Projet SIM215 : Partie 1

Question 1

On supprime les valeurs égales à '.' et on vérifie par la suite que tous les salaires sont positifs avec le booléen wagepositif.
On obtient alors un tableau de 428 lignes et de 22 colonnes.

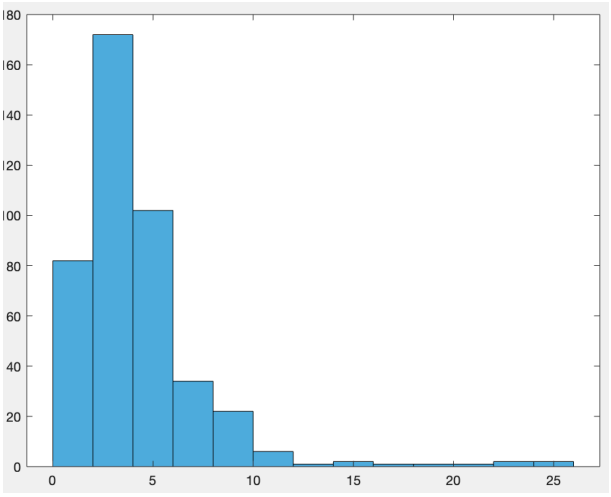
Question 2

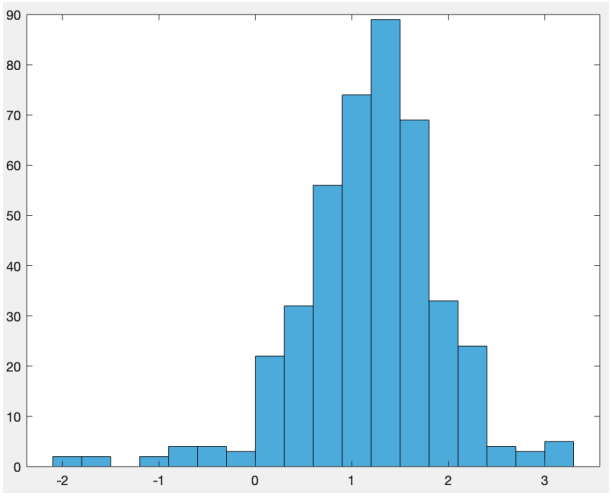
Après avoir séparé les différentes variables en sup et inf grâce à la médiane, on trouve les différentes statistiques descriptives qu'on place dans le tableau suivant :

	Mean	Max	Min	Std	Cov	Corrcoef avec huswage
Wage	4,1777	25	0,1282	3,3103	10,9580	0,0887
Wagesup	4,8968	25	0,1616	4,0416	16,3346	0,1023
Wageinf	3,4585	18,2670	0,1282	2,1433	4,5936	0,0169
Age	41,9720	60	30	7,7211	59,6151	0,2159
Agesup	42,2757	59	30	7,3888	54,5950	0,1576
Ageinf	41,6682	60	30	8,0455	64,7298	-0,0291
Educ	12,6589	17	5	2,2854	5,2229	0,3030
Educsup	13,2430	17	5	2,3590	5,5651	0,2617
Educinf	12,0748	17	6	2,0542	4,2197	-0,0127

Question 3

Histogramme de wage et lwage :





On remarque que wage et lwage suivent une loi gaussienne, les valeurs sont naturellement plus centrés (variance plus petite) dans le cas de lwage. En effet, le log agit comme un pourcentage, il sera plus commode par la suite de faire des études statistiques sur lwage plutôt que wage.

Question 4

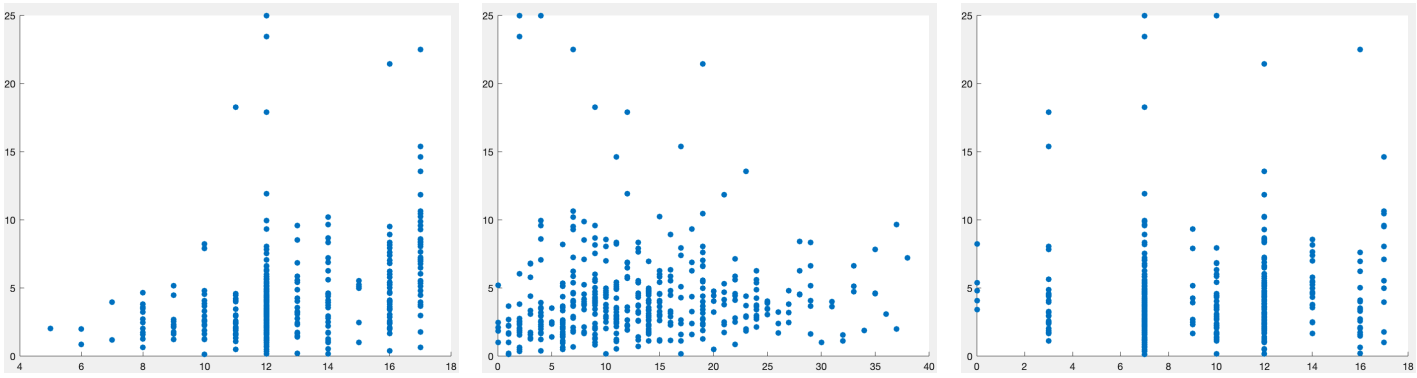
Matrice de corrélation entre motheduc et fatheduc :

```
1.0000    0.5541
0.5541    1.0000
```

Le salaire du père et de la mère sont fortement positivement corrélés (0.5541). En effet, deux personnes du même milieu social auront plus tendance à se marier que deux personnes de milieux sociaux différents. Il y aura sans doute un problème de multicollinéarité si on utilise ces deux variables explicatives. On utilisera donc par la suite seulement fatheduc.

Question 5

Nuages de points du salaire (ordonnée) en fonction de l'education, de l'expérience et du salaire du père (abscisse) de gauche à droite :



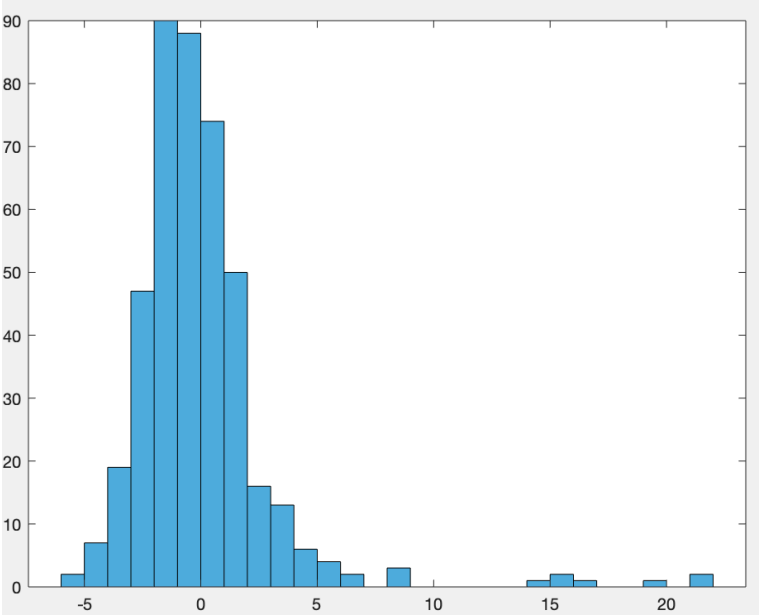
Pour le cas de l'expérience et du salaire du père, il est difficile de remarquer une répartition linéaire du salaire, on ne voit pas une corrélation claire entre wage d'une part et exper, fatheduc d'autre part. Il n'y a pas d'effet toutes choses égales par ailleurs.
Cependant dans le cas de l'éducation, il y a une relation linéaire entre salaire et éducation. Plus le nombre d'années d'éducation augmente, plus le salaire est élevé, il y a ainsi une corrélation claire entre wage et educ. Il s'agit d'un effet toutes choses égales par ailleurs.

Question 6

L'hypothèse fondamentale pour avoir des estimateurs non biaisés est que les résidus et les variables explicatives soient indépendants (u et X).
Le biais des variables omises est la situation où des variables non mesurées (voire mesurables) impactent la variable observée. Cela remet en cause l'hypothèse fondamentale précédente.

Question 7

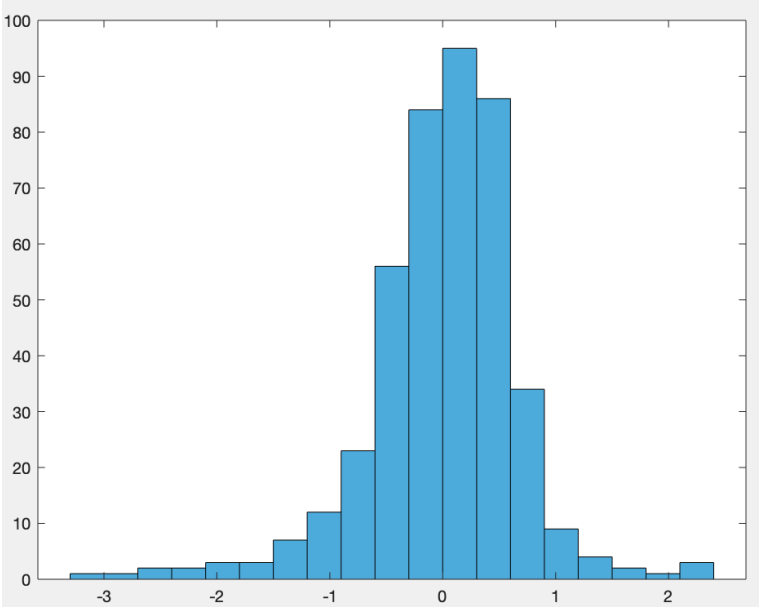
Histogramme des résidus après régression des moindres carrées de wage en fonction de différentes variables explicatives :



L'histogramme est centré autour de 0. Cependant on remarque certaines données marginales qui peuvent fausser notre modélisation (autour de 15/20).

Question 8

Histogramme des résidus après régression des moindres carrées de lwage en fonction de différentes variables explicatives :



L'histogramme est encore plus centré autour de 0 qu'avant (c'est normal, on regarde lwage au lieu de wage, la variance est plus petites). L'impact des données marginales est énormément minimisé, cela permettra de faire une bonne modélisation en utilisant lwage au lieu de wage.

Question 9

on accepte l'hypothèse de non significativité de nwifeinc à 1%.
on accepte l'hypothèse de non significativité de nwifeinc à 5%.
on accepte l'hypothèse de non significativité de nwifeinc à 10%.
p-value : 0.1434
La p-value confirme bien nos acceptations précédentes.

Question 10

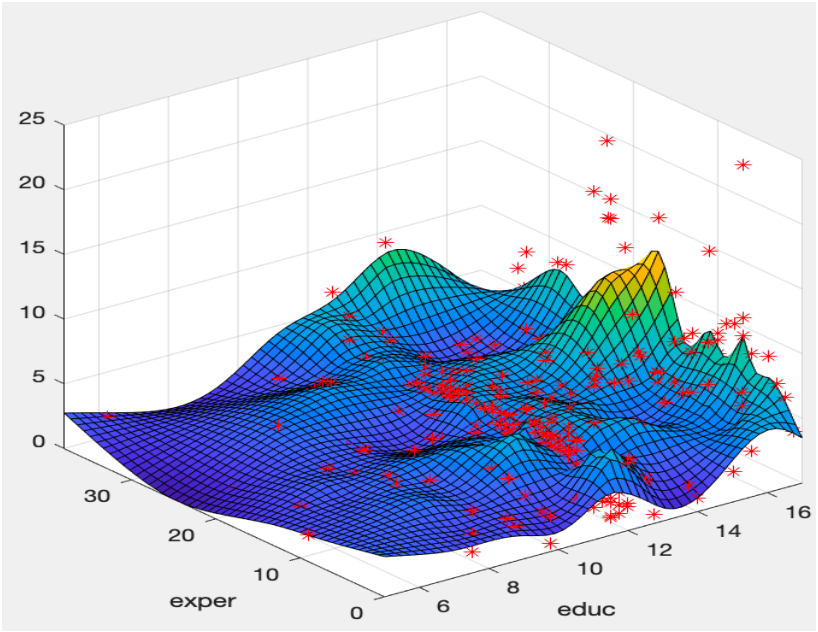
on accepte l'hypothèse que coefnwifeinc=0.01 à 5%.
p-value : 0.1252

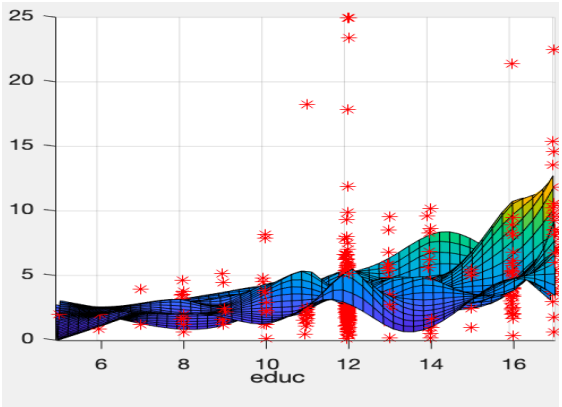
Question 11

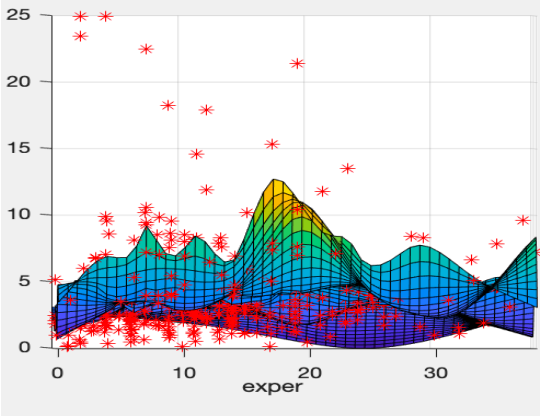
on accepte l hypothèse jointe que coefnwifeinc = 0.01 et que coefcity = 0.05 à 5%.
p-value : 0.2637

Question 12

Modélisation 2D du salaire en fonction de l'éducation et de l'expérience grâce à gridift :







Cela confirme bien que le salaire augmente si l'éducation augmente et qu'on ne peut pas dire grand chose quant au lien entre expérience et salaire.

Question 13

On accepte l hypothèse que kidsge6 = kidslt6 à 5%.
p-value : 0.6204
En effet, on imagine bien que l'âge de l'enfant a peu d'importance. C'est plus la présence ou non d'enfants qui importe et qui a du poids lors de la regression du salaire.

Question 14

Après test d'hétéroscédasticité de forme linéaire, on obtient la p-valeur suivante :
p-value : 0.1477
On corrige alors le problème via kidslt et on obtient alors la p-valeur suivante :
p-value : 0.4721
Dans le second cas, la p-valeur est bien plus importante. Notre modèle est bien plus acceptable qu'avant.

Question 15

1ère partie de la question :
p-value : 0.9687

2ème partie de la question :
P-values :

0.0405
0.0116
0.7949
0.8422
0.7853
0.5528
0.5025
0.0000
0.0087
0.3550
0.5528
0.3806
0.6026
0.0000
0.9011
0.0013
0.0507
0.9989
0.6794

Question 16

p-values :

1.0e-12 *

0.5873

0

0

Les p-valeurs sont extrêmement proches de 0.

Projet d'économétrie - SIM215

Partie 2. Séries temporelles

1. Il n'y a pas de valeurs manquante dans le fichier 'quarterly.xls'. On procède à sa lecture grâce à la commande :

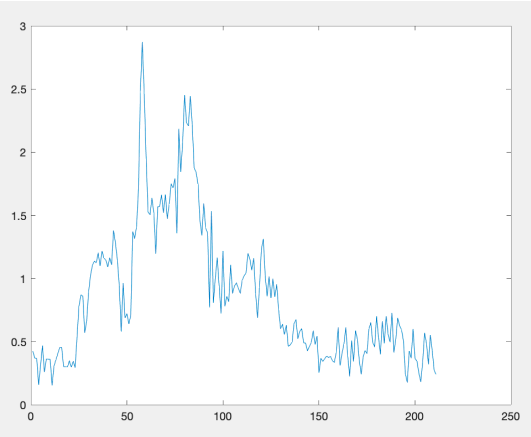
xlsread(quarterly.xls)

2. Taux d'inflation en fonction du temps. On note *inflation* le taux d'inflation en fonction du temps. Il est donné par :

$$inflation(t) = \frac{CPI(t) - CPI(t-1)}{CPI(t)}$$ où CPI est l'indice des prix à la consommation.

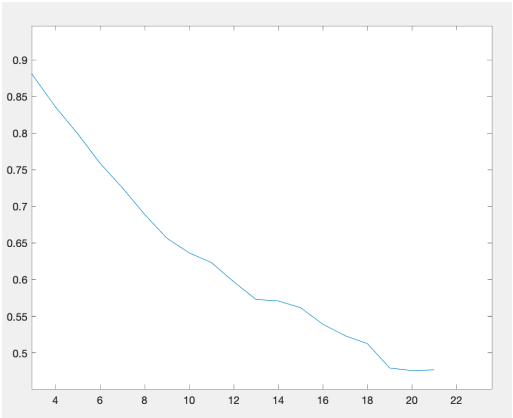
```
inflation = []
CPI = quarterly(:,9)
for i = 1:211
    inflation(i+1) = (CPI(i+1)-CPI(i))/CPI(i+1)
end
```

On obtient la distribution suivante en fonction du temps :

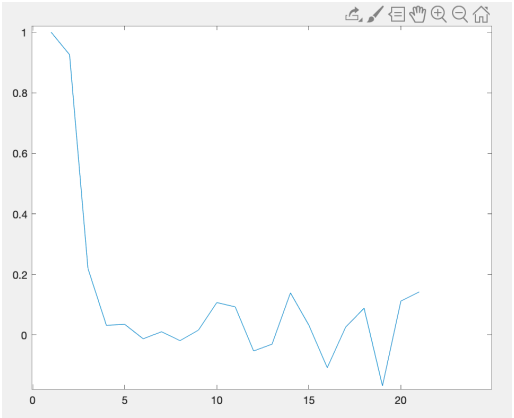


Il s'agit bien d'une série temporelle, mais sans motif répétitif.

3. Autocorrélogramme :



Autocorrélogramme partiel :



```
autocorrelogramme = autocorr(inflation)
```

```
partiel = parcorr(inflation)
```

D'après
l'autocorrélogramme partiel, il n'y a pas de lien entre l'inflation et la la durée, contrairement à ce que l'aucorrélogramme laisse à penser.

4. Un processus est stationnaire si le futur et le passé se ressemblent, ie si les statistiques sont invariantes dans le temps.
Un processus ergodique est un processus stochastique qui oublie les conditions initiales, ie pour lequel les statistiques peuvent être étudiées par un seul échantillon suffisamment long.
L'intérêt d'avoir un processus stationnaire et ergodique réside dans le fait d'avoir une invariance des statistiques dans le temps d'une part, et la possibilité de déduire les statistiques à partir d'une réalisation d'autre part. Si deux processus ne sont ni stationnaires ni ergodiques, alors on pourrait obtenir une fausse corrélation entre les deux, d'où le terme de « spurious ». On pourrait obtenir une régression montrant que les deux processus sont corrélés alors qu'en réalité ils ne le sont pas.

5. On choisit une modélisation auto-régressive d'ordre 3 pour l'inflation, car c'est la valeur renvoyée par le critère d'information bayésien.

$$inflation(t) = \phi_0 + \phi_1 inflation(t-1) + \phi_2 inflation(t-2) + \phi_3 inflation(t-3) + \epsilon_t$$

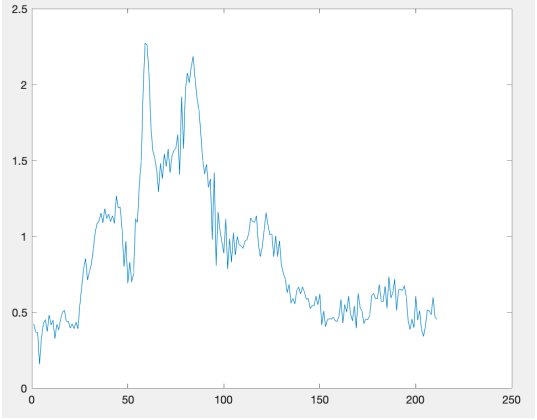
```
ar(inflation, 3)
```

On obtient les valeurs suivantes pour les coefficients du filtre :

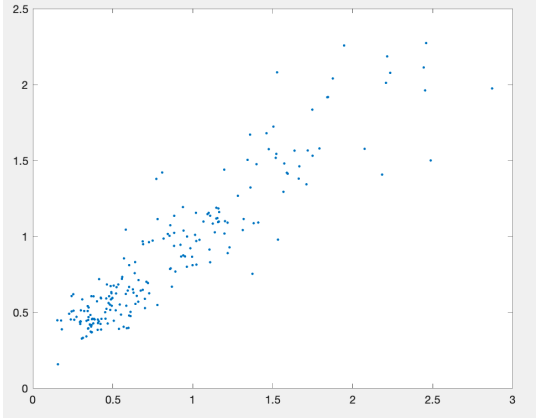
ϕ_0	0.1365
ϕ_1	0.5827
ϕ_2	-0.0184
ϕ_3	0.2978

On trace ensuite d'une part la modélisation, d'autre part la corrélation avec l'inflation.

Modélisation AR(p=3) de l'inflation :

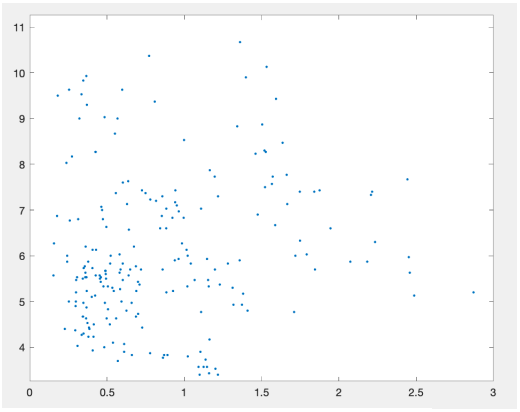


Corrélation avec inflation :

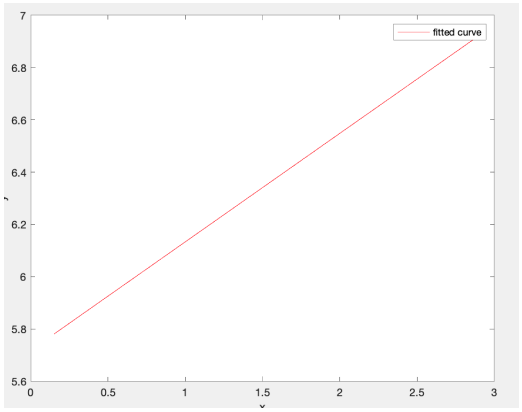


La corrélation étant linéaire, la modélisation est donc valable.

6. Courbe de Philipps :

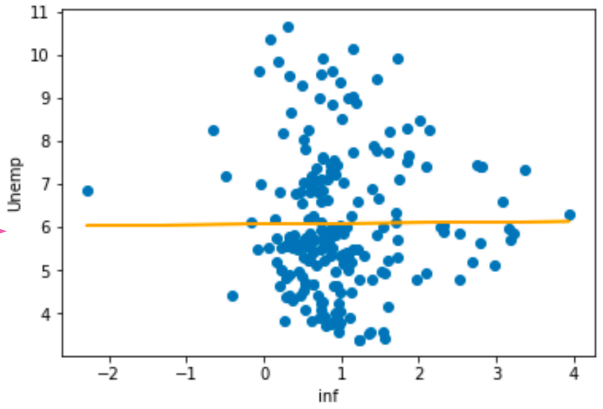


```
phillips = plot(inflation, unemp, '.')
```




```
[fit2, gof2, fitinfo2] = fit(inflation.', unemp, 'a+b*x')
```

```
const = ones(n,1);

y = unemp;
X = [const,inflation];
[n,k]=size(X);
beta=inv(X'*X)*X'*y;
u=y-X*beta;

x = inf;
y = const*beta(1)+inf*beta(2);
figure;
hold on
plot(x,y)
scatter(inflation,unemp);
title('courbe de Philips');
hold off
xlabel('Inflation');
ylabel('Unemployement');
```



On obtient les coefficients suivants :

```
>> [fit2, gof2, fitinfo2] = fit(inflation.', unemp, 'a+b*x')
Warning: Start point not provided, choosing random start point.
> In curvefit.attention/Warning/throw (line 30)
  In fit>iFit (line 307)
  In fit (line 116)

fit2 =

     General model:
     fit2(x) = a+b*x
     Coefficients (with 95% confidence bounds):
       a =       5.718  (5.315, 6.121)
       b =      0.4152  (0.01967, 0.8108)

gof2 =

  struct with fields:

          sse: 535.8724
      rsquare: 0.0201
          dfe: 209
   adjrsquare: 0.0154
         rmse: 1.6012


fitinfo2 =

  struct with fields:

          numobs: 211
        numparam: 2
       residuals: [211×1 double]
        Jacobian: [211×2 double]
        exitflag: 1
   firstorderopt: 3.6180e-07
      iterations: 1
       funcCount: 6
    cgiterations: 0
       algorithm: 'trust-region-reflective'
        stepsize: 4.8206
         message: 'Success. Fitting converged to a solution.'
```

7. Pour tester l'autocorrélation des erreurs, on effectue le test de Durbin-Watson.
La statistique de Durbin-Watson et l'estimateur sont donnés par :

$$DW = \frac{\sum_{k=0}^{n} (\epsilon_k - \epsilon_{k-1})^2}{\sum_{k=0}^{n} \epsilon_k^2} \quad \text{et} \quad \hat{\rho} = \frac{\sum_{k=0}^{n} \epsilon_k \epsilon_{k-1}}{\sum_{k=0}^{n} \epsilon_k^2}$$

Le test statistique est donné par :

$$H_0 : \hat{\rho} = 0$$

On obtient : $\hat{\rho} = 0.9797$.

Ainsi, puisque $\hat{\rho}$ est proche de 1, DW tend lui vers 0, ce qui implique une très forte autocorrélation positive dans les résidus.
Conclusion : on accepte l'hypothèse H_0 au seuil de 5%.

8. Pour corriger l'autocorrélation des erreurs, on a recours à la méthode des moindres carrés généralisés.
On obtient les coefficients suivants :

```
>> coefficients = fgls(inflation., unemp, 'display', 'final')

OLS Estimates:

       |  Coeff    SE
------------------------
 Const | 5.7178  0.2043
 x1    | 0.4152  0.2007

FGLS Estimates:

       |  Coeff     SE
-------------------------
 Const |  6.3448  0.9099
 x1    | -0.1685  0.1095

coefficients =

    6.3448
   -0.1685
```

9. On teste la stabilité de la relation inflation-chômage sur deux sous périodes : p1 = 1:105 et p2 = 106:211.

```
>> coeff_p1 = fgls(inflation_p1., unemp_p1, 'display', 'final')

OLS Estimates:

       |  Coeff    SE
------------------------
 Const | 5.3370  0.3362
 x1    | 0.6865  0.2599

FGLS Estimates:

       |  Coeff     SE
-------------------------
 Const |  6.2620  0.8838
 x1    | -0.1646  0.1435

coeff_p1 =

    6.2620
   -0.1646
```

```
>> coeff_p2 = fgls(inflation_p2., unemp_p2, 'display', 'final')

OLS Estimates:

       |  Coeff     SE
-------------------------
 Const |  6.3582  0.3676
 x1    | -0.5584  0.5769

FGLS Estimates:

       |  Coeff     SE
-------------------------
 Const |  6.9851  1.2282
 x1    | -0.1622  0.1903

coeff_p2 =

    6.9851
   -0.1622
```

On calcule ensuite les valeurs pour SSR0, SSR1, SSR2.

```
%%Question 9
inflation_p1 = inflation(1:105);
inflation_p2 = inflation(106:211);

unemp_p1 = unemp(1:105);
unemp_p2 = unemp(106:211);

coeff_p1 = fgls(inflation_p1.', unemp_p1, 'display', 'final');
coeff_p2 = fgls(inflation_p2.', unemp_p2, 'display', 'final');

X_1 = [ones(105, 1), unemp_p1];
[n,k] = size(X_1);
beta = inv(X_1'*X_1)*X_1'*inflation_p1;
u = inflation_p1 - X_1*beta;
SSR0=u'*u;


X_2 = [ones(106, 1), unemp_p2];
[n,k] = size(X_2);
beta = inv(X_2.'*X_2)*X_2.'*inflation_p2;
v = inflation_p2 - X_2*beta;
SSR1=v'*v;


X = [ones(212, 1), unemp];
[n,k] = size(X_2);
beta = inv(X.'*X)*X.'*inflation;
w = unemp - X*beta;
SSR2=w'*w;
```

On en déduit la valeur de F :

$$F = \frac{SSR0 - (SSR1 + SSR2)/k}{(SSR1 + SSR2)(n - 2k)}.$$

```
[n,k] = size(inf);
k = 2;
F = ((SSR0-(SSR1+SSR2))/k)/((SSR1+SSR2)*(n-2*k));
fprintf('The F value is : %d \n',F);

pval=fdis_prb(F,2,n-2*k);
fprintf('The p-value is : %d \n',pval);
```

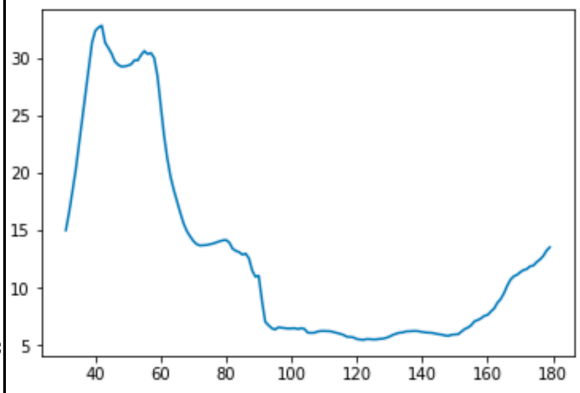
SSR0	549.9226
SSR1	303.5958
SSR2	245.9327
F	1.7237.10^-6
pvaleur	0.999998

Les valeurs sont proches, ie le test est vérifié donc la relation est stable.

10. On effectue les tests de changement de structure de Chow puisque $SSR0 > SSR1 + SSR2$. On obtient le booléen 0 qui signifie que le test est faux. On code une boucle while afin de récupérer la première valeur pour laquelle le test affiche 1. On obtient $bp = 42$.
On trouve Chow = 1, indiquant que les coefficients sont stables.

```
  i=0
while chow = 0
    i = = i +1
    [chow, pValue, stat, cValue] = chowtest(inflation.', unemp, i)
    disp(i)
return i
```

```
%Tracé de la courbe des maximums des F
[n,k] = size(inflation);
n_min = fix(0.15*n);
n_max = fix((1-0.15)*n);
F = [];
List = [];
for i = n_min:1:n_max+1
   const = ones(n,1);
   X = [const];
   X_top = X(1:i);
   X_bot = X(i:212);
   Unemp_top = unemp(1:i) ;
   Unemp_bot = unemp(i:212) ;
   beta_top = inv(X_top'*X_top)*X_top'*Unemp_top;
   beta_bot = inv(X_bot'*X_bot)*X_bot'*Unemp_bot;
   u_bot = Unemp_bot - X_bot*beta_bot ;
   u_top = Unemp_top - X_top*beta_top ;
   SSR_top =u_top'*u_top ;
   SSR_bot =u_bot'*u_bot ;
   Fn = ((SSR0-(SSR_top+SSR_bot))/k)/((SSR_top+SSR_bot)/(n-2*2));
   F = [F Fn];
   List = [List i];
end
figure;
hold on
plot(List, F);
hold off
Top = List(find(F==max(F)))
```



Maximum des F

11. On calcule les différents délais d'ordre 1 à 4 de l'inflation et du chômage, puis on effectue une régression de Unemp en fonction des variables explicatives.

Inf2	0.0830
Inf3	-0.1043
Inf4	-0.0587
Inf5	0.1282
Unemp2	1.6986
Unemp3	-0.6869
Unemp4	-0.1180
Unemp5	0.0954

```
%%Question 11
[n,k]=size(data);
unemp1 = unemp(1:n-5);
unemp2 = unemp(2:n-4);
unemp3 = unemp(3:n-3);
unemp4 = unemp(4:n-2);
unemp5 = unemp(5:n-1);

inf1 = inflation(1:n-5);
inf2 = inflation(2:n-4);
inf3 = inflation(3:n-3);
inf4 = inflation(4:n-2);
inf5 = inflation(5:n-1);

y = unemp1;
X = [inf2,inf3,inf4,inf5,unemp2,unemp3,unemp4,unemp5];
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;
```

Enfin, on effectue le test de Granger :

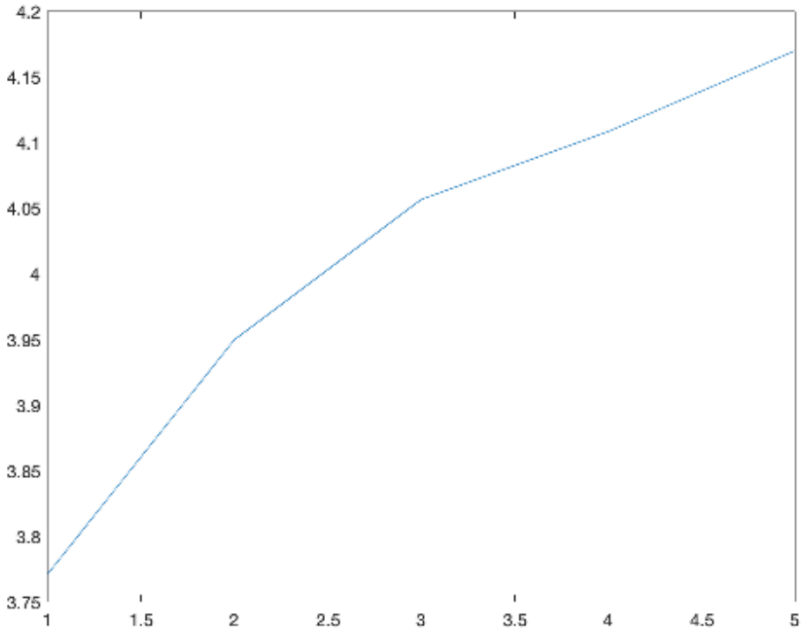
```
granger_cause(inflation, unemp, 0.05, 212);
```

On obtient une p-valeur de 0.0054 donc on peut rejeter l'hypothèse de causalité.

Conclusion : le chômage n'est pas la cause de l'inflation.

12. On représente les délais distribués graphiquement :

```
%%Question 12
inf1 = inflation(1:n-5);
inf2 = inflation(2:n-4);
inf3 = inflation(3:n-3);
inf4 = inflation(4:n-2);
inf5 = inflation(5:n-1);

unemp1 = unemp(1:n-5);

%Cas 1
y = unemp1;
X = [inf1];
beta=inv(X'*X)*X'*y;
One = sum(beta);
fprintf('The 1st value is : %d \n',One);

%Cas 2
y = unemp1;
X = [inf1,inf2];
beta=inv(X'*X)*X'*y;
Two = sum(beta);
fprintf('The 2nd value is : %d \n',Two);

%Cas 3
y = unemp1;
X = [inf1,inf2,inf3];
beta=inv(X'*X)*X'*y;
Three = sum(beta);
fprintf('The 3rd value is : %d \n',Three);

%Cas 4
y = unemp1;
X = [inf1,inf2,inf3,inf4];
beta=inv(X'*X)*X'*y;
Four = sum(beta);
fprintf('The 4th value is : %d \n',Four);

%Cas 5
y = unemp1;
X = [inf1,inf2,inf3,inf4,inf5];
beta=inv(X'*X)*X'*y;
Five = sum(beta);
fprintf('The 5th value is : %d \n',Five);

figure;
y = [One,Two,Three,Four,Five];
plot(y);
```



Plus on rajoute de délais, plus le poids est fort, ie, plus il y a d'information sur le passé, plus le poids est fort.

13. On retrouve la même p-valeur que précédemment. Donc le chômage n'est pas la Granger cause de l'inflation.

```
%%Question 13
unemp1 = unemp(1:n-5);
unemp2 = unemp(2:n-4);
unemp3 = unemp(3:n-3);
unemp4 = unemp(4:n-2);
unemp5 = unemp(5:n-1);

inf1 = inflation(1:n-5);
inf2 = inflation(2:n-4);
inf3 = inflation(3:n-3);
inf4 = inflation(4:n-2);
inf5 = inflation(5:n-1);

y = inf1;
X = [inf2,inf3,inf4,inf5,unemp2,unemp3,unemp4,unemp5];
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;
```

Inf2	0.5114
Inf3	-0.0265
Inf4	0.3309
Inf5	-0.0016
Unemp2	-0.1492
Unemp3	0.1099
Unemp4	-0.2420
Unemp5	0.3116