MS BigData - SES722

Avril 2019

Projet d'économétrie

Mohamed Dhaoui, Matthieu Roussel

1. Séries temporelles

1.1. Importer les données du fichier quarterly.xls

```
[data, text] = xlsread("quarterly.xls");
idx=2:213;
dates = datenum(text(idx,1),'yyyyQQ');
TF = sum(isnan(data))
```

	DATE	FFR	Tbill	Tb1yr	r5	r10	PPINSA	Finished	CPI	CPICORE	M1NSA	M2SA	M2NSA	Unemp	IndProd	RGDP	Potent	Deflator	Curr
0	1960Q1	3.93	3.87	4.57	4.64	4.49	31.67	33.20	29.40	18.92	140.53	896.1	299.40	5.13	23.93	2845.3	2824.2	18.521	31.830
1	1960Q2	3.70	2.99	3.87	4.30	4.26	31.73	33.40	29.57	19.00	138.40	903.3	300.03	5.23	23.41	2832.0	2851.2	18.579	31.862
2	1960Q3	2.94	2.36	3.07	3.67	3.83	31.63	33.43	29.59	19.07	139.60	919.4	305.50	5.53	23.02	2836.6	2878.7	18.648	32.217
3	1960Q4	2.30	2.31	2.99	3.75	3.89	31.70	33.67	29.78	19.14	142.67	932.8	312.30	6.27	22.47	2800.2	2906.7	18.700	32.624
4	1961Q1	2.00	2.35	2.87	3.64	3.79	31.80	33.63	29.84	19.17	142.23	948.9	317.10	6.80	22.13	2816.9	2934.8	18.743	32.073
5	1961Q2	1.73	2.30	2.94	3.62	3.79	31.47	33.33	29.83	19.23	141.40	966.4	320.97	7.00	23.00	2869.6	2962.9	18.785	32.131
6	1961Q3	1.68	2.30	3.01	3.90	3.98	31.50	33.33	29.95	19.32	142.00	982.7	326.50	6.77	23.74	2915.9	2991.3	18.843	32.699
7	1961Q4	2.40	2.46	3.10	3.84	3.97	31.53	33.37	29.99	19.37	146.63	1000.0	334.70	6.20	24.57	2975.3	3019.9	18.908	33.421
8	1962Q1	2.46	2.72	3.21	3.84	4.02	31.70	33.53	30.11	19.44	146.37	1020.7	341.17	5.63	24.94	3028.7	3048.7	19.020	33.136
9	1962Q2	2.61	2.72	3.02	3.63	3.87	31.53	33.43	30.22	19.51	145.33	1042.3	346.23	5.53	25.18	3062.1	3078.0	19.047	33.468

Pas de valeurs manquantes constatées.

Ci-dessous les statistiques descriptives des champs Unemp et CPI :

	CPI	Unemp
count	212.000000	212.000000
mean	113.182028	6.081981
std	65.310267	1.614394
min	29.400000	3.400000
25%	43.705000	4.970000
50%	109.635000	5.715000
75%	167.507500	7.130000
max	231.280000	10.670000

On voit que l'écart-type de la variable CPI est élevé avec une grande différence entre le min, la moyenne et le max .

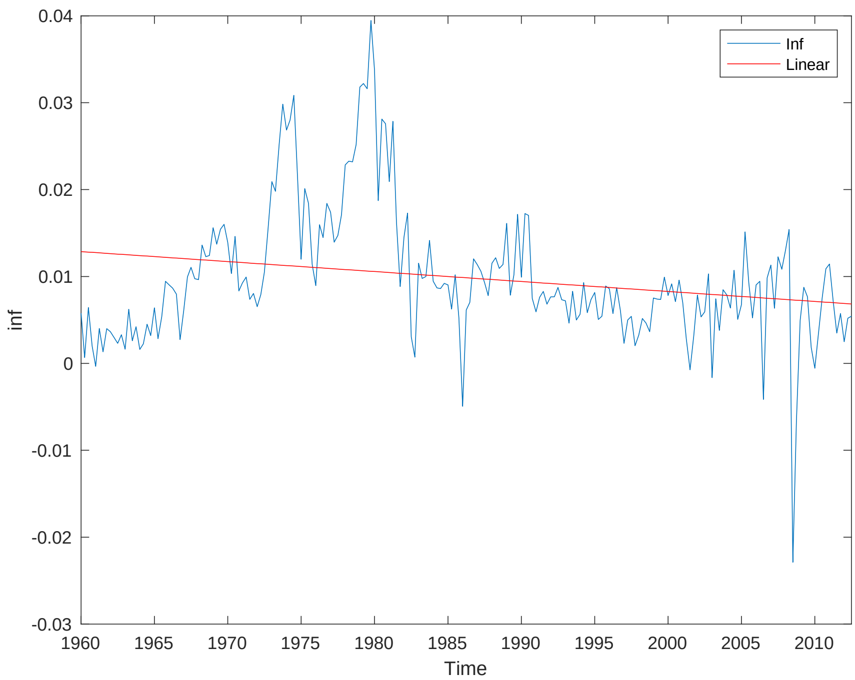
1.2. Calculer inf, le taux d'inflation à partir de la variable CPI. Faire un graphique dans le temps de inf. Commentez.

La variable CPI correspond à l'index des prix à la consommation.

Sans plus de précision, nous partirons sur un taux d'inflation trimestriel (pas de multiplication par 4).

$$Inflation\ (t) = \frac{CPI(t) - CPI(t-1)}{CPI(t-1)}$$

```
idx=2:212;
% Calcul du taux d?inflation
inf=(data(idx,8)./ data(idx-1,8) - 1);
idx=1:211;
ts1 = timeseries(inf,datestr(dates(idx,1)));
ts1.Name = 'Inflation Rate';
figure(1)
plot(ts1)
b = polyfit((dates(idx,1)),inf, 1);
fr = polyval(b, (dates(idx,1)));
ts2 = timeseries(fr,datestr(dates(idx,1)));
hold on
plot(ts2, '-r')
hold off
datetick('x','yyyy','keeplimits')
h1 = legend('Inf', 'Linear')
xlabel('Year')
ylabel('inf')
```



Commentaires :

Nous observons ici l’évolution trimestrielle du taux d’inflation sur une période de 50 ans. L’analyse visuelle de la série ne permet pas de statuer sur la stationnarité de la série (tendance légèrement décroissante, autour de 1%). Cependant, nous pouvons distinguer plusieurs périodes :

- Période de plus forte inflation jusqu’au troisième trimestre 1981
- Inflation modérée depuis, avec cependant
  - Une forte volatilité dans les années 80
  - Une variabilité croissante de 1991 à la crise de fin 2008, clairement mise en valeur au 4ème trimestre 2008

La stationnarité de la série est difficile à évaluer graphiquement, malgré les périodes pré-listées qui pourraient faire penser à une série non stationnaire.

1.3. Interpréter l'autocorrélogramme et l'autocorrélogrammes partiels de inf. Quelle est la différence entre ces deux graphiques ?

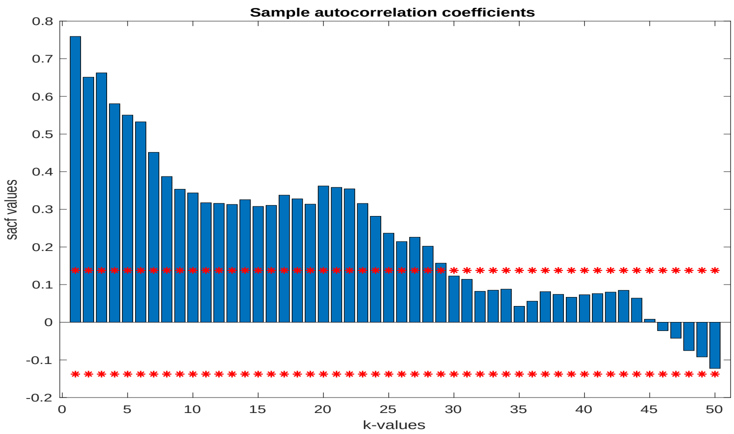
L'estimation des modèles ARIMA nécessite l'examen des fonctions d'auto-corrélation et d'auto-corrélation partielle. L'auto-corrélation est la corrélation d'une série avec elle-même, selon un décalage (lag) défini. L'auto-corrélation de décalage 0 est par définition égale à 1. La fonction d'auto-corrélation fait correspondre à chaque décalage l'auto-corrélation correspondante. La corrélation partielle entre deux variables est la quantité de corrélations qui n'est pas expliquée par les relations de ces variables avec un ensemble spécifié d'autres variables. Supposons par exemple que l'on réalise la régression de Y sur trois variables X1, X2 et X3. La corrélation partielle entre Y et X3 contrôlant X1 et X2 est la quantité de corrélation entre Y et X3 qui n'est pas expliqué par leurs relations communes avec X1et X2. Elle peut être calculée comme la racine carrée du gain de variance expliquée obtenu en ajoutant X3 à la régression de Y sur X1 et X2.

Dans le cas des séries temporelles, la corrélation partielle de décalage k est la corrélation entre yt et yt-k, contrôlant l'influence des k-1 valeurs interposées.

Ci-dessous les courbes d'autocorrélation et d'autocorrélation partielle obtenues sur l'inflation

Autocorrélogramme

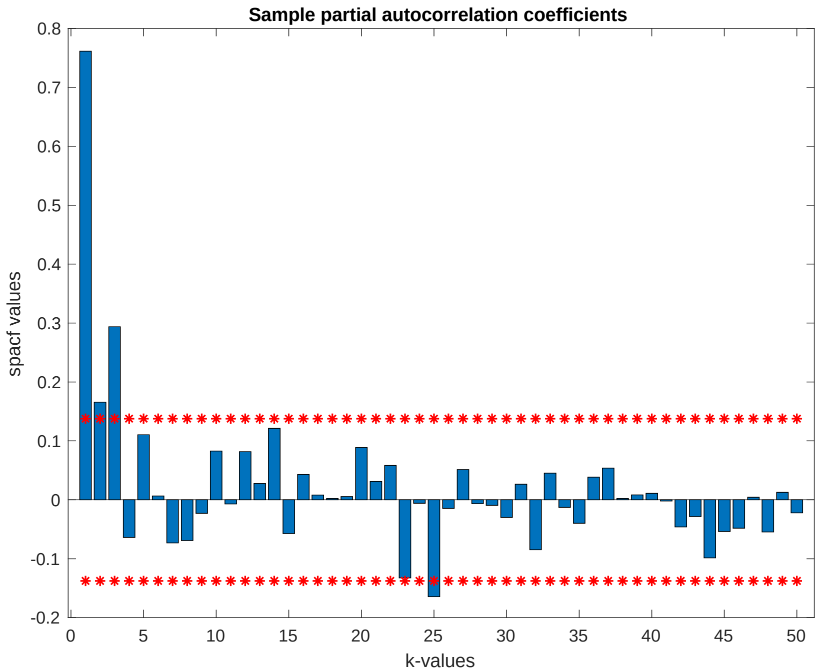
```
acf=sacf(inf,50);
```



Là encore, l'autocorrélogramme décroit mais plutôt lentement, là où les séries stationnaires affichent généralement une nette décroissance exponentielle.

Autocorrélogrammes partiels

```
pacf=spacf(inf,50)
```



Commentaires :

Le graphique des autocorrélations partielles est caractéristique d'un modèle auto-régressif, montrant que seuls les 3 premiers coefficients sont significatifs, ie que p_max = 3. Un pic est également visible autour de 25 trimestres, qui pourrait caractériser des cycles de 6 ans.

1.4. Quelle est la différence entre la stationnarité et l'ergodicité ? Pourquoi a-t-on besoin de ces deux conditions. Expliquez le terme "spurious regression".

Avant d'aborder la différence entre la stationnarité et l'ergodicité, rappelons les définitions des deux concepts :

Stationnarité

On dit qu'un processus est stationnaire si ses caractéristiques statistiques ne varient pas le long du temps. On définit principalement deux types de stationnarité.

- Stationnarité au sens strict : Un processus est stationnaire au sens strict si pour toute valeur de N, sa caractérisation d'ordre N est invariante par rapport à une translation de l'axe des temps :
- Stationnarité au sens large (ou de deuxième ordre) : Un processus aléatoire est stationnaire au sens large si sa moyenne est constante, et sa fonction d'autocovariance ne dépend pas de l'origine du temps.

La condition nécessaire de stationnarité est que la moyenne et la variance de la série soient constantes. Elle implique donc que le graphe de la série en fonction du temps montre un niveau moyen à peu près constant et des fluctuations à peu près de même ampleur autour de la moyenne supposée, quelle que soit la date autour de laquelle on examine la série.

Ergodicité

Cette notion fondamentale vient de la considération suivante : nous savons à présent qu'une série temporelle est une réalisation particulière d'un processus et donc que chaque observation est l'une des réalisations de la variable aléatoire indicée correspondante. Comment, alors, calculer l'espérance, la variance, la fonction d'autocorrélation du processus alors que nous savons qu'il nous faut connaître beaucoup plus d'un point par variable aléatoire ?

En d'autres termes, comme on ne peut faire de calcul statistique sur un cas, il faut trouver un autre moyen. On dira, alors, qu'un processus stationnaire est ergodique si l'on peut calculer l'ensemble de ses caractéristiques (moyenne, variance, fonction d'autocorrélation) à partir d'une seule trajectoire du processus, c'est-à-dire à partir d'une observation du processus et, par conséquent, de façon pratique, à partir de la série temporelle observée suffisamment longtemps. En bref, on décide que la série observée est typique du processus.

Ergodicité et stationnarité sont deux notions voisines mais distinctes.

Etant donné un processus stochastique X(n), n entier, pour calculer la moyenne de X(1), on peut procéder de deux manières différentes : soit on prend plusieurs réalisations de ce processus, qui donnent chacune une valeur de X(1), et on fait la moyenne de ces valeurs ; soit on prend un seul processus, et on fait la moyenne de X(1), X(2), X(3), etc. Le résultat que l'on veut est le premier, mais dans certains cas, les deux coïncident : on dit alors que le processus est ergodique. Les processus ergodiques sont intéressants, car généralement, on ne dispose que d'une seule réalisation.

Spurious Regression

La régression fallacieuse désigne une situation dans laquelle l'utilisation de séries temporelles non stationnaires dans une régression linéaire fait apparaître des résultats erronés, trop optimistes, qui font croire à une relation entre les variables alors que ce n'est pas le cas.

1.5. Proposer une modélisation AR(p) de inf, en utilisant tous les outils vus au cours.

1. Test de stationnarité

Afin d'être certain de pouvoir proposer un modèle auto-régressif, menons le test de recherche de racine unitaire de Dickey-Fuller :

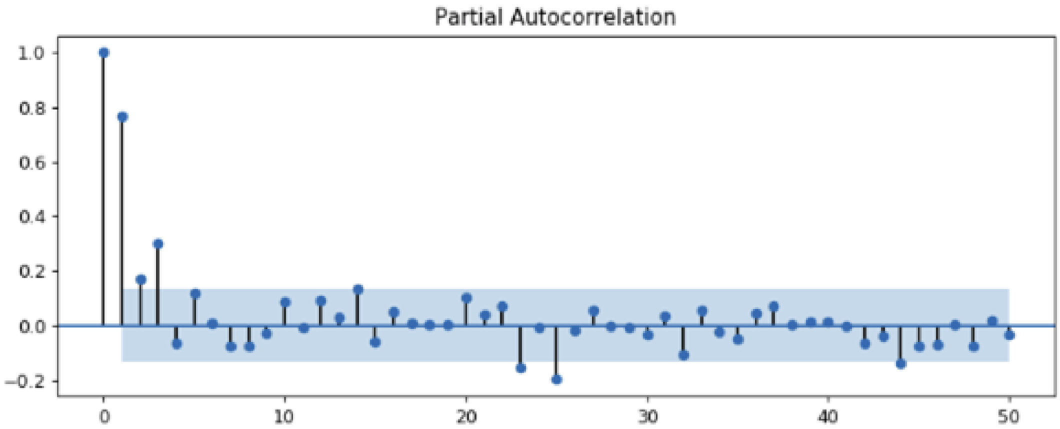
- H0 : présence d'une racine unitaire
- H1 : H0 est fausse

```
[n,k]= size(inf)
y_=inf(2: n)
y_lag =inf(1:n -1)
Delta_y =y_ - y_lag
[n,k]= size ( Delta_y )
X=[ones(n,1),y_lag]
[n,k]= size (X)
beta = inv (X'*X)*X'* Delta_y
u= Delta_y -X* beta ;
sig2 =u'*u/(n-k);
std = sqrt ( diag ( sig2 * inv (X'*X )))
t= beta ./ std
p=tdis_prb(t(1),n-k)
```

La statistique du test vaut -5.292 pour une probabilité critique de 4.622e-05, conduisant au rejet de H0. L'hypothèse d'un modèle autorégressif stable d'ordre 1 est donc plausible. La série est alors considéré comme stationnaire

2. Determination visual du parameter p

Pour déterminer le paramètre de la régression AR , on peut commencer par visualiser la courbe des autocorrélations partielles pour déterminer l'indice à partir duquel les legs deviennent non significatifs .



A partir du graphe , on voit que les lags deviennent peu significatifs à partir de l'ordre 3 .

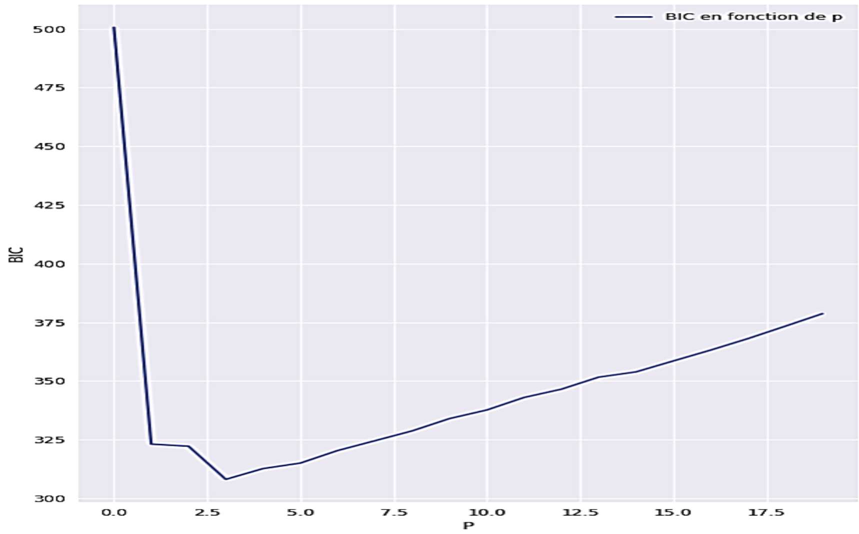
3. Determination de p avec BIC

Une autre manière de déterminer p est d'utiliser le BIC (Bayesian Information Criterion) : Le critère d'information bayésien qui est un critère d'information dérivé du critère d'information d'Akaike proposé par Gideon Schwarz en 1978. Il s'écrit :

$$BIC = -2 \ln(L) + k * \ln(N)$$

avec L la vraisemblance du modèle estimée, N le nombre d'observations dans l'échantillon et k le nombre de paramètres libres du modèle1.

Ci-dessous le graphe des BIC en fonction du paramètres p testés :



On voit que le paramètre p qui minimise le BIC est 3 . De ce fait , on peut utiliser cet ordre pour notre modèle AR.

4. Test de Ljung-Box

- H0 : non correlation des données
- H1 : H0 fausse

```
[qstat , pval]= qstat2 (inf ,1)
```

Le test a une statistique de 123.36 et une probabilité critique associée de 0, conduisant à rejeter fortement H0. La série est donc auto corrélée.

5. *Recherche de p par le critère d'Akaike*

Nous allons comparer les AIC pour les 3 valeurs de p candidates.

$$AIC = ln(\hat{\sigma}^2) + \frac{2p}{T},$$

```
[n,k]=size(inf);
% Test of AR1 :
y_=inf(2: n);
y_lag =inf(1:n -1);
X= y_lag;
[n,k]= size (X);
beta = inv (X'*X)*X'* y_;
u=y_ -X* beta ;
sig2 =u'*u/(n-k);
AIC1 = log(sig2) + (2.*(1))./ n
% Test of AR2 :
```

```
y_=inf(3: n);
y_lag =inf(2:n -1);
y_lag2 =inf(1:n -2);
X=[ y_lag y_lag2 ];
[n,k]= size (X);
beta = inv (X'*X)*X'* y_;
u=y_ -X* beta ;
sig2 =u'*u/(n-k)
AIC2 = log( sig2 ) + (2.*(2))./ n
% Test of AR3 :
y_=inf(4: n);
y_lag =inf(3:n -1);
y_lag2 =inf(2:n -2);
y_lag3 =inf(1:n -3);
X=[ y_lag y_lag2 y_lag3];
[n,k]= size (X);
beta = inv (X'*X)*X'* y_;
u=y_ -X* beta ;
sig2 =u'*u/(n-k);
AIC3 = log( sig2 ) + (2.*(3))./ n
```

Les résultats sont les suivants :

- AIC(1) = -7.7267
- AIC(2) = -7.7642
- AIC(3) = -7.8616

AIC(3) < AIC(2) < AIC(1), nous choisirons p = 3 pour le modèle AR(p)

6. Test de Dickey-Fuller augmenté pour vérifier la stationnarité d'un AR(3)

Test mené :
$H_0 : \rho = 1$ versus $H_1 : \rho < 1$.

```
% test de Dickey-Fuller augmentÈ pour vérifier la stationnaritÈ d?un
AR(3).
[n,k]=size(inf);
y_=inf(2: n);

y_=inf(2:n);
y_lag=inf(1:n-1);
Delta_y=y_-y_lag;
[n,k]=size(Delta_y);
Delta_y_=Delta_y(1:n-3);
Delta_y_lag=Delta_y(2:n-2);
Delta_y_lag2=Delta_y(3:n-1);
Delta_y_lag3=Delta_y(4:n);
y_lag_=inf(1:n-3);
[n,k]=size(Delta_y_);
X=[ones(n,1) y_lag_ Delta_y_lag Delta_y_lag2 Delta_y_lag3];
[n,k]=size(X);
beta=inv(X'*X)*X'*Delta_y_;
u=Delta_y_-X*beta;
sig2=u'*u/(n-k);
std=sqrt(diag(sig2*inv(X'*X)));
t=beta./std
p= tdis_prb (t,n-k)
```

Les p-values associées aux hypothèses sont les suivantes :

- P=1 => p_val = 0.0000 => Rejet de H0 => processus autorégressif stable d'ordre p = 1 « accepté »
- P=2 => p_val = 0.0000 => Rejet de H0 => processus autorégressif stable d'ordre p = 2 « accepté »
- P=3 => p_val = 0.0000 => Rejet de H0 => processus autorégressif stable d'ordre p = 3 « accepté »
- P=4 => p_val = 0.251 => H0 acceptée => la série n'est pas un processus autorégressif stable d'ordre p = 4

Nous proposerons donc une modélisation AR (3)

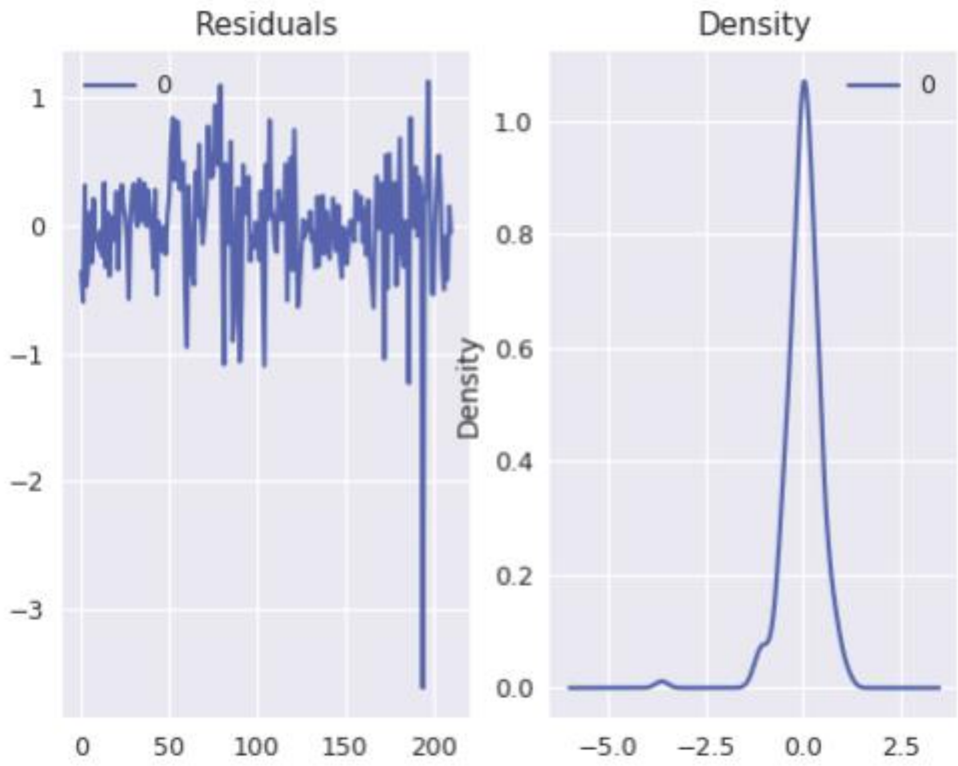
1.6. Estimer le modèle de la courbe de Philips qui explique le taux de chômage (Unemp) en fonction du taux d'inflation courant et une constante.

```
% Variables
idx=2:212;
unemp=table(idx,13);
% Model
y=unemp;
[n,k]=size(y);
X=[ones(n,1),inf];
[n,k]=size(X);
% estimation of parameters
beta=inv(X'*X)*X'*y
% ecarts-types
u=y-X*beta;
sig2=u'*u/(n-k);
std=sqrt(diag(sig2*inv(X'*X)));
t=beta./std
p_val = tdis_prb(t,n-k)
results=[beta std t p_val]
rowNames = {'cste','inf'};
colNames = {'beta','std','t','p_value'};
sTable =
array2table(results,'RowNames',rowNames,'VariableNames',colNames)
```

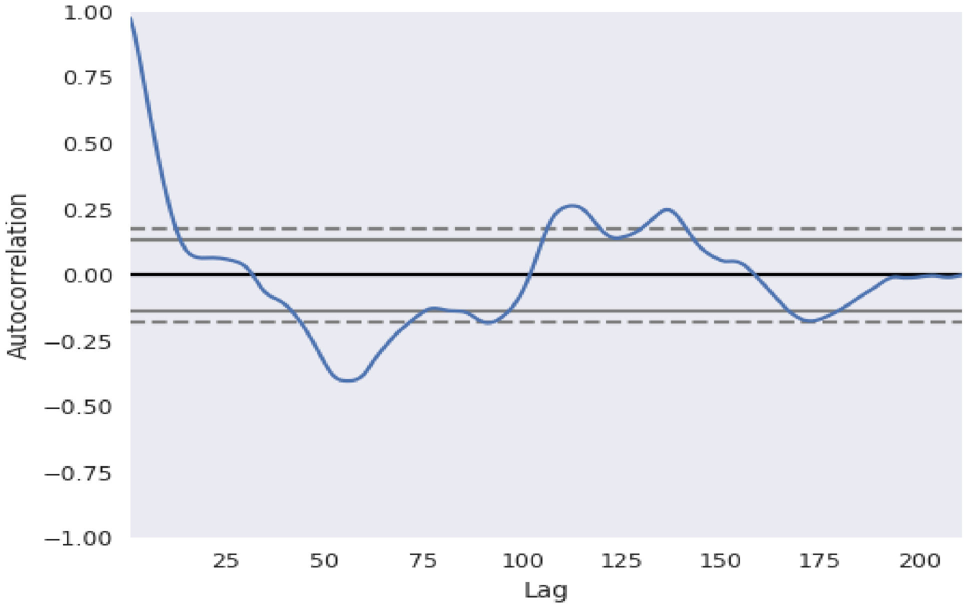
	beta	std	t	p_value
cste	6.0708	0.18081	33.576	0
inf	1.5908	14.44	0.11016	0.91238

1.7. Tester l'autocorrélation des erreurs.

Essayons de regarder la distribution des résidus :



Comme on peut le voir, la densité a une forme gaussienne, mais ceci n'est pas suffisant. Essayons de regarder la courbe de l'autocorrélation



La courbe d'autocorrélation ne descend pas rapidement vers 0 et il y a des coefficients d'autocorrélation significatifs au niveau de lag 50 et 110. Cela est un signe d'une autocorrélation des erreurs.
Essayons maintenant de quantifier le test d'autocorrélation en faisant tester de Durbin Watson

Pour tester l'hypothèse nulle, on calcule la statistique de Durbin Watson

$$DW = \frac{\sum_{t=2}^{n} (\hat{u}_t - \hat{u}_{t-1})^2}{\sum_{t=1}^{n} \hat{u}_t^2}$$

De par sa construction, cette statistique varie entre 0 et 4 . Afin de tester l'hypothèse nulle, Durbin et Watson (1950) ont tabulé les valeurs critiques de DW au seuil de 5%, en fonction de la taille de l'échantillon T et du nombre de variables explicatives (k). La lecture de la table permet de déterminer 2 valeurs d1 et d2 comprises entre 0 et 2 , à partir de n et p où

p : nombre de variables explicatives (constante exclue)
n : nombre d'observations

Dans notre cas, et en se basant sur les valeurs critiques dans le tableau de Watson, on peut utiliser le critère ci-dessous pour tester l'autocorrélation :

Durbin-Watson Test	**Description**
≈ 2	Pas d'autocorrélation
Très proche de 0	Auto-corrélation positive
Très proche de 4	Auto-corrélation négative

Pour notre cas , DW = 0.043 et est loin de 2 , de ce fait on rejette le test et on peut confirmer à 5% que les résidus sont corrélés

Une autre façon de vérifier l'autocorrélation est d'estimer un modèle AR(1) des résidus et tester si le coefficient d'autocorrélation est nul (la pente) :

Les hypothèses à tester sont les suivantes :

- H0 : le coefficient d'autocorrélation des erreurs est nul
- H1 : H0 est fausse

```
u_=[u(2:n)]
u_lag= [u(1:n-1)]
y=u_
X=[u_lag]
[n,k]=size(X)
rho=inv(X'*X)*X'*y
u=y-X*rho
sig2=u'*u/(n-k)
std=sqrt(diag(sig2*inv(X'*X)))
t=rho./std
p=tdis_prb(t,n-k)
```

La statistique du test vaut 67.50 et suit sous H0 une loi de Student à 210 – 1 ddl. La probabilité critique associée vaut 0, conduisant au rejet de H0 (absence d'autocorrélation des erreurs).

1.8. Corriger l'autocorrélation des erreurs par la méthode vue en cours.

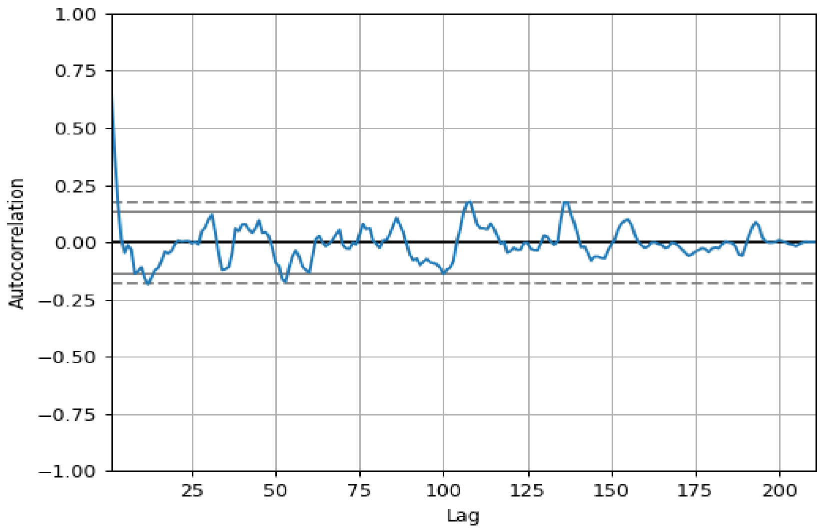
Nous allons tenter de corriger l'autocorrélation des erreurs par la méthode des MCG.

```
% Variables
idx=2:212;
unemp=table(idx,13);
% Model
y=unemp;
[n,k]=size(y);
X=[ones(n,1),inf];
[n,k]=size(X);
y_=[y(2:n)];
y_lag= [y(1:n-1)];
new_y=[sqrt(1-rho^2)*y(1) ; y_-rho*y_lag];
X_=[X(2:n,:)];
X_lag= [X(1:n-1,:)];
new_X=[sqrt(1-rho^2)*X(1,:); X_-rho*X_lag];
% Nouveau modËle sur donnÈes corrigÈes
[n,k]=size(new_X)
beta=inv(new_X'*new_X)*new_X'*new_y;
u=new_y-new_X*beta;
sig2=u'*u/(n-k);
std=sqrt(diag(sig2*inv(X'*X))) ;
t=beta./std;
p_val = tdis_prb(t,n-k);
results=[beta std t p_val];
rowNames = {'cste','inf'};
colNames = {'beta','std','t','p_value'};
sTable =
array2table(results,'RowNames',rowNames,'VariableNames',colNames)
% Tests d'autocorrÈlation
u_=[u(2:n)];
u_lag= [u(1:n-1)];
y=u_;
X=[u_lag];
[n,k]=size(X);
rho2=inv(X'*X)*X'*y;
u=y-X*rho2;
```

```
sig2=u'*u/(n-k);
std=sqrt(diag(sig2*inv(X'*X)));
t=rho2./std
p=tdis_prb(t,n-k)
```

	beta	std	t	p_value
cste	6.3113	0.037241	169.47	0
inf	-9.9301	2.9743	-3.3387	0.00099661

Ci-dessous la nouvelle courbe d'autocorrélation des nouveaux résidus :



Visuellement, on peut dire que qu'on a réussi à réduire significativement l'auto-corrélation des erreurs . Essayons de refaire le test d'autocorrélation sur les nouveaux résidus :

La statistique du test vaut 11.86 avec une probabilité critique associée toujours < à 5% , conduisant au rejet de H0. De ce fait, il reste encore de l'autocorrélation. Si on fait le test de Durbin Watson, on obtiens une statistique de test F_durbin_watson test = 0.9779 , ce qui n'est pas concluant pour dire qu'on a éliminé l'autocorrélation .

Nous avions déterminé que le nombre de retards à prendre en compte était de 3 et nous n'avons ici pris en compte qu'un seul retard… Un modèle dynamique s'impose.

1.9. Tester la stabilité de la relation chômage-inflation sur deux sous-périodes de taille identique.

La longueur de la série étant de 211 observations, nous couperons à 105 observations pour la première période et 105 pour la deuxième.

Méthode utilisée :

$$F = 1 + \frac{[SSR - (SSR1 + SSR2)]}{SSR1 + SSR2} . \frac{[n - 2\,(k + 1]}{k + 1}$$

La statistique suit sous H0 une loi de Fisher à (1+1) et (n – 2 (1+1)) ddl

```
% Variables
idx=2:212;
unemp=data(idx,13);
% Calculate SSR for original model
y=unemp;
[n0,k]=size(y);
X=[ones(n0,1),inf];
[n0,k]=size(X);
beta=inv(X'*X)*X'*y;
u=y-X*beta;
SSR0 = u'*u;
% Select first 105 obs
% Variables
unemp_1=unemp(1:105,1);
inf_1=inf(1:105,1);
% Calculate SSR for sel1
y=unemp_1;
[n1,k]=size(y);
X=[ones(n1,1),inf_1];
[n1,k]=size(X);
beta=inv(X'*X)*X'*y;
u=y-X*beta;
SSR_1 = u'*u
% Select next 106 obs
unemp_2=unemp(106:211,1);
inf_2=inf(106:211,1);
% Calculate SSR for sel2
y=unemp_2;
[n2,k]=size(y);
X=[ones(n2,1),inf_2];
[n2,k]=size(X);
beta=inv(X'*X)*X'*y;
u=y-X*beta;
SSR_2 = u'*u

% Calculate F statistic
SSR_sum = SSR_1 + SSR_2
F = ((SSR0 - SSR_sum) / SSR_sum) * ((n0-2*k)/k)
fdis_prb(F,k,n0-2*k)
```

La valeur obtenue pour la statistique est 1.942 et la probabilité critique associée est de 0.146. On ne rejettera pas la stabilité de la relation chômage/inflation entre les deux périodes.

Une autre façon de tester la stabilité de la relation est de vérifier s'il y a une différence significative entre les coefficients des modèles de chaque période :

« H0 : Il n'y a pas de différence entre les pentes des modèles »

- z= 1.9711
- la p-value = 0.0587

A 5% près, on n'accepte pas l'hypothèse H0 et on peut dire qu'il n'y a pas une différence entre les coefficients. On refait le même test pour les deux intercepts :

- z=-0.7392
- pvalue =0.459

On peut dire qu'il n'y a pas une différence significative entre les coefficients des deux modèles.
Il faut noter que le choix des périodes influence les résultats de test , dans la mesure où si on choisi deux période de taille 50 , une au début de la série et une autre contenant le point de la crise de 2008 , on obtient des modèles statistiquement différents .

1.10. Faites les tests changement de structure de Chow et détecter le point de rupture.

1. Test de Chow

Dans l'analyse des modèles temporelles, la relation entre l'endogène et les exogènes peut être modifiée au cours du temps, à partir d'une date donnée. La régression sur l'ensemble de la période est différente de la régression sur les sous - périodes. Le but de test de chow est de détecter ce changement de structure

Principe :

On estime 3 modèles de régressions :

(1) Une régression contrainte = régression sur l'ensemble de l'échantillon ==> on obtient une somme des carrés des résidus (SCR)

(2) Deux régressions non-contraintes = régressions sur les deux sous-échantillons définies sur les dates==> on calcule deux sommes des carrés de résidus S1 et S2

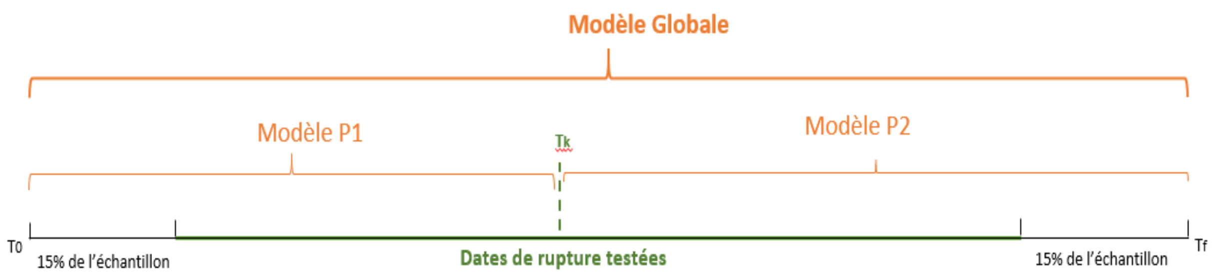
L'hypothèse nulle du test de Chow nous dit que les coefficients de régressions sont égaux entre les deux périodes

Soient SCR la somme des carrés des résidus estimés du modèle initial, S1 la somme des carrés des résidus estimés du premier groupe, et S2 la somme des carrés des résidus estimés du groupe 2. Les valeurs N1 et N2 représentent le nombre d'observations dans chaque groupe et k est le nombre total de paramètres à estimer (3 dans ce cas). Alors la statistique du test de Chow est égale à :

$$F = \frac{(S_C - (S_1 + S_2))/k}{(S_1 + S_2)/(N_1 + N_2 - 2k)}.$$

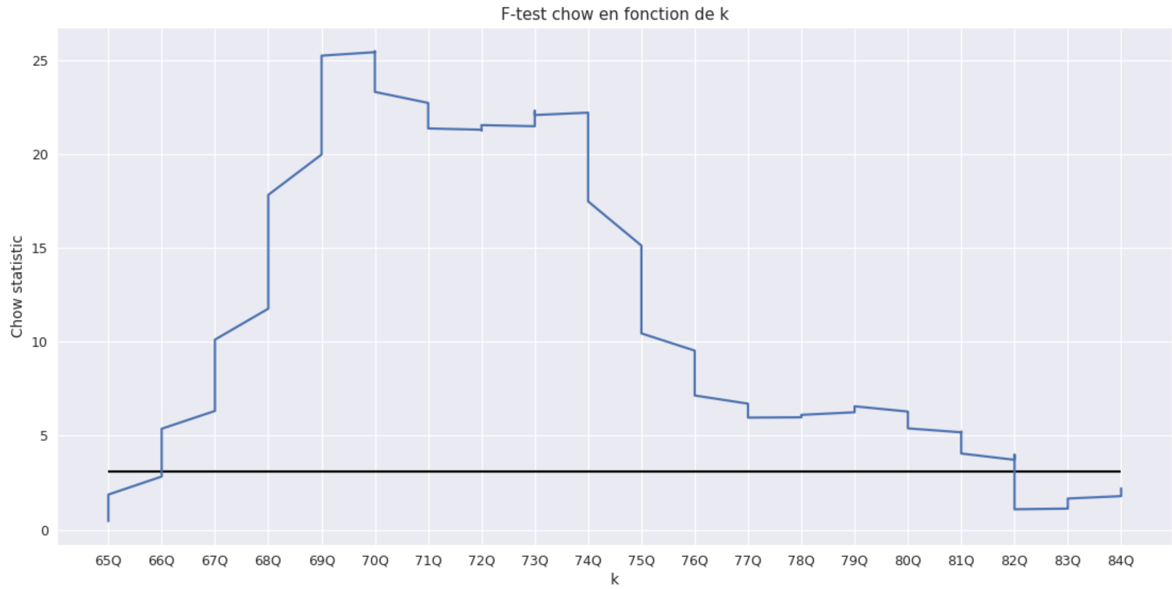
La statistique du test suit une loi de Fisher(v1,v2) avec ν1 = k et ν2 = N1 + N2 - 2k .
Pour la détermination de point de rupture, on a procédé comme suit :

Comme illustré dans le graphe ci-dessous, on choisit à chaque itération k une date de rupture Tk , on entraine un premier modèle avec l'échantillon des dates situées entre [T0 , Tk] , un deuxième modèle sur la période [Tk , Tf] et un troisième modèle global sur la période [T0,Tf] . On fait ensuite le test de Chow entre les trois périodes et on récupère la Statistique de test. Le point de rupture est la date ayant la statistique de test la plus grande.




```
[n,k]=size(inf);
tau0=floor(0.15*n);
tau1=floor(0.85*n);
all_chows_F= [];
all_chows_pv= [];
all_chows_t= [];
y_=inf(2:n);
y_lag=inf(1:n-1);
x_=inf(2:n);
x_lag=inf(1:n-1);
k=4;
X=[ones(n,1),inf];
for t=tau0:tau1
    all_chows_F(end+1) = stat;
    all_chows_pv(end+1) = pValue;
    all_chows_t(end+1) = t;
end
[QLR, tau]=max(all_chows_F)
all_chows_F(tau)
all_chows_t(tau)
% Q2 1970
```

Ci-dessous le graphe de la statistique de Chow en fonction de la date de rupture testée, le max correspond à la date **1970 Quarter 2**



La point de rupture est donc 1970 Quarter 2

1.11. Estimer la courbe de Philips en supprimant l'inflation courante des variables explicatives mais en ajoutant les délais d'ordre 1, 2, 3 et 4 de l'inflation et du chômage. Faire le test de Granger de non causalité de l'inflation sur le chômage. Donnez la p-valeur.

1. Modèle avec capture des effets dynamiques

```
y=unemp;
[n,k]=size(unemp);
X=[inf,unemp];
y=unemp(5:n);
X_lag= [X(4:n-1,:)];
X_lag2= [X(3:n-2,:)];
X_lag3= [X(2:n-3,:)];
X_lag4= [X(1:n-4,:)];
X=[X_lag X_lag2 X_lag3 X_lag4];
[n,k]=size(X);
beta=inv(X'*X)*X'*y ;
u=y-X*beta;
sig2=u'*u/(n-k) ;
std=sqrt(diag(sig2*inv(X'*X)));
t=beta./std;
sig2=u'*u/(n-k);
std=sqrt(diag(sig2*inv(X'*X)));
```

```
t=beta./std;
p_val = tdis_prb(t,n-k);
results=[beta std t p_val];
rowNames = {'inf(t-1)', 'unemp(t-1)','inf(t-2)', 'unemp(t-2)','inf(t-3)', 'unemp(t-3)','inf(t-4)', 'unemp(t-4)'};
colNames = {'beta','std','t','p_value'};
sTable =
array2table(results,'RowNames',rowNames,'VariableNames',colNames)
```

	beta	std	t	p_value
inf(t-1)	4.3245	3.7391	1.1566	0.24884
inf(t-2)	-2.3269	4.1203	-0.56473	0.57289
inf(t-3)	7.2787	4.0145	1.8131	0.071324
inf(t-4)	1.3123	3.7807	0.34709	0.72889
unemp(t-1)	1.6155	0.070915	22.781	0
unemp(t-2)	-0.6537	0.13493	-4.8448	2.5446e-06
unemp(t-3)	-0.0018916	0.13595	-0.013914	0.98891
unemp(t-4)	0.021504	0.0692	0.31076	0.75631

2. Test de Granger de non causalité de l'inflation sur le chômage

On cherche à tester :

- H0 : l'inflation n'a pas d'effet sur le chômage
- H1 : H0 est fausse

On va neutraliser l'inflation dans le modèle contraint…

```
SSR0=u'*u
X=X(:,[2,4,6,8])
beta1=inv(X'*X)*X'*y
u1=y-X*beta1
SSR1=u1'*u1
F=((SSR1-SSR0)/SSR0)*((n-k)/4)
p=fdis_prb(F,4,n-k)
```

La statistique du test suit une loi de Fisher à n-8 et 4 ddl. Sa valeur est 5.135 pour une probabilité critique de 5.87e-04. On rejettera H0, l'absence d'influence de l'inflation sur le chômage.

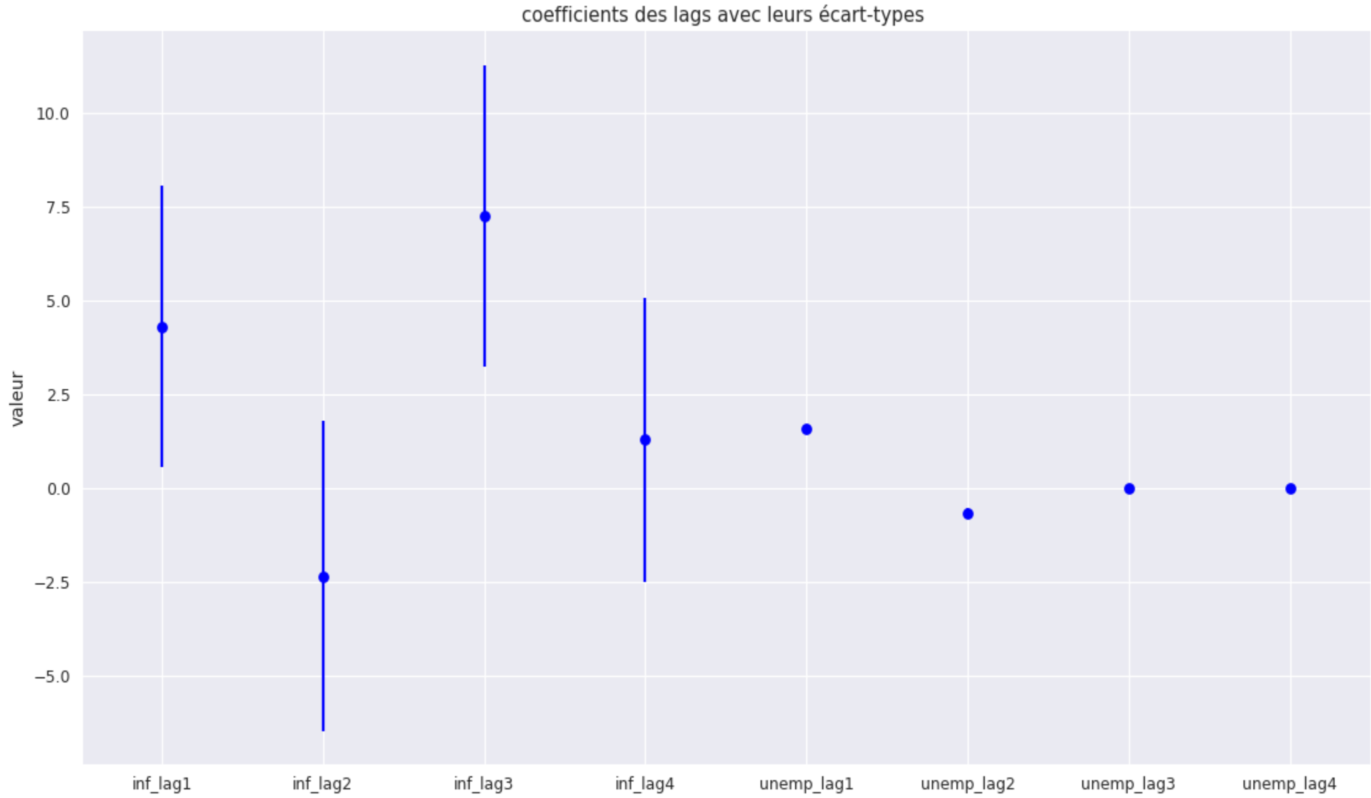
On a vérifié également que la somme des résidus et le critère AIC du modèle contraint sont plus élevés que le modèle non contraint, ce qui confirme aussi l'importance de l'inflation dans la modélisation du chômage

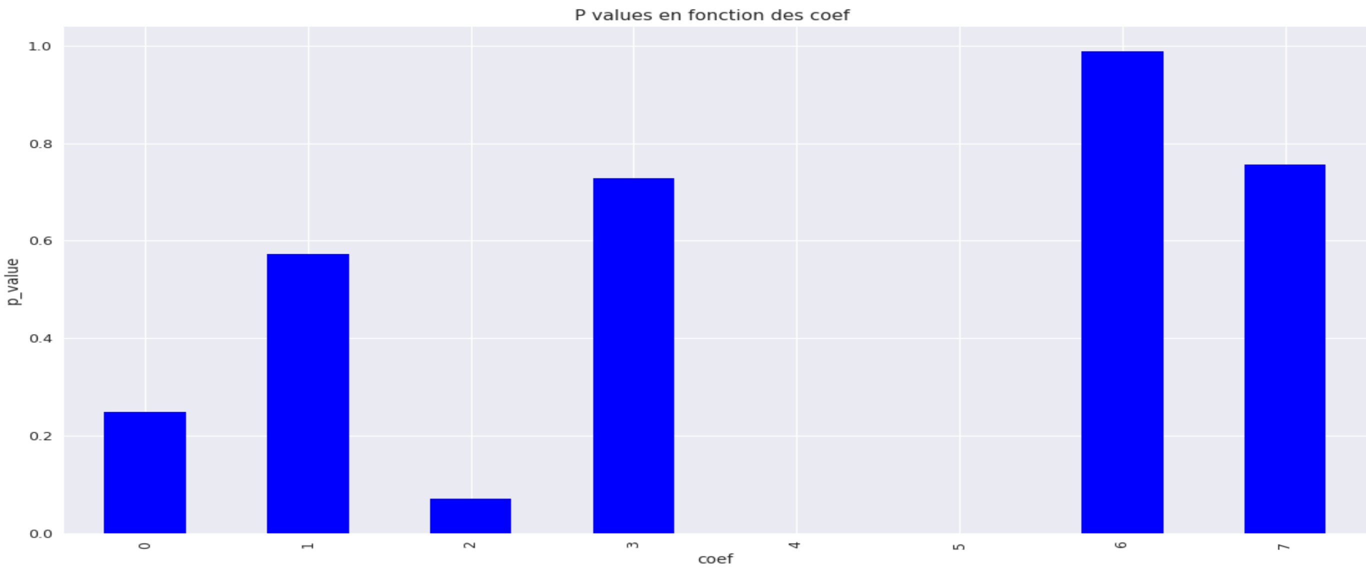
	Restricted	**Unrestricted**
SSR	12.476148	11.587384

	Restricted	**Unrestricted**
AIC	15.998228	8.70057

1.12. Représentez graphiquement les délais distribués et commentez. Calculer l'impact à long de terme de l'inflation sur le chômage.

On présente ci-dessous les coefficients des lags avec une bande représentant leurs écarts-types (graphe 1) puis la p-value de chaque coefficient (graphe 2)





A partir de ces deux graphes, on peut dire que, parmi les variables de l'inflation, c'est uniquement le lag 3 qui a une influence significative (à 10%) sur le chômage , mais pour décider si on peut supprimer l'effet des lags 1 , 2 et 4 de l'inflation , il faudra faire le test de rapport de vraisemblance :

- Modèle non contraint : modèle global avec tous les lags
- Modèle contraint : modèle global en supprimant le lag 1, lag 2 et lag 4 de l'inflation

« H0 : les deux modèles sont identiques »

Les résultats de test de rapport de vraisemblance sont les suivantes :

F-test= 0.663963 p-value = 0.575119

La p-value du test est très élevé (> 10%) , de ce fait on ne rejette pas l'hypothèse H0 et on peut enlever ces lags du modèle .

<u>Calcul de l'effet à long terme de l'inflation sur le chômage</u>

Pour calculer l'effet à long terme de l'inflation sur le chômage, il suffit de sommer les coefficients des lags de chômage estimés avec le modèle globale :

Effet_inflation = unemp_lag1+ unemp_lag2+ unemp_lag3+unemp_lag4

=4.32+-2.32+7.27+1.31

Effet_inflation = 10.58 %

Donc on peut dire que l'effet à long terme de l'inflation sur le chômage est 10.58% . Par contre cette formue ne tient pas compte de la significativité des coefficients. En effet, on a vu que c'est uniquement le lag 3 de l'inflation qui a une influence significative sur le chômage , de ce fait une façon plus robuste de calculer l'effet à long terme de l'inflation est juste de prendre le coefficient du lag_3 de l'inflation :

Effet_inflation_robuste = 7.27%