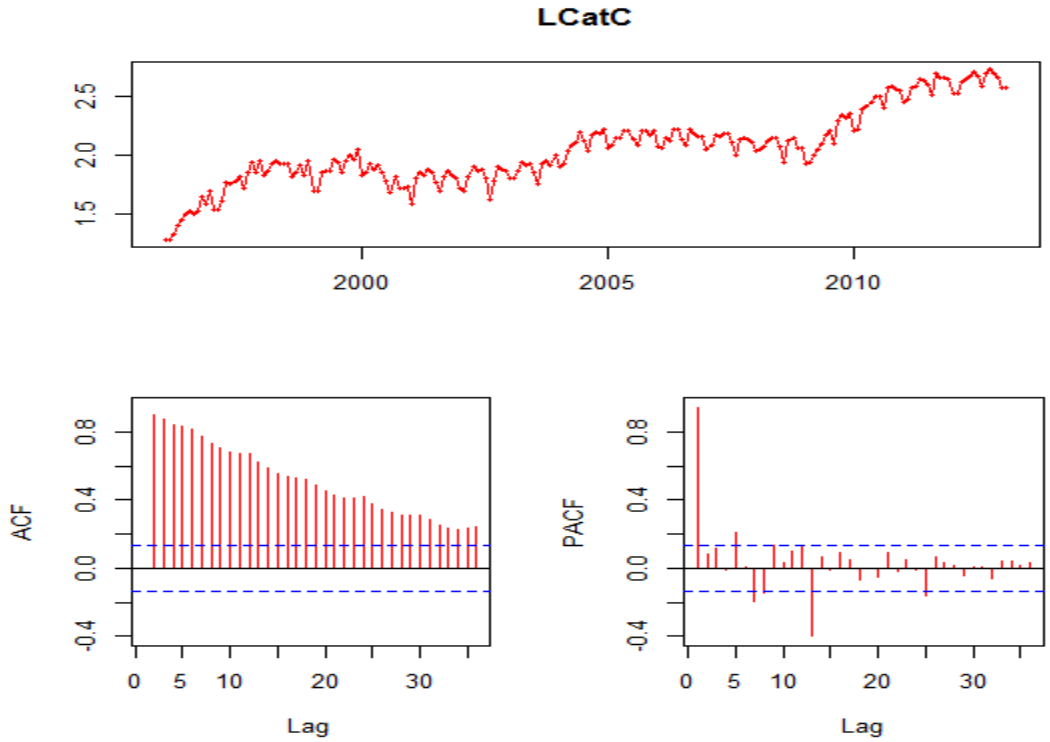
Le graphique de la série CatC (nombre de demandeurs d'emploi catégorie C) fait ressortir une tendance globale à la hausse avec une saisonnalité, Il semble donc que la série soit non stationnaire.

Série transformée :

Afin de stabiliser la série, on lui applique une transformation logarithmique qui offre les avantages suivants :

- Minimise l'influence des effets du temps sur la série.
- Réduire le nombre d'étapes pour aboutir à une série stationnaire.
- Permet de ne pas perdre l'information sur les premières valeurs de la série.

5.1 Etude de la stationnarité de la série LCatC :



De plus l'analyse visuelle du plot montre la présence d'une tendance.

L'autocorrélogramme de la série LCatC permet de se rendre compte rapidement que la série n'est pas stationnaire

La fonction d'auto corrélation partielle : il y a des pics significatifs, correspond au retard N°1, 5,7 et le pic N°12 ce qui implique que cette série est affectée par une saisonnalité d'ordre 1.

La fonction d'auto corrélation : ces termes décroissent lentement vers 0, ce qui implique qu'il y a l'effet de la tendance, donc la série (LCat) est apriori non stationnaire

Le Test KPSS (Kwiatkowski, Phillips, Schmidt,Shin)

L'hypothèse nulle de ce test est celle de la stationnarité (autour d'une constante ou d'une tendance déterministe linéaire).

- Cas 1 : $y_t = r_t + \varepsilon_t$ avec ε_t ~ I(0) et $r_t = r_{t-1} + u_t$ et u_t ~BB(0, σ_u^2)
- Cas 2 : $y_t = Bt + r_t + \varepsilon_t$ avec ε_t ~ I(0) et $r_t = r_{t-1} + u_t$ et u_t ~BB(0, σ_u^2)

L'hypothèse H_0 de stationnarité peut alors être formulée sous la forme :

H_0 :$\sigma_u^2 = 0$. Elle correspond aux deux cas suivants:

- Cas 1 : $y_t = r_0 + \varepsilon_t$ avec ε_t ~ I(0)
- Cas 2 : $y_t = Bt + r_0 + \varepsilon_t$ avec ε_t ~ I(0)

La statistique de test utilisée correspond à la statistique du test du score lorsque les ε_t sont i.i.d de loi $\mathbb{N}(0, \sigma_u^2)$.

Cependant, elle est corrigée de façon à tenir compte de l'autocorrélation des ε_t dans le cas général.

La procédure employée est alors la suivante:

- on régresse y_t sur une constante (cas 1) ou sur une constante et un trend (cas 2) et on calcule les résidus $\widehat{u_t}$ de la régression ($\widehat{u_t} = y_t + \bar{y}$) dans le cas 1, $\widehat{u_t} = \hat{a} + \widehat{B_t}$ dans le cas 2);
- on calcule :

$$S_t = \sum_{k=1}^{\widehat{t}} \widehat{u_k}$$

Et

$$w_{TK}^2 = \widehat{\gamma_u(0)} + 2\sum_{k=1}^{K}(1 - \frac{k}{K+1}) * \widehat{\gamma_u(k)}$$

Avec, comme précédemment : ∀ k≥0, $\widehat{\gamma_u(k)} = \frac{1}{T}\sum_{t=k+1}^{T} \widehat{u_t} * \widehat{u_{t-k}}$ et k de l'ordre de $\sqrt{T}$.

- la statistique de test est $\eta = \frac{\frac{1}{T}\sum_{t=1}^{T} S_t^2}{w_{TK}^2}$

La loi limite de η est tabulée dans le cas 1 (η_u dans la table) et dans le cas 2 (η_t Dans la table).

On refuse $H_0 : \sigma_u^2 = 0$ au seuil a lorsque la valeur obtenue de η est supérieure à la valeur critique correspondante.

Pour confirmer la non stationnarité de la série, on applique ce test à la série LCatC :

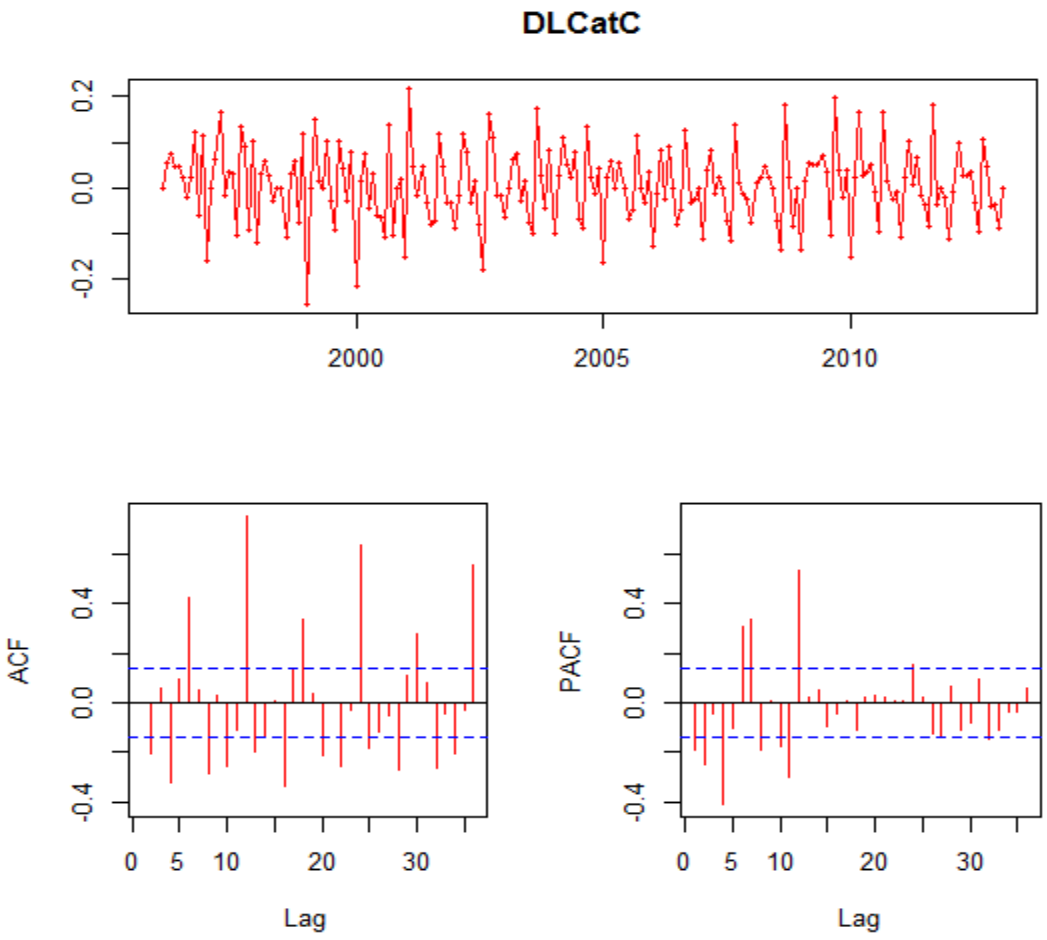
```
>kpss.test(LCatC)
```

```
        KPSS Test for Level Stationarity
data:  LCatC
KPSS Level = 4.1991, Truncation lag parameter = 3, p-value = 0.01
Message d'avis :
In kpss.test(LCatC) : p-value smaller than printed p-value
```

la p-valeur est inférieure à 0.05, donc on rejette l'hypothèse H_0 de la stationnarité de la série, il convient de différencié la série pour la rendre stationnaire.

5.2 Etude de la stationnarité de la série DLCatC :

$$\boldsymbol{DLCatC = LCatC_t - LCatC_{t-1}}$$



Sur le graphique ci-dessus, la série semble stationnaire, pour les autocorrélogramme de la série ;

La fonction d'auto corrélation partielle : il y a des pics significatifs, on remarque aussi que le retard N°12 est significatif ce qui implique que cette série reste toujours affectée par une saisonnalité d'ordre 1, (on applique Un lissage exponentiel par ratio de moyenne mobile avec une approche additive pour pré-blanchire la série).

La fonction d'auto corrélation : présence de pic significatifs.

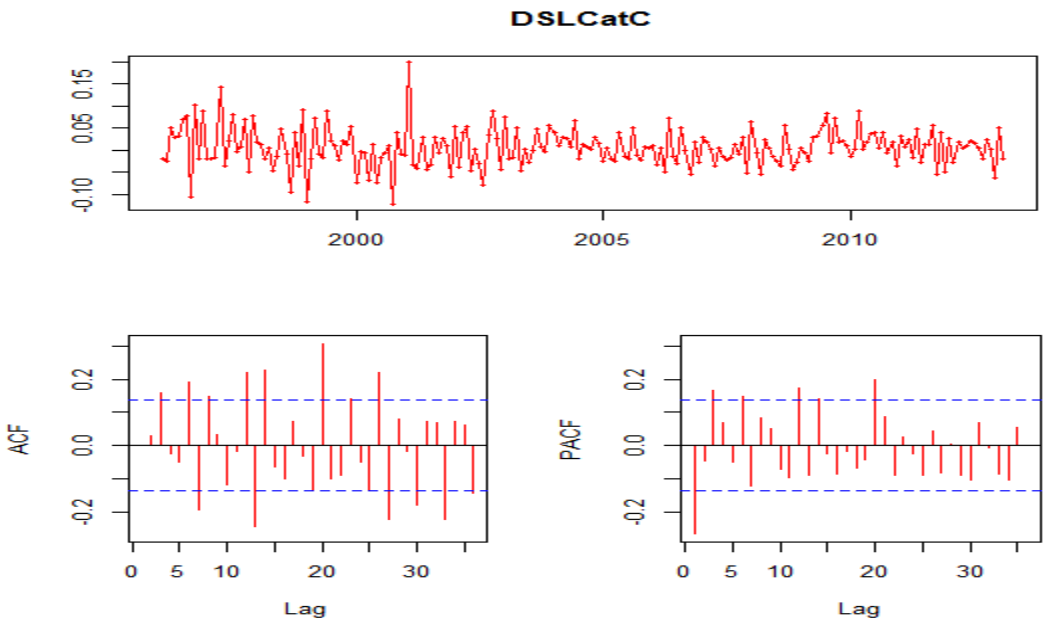
Test de stationnarité de la série DLCatC :

```
>kpss.test(DSLCatC)
```

```
        KPSS Test for Level Stationarity
data:  DSLCatC
KPSS Level = 0.1019, Truncation lag parameter = 3, p-value = 0.1
Message d'avis :
In kpss.test(DSLCatC) : p-value greater than printed p-value
```

La p- valeur est egale à 0.1 supérieure à 0.05 , donc on ne rejette pas l'hypothèse H_0, on accepte l'hypothèse de la stationnarité de la série DSLCatC.

La série Désaisonnalisée et différenciée $DSLCatC = DSLCatC_t - DSLCat_{t-1}$ a l'allure suivante,



La série ainsi formée semble stationnaire. A titre comparatif, la série obtenue en différenciant 2 fois donne des résultats ne semblant pas significativement différents, la même chose pour la série différenciée à l'ordre 12.

5.3 Vérification des différentes hypothèses :

5.3.1 Modélisation de la série désaisonnalisée :

Compte tenu de l'allure des Autocorrélogramme de D12SLCatE, nous pouvons penser modéliser la série X_t par un processus ARMA (p; q)

>each(DSLCatC)

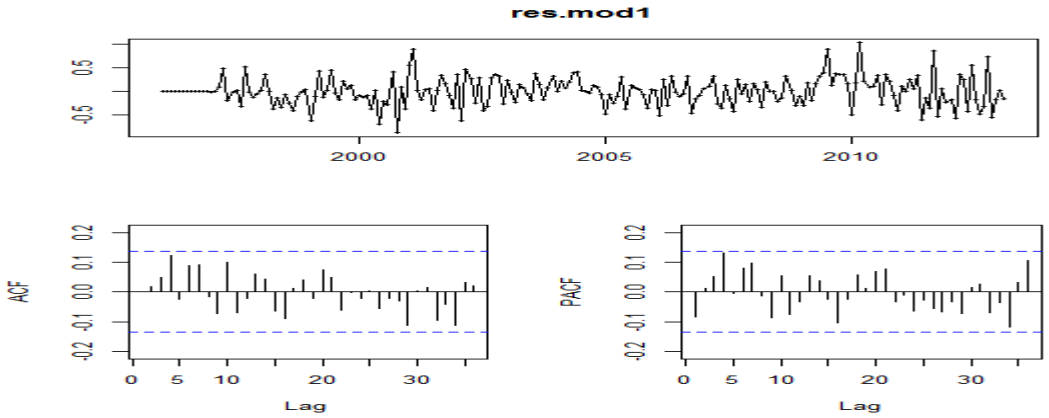
```
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x o x o o x x x o o o  x  x  x
1 x o x o o o o x o o o  o  o  x
2 x x o o o o o o o o o  o  o  o
3 x x o x o o o o o o o  o  o  o
4 x o x o o o o o o o o  o  o  o
5 x x x o o o o o o o o  o  o  o
6 x o x o o o o o o o o  o  o  o
7 x o x o o o o o o o o  o  o  o
```

a- Estimation des paramètres d'une modélisation ARIMA(2,1,2)(0,1,1)[12]

L'estimation donne les résultats suivants :

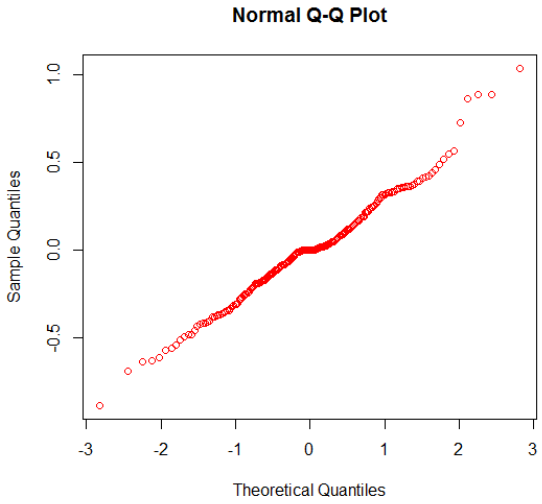
>mod1

```
Series: x
ARIMA(2,1,2)(0,1,1)[12]

Coefficients:
          ar1      ar2     ma1     ma2     sma1
      -1.1545  -0.9994  1.1390  0.9996  -0.5685
s.e.   0.0066   0.0019  0.0290  0.0297   0.0637

sigma^2 estimated as 0.09757:  log likelihood=-55.08
AIC=120.15   AICc=120.6   BIC=139.73
```



D'après les autocorrélogramme de la fonction d'autocorrélation et de celle d'autocorrélation partiel, on remarque que tous les pics sont significatifs. Le test KPSS sur la stationnarité des résidus est accepté, reste à tester la normalité, la blancheur et l'Hétéroscédasticité des résidus.

a.1 La normalité des résidus :



Bien que le test de Shapiro Wilk et Kolmogorov-Smirnov rejette l'hypothèse de la normalité des résidus, l'analyse visuelle du tracée QQ plot montre un assez bon ajustement à la loi normale

a.2 les statistiques de Portmanteau

La statistique classique du test de Portmanteau est proposée par Box et Pierce

$$Q_{BP} = n\sum_{j=1}^{m}\widehat{\rho}_j^2$$

Cette statistique est étudiée puis développée en d'autres statistiques (2.1) et (2.2)

a.2.1 Test d'autocorrélation des résidus (*Test* de *Ljung-Box*) :

La statistique Q(m) de Ljung-Box permet de tester l'hypothèse d'indépendance sérielle d'une série (ou que la statistique Q(m) de Ljung-Box permet de tester l'hypothèse d'indépendance sérielle d'une série (ou que la série est bruit blanc). Plus spécifiquement cette statistique teste l'hypothèse que les m coefficients d'autocorrélation sont nuls. Elle est basée sur la somme des autocorrélations de la série et elle est distribuée selon une loi Chi-carrée avec m degrés de liberté.

L'hypothèse nulle est:

$$\rho_1 = \cdots \ldots \ldots \ldots \ldots \ldots = \rho_m = 0$$

La statistique du test est:

$$Q(m) = n(n+2)\sum_{j=1}^{m}\frac{\hat{\rho}_j^2}{n-j}$$

Ou ; n est la taille de l'échantillon, $\hat{\rho}_j^2$ est l'autocorrélation au retard j , et m c'est le nombre des retards testés .

La région critique du rejet de cette statistique est définie comme suit ;

$$Q > x_{1-\alpha}^2$$

Ou $x_{1-\alpha}^2$ est le (a-quantile) d'une loi de khi deux à m degré de liberté.

Sur le logiciel R :

```
> Box.test(res.mod1, lag=30, type="Ljung")
```

```
        Box-Ljung test

data:  res.mod1

X-squared = 25.1185, df = 30, p-value = 0.7193
```

La p-valeur est supérieure à 0.05, donc on ne rejette pas l'hypothèse nulle de la blancheur des résidus ce qui signifie que les résidus suivent un bruit blanc.

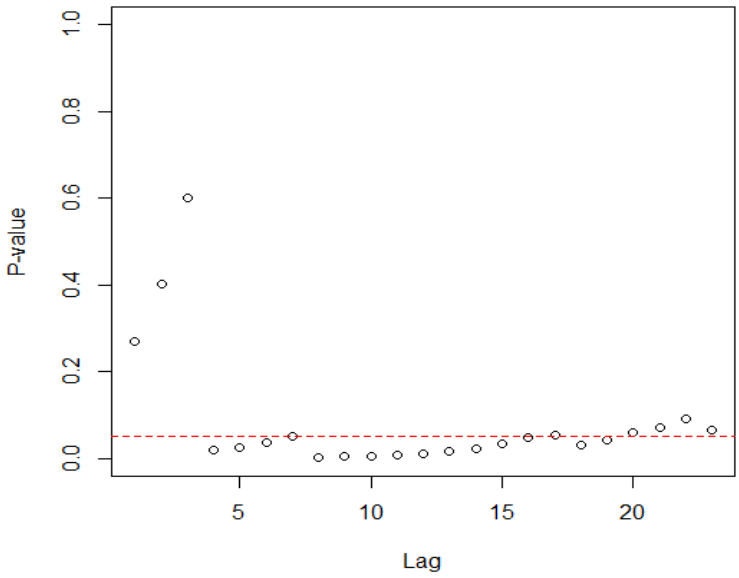
a.2.2 Le test McLeod-Li :

C'est une autre modification de la statistique Q_{BP} , la statistique Q_{LM} est :

$$Q_{LM} = Q_{BP} + \frac{m(m+1)}{2n} = \frac{m(m+1)}{2n} + n\sum_{j=1}^{m}\hat{\rho}_j^2$$

Résultats de l'application de ce test sur le logiciel R :

```
> print(McLeod.Li.test(object=mod2))
```

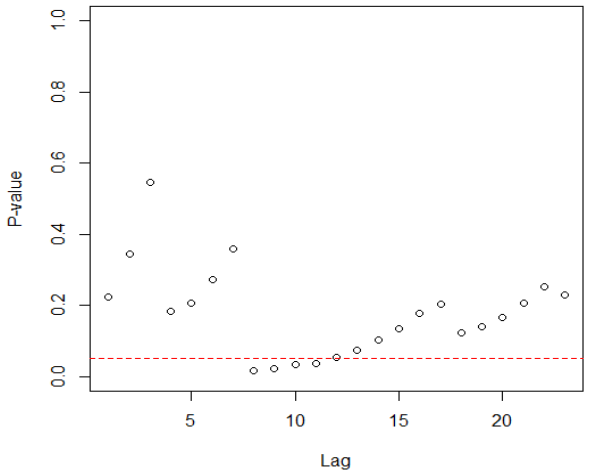


Les p valeurs relatives à ce test ne sont pas toute inférieures à 0.05, d'où le rejet de l'hypothèse de H_0. Les mêmes résultats sont visibles, si on applique ce test aux résidus de notre modèle.

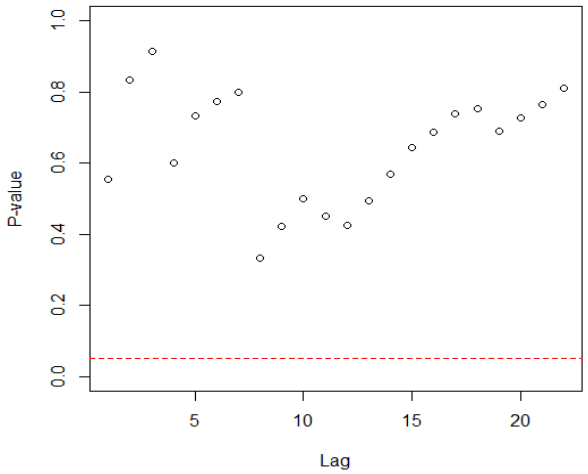
b- Estimation des paramètres d'une modélisation ARIMA(3,1,4)(0,1,1)[12] et ARIMA(3,1,12)(0,1,1)[12]

On essaye d'estimer manuellement d'autres paramètres d'une modélisation d'une **ARIMA(3,1,4)(0,1,1)[12], et ARIMA (3,1,12)(0,1,1)[12].**

Les résultats obtenus des tests de normalité et d'autocorrélation des erreurs sont satisfaisants, ainsi que le test de McLeod-Li pour le modèle 3 ou les p valeurs sont au-dessus de 0.05.



ARIMA(3,1,4)(0,1,1)[12]



ARIMA(3,1,12)(0,1,1)[12]

5.4 Sélection du modèle finale :

On a vu dans la partie théorique qu'il existe plusieurs méthodes et critères pour sélectionner un modèle de prévision, dans notre cas, nous nous appuyons sur le critère d'Akaike (AIC) pour sélectionner le modèle approprié.

MODELE	mod1	mod2	mod3
Critère AKAIKE	122.1502	123.432	147.9978

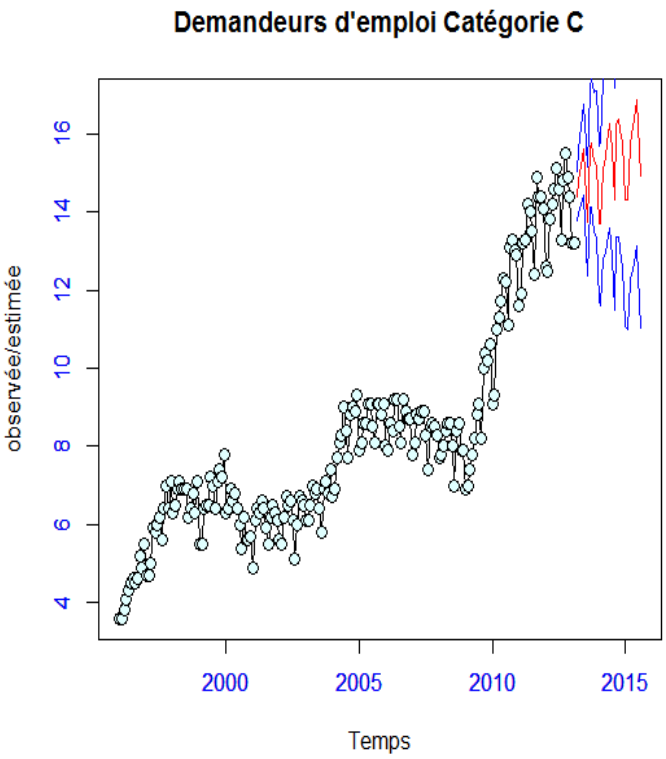
Le critère AKAIKE le plus petit est celui d'u ARIMA(2,1,2)(0,1,1)[12], cependant le test de McLeod-Li a montré qu'il existe une Hétéroscédasticité des résidus , donc à la fin , on retient le modèle 3 , **ARIMA(3,1,12)(0,1,1)[12],** mathématiquement ce modèle s'écrit de la manière suivante :

$$(1 - 0.0104B + 0.0833B^2 - 0.4463B^3)(1 - B)\,\boldsymbol{CatC_t} = (1 - 0.1875B + 0.0902B^2 - 0.3482B^3 - 0.0025B^4 - 0.0063B^5 + 0.0875B^6 - 0.0713B^7 + 0.1531B^8 - 0.1160B^9 + 0.0033B^{10} - 0.0001B^{11} - 0.6021B^{12})(1 + 0.0219B^{12})\,\boldsymbol{\varepsilon_t}$$

5.4.1 Prédiction des demandeurs d'emploi CatC :

Nous avons modélisé la série sur un intervalle d'apprentissage , on considère maintenant la prédiction de la série à l'horizon 15 (mois) en nous basant sur le modèle combiner de la moyen et de l'Ecart type , les intervalles de confiance sont calculés à 95%.

	Prévisions	Borne sup	Borne inf
avr-13	14796	15604	13988
mai-13	15200	16161	14239
juin-13	15599	16729	14470
juil-13	15002	16256	13747
août-13	13749	15114	12383
sept-13	15643	17147	14138
oct-13	15771	17380	14162
nov-13	15299	17035	13562
déc-13	15205	17054	13357
janv-14	13698	15642	11754
févr-14	13702	15748	11655
mars-14	15114	17349	12878
avr-14	15367	17730	13003
mai-14	15753	18255	13250



Sur le graphique ci-dessus nous remarquons que le nombre de demandeurs d'emploi catégorie C va encore augmenter dans les quinze prochains mois.

5.5 Application de la méthodologie de Box Jenkins sur le nombre de demandeurs d'emploi CatA :

Avec la même méthode, nous avons estimés le nombre des demandeurs d'emploi catégorie A **(voir l'annexe 1. 2)**

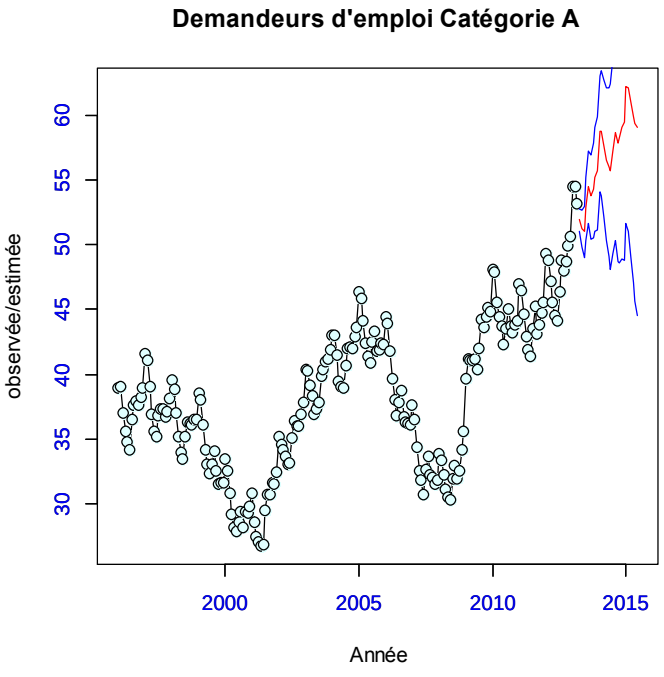
Le modèle finale retenu est un **ARIMA(4,1,3)(0,1,3)[12],** mathématiquement , il s'écrit :

$$(1 - 0.4697B - 0.0370B^2 + 0.8809B^3 + 0.1671B^4)(1 - B)\, \boldsymbol{CatA_t} = (1 + 0.3607B + 0.0693B^2 - 0.7946B^3)(1 - 0.5893B^{12} - 0.0386B^{24} + 0.1099B^{36})\, \boldsymbol{\varepsilon_t}$$

5.5.1 Prédiction des demandeurs d'emploi CatA :

Nous avons modélisé la série sur un intervalle d'apprentissage , on considère maintenant la prédiction de la série à l'horizon 15 (mois) en nous basant sur le modèle combiner de la moyen et de l'Ecart type , les intervalles de confiance sont calculés à 95%.

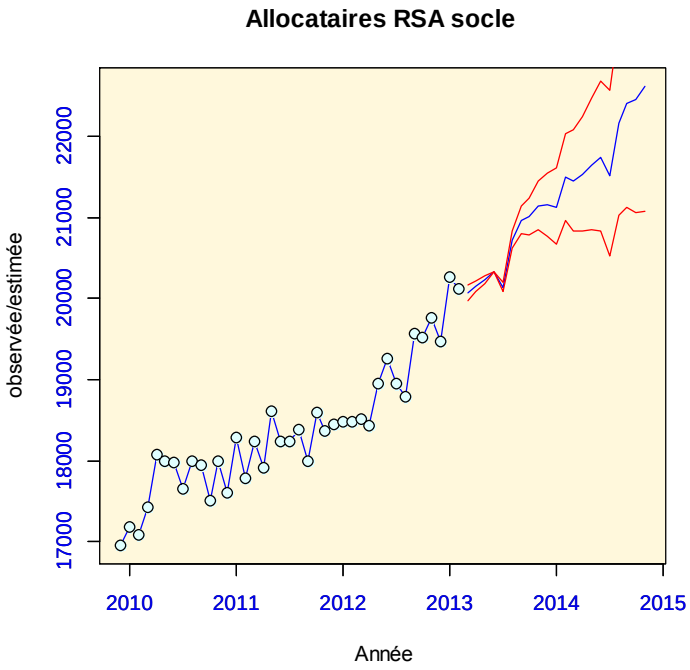
	prévision	Borne sup	Borne inf
avr-13	51939	51080	52784
mai-13	51281	49858	52665
juin-13	51032	49000	52985
juil-13	52817	50305	55215
août-13	54504	51633	57232
sept-13	53759	50416	56906
oct-13	54261	50487	57789
nov-13	55211	51054	59076
déc-13	55683	51158	59867
janv-14	58736	54094	63036
févr-14	58791	53769	63417
mars-14	57647	52125	62685
avr-14	56552	50343	62145
mai-14	55989	49103	62116
juin-14	55724	48131	62400



5.6 Prévision du nombre d'allocataires RSA socle :

Pour obtenir les prévisions du nombre d'allocataires RSA socle payé par le conseil général du Bas Rhin, nous n'avons qu'à remplacer le nombre des demandeurs d'emploi CatA et CatC estimés dans l'équation (1.1), les résultats représentant cette évolution sont :

	Borne sup	Borne sup	Borne inf
avr-13	19998	19545	20458
mai-13	20041	19342	20658
juin-13	20153	19201	20942
juil-13	20228	19416	21151
août-13	20032	19637	21051
sept-13	20645	19435	21831
oct-13	20785	19446	22095
nov-13	20764	19541	22191
déc-13	20811	19558	22349
janv-14	20749	20055	22346
févr-14	20760	19999	22465
mars-14	21125	19720	23024
avr-14	21031	19422	23087
mai-14	21086	18947	23316
juin-14	21243	18935	23647



Conclusion générale

Après l'estimation et les différents tests effectués, le modèle a été jugé valide et acceptable statistiquement et économiquement. Ce modèle comporte Le nombre des demandeurs d'emploi catégorie A et catégorie C comme variable explicative à 92 % de la variation du nombre d'allocataires RSA socle

Certes, il aurait été plus intéressant de travailler avec d'autres variables qui pourrons rajouter un plus à la fiabilité de l'estimation comme le taux de pauvreté, les salaires à bas revenus , mais le manque des données mensuellse et même parfois annuelle a limité notre travail sur les demandeurs d'emploi qui toutefois représentent plus de 45 % de la totalité des bénéficiaires du RSA payés par le conseil général du Bas Rhin.

La même démarche a été appliquée pour estimer l'enveloppe budgétaire consacrée par le conseil général du Bas Rhin aux bénéficiaires du RSA socle et activité.

Bibliographie :

- Livre vert vers un revenu de solidarité active
- Etat des lieux de l'insertion et de la précarité
- Programme départementale d'insertion 2010-2013
- Yves Aragon, Séries temporelles avec R, Méthodes et cas, 2011
- Régis Bourbonnais et Michel Terraza, Analyse des séries temporelles, application à l'économie et à la gestion, manuel et exercices corrigés, 2010.
- RUEY S. TSAY, Analysis of Financial Time Series, Second Edition,2005
- ARTHUR CHARPENTIER , cours des série temporelles théories et applications, volume 1.
- J.Johnston Méthodes économétriques 3eme Edition economica (paris 1985)
- Régis Bourbonnais "économétrie " 3eme édition dumond
- Régression, Cours de deuxième année de master, Bernard Delyon, mai 2013
- Cours de Segolène Geffray ,2013

LES SITES INTERNETS :

- www.insee.fr
- http://freakonometrics.blog.free.fr
- www.wikipedia.org

Logiciel utilisé :

- **R**

Annexe 1

Partie du Code R

Régression linéaire généralisée

```
library(bestglm); library(car)

data<-read.table("C:/Users/hsadaoui/Desktop/données_DE.txt",header=TRUE)

attach(data)

mod1<-glm(RSA_S~CatA+CatC+CatE,family="poisson")

mod2<-glm(RSA_SA~CatA+CatB+CatC+CatD+CatE,family="poisson")

summary(mod1) ; summary(mod2)

########## regression linéaire généralisée modele général ########

# Les résidus de Pearson

res1<-residuals(mod1,type="pearson"); res2<-residuals(mod2,type="pearson")

# La déviance résiduelle

resd1<-residuals(mod1, type="deviance") ; resd2<-residuals(mod2, type="deviance")

# Les résidus studentisés

rst1<-rstudent(mod1); rst2<-rstudent(mod2)

# Graphiques des résidus

residualPlots(mod1, layout=c(2, 3)) ; residualPlots(mod2, layout=c(2, 3))

# Les mesures d'influences

influenceIndexPlot(mod1, vars=c("Cook", "hat"), id.n=3)

influenceIndexPlot(mod2, vars=c("Cook", "hat"), id.n=3)

# Supression de ces mesures et comparaison avec le nouveau model

compareCoefs(mod1, update(mod1, subset=-c(22, 38,39)))

compareCoefs(mod2, update(mod1, subset=-c(26, 30,35)))

# Collinearity and Variance Inflation Factors

vif(mod1); vif(mod2)

# selection ascendante

model_asc<-step(mod1,direction="forward")

# selection déscendante

model_des<-step(mod1,direction="backward")

# selection progressive

model_pro<-step(mod1,direction="both")
```

```
# selection automatique du bon modèle

Y<-RSA_S

x<-data.frame(CatA,CatB,CatC,CatD,CatE)

Xy<-data.frame(x,RSA_S)

modelF<-bestglm(Xy,family=poisson)
```

2 Série chronologique (Catégorie A)

```
library(urca); library(tseries); library(lmtest); library(forecast); library(TSA); library(fUnitRoots);
library(nlme); library(CADFtest); library(nlme); library(CADFtest)

# Demandeurs d'emploi "Catégorie A"  période 1996-2013

#chargement des données

CatA<-read.table("C:/Users/hsadaoui/Desktop/Prévision/CatA.txt",header=T)

data<-data.frame(CatA)

attach(data)

CatA<-ts(data,frequency=12,start=c(1996,1))

#décomposition additive par moyenne mobile

CatA.decomp<- decompose(CatA, "additive")

attach(CatA.decomp); plot(CatA.decomp)

## utilisons la méthode de Holt-Winters pour effectuer de la prévision

# lissage exponentiel de Holt-Winters avec tendance avec saisonnalité  avec choix automatique de alpha, beta
et gamma  de façon à minimiser le one-step ahead prediction error

CatA.hw4<- HoltWinters(CatA,seasonal="additive")

CatA.p4<- predict(CatA.hw4, 22, prediction.interval = TRUE)

plot(CatA.hw4,CatA.p4,xlab="Temps",main="Demandeurs d'emploi catégorie A")

axis(1, col.axis="blue"); axis(2, col.axis="blue")

CatA.hw4$SSE

# la Série des logarithmes

par(mfrow=c(2,2))

LCatA<-log(CatA)

plot(log(CatA),ylab="log(CatA)")

plot(aggregate(log(CatA))/12) #série annuelle moyennée

LCatA.decomp<-decompose(log(CatA))

seasonplot(LCatA)

boxplot(LCatA~cycle(LCatA))  # pour voir si un effet saisonnier semble se dégager

plot(LCatA.decomp)

pacf(LCatA); acf(LCatA)

acf(LCatA.decomp$random[7:201])

#  la série résiduelle (même recentrée) n'est pas un bruit blanc, il semble persister une saisonnalité
```

```
## prédiction à l'aide de la méthode de lissage exponentielle de Holt-Winters sur la série des log

m<-HoltWinters(LCatA); p<-predict(m, 12, prediction.interval = TRUE)

 plot(m, p)

# cherchons un modèle difference-stationary sur la série des log(CatA)

# cherchons à différencier la série pour la rendre stationnaire

ndiffs(LCatA) ; nsdiffs(LCatA)

DLCatA<-diff(LCatA)

plot(DLCatA,col="blue")

tsdisplay(DLCatA)    # il y'a toujours une saisonnalité des observations qui apparait

CatA.decomp<- decompose(CatA, "additive"); attach(CatA.decomp)

DLSCatA<-DLCatA-seasonal

tsdisplay(DLSCatA)

kpss.test(DLSCatA)

eacf(DLCatA)

## Identification manuelle du modèle

mod<-arima(CatA,order=c(3,1,4),seasonal=list(order=c(0,1,3),period=12))

summary(mod)

r<-residuals(mod); tsdisplay(r)

print(McLeod.Li.test(y=r)) ; print(McLeod.Li.test(object=mod))

mod1<-arima(CatA,order=c(4,1,3),seasonal=list(order=c(0,1,3),period=12))

summary(mod1)

r1<-residuals(mod1); tsdisplay(r1)

print(McLeod.Li.test(y=r1)); print(McLeod.Li.test(object=mod1))


mod2<-arima(CatA^2,order=c(7,1,3),seasonal=list(order=c(0,1,3),period=12))

summary(mod2)

r2<-residuals(mod2); tsdisplay(r2)

print(McLeod.Li.test(y=r2)) ; print(McLeod.Li.test(object=mod2))

## Prévision

Prev.CatE<-predict(mod,n.ahead=15,se.fit=TRUE)

attach(Prev.CatE)

ts.plot(CatA,Prev.CatE$pred,xlab="Temps",ylab="observée/estimée", main="Demandeurs d'emploi catégorie
A",gpars=list(col=c('black','red')))

lines(Prev.CatE$pred+2*Prev.CatE$se,col='blue'); lines(Prev.CatE$pred-2*Prev.CatE$se,col='blue')

points(data1, pch=21, bg="lightcyan", cex=1.25)

axis(1, col.axis="blue") ; axis(2, col.axis="blue")
```

Annexe 2

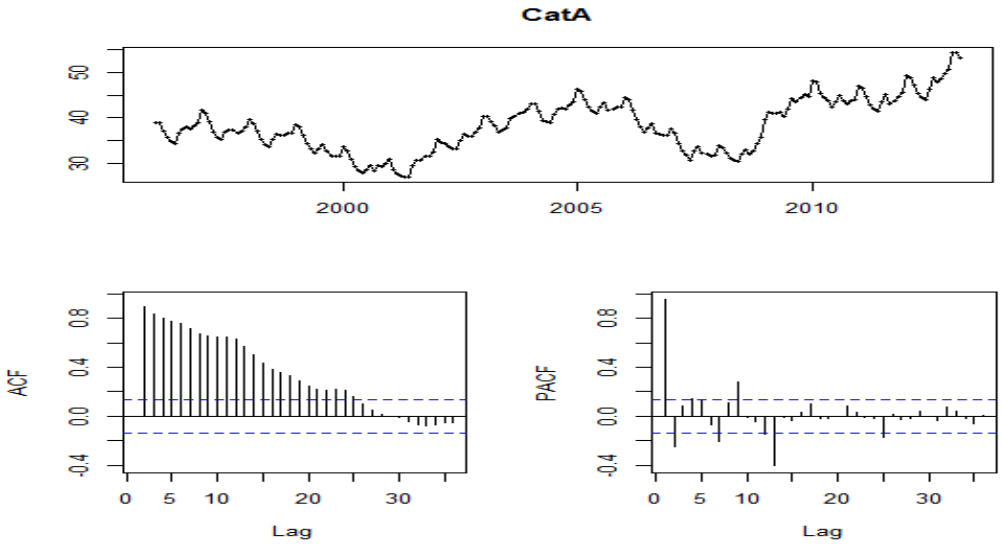
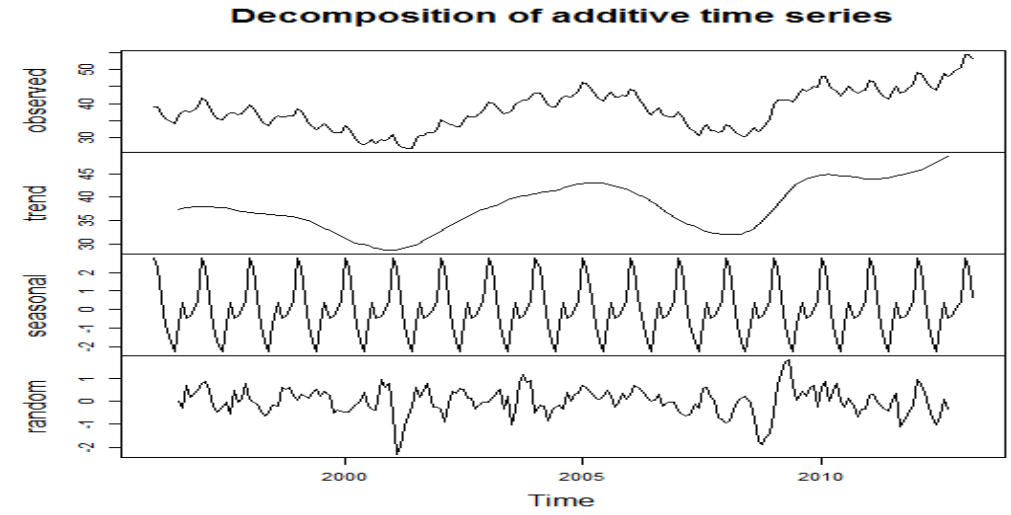


Figure N°1 : Graphique et corrélogrammes de la série CatA



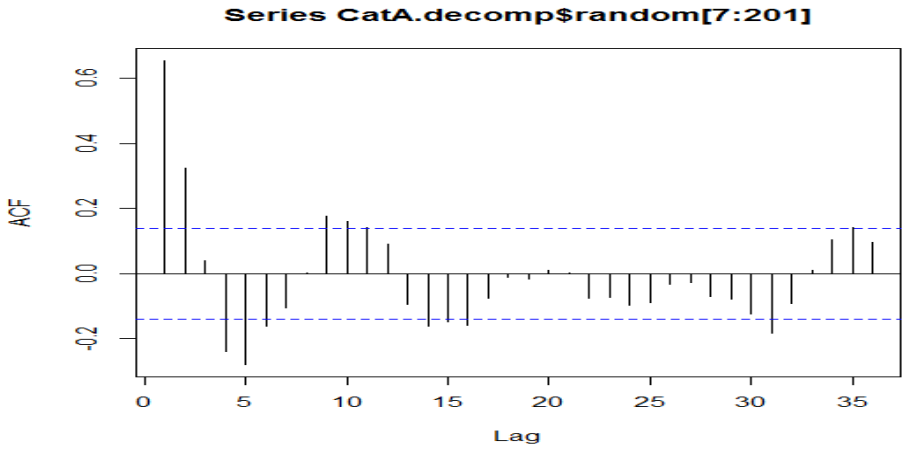


Figure N°2: Graphique et corrélogrammes de la série annuelle aggrégée

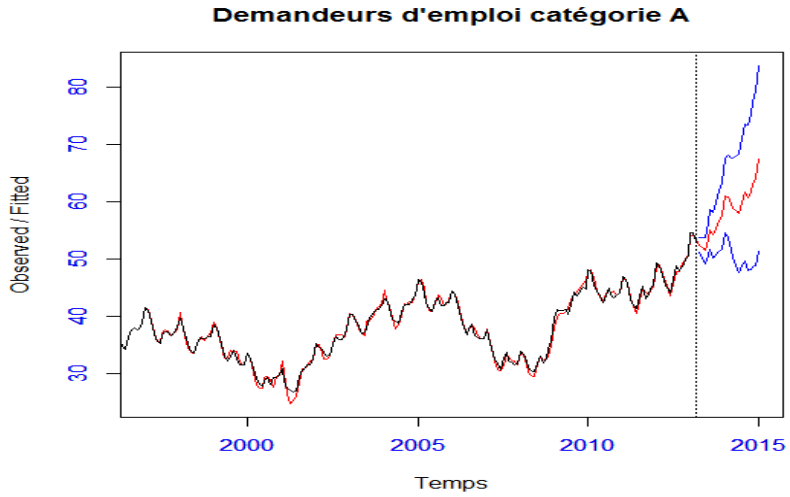


Figure N°3: lissage exponentiel de Holt-Winters avec tendance avec saisonnalité avec choix automatique de alpha, beta et gamma

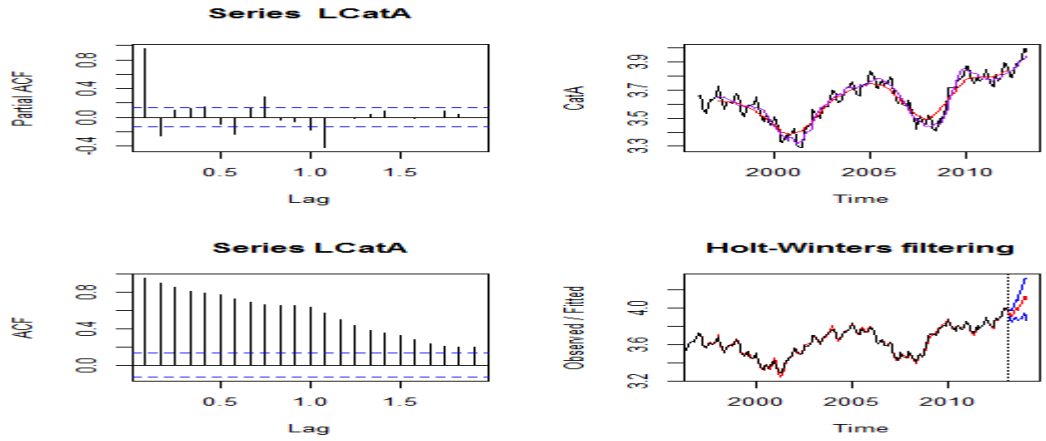


Figure N°4 : Graphique et corrélogrammes de la série DLCatA

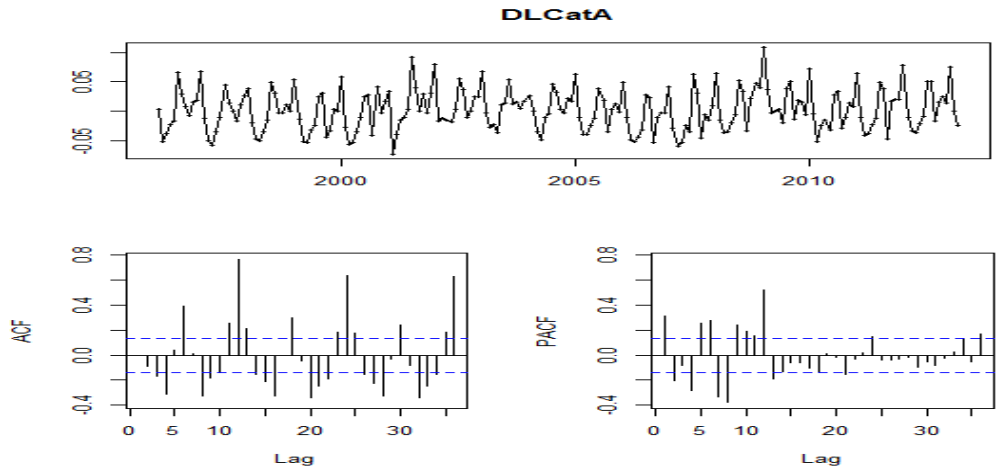


Figure N°5 : Graphique et corrélogrammes de la série DLCatA

```
> kpss.test(DLSCatA)

        KPSS Test for Level Stationarity

data:  DLSCatA
KPSS Level = 0.0139, Truncation lag parameter = 3, p-value = 0.1

Message d'avis :
In kpss.test(DLSCatA) : p-value greater than printed p-value
> eacf(DLCatA)
AR/MA
  0 1 2 3 4 5 6 7 8 9 10 11 12 13
0 x o x x o x o x x o x  x  x  x
1 x x o x o x o x o o o  x  x  x
2 x x o x o o o x o o o  x  x  x
3 x x x x o o o x x o o  x  x  x
4 x x o o o o o o o o o  x  x  x
5 x x o o o o o o o o o  x  x  o
6 x x o x o o o o o x o  x  x  x
7 x x x o o o o o o o o  x  x  x
> mod<-arima(CatA,order=c(3,1,4),seasonal=list(order=c(0,1,3),period=12))
> summary(mod)
Series: x
ARIMA(3,1,4)(0,1,3)[12]

Coefficients:
         ar1      ar2     ar3      ma1     ma2      ma3      ma4     sma1
      0.8877  -0.8025  0.7133  -0.6470  0.9719  -0.6571  -0.0216  -0.5432
s.e.  0.0926   0.0864  0.0686   0.1198  0.1111   0.1167   0.1017   0.0792
         sma2     sma3
      -0.2940   0.1337
s.e.   0.0879   0.0772

sigma^2 estimated as 0.288:  log likelihood=-163.16
AIC=346.32   AICc=347.77   BIC=382.26

Training set error measures:
          ME        RMSE         MAE          MPE        MAPE         MASE
0.008809951 0.519600161 0.382332675 0.036170080 1.024979472 0.354617429
```

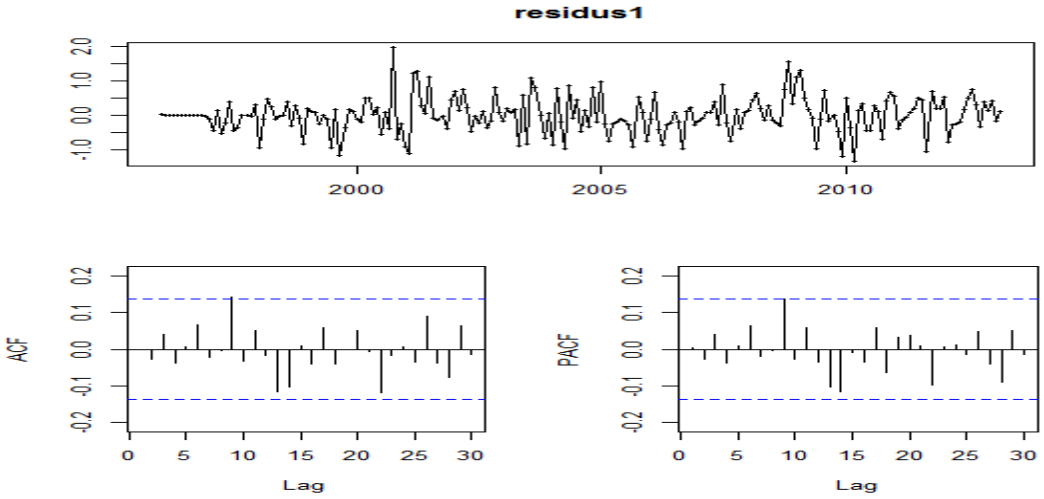
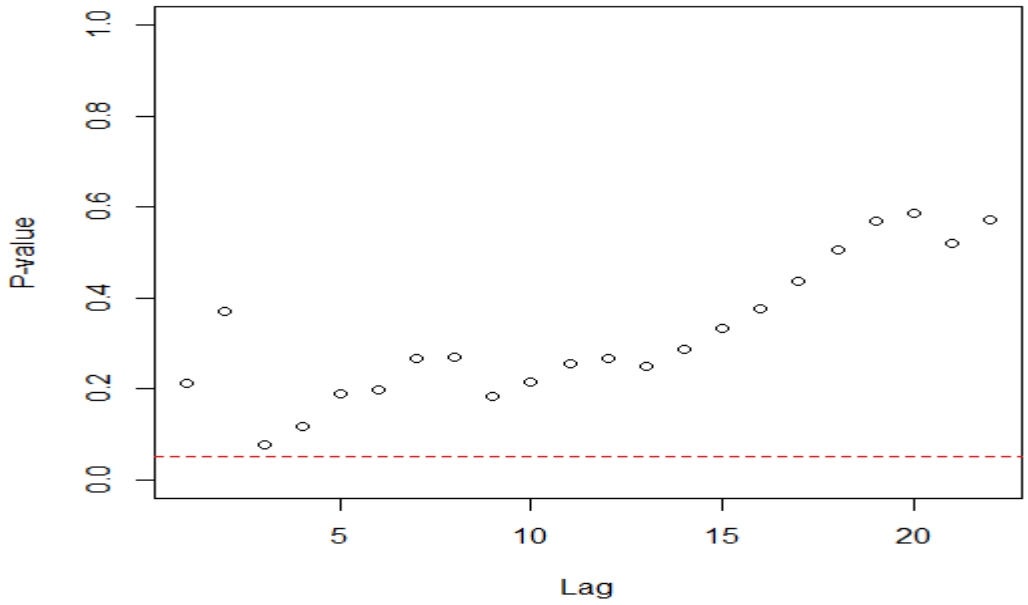


Figure N°6 : Graphique et corrélogrammes de la série des résidus

Etude de l'Hétéroscédasticité des résidus du modèle 1 (mod1) :




```
> mod1<-arima(CatA,order=c(4,1,3),seasonal=list(order=c(0,1,3),period=12))
> mod2<-arima(CatA^2,order=c(7,1,3),seasonal=list(order=c(0,1,3),period=12))
> summary(mod1)
Series: x
ARIMA(4,1,3)(0,1,3)[12]

Coefficients:
         ar1      ar2     ar3      ar4      ma1     ma2      ma3     sma1
      0.5191  -0.1973  0.7159  -0.2004  -0.2802  0.3806  -0.5670  -0.6561
s.e.  0.2221   0.1999  0.1276   0.1125   0.2119  0.1958   0.1197   0.0884
         sma2    sma3
      -0.2711  0.1783
s.e.   0.0946  0.0826

sigma^2 estimated as 0.2998:  log likelihood=-164.11
AIC=348.22   AICc=349.67   BIC=384.16

Training set error measures:
        ME       RMSE        MAE         MPE       MAPE       MASE
0.01043487 0.53009987 0.39535832 0.03741138 1.06148888 0.36669885
> summary(mod2)
Series: x
ARIMA(7,1,3)(0,1,3)[12]

Coefficients:
         ar1      ar2     ar3     ar4     ar5     ar6      ar7      ma1     ma2
      0.3018  -0.3620  0.1513  0.0016  0.1338  0.1894  -0.0250  -0.0232  0.5517
s.e.  1.1775   0.4242  0.4820  0.2734  0.0886  0.1781   0.2296   1.1750  0.2436
         ma3     sma1     sma2    sma3
      0.0643  -0.4988  -0.2691  0.1253
s.e.  0.6769   0.0921   0.0889  0.0806
```

Selection du bon modèle :

```
> AIC(mod)
[1] 348.3179
> AIC(mod1)
[1] 350.2158
> AIC(mod2)
[1] 2065.29
```

Figure N°8 : Prévision du nombre des demandeurs d'emploi CatA par la méthode de Box-Jenkins