

# Projet d'économétrie

## Partie 1. Régression

1. Le dataset comprend 22 variables :

1. inlf	=1 if in labor force, 1975
2. hours	hours worked, 1975
3. kidslt6	# kids < 6 years
4. kidsge6	# kids 6-18
5. age	woman's age in yrs
6. educ	years of schooling
7. wage	estimated wage from earns., hours
8. repwage	reported wage at interview in 1976
9. hushrs	hours worked by husband, 1975
10. husage	husband's age
11. huseduc	husband's years of schooling
12. huswage	husband's hourly wage, 1975
13. faminc	family income, 1975
14. mtr	fed. marginal tax rate facing woman
15. motheduc	mother's years of schooling
16. fatheduc	father's years of schooling
17. unem	unem. rate in county of resid.
18. city	=1 if live in SMSA
19. exper	actual labor mkt exper
20. nwifeinc	(faminc - wage*hours)/1000
21. lwage	log(wage)
22. expersq	exper^2

On compte initialement 753 observations, réduites à 428 observations après avoir supprimé les observations pour lesquelles la variable wage est inférieure ou égale à 0 :

```
number of variables in the mroz dataset: 22
number of observations in the initial dataset: 753
number of observations after trimming non-positive wage values: 428
```

2. On calcule les statistiques descriptives pour trois groupes : toutes les femmes, femmes dont le revenu du mari est au-dessus du 65ème percentile, et femmes dont le revenu du mari est en-dessous du 65ème percentile :

descriptive statistics for all women:

" "	"age"	"educ"	"wage"
"count"	"428"	"428"	"428"
"mean"	"41.972"	"12.6589"	"4.17768"
"median"	"42"	"12"	"3.4819"
"std"	"7.7211"	"2.2854"	"3.3103"
"min"	"30"	"5"	"0.1282"
"25%"	"35"	"12"	"2.2584"
"50%"	"42"	"12"	"3.4819"
"75%"	"47.5"	"14"	"4.9726"
"max"	"60"	"17"	"25"

descriptive statistics for women whose husband's income is more than 65% of the sample:

" "	"age"	"educ"	"wage"
"count"	"428"	"428"	"428"
"mean"	"42.6776"	"13.4539"	"5.12859"
"median"	"43"	"12"	"4.03955"
"std"	"7.3373"	"2.3553"	"4.2957"
"min"	"30"	"5"	"0.2137"
"25%"	"36"	"12"	"2.5685"
"50%"	"43"	"12"	"4.03955"
"75%"	"48"	"16"	"6.36375"
"max"	"59"	"17"	"25"

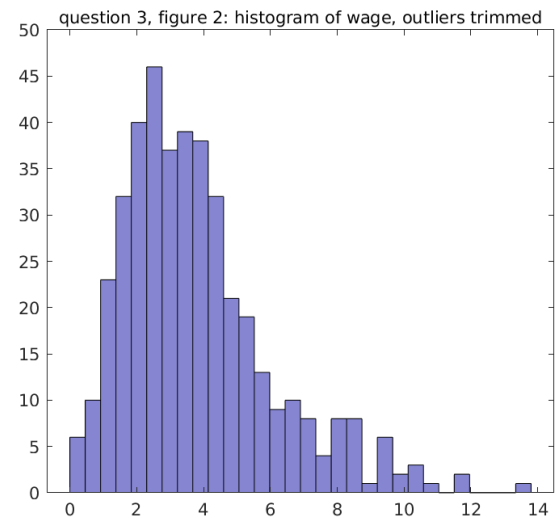
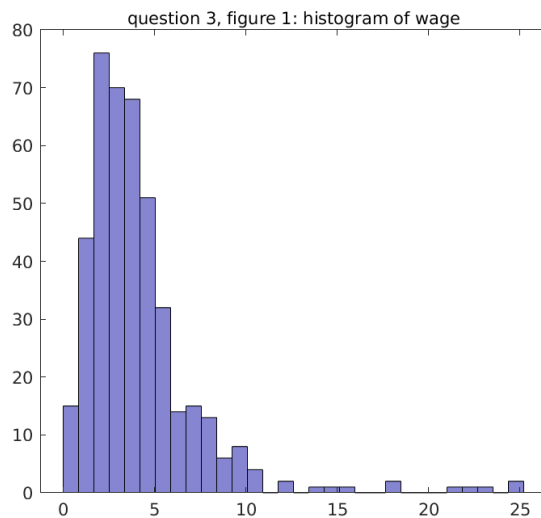
descriptive statistics for women whose husband's income is less than 65% of the sample:

" "	"age"	"educ"	"wage"
"count"	"428"	"428"	"428"
"mean"	"41.5833"	"12.221"	"3.654"
"median"	"41"	"12"	"3.1697"
"std"	"7.9107"	"2.1265"	"2.4713"
"min"	"30"	"6"	"0.1282"
"25%"	"35"	"12"	"2.1353"
"50%"	"41"	"12"	"3.1697"
"75%"	"47"	"12.5"	"4.51755"
"max"	"60"	"17"	"22.5"

La principale conclusion est que l'âge, le niveau d'éducation et de revenu de l'épouse sont corrélés à ceux du mari. Les femmes dont le revenu du mari est inférieur au 65ème percentile sont en moyenne plus jeunes et ont un niveau d'éducation et de revenu moindre que celui du groupe total. Inversement, Les femmes dont le revenu du mari est supérieur au 65ème percentile sont en moyenne plus âgées et ont un niveau d'éducation et de revenu supérieur à celui du groupe total. La conclusion est robuste à l'utilisation de la médiane au lieu de la moyenne, et l'écart-type apparaît constant entre les groupes.

D'un point de vue interprétatif, on conclut surtout que les mariages se font entre profils similaires. Les individus tendent à épouser un partenaire d'âge, de niveau d'éducation et de revenu similaires au leur. D'un point de vue économétrique, on remarquera que les variables age, éducation et salaire des femmes sont corrélées à la variable salaire du mari, entraînant de potentiels problèmes de colinéarité lors de l'estimation.

3. On observe que le salaire est positivement asymétrique (positively skewed), la distribution étant d'avantage concentrée sur les valeurs inférieures à la moyenne. On observe aussi que la série comprend un certain nombre de valeurs extrêmes (outliers), consistant en salaire supérieurs à 15 dollars. On peut supprimer des outliers pour obtenir une série plus représentative. On note toutefois que l'asymétrie demeure.

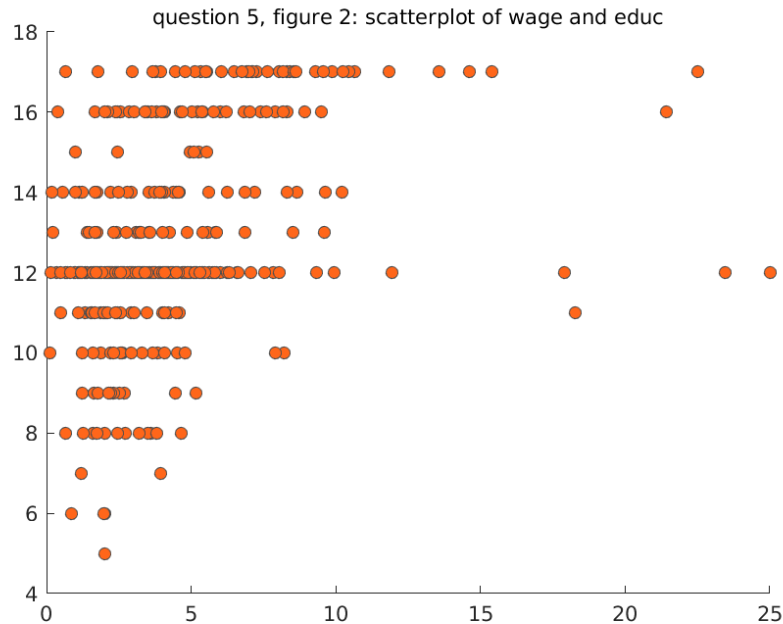


4. On obtient les résultats suivants :

```
correlation coefficient between motheduc and fatheduc: 0.55406
```

Le coefficient de corrélation entre motheduc et fatheduc est de 0,55 environ. Le problème de multicollinéarité se pose lorsque le coefficient de corrélation entre deux variables approche de 1. Les variables en question sont alors quasi linéairement dépendantes, ce qui amène le déterminant de la matrice  $X'X$  à approcher 0. Il devient alors difficile d'inverser la matrice  $X'X$ , et l'estimateur MCO est estimé de manière très imprécise. La variance des estimateurs devient alors très large, et en conséquence il en va de même des intervalles de confiance. Ici la valeur de la corrélation est de 0.55, on n'a donc pas à proprement parler de colinéarité entre motheduc et fatheduc. Il faut toutefois garder à l'esprit qu'en raison de la corrélation significative qui existe entre ces deux variables, le modèle peut avoir plus de difficultés à séparer leurs effets respectifs, et que la précision des coefficients se trouvera amoindrie.

5. Le graphique suggère une corrélation entre wage et educ. L'effet observé n'est toutefois pas un effet *ceteris paribus* (« toute chose égale par ailleurs ») en raison de possibles variables non observées. Par exemple, le salaire est typiquement déterminé au moins en partie par les aptitudes intellectuelles innées d'une personne. Or ces mêmes aptitudes déterminent aussi le niveau d'éducation, les personnes les plus douées étant naturellement enclines à poursuivre leurs études, quand les personnes moins douées les arrêteront de manière précoce. La corrélation observée entre wage et educ omet donc la variables aptitude intellectuelle innées, et n'est pas *ceteris paribus*.



6. L'hypothèse fondamentale qui garantit des estimateurs non biaisés est l'hypothèse qui veut que la variable non observée  $u$  (le résidu) soit de moyenne nulle, et surtout qu'elle soit non corrélée avec les régresseurs  $X$ . De ce fait, la moyenne conditionnelle est égale à la moyenne inconditionnelle :

$$E(u | X) = E(u) = 0$$

Si l'on omet une variable dans le modèle, et que cette variable possède un pouvoir explicatif, alors l'estimateur des moindres carrés sera biaisé. Admettons par exemple que l'on analyse le taux de réussite à un examen national uniquement en fonction du taux d'attribution de bourses scolaires. On omet dans ce cas la variable revenu, qui va alors faire partie du terme d'erreur. Or cette variable est clairement corrélée à l'attribution d'une bourse scolaire (donc  $E(u|X) \neq 0$ ), et elle affecte également la réussite aux examens, les ménages plus pauvres ayant un capital éducatif plus faible. L'estimateur MCO ne rendra pas alors l'effet correct de l'attribution d'une bourse, puisque le terme d'erreur inclura une variable explicative. On peut aussi montrer ce biais de façon algébrique :

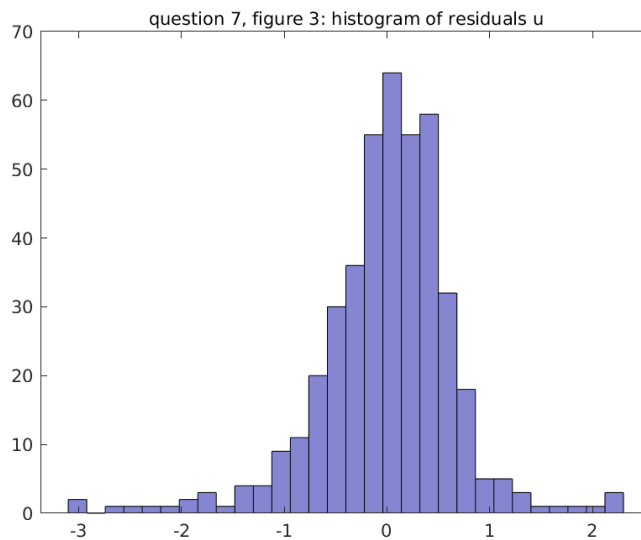
$$\begin{aligned} E(\beta|X) &= E((X'X)^{-1}X'y|X) = E((X'X)^{-1}X'(X\beta + u)|X) = E((X'X)^{-1}(X'X)\beta|X) + E((X'X)^{-1}X'u|X) \\ &= \beta + (X'X)^{-1}X'E(u | X) \end{aligned}$$

On peut constater que si  $E(u|X) \neq 0$ , un terme de biais apparaît à droite de l'espérance.

7. L'estimateur de  $\beta$  et les écarts-type associés sont donnés par :

beta estimates:		coefficient standard deviations:	
constant:	-0.39898	constant:	0.20705
city:	0.035268	city:	0.070178
educ:	0.10225	educ:	0.015102
exper:	0.015488	exper:	0.004487
nwifeinc:	0.0048827	nwifeinc:	0.0033307
kidslt6:	-0.045303	kidslt6:	0.085308
kidsgt6:	-0.011704	kidsgt6:	0.026944

L'histogramme des résidus du modèle est donné par :



On remarque que l'histogramme semble centré en 0. Sa forme globale rappelle la forme en cloche caractéristique de la distribution normale, donc postuler une distribution normale centrée en 0 pour les résidus semble raisonnable. Toutefois la distribution semble légèrement (négativement) asymétrique. On pourrait donc s'intéresser à cette possible asymétrie et tester formellement la normalité de la distribution avec un test de type Jarque-Berra.

8. Les valeurs des statistiques de student et le p-values associées aux coefficients du modèle sont données par :

t-stats for the two-sided t-test:

```
constant: 1.9269
city: 0.50255
educ: 6.7706
exper: 3.4517
nwifeinc: 1.466
kidslt6: 0.53105
kidsgt6: 0.43436
critical values at 10%, 5% and 1% are respectively 1.6485, 1.9656, and 2.5876
```

p-values for the two-sided t-test:

```
constant: 0.054659
city: 0.61555
educ: 4.3245e-11
exper: 0.00061337
nwifeinc: 0.14341
kidslt6: 0.59566
kidsgt6: 0.66425
```

Pour la variable `nwifeinc`, on pose  $H_0 : \beta_{nwifeinc} = 0$  et  $H_1 : \beta_{nwifeinc} \neq 0$ . La valeur de la statistique de student est de 1.466. Comparée aux valeurs critiques de 1.648, 1.965 et 2.587 à 10 %, 5 % et 1 % respectivement, on conclut qu'on ne peut pas rejeter l'hypothèse nulle, quel que soit le seuil de significativité. On obtient logiquement la même conclusion en regardant la p-value de la variable `nwifeinc` (p-value de 0.143 supérieure à 10 %, 5 % et 1 %). On peut donc conclure que la variable `nwifeinc` n'est pas utile pour prédire la variable `wage`. Cela est logique puisque `nwifeinc` (non-wife

income) représente les revenus autres que ceux de l'épouse. Or typiquement, on peut attendre que le salaire de l'épouse dépende de facteurs personnels à l'épouse comme son expérience ou son niveau d'éducation. Logiquement, les variables indépendantes de l'épouse n'affectent pas son niveau de salaire.

En regardant les autres p-values, on constate que les variables city, kidslt6 et kidsgt6 ne sont pas significatives non plus, même à 10 %. La constante est significative à 6 %. Les variables educ et exper sont en revanche significatives même à un seuil de 1 %: elles contribuent donc certainement à déterminer le niveau de salaire de l'épouse.

9. La statistique de student et la p-value associées au test sur la variable nwifeinc sont données par :

```
t-stat for the two-sided t-test:    1.5364
p-value for the two-sided t-test:    0.12519
```

Pour la variable nwifeinc, on pose maintenant  $H_0 : \beta_{nwifeinc} = 0.01$  contre  $H_1 : \beta_{nwifeinc} \neq 0.01$ . La valeur de la statistique de student est de 1.536. Comparée à une valeur critique à 5 % de 1.965, on ne rejette pas l'hypothèse nulle. De même, la p-value de 0.125 indique qu'on ne rejette pas l'hypothèse nulle à un seuil de 5 %.

10. Pour tester une hypothèse jointe, on utilise un test de Fisher basé sur la statistique :

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)} \sim F(q, n-k-1)$$

On peut montrer (voir Johnston et Dinardo, Econometrics Methods, 4th edition, p97) que cette statistique est équivalente à :

$$F = \frac{(R\beta - r)' [R(X'X)^{-1}R']^{-1} (R\beta - r)}{u'u/(n-k-1)} \sim F(q, n-k-1)$$

où on exprime la contrainte testée par le système linéaire  $R\beta = r$ . On teste donc l'hypothèse jointe  $H_0 : R\beta = r$  contre l'alternative  $H_1 : R\beta \neq r$ . Pour tester l'hypothèse jointe  $\beta_{nwifeinc} = 0.01$  et  $\beta_{city} = 0.05$ , on définit :

$$R = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad \text{et} \quad r = \begin{pmatrix} 0.01 \\ 0.05 \end{pmatrix}$$

On obtient alors les résultats suivants :

```
F-stat for the joint hypothesis:    1.3371
critical value at 5%:              3.0172
p-value for the joint hypothesis:    0.26373
```

La statistique (1.337) est inférieure à la valeur critique (3.017). De même, la p-value (0.263) est supérieure à 5 %. On conclut qu'on ne rejette pas l'hypothèse nulle.

11. On procède identiquement à la question 10. L'hypothèse jointe à traiter implique :

$$R = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 \end{pmatrix} \quad \text{et} \quad r = \begin{pmatrix} 0.1 \\ 0.1 \end{pmatrix}$$

On obtient alors les résultats suivants :

F-stat for the joint hypothesis: 0.92271  
critical value at 5%: 3.0172  
p-value for the joint hypothesis: 0.39824

La statistique (0.922) est inférieure à la valeur critique (3.017). De même, la p-value (0.398) est supérieure à 5 %. On conclut qu'on ne rejette pas l'hypothèse nulle.

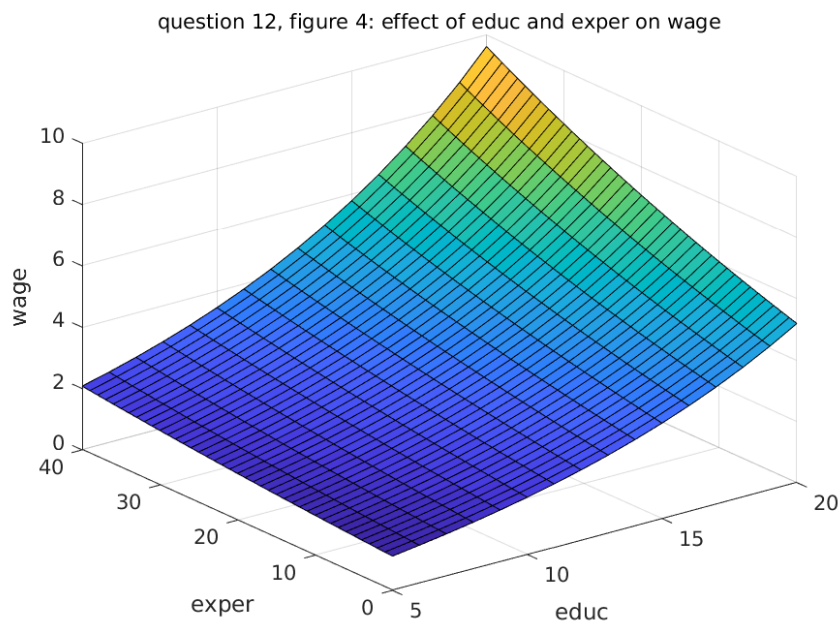
12. On note ici que l'on a estimé notre modèle non pas sur wage, mais sur  $\log(\text{wage})$ , ce qui induit une non-linéarité de l'effet. En particulier, on peut montrer aisément qu'avec la formulation log-linéaire adoptée, on obtient:

$$\begin{aligned} \% \Delta \text{ wage} &\approx (100 \cdot \beta_{\text{educ}}) \Delta \text{educ} \\ \% \Delta \text{ wage} &\approx (100 \cdot \beta_{\text{exper}}) \Delta \text{exper} \end{aligned}$$

Pour obtenir la valeur de la variable wage, on utilise simplement (omettant l'effet des variables autres que educ et exper):

$$\text{wage} = \exp(\log(\text{wage})) = \exp(\text{constant} + \beta_{\text{educ}} \text{educ} + \beta_{\text{exper}} \text{exper})$$

On obtient alors la représentation graphique suivante :



On observe une surface convexe et non un plan, ce qui montre bien l'effet non-linéaire des variables educ et exper. L'effet exponentiel implique que plus les variables educ et exper prennent une valeur large, plus la bénéfice marginal en termes de salaire est important.

13. On procède encore une fois de façon similaire à la question 10. L'hypothèse jointe à traiter implique :

$$R = (000001 - 1) \quad \text{et} \quad r = (0)$$

On obtient alors les résultats suivants :

```
F-stat for the joint hypothesis:    0.1382
critical value at 5%:              3.8636
p-value for the joint hypothesis:    0.71027
```

La statistique (0.1382) est inférieure à la valeur critique (3.8636). De même, la p-value (0.710) est supérieure à 5 %. On conclut qu'on ne rejette pas l'hypothèse nulle.

14. On procède au test d'hétéroscédasticité de forme linéaire basé sur le test de Fisher. On utilise pour cela la régression :

$$u^2 = \delta_0 + \delta_1 x_1 + \delta_2 x_2 + \delta_3 x_3 + \delta_4 x_4 + \delta_5 x_5 + \delta_6 x_6 + v$$

Pour le test de Fisher, on suit encore une fois la méthodologie de la question 10. La statistique de Fisher est donnée par :

$$F = \frac{(R\delta - r)'[R(X'X)^{-1}R']^{-1}(R\delta - r)}{v'v/(n-k-1)} \sim F(q, n-k-1)$$

avec:

$$R = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \quad \text{et} \quad r = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}$$

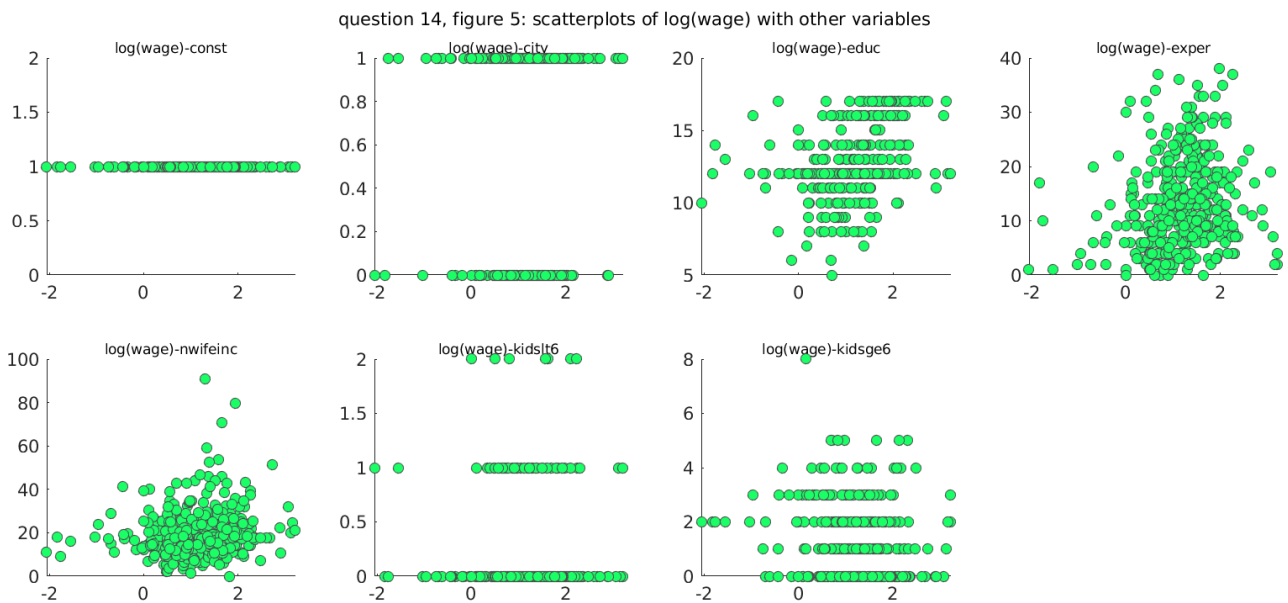
On obtient alors les résultats suivants :

```
F-stat for linear heteroskedasticity:    2.0087
critical value at 5%:                    2.1201
p-value for the joint hypothesis:         0.063341
```

Avec une p-value de 0.063, on ne rejette pas l'hypothèse nulle d'homoscédasticité à 5 %. Bien que le test ne passe pas strictement à 5 %, il rejette l'hypothèse nulle à un seuil très proche (7%), ce qui suggère la présence d'hétéroscédasticité dans ce modèle.

On cherche maintenant à identifier les potentielles sources d'hétéroscédasticité du modèle. Pour cela, le plus simple est de tracer un nuage de points entre la variable expliquée  $\log(\text{wage})$  et chaque variable explicative du modèle.





On fait deux constats. Premièrement, les variables educ, exper et nwifeinc semblent affecter non linéairement la variable log(salaire). Deuxièmement, les variables city, kidslt6 et kidsge6 sont entrées en tant que variables quantitatives, alors qu'elles seraient plus adaptées sous une formulation de variables binaires. On identifie ainsi deux sources possibles d'hétéroscédasticité.

Sur la base des ces conclusions, on peut proposer deux stratégies de correction :

- une solution simple, qui consiste à passer les variables educ, exper et nwifeinc en log afin d'obtenir un modèle log-log (qui généralement ne souffre pas d'hétéroscédasticité), et les variables city, kidslt6 et kidsge6 en variables binaires. On prendra l'estimateur MCO traditionnel sur ces nouvelles variables, et on vérifiera si l'hétéroscédasticité a disparu.
- une solution plus sophistiquée, qui consiste à utiliser un estimateur GLS avec les variables d'origine. On utilisera cet estimateur pour comparer les écarts-type des coefficients estimés avec ceux obtenus à la question 7.

On applique d'abord la première stratégie. Après transformations des variables en log et variables binaires, on obtient les estimateurs suivants :

beta estimates for the transformed model:		coefficient standard deviations:	
constant:	-2.6082	constant:	0.57889
log(educ):	1.1968	log(educ):	0.18388
log(exper):	0.2105	log(exper):	0.052214
log(nwifeinc):	0.12127	log(nwifeinc):	0.070011
city:	0.0090407	city:	0.07137
kidslt6_0:	0.10456	kidslt6_0:	0.25668
kidslt6_1:	0.10829	kidslt6_1:	0.27401
kidsgt6_0:	-0.16938	kidsgt6_0:	0.24836
kidsgt6_1:	-0.12033	kidsgt6_1:	0.24962
kidsgt6_2:	-0.26965	kidsgt6_2:	0.24932
kidsgt6_3:	-0.32712	kidsgt6_3:	0.25359
kidsgt6_4:	0.047768	kidsgt6_4:	0.28986

On opère ensuite le test d'hétéroscédasticité linéaire, en suivant la même méthode que précédemment :

F-stat for linear heteroskedasticity:	1.7206
critical value at 5%:	1.8117
p-value for the joint hypothesis:	0.066539

On note que le passage au modèle en log-log n'a pas amélioré le problème d'hétéroscédasticité. La p-value du test d'hétéroscédasticité linéaire est virtuellement similaire à celle du modèle initial, et suggère toujours la présence d'hétéroscédasticité.

On étudie donc maintenant la seconde solution : l'estimateur GLS. Si la forme de l'hétéroscédasticité était connue, on pourrait utiliser la méthode des moindres carrés pondérés. Ici, il est plus prudent de partir du principe que la forme de l'hétéroscédasticité est inconnue. Dans ce cas, on peut postuler une forme flexible :

$$\text{var}(u|X) = \sigma^2 \exp(\delta_0 + \delta_1 x_1 + \delta_2 x_2 + \dots + \delta_k x_k)$$

Cela équivaut à estimer la régression :

$$\log(u^2) = a_0 + \delta_1 x_1 + \delta_2 x_2 + \dots + \delta_k x_k + e$$

On prend ensuite les prédictions  $g$  de ce modèle, et calcule les poids  $h = \exp(g)$ . On peut alors utiliser l'estimateur des moindres carrés pondérés. On obtient les résultats suivants :

beta estimates with GLS:		coefficient	standard deviations:
constant:	-0.36569	constant:	0.18747
city:	0.019564	city:	0.064052
educ:	0.10145	educ:	0.013841
exper:	0.012917	exper:	0.0039044
nwifeinc:	0.0066668	nwifeinc:	0.0030834
kidslt6:	-0.053652	kidslt6:	0.095448
kidsgt6:	-0.019801	kidsgt6:	0.028285

Pour rappel, les valeurs du modèle de base sont :

betahat estimates:		coefficient	standard deviations:
constant:	-0.39898	constant:	0.20705
city:	0.035268	city:	0.070178
educ:	0.10225	educ:	0.015102
exper:	0.015488	exper:	0.004487
nwifeinc:	0.0048827	nwifeinc:	0.0033307
kidslt6:	-0.045303	kidslt6:	0.085308
kidsgt6:	-0.011704	kidsgt6:	0.026944

Sur les sept coefficients du modèle, on remarque cinq ont un écart-type plus petit avec le modèle GLS. Cela est attendu puisque l'estimateur GLS donne moins de poids aux observations avec une variance de résidus plus grande. On obtient alors des estimateurs plus précis. Les deux coefficients dont la variance est légèrement augmentée (kidslt6 et kidsgt6) ne sont pas significatifs (voir question 7) et n'ont pas de pouvoir explicatif sur  $\log(\text{wage})$ . Il est donc possible qu'ils gagnent en variance, puisqu'ils captent principalement du bruit.

15. On teste un possible changement de structure entre les femmes de plus de 43 ans et les autres. Pour cela, on crée une variable binaire qui prend une valeur de 1 pour les femmes de plus de 43 ans, et on crée une variable d'interactions pour chaque variable explicative du modèle. On procède ensuite à un test de Fisher sur les coefficients des variables d'interactions. On obtient alors les résultats suivants :

beta estimates with structural change:		coefficient p-values:	
constant:	-0.54118	constant:	0.082231
city:	0.084998	city:	0.34674
educ:	0.11673	educ:	3.8853e-07
exper:	0.020163	exper:	0.0096529
nwifeinc:	0.0025589	nwifeinc:	0.59031
kidslt6:	-0.093295	kidslt6:	0.30946
kidsgt6:	-0.023323	kidsgt6:	0.51104
constant_d:	0.21798	constant_d:	0.60523
city_d:	-0.10384	city_d:	0.4753
educ_d:	-0.03146	educ_d:	0.30703
exper_d:	-0.0023257	exper_d:	0.81318
nwifeinc_d:	0.0067106	nwifeinc_d:	0.32564
kidslt6_d:	0.084172	kidslt6_d:	0.81268
kidsgt6_d:	-0.0015603	kidsgt6_d:	0.98045

On procède ensuite à un test de Fisher pour tester l'hypothèse jointe de nullité des coefficients d'interaction . On obtient les résultats suivants :

F-stat for structural change:	0.82604
critical value at 5%:	2.0317
p-value for the joint hypothesis:	0.56604

On tire deux conclusions. Premièrement, les p-values des coefficients individuels sur les variables d'interactions sont toutes très élevées, impliquant qu'aucun de ces coefficients ne semble significatif à titre individuel. On peut donc déjà douter de la présence d'un changement structurel. Toutefois, seul le test de Fisher permet d'établir formellement la présence d'un changement. On constate que la statistique (0.826) est très inférieure à la valeur critique (2.0317) : on ne rejette donc pas l'hypothèse nulle d'absence de rupture. La p-value (0.56>0.05) confirme ce résultat. On conclut donc qu'il n'y a pas de changement structurel lié à l'âge des femmes.

16. On intègre maintenant la variable huseduc. Au modèle de la question 7. Celle-ci représente le nombre d'années d'éducation du mari.

beta estimates with huseduc:		coefficient p-values:	
constant:	-0.34941	constant:	0.096795
city:	0.048886	city:	0.49026
educ:	0.11549	educ:	2.9844e-10
exper:	0.015174	exper:	0.00079008
nwifeinc:	0.0057784	nwifeinc:	0.089034
kidslt6:	-0.040547	kidslt6:	0.63473
kidsgt6:	-0.011601	kidsgt6:	0.66667
huseduc:	-0.018999	huseduc:	0.16922

On note un effet négatif de la variable huseduc sur le salaire de l'épouse. Il n'y a pas d'explication évidente à ce résultat. On peut éventuellement envisager un effet substitution : un mari avec un haut niveau d'étude occupera un poste à haut niveau de compétence et d'engagement, ce qui pousserait par compensation l'épouse à freiner ses propres ambitions pour assumer la gestion du ménage, la conduisant à un emploi moins payé. L'effet n'est toutefois pas significatif (p-value de 0.169).

On cherche maintenant à déterminer si la variable huseduc peut devenir plus significative en la décomposant en 4 variables binaires. Le minimum de la série est 4, le maximum 17. Si l'on part du principe que les données sont américaines, alors on peut décomposer la série en 4 niveaux d'éducation : 4-8 (eighth grade ou moins, soit fin du collège en France), 9-12 (twelfth grade ou moins, soit fin du lycée), 13-15 (bachelor, soit niveau licence), et 16 et plus, ce qui correspond à un

niveau Master ou plus. On crée ces variables binaires en omettant la dernière catégorie pour éviter la multi-colinéarité (on maintient la constante dans la régression). On obtient alors les résultats suivants :

beta estimates with huseduc:	coefficient	p-values:
constant:	-0.81512	0.0049719
city:	0.044898	0.52648
educ:	0.12027	4.3191e-11
exper:	0.015861	0.00049106
nwifeinc:	0.007105	0.039655
kidslt6:	-0.025624	0.76434
kidsgt6:	-0.012531	0.64282
huseduc_1:	0.14238	0.29719
huseduc_2:	0.22686	0.019459
huseduc_3:	0.068939	0.52407

On procède ensuite à un test de Fisher pour tester l'hypothèse jointe de nullité des coefficients d'éducation du mari . On obtient les résultats suivants :

F-stat for structural change:	2.0424
critical value at 95%:	2.6262
p-value for the joint hypothesis:	0.10732

On tire les conclusions suivantes. Premièrement, les coefficients liés à l'éducation du mari ne sont toujours pas significatifs, sauf pour huseduc\_2 qui correspond à la variable binaire de l'éducation à la fin du lycée. On note également que l'impact de la variable huseduc cesse d'être négatif, et devient positif dans tous les cas. On peut alors interpréter l'effet significatif de huseduc\_2 comme un effet de sélection sociale. Il existe une différence fondamentale de profil entre les femmes qui épousent un homme n'ayant pas atteint la fin du lycée, et une femme qui épouse un homme ayant fait des études supérieures. Ces dernières appartiennent généralement elles-mêmes à une classe sociale plus élevée, ont un meilleur niveau d'étude, ce qui se reflète dans le salaire.

Deuxièmement, le test de Fisher ne parvient pas à rejeter l'hypothèse jointe de nullité des coefficients d'éducation du mari. La valeur du test (2.042) est inférieure à la valeur critique au seuil de 5 % (2.626). De même, la p-value ne rejette pas l'hypothèse nulle au seuil de 5 % ( $0.107 > 0.05$ ). Elle ne parvient pas non plus à rejeter l'hypothèse à un seuil de 10 %, ce qui suggère clairement que l'impact de l'éducation du mari sur le salaire de l'épouse n'est pas significatif. En cela, on rejoint le premier modèle de la question 16, qui ne trouvait pas d'effet significatif de huseduc dans son ensemble.

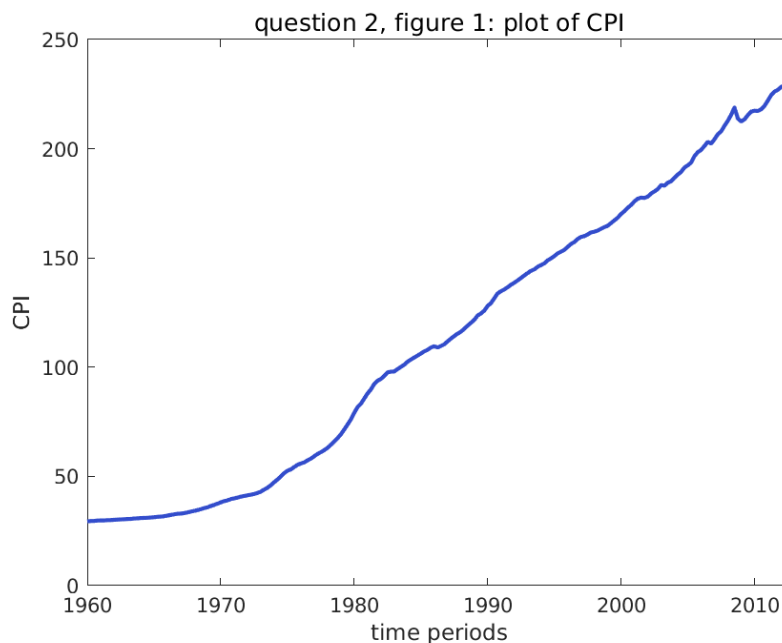
## Partie 2. Séries temporelles

### 1. Un regard préliminaire sur les données montre :

```
number of variables in the quarterly dataset: 18  
number of observations in the quarterly dataset: 212  
number of missing observations: 0  
There are no missing observations, hence no need for pre-treatment of the dataset.
```

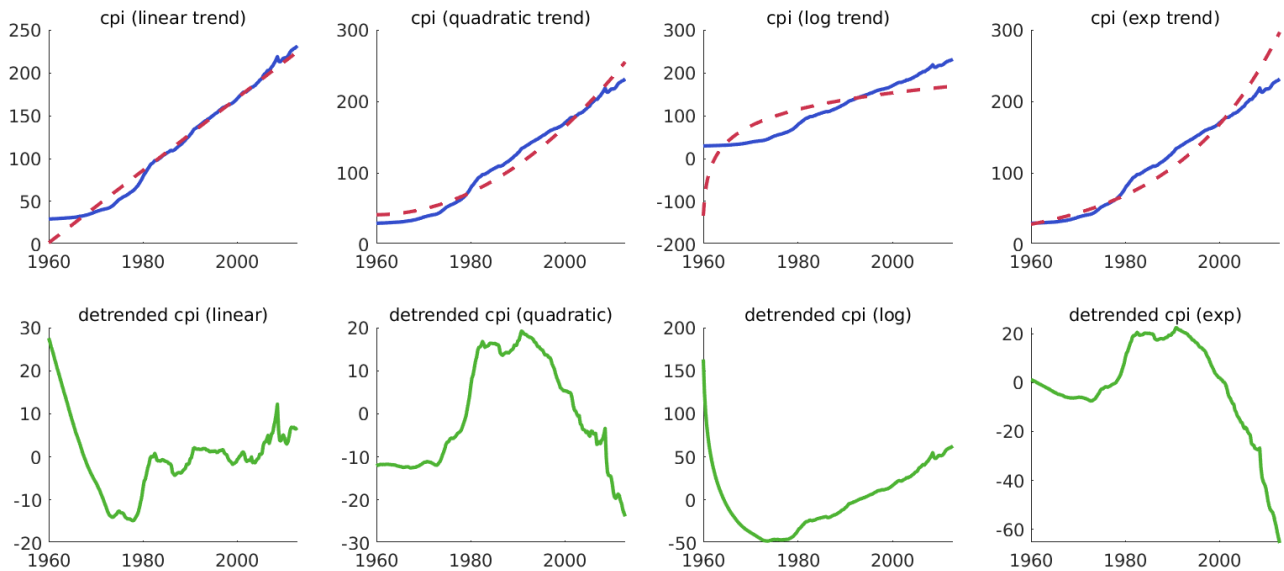
L'ensemble de données ne contient pas de valeurs manquantes. Il n'y a donc pas besoin de pré-traitement pour corriger ce problème.

### 2. On commence par observer la série :



La tendance semble quadratique sur la période 1960-1980, mais plutôt linéaire sur la période 1980-2012. Il n'est donc pas évident de déterminer la meilleure forme fonctionnelle. On va donc tester quatre spécifications possible des tendances. Pour cela, on crée une série de périodes, puis on la passe en carré et logarithme. Pour la tendance exponentielle, on passe la série `cpi` en `log` pour ensuite régresser `log_cpi` `log` sur une tendance linéaire. On procède donc à un ajustement sur quatre tendances possibles. Les résultats sont les suivants :

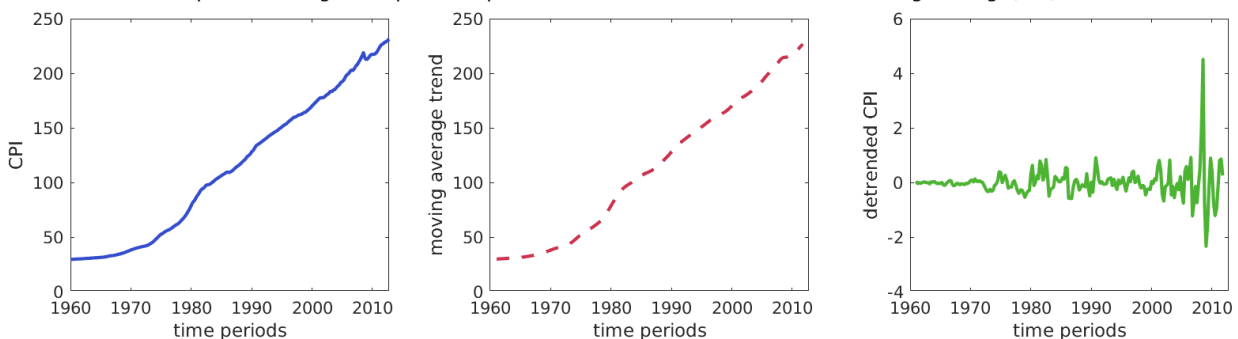
question 2, figure 2: plots of different trends for the cpi series



Sur le graphique, les courbes bleues représentent la série de données originale, la série rouge pointillée la tendance, et la série verte la série une fois la tendance enlevée. On observe que la tendance linéaire produit de mauvais résultats sur la première période de l'échantillon (1960-1980). La tendance logarithmique est clairement inadaptée. Les tendances quadratiques et exponentielles correspondent mieux à la forme des données, et de ces deux tendances, c'est la tendance quadratique qui semble la plus proche de la série. On retient donc la tendance quadratique, ce qui produit la série verte sur la second ligne du graphique. Celle-ci est stationnaire, mais son comportement semble erratique. On pourra donc envisager une autre approche pour stationnariser la série.

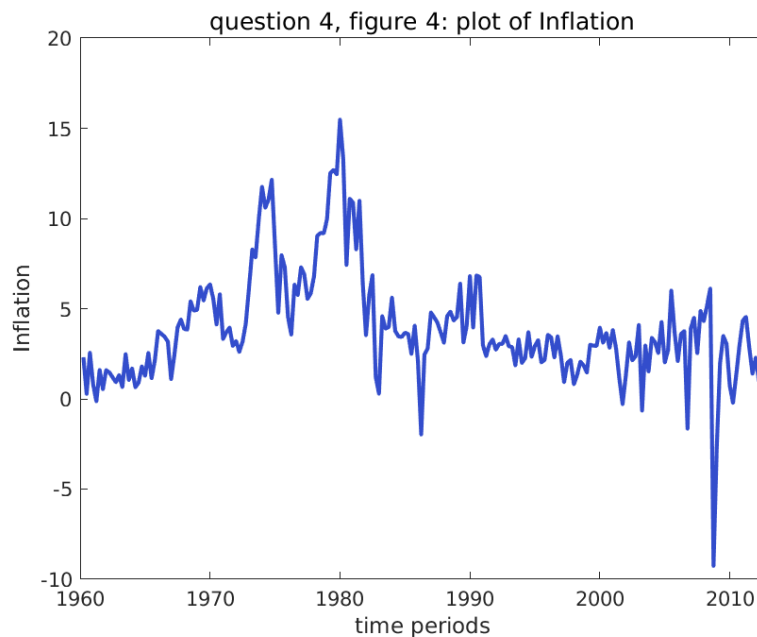
3. On applique à la série CPI une moyenne mobile centrée 5×5 pour la stationnariser. Notons qu'avec ce processus, on perd 8 observations en tout (4 à la fin et 4 au début). On obtient alors :

question 3, figure 3: plots of cpi: actual, trend and detrended with moving average(5,5)



On tire deux conclusions. Tout d'abord, la tendance obtenue par moyenne mobile (en rouge) est proche de la série d'origine. Le lissage a toutefois permis d'éliminer quelques cycles faibles (comme le creux de 2008 suite à la crise). Ensuite, on note l'impact de ce type de tendance en lissage sur la série obtenue après stationnarisation (en vert). Comme la série d'origine a très peu de cycles en début de période, on note que la série stationnarisée connaît peu de fluctuations sur cette même période. A l'inverse, la fin de la période (2007-2012) correspond à la crise, ce qui laisse apparaître des cycles marqués sur la fin de l'échantillon.

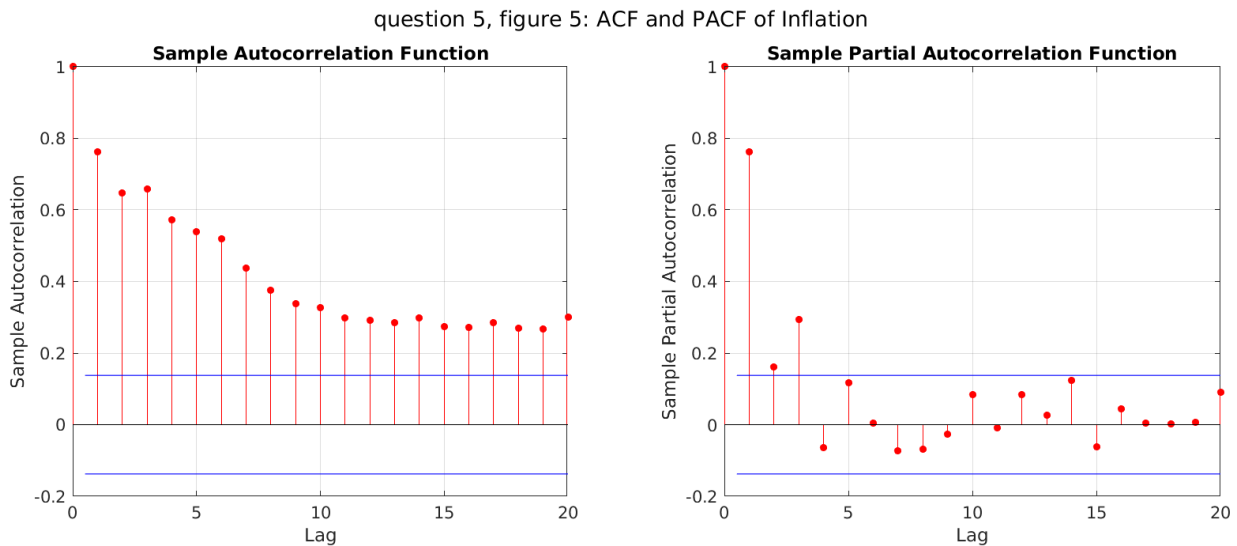
4. Pour calculer le taux d'inflation à partir du CPI, on passe la variable en logarithme, puis on prend la différence première (la différence du log étant à peu près égale au taux de croissance). On obtient alors un taux d'inflation trimestriel, qu'on peut optionnellement multiplier par 4 pour obtenir un taux annualisé. On retiendra ici cette option. On obtient alors :



On note que l'inflation a connu des évolutions très différentes au cours du temps. Ces périodes correspondent étroitement aux gouverneurs de la FED, ainsi qu'aux événements économiques marquants de l'histoire des Etats-Unis. On distingue :

- la Grande Inflation (Great Inflation) de 1965-1982, caractérisée par une forte hausse du niveau de prix. Cette inflation soutenue est due à la politique de la FED qui a mené sur la période une politique monétaire expansionniste avec l'objectif d'exploiter la courbe de Phillips pour stabiliser l'activité économique.
- une période de décélération progressive de l'inflation sur la période 1982-2007. Cette décélération est due à un changement de politique de la FED, qui suite à la Grande Inflation a adopté une cible d'inflation explicite (Inflation Targetting). On note une période de chute rapide jusqu'en 1987 sous le mandat de Paul Volcker, puis une stabilisation relative sous les mandats d'Alan Greenspan. Cette période de stabilisation macroéconomique est connue sous le nom de Grande Modération (Great Moderation).
- une période de plus forte volatilité à partir de 2007, avec un fort creux en 2008 qui correspond au plus fort de la crise financière débutée en 2007. Cette période correspond au mandat de Ben Bernanke, et la FED durant cette période a essentiellement cherché à concilier le contrôle de l'inflation avec le retour à un niveau d'activité économique optimal.

5. Les fonctions d'autocorrélation et d'autocorrélation partielles de la série Inf dont données par :



La fonction d'autocorrélation indique la corrélation brute qui existe dans une série entre la période  $t$  et la période  $t+h$ . Formellement, elle est définie comme :

$$\rho = \frac{\gamma(h)}{\gamma(0)} \quad , \quad \gamma(h) = \text{Cov}(X_{t+h}, X_t) = E[(X_{t+h} - \mu)(X_t - \mu)]$$

La fonction d'autocorrélation partielle indique la corrélation qui existe dans une série entre la période  $t$  et la période  $t+h$ , une fois enlevé l'effet des corrélations des périodes  $t+1, \dots, t+h-1$ . Formellement, elle correspond au coefficient  $\phi_{kk}$  de la régression :

$$X_t = \phi_{1k} X_{t-1} + \phi_{2k} X_{t-2} + \dots + \phi_{kk} X_{t-k} + \text{error}_t$$

6. La stationnarité est liée à l'idée de distribution des données. Une série temporelle  $\{x_t\}$  est stationnaire au sens stricte si la distribution jointe de  $\{x_t, x_{t+1}, \dots, x_T\}$  est la même que celle de  $\{x_{t+h}, x_{t+1+h}, \dots, x_{T+h}\}$ . Autrement dit, on peut décaler la série de  $h$  période, et la distribution de la série se trouvera inchangée. Derrière la stationnarité se trouve donc l'idée d'invariance de la distribution au cours des périodes.

Un processus est stationnaire au sens faible (ou en covariance) si pour tout  $t > 1$ ,  $E(X_t)$  est constante et  $E(X_t^2) < \infty$ ,  $\text{var}(X_t)$  est constante,  $\text{cov}(X_t, X_{t+h})$  est constante et ne dépend pas de  $t$ .

L'ergodicité est liée aux propriétés asymptotiques de la série temporelle. Elle implique que l'autocorrélation entre  $X_t$  et  $X_{t+h}$  tend vers 0 lorsque  $h$  tend vers l'infini. Autrement dit, le processus finit par oublier la condition initiale  $X_t$  à la période  $X_{t+h}$  si  $h$  est suffisamment grand. L'ergodicité implique en conséquence qu'à long terme, la moyenne temporelle soit égale à la moyenne spatiale.

Stationnarité et ergodicité ne sont pas équivalents. Considérons par exemple le processus :

$$X_t = Y + Z_t \quad , \quad Y \sim N(\mu, \sigma^2) \quad , \quad Z_t \sim N(0, 1)$$

Le processus est clairement stationnaire au sens faible, avec notamment une espérance égale à  $E(X_t) = \mu$ .



Il n'est en revanche pas ergodique. En effet, sa moyenne empirique est donnée par :

$$\hat{\mu} = \frac{1}{T} \sum_{t=1}^T X_t = \frac{1}{T} \sum_{t=1}^T (Y + Z_t) = Y + \frac{1}{T} \sum_{t=1}^T Z_t \rightarrow Y \text{ puisque } \frac{1}{T} \sum_{t=1}^T Z_t \rightarrow 0 \text{ a.s.}$$

Or  $Y \neq \mu$  (sauf sur un ensemble nul), de sorte que le processus ne convergera pas vers  $\mu$ .

Il est nécessaire d'avoir à la fois la stationnarité et l'ergodicité. La stationnarité garantit que quelques soient les observations, elles présenteront les mêmes propriétés théoriques. L'ergodicité garantit qu'avec suffisamment d'observations on peut atteindre les propriétés théoriques du processus, et qu'ainsi une seule trajectoire permet d'analyser le processus. Elle permet notamment de pouvoir utiliser la loi des grands nombres et le théorème de limite centrale.

Une « spurious regression » est une régression qui présente une très forte corrélation entre deux séries temporelles, alors même que celles-ci ne présentent aucun lien réel. Par exemple, on peut trouver une très forte corrélation entre la dépense publique pour la recherche scientifique et le nombre de suicides par pendaison au USA. Cette forte corrélation (erronée) est due à la présence d'une tendance (stochastique ou non) commune aux deux séries. Pour éviter ce problème de spurious regression, on peut soit enlever la tendance commune des séries au préalable, soit la modéliser explicitement dans la régression.

7. On réalise un test Augmented Dickey-Fukker pour inf, en utilisant le critère AIC pour déterminer le nombre de retards. On obtient les résultats suivants :

```
The minimum AIC value for the ADF test is 850.6507, which corresponds to 5 lags
the p-value for the ADF test with drift is 0.09465
this is larger than the 5% threshold, hence do not reject the null and conclude that the series is non-stationary
```

On choisit une spécification du test ADF avec un terme de drift (une constante). On teste d'abord un nombre de retards de 1 à 5, et le critère AIC sélectionne 5 retards. Le test est ensuite mené avec 5 retards. Il produit une p-value de 0.094, qui est donc supérieure au seuil de 0.05. On ne rejette pas l'hypothèse nulle, et on conclut que la série d'inflation n'est pas stationnaire.

Il faut toutefois nuancer ce résultat. D'abord, le test rejette l'hypothèse nulle à un seuil de 10 %, qui reste un niveau raisonnable et fréquemment employé. Il y a donc une ambiguïté sur la conclusion du test, selon le seuil choisi. Ensuite, ce résultat est typique des séries économiques et financières (taux d'intérêt, de change, d'inflation) qui présentent souvent cette ambiguïté. Il faut donc rester prudent sur la conclusion du test et conclure qu'on ne peut pas écarter la possibilité d'une série stationnaire.

Enfin, on notera aussi une différence intéressante entre implémentations logicielles. Les résultats ci-dessus sont obtenus en Matlab. Ci-dessous, on communique les résultats fournis en Python. Les deux tests sont réalisés sur la même série de données, avec le même modèle (random walk with drift, ou marche aléatoire avec dérive), et le même critère (Akaike Information Criterion). Pourtant, les deux procédures ne sélectionnent pas le même nombre de retards (5 pour Matlab, et 2 pour Python), et fournissent aussi des p-value très différentes (0.094 pour Matlab, 0.043 pour Python). Dans la mesure où les implémentations des deux logiciels sont « boîte noire » (le détail du code n'est pas disponible), il est difficile de déterminer où se situe le problème. De plus, comme Python rejette l'hypothèse nulle de marche aléatoire à un seuil de 5 % (p-value du test = 0.043), on pourra conclure à la stationnarité de la série, malgré l'ambiguïté du résultat de Matlab.

### Question 7

Faire le test Augmented Dickey Fuller pour inf en utilisant le critère AIC pour déterminer le nombre de lags à inclure. Commenter

```
Entrée [47]: # réalisation du test ADF en utilisant le critère AIC avec une constante
adf_inf_aic = adfuller(df['inf'].dropna(), autolag='AIC', regression='c')
adf_inf_aic
```

```
Out[47]: (-2.919055814337232,
0.043176516871547096,
2,
208,
{'1%': -3.4621857592784546,
'5%': -2.875537986778846,
'10%': -2.574231080806213},
817.4131078108454)
```

```
Entrée [48]: print("Pvalue : ", adf_inf_aic[1])
print("Lags inclus ", (adf_inf_aic[2]))
```

```
Pvalue : 0.043176516871547096
Lags inclus 2
```

L'hypothèse nulle est que la série a été générée par un processus présentant une racine unitaire, et donc, qu'elle n'est pas stationnaire. On rejette cette hypothèse. Par ailleurs, le nombre de lags à inclure est 2.

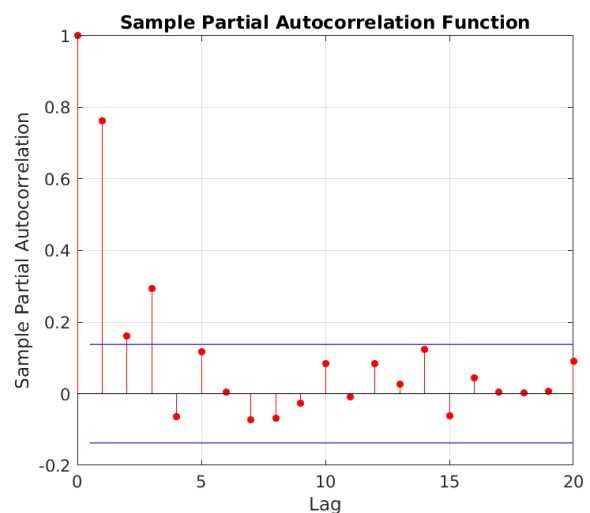
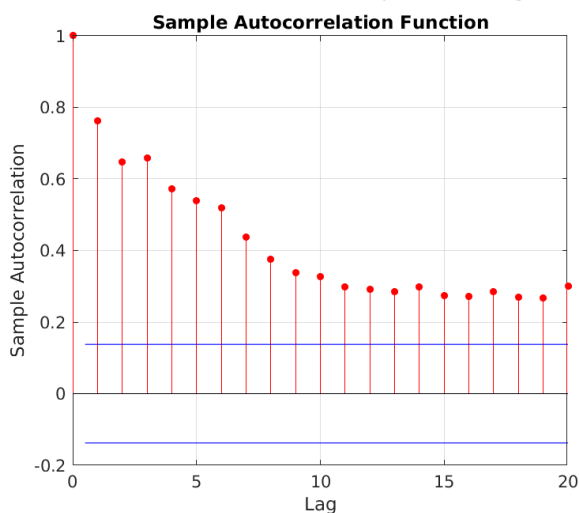
### Question 9

8. On détermine d'abord le nombre de retards  $p$  à inclure dans le modèle  $AR(p)$ . Plusieurs stratégies sont pour cela possibles :

- utiliser les autocorrélations empiriques et les comparer aux autocorrélations théoriques des modèles  $ARMA(p, q)$ .
- utiliser des critères d'information tels que l'AIC ou le BIC.

On considère d'abord les autocorrélations empiriques. Celles-ci ont déjà été estimées à la question 5. On les affiche à nouveau par commodité.

question 5, figure 5: ACF and PACF of Inflation



Ces fonctions sont cohérentes avec un modèle  $AR(p)$ . En effet, la fonction d'autocorrélation décroît de manière progressive, ce qui suggère un ordre  $MA(q) = 0$ . En revanche, la fonction d'autocorrélation partielle montre une chute rapide des coefficients, et seuls les trois premiers retards sont significatifs. Cela suggère un ordre  $AR(p)=3$ . Les fonctions d'autocorrélation empiriques suggèrent donc un  $AR(3)$ .

On utilise maintenant les critères d'information AIC et BIC. Ceux-ci sont définis respectivement par :

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

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

On réalise ces tests pour des valeurs de retards entre 1 et 10. On obtient les résultats suivants :

The results for the AIC and BIC information criteria are:

" "	"AIC"	"BIC"
"1"	"1.3942"	"1.4356"
"2"	"1.3764"	"1.4595"
"3"	"1.3017"	"1.4268"
"4"	"1.3131"	"1.4806"
"5"	"1.3161"	"1.5262"
"6"	"1.336"	"1.5891"
"7"	"1.3494"	"1.6458"
"8"	"1.3637"	"1.7036"
"9"	"1.3829"	"1.7668"
"10"	"1.3957"	"1.8239"

The minimum value for AIC is 1.3017, which corresponds to 3 lags

The minimum value for BIC is 1.4268, which corresponds to 3 lags

Les deux critères sélectionnent 3 retards. Au final, quelle que soit la stratégie retenue, on conclut à trois retards. On proposera donc pour inf un modèle AR(3).

9. On estime le modèle de la courbe de Phillips qui explique la variable unemp par une constante et le taux d'inflation courant. On obtient les résultats suivants :

beta estimates:

constant: 6.0721

inf: 0.0036744

coefficient standard deviations:

constant: 0.18136

inflation: 0.036566

t-stats for the two-sided t-test:

constant: 33.4808

inflation: 0.10049

critical values at 10%, 5% and 1% are respectively 1.6522, 1.9714, and 2.5996

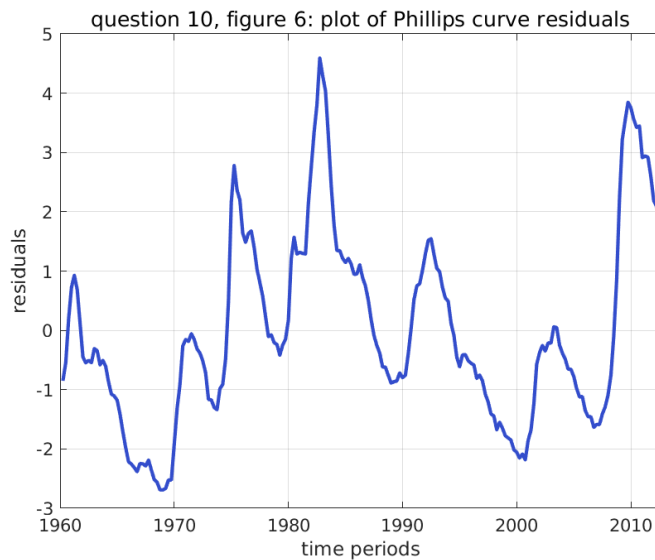
p-values for the two-sided t-test:

constant: 0

inflation: 0.92005

On note que la relation n'a pas le sens attendu. Celle-ci est supposée être négative (à court terme, laisser l'inflation augmenter permet de réduire le chômage). Or elle est positive selon le modèle : le coefficient sur l'inflation a une valeur de 0.003, ce qui implique que l'inflation provoque la hausse du taux de chômage. On peut produire deux remarques. Tout d'abord, la relation est en réalité inexistante, puisque le modèle affiche une p-value de 0.92 > 0.05 pour le coefficient associé à l'inflation. Celui-ci n'est donc pas significatif. Ensuite, on peut noter qu'il est normal que la relation ne soit pas significative. C'est en effet la conclusion qu'a atteinte la FED elle-même au début des années 1980, puisque l'économie américaine conjugait une inflation soutenue avec un fort taux de chômage. La stabilité de la courbe de Phillips fut alors contestée, ce qui poussa la FED à modifier sa politique monétaire.

10. On commence par une analyse graphique de la série des résidus :



On constate que les résidus ont l'air clairement autocorrélés : les valeurs positives ont tendance à succéder aux valeurs positives, les valeurs négatives à succéder aux valeurs négatives.

Pour tester formellement l'autocorrélation des résidus, on peut procéder de deux manières :

- estimer le modèle par MCO, obtenir les résidus  $u$ , puis estimer la corrélation des résidus avec leur premier retard par la régression :  $u_t = \rho u_{t-1} + e_t$ . On réalise ensuite un test classique de student sur l'hypothèse nulle  $H_0: \rho = 0$ .

- utiliser le test de Durbin-Watson. La statistique de test est définie comme :

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

On procède d'abord au test d'autocorrélation des résidus. On obtient les résultats suivants :

rho estimate:	0.97994
coefficient standard deviation:	0.01451
t-stat for the two-sided t-test:	67.5366
critical values at 10%, 5% and 1%:	1.6522, 1.9714, and 2.5996
p-value for the two-sided t-test:	0

La statistique de student a une valeur de 67.53, pour une valeur critique de 2.59 à un seuil de 1 %. On rejette donc l'hypothèse nulle d'absence d'autocorrélation, même à un seuil de 1 %. La p-value de 0 confirme cette conclusion. Les résidus sont clairement autocorrélés.

On procède maintenant au test de Durbin-Watson. On obtient les résultats suivants :

```
DW stat:      0.044214
dL:           1.653
dU:           1.693
```

La décision de rejet ou non de l'hypothèse nulle est basée sur deux valeurs critiques : dL et dU. Si la statistique est supérieure à dU, on ne rejette pas l'hypothèse nulle. Si la statistique est inférieure à dL, on rejette l'hypothèse nulle et conclut à la présence d'autocorrélation positive de premier ordre. Dans les autres cas, le test est inconclusif. Les valeurs dL et dU dépendent du nombre de variables explicatives du modèle (ici  $k=2$ ), et de la taille de l'échantillon (ici  $T=212 \approx 200$ ). On utilise les valeurs critiques calculées par Savin-White (voir Johnston et Dinardo, *Econometric Methods*, 4ème édition, Appendix D). Ici, la valeur de la statistique est de 0.044, très inférieure à  $dL = 1.653$ . On rejette donc sans hésitation l'hypothèse nulle et conclut que les résidus sont caractérisés par l'autocorrélation d'ordre 1.

11. On corrige l'autocorrélation des erreurs en utilisant la méthode des moindres carrés généralisés faisables. Pour cela, on suit la méthode suivante :

- on estime le modèle initial par MCO :  $y_t = \text{const} + \beta x_t + u_t$
- on obtient les résidus empiriques  $u_t$  et on procède à la régression :  $u_t = \rho u_{t-1} + e_t$
- on transforme les variables initiales selon  $\tilde{z}_t = z_t - \rho z_{t-1}$ , puis estime la régression :  
 $\tilde{y}_t = (1 - \rho)\text{const} + \beta \tilde{x}_t + e_t$

On procède à l'estimation et obtient les résultats suivants :

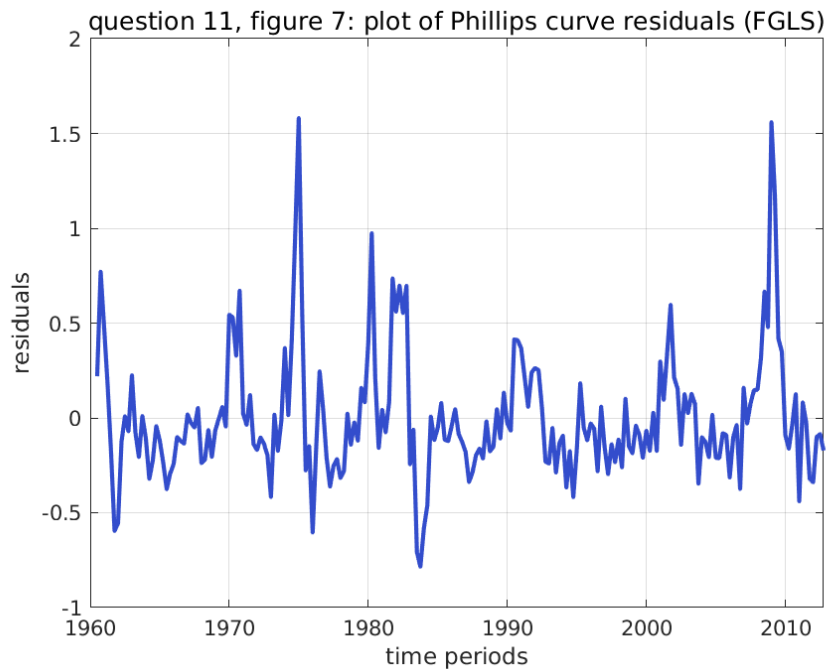
```
beta estimates (FGLS):
constant:    6.3099
inf:         -0.024576

coefficient standard deviations:
constant:    0.94824
inflation:   0.010974

t-stats for the two-sided t-test:
constant:    6.6544
inflation:   2.2396
critical values at 10%, 5% and 1% are respectively 1.6522, 1.9714, and 2.5996

p-values for the two-sided t-test:
constant:    2.4527e-10
inflation:   0.026173
```

On tire plusieurs conclusions. Tout d'abord, le coefficient sur la variable inflation a retrouvé le bon signe : il est maintenant négatif (-0.024), en accord avec la théorie de la courbe de Phillips. Ensuite, ce coefficient est devenu significatif, et la p-value de 0.026 permet de rejeter l'hypothèse nulle à un seuil de 5 %. Enfin, la méthode FGLS semble avoir permis d'éliminer l'autocorrélation des résidus, comme on peut le constater sur le graphique suivant.



Pour la suite du projet, on reviendra à un estimateur OLS classique afin de ne pas interférer avec les autres procédures considérées.

12. Pour détecter une possible rupture, on procède à un test de Chow. Ici, on partage l'échantillon en deux sous-périodes identiques. Comme l'échantillon comporte 211 données, on le partage en un premier sous-échantillon de taille 105 (soit de 1960q2 à 1986q2), et un second de taille 106 (soit de 1986q3 à 2012q4). On réalise ensuite la régression suivante :

$$y_t = \text{const} + \gamma_0 \delta(t > \tau) \text{const} + \beta x_t + \gamma_1 \delta(t > \tau) x_t + u_t$$

Le test de Chow revient à faire un test de Fischer sur l'hypothèse jointe :  $H_0: \gamma_0 = \gamma_1 = 0$ . On procède pour cela de la même façon qu'en partie 1, question 10. On obtient les résultats suivants :

F-stat for the joint hypothesis:	1.9364
critical value at 5%:	3.0395
p-value for the joint hypothesis:	0.14683

On constate que la valeur critique du test (1.9364) est inférieure à la valeur critique de 3.039. On ne rejette donc pas l'hypothèse nulle d'égalité des coefficients sur les deux périodes. La p-value de 0.14 confirme ce résultat, l'hypothèse nulle n'étant rejetée qu'à un seuil de 15 %.

13. On procède maintenant à un test QLR pour chercher la meilleure date possible de rupture. Pour cela, on procède séquentiellement en examinant toutes les dates de rupture possibles sur la fenêtre  $\{\tau_0, \tau_1\}$ , où  $\tau_0$  et  $\tau_1$  représentent les périodes à 15 % du début et de la fin de l'échantillon, respectivement. On conserve la statistique F qui possède la valeur la plus élevée. Comme on travaille sur un maximum, il faut ajuster en conséquence les valeurs critiques du test (voir slide 26, cours 7 avec  $q=2$  restrictions).

On obtient les résultats suivants :

```
max F-stat for the joint hypothesis (QLR test):    27.6928
critical value at 10%:                            5
critical value at 5%:                             5.86
critical value at 1%:                             7.78
```

The optimal period for the structural break is 1970.5

Contrairement à la question précédente, le test QLR rejette largement l'hypothèse nulle d'absence de rupture. Avec une valeur de test de 27.69, on rejette l'hypothèse nulle d'absence de rupture aux seuils de 10 %, 5 % et même 1 %. La rupture est clairement acceptée. La date optimale de rupture est identifiée en 1970q3.

14. On estime maintenant le modèle suivant :

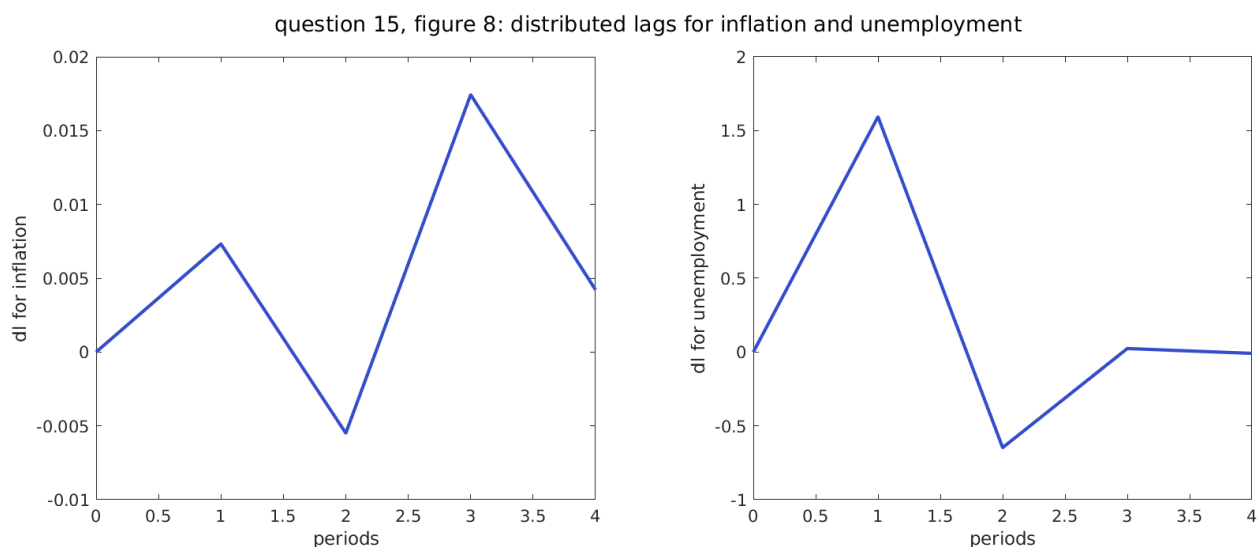
$$\begin{aligned} unemp_t = & const + \beta_1 inf_{t-1} + \beta_2 inf_{t-2} + \beta_3 inf_{t-3} + \beta_4 inf_{t-4} \\ & + \alpha_1 unemp_{t-1} + \alpha_2 unemp_{t-2} + \alpha_3 unemp_{t-3} + \alpha_4 unemp_{t-4} + u_t \end{aligned}$$

On conduit ensuite un test de causalité de Granger de l'inflation sur le chômage. Celui-ci revient à tester l'hypothèse nulle jointe  $H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ . On procède à un test de Fisher classique (voir question 10, partie 1), et on obtient les résultats suivants :

```
F-stat for the joint hypothesis (no Granger causality):    3.7693
critical value at 5%:                                       2.4173
p-value for the joint hypothesis:                          0.0055986
```

La valeur du test (3.76) est supérieure à la valeur critique (2.41). On rejette donc l'hypothèse nulle d'absence de causalité de Granger de l'inflation sur le chômage. La p-value de  $0.005 < 0.05$  confirme ce résultat.

15. On représente graphiquement les délais distribués (on s'appuie pour cette question sur le traitement de Wooldridge, « Introductory econometrics, a modern approach », pages 346-348). On obtient :



On note que le chômage (graphique de droite) s'explique principalement par son passé immédiat. En effet, l'impact le plus élevé se produit à la période qui suit le choc, puis devient légèrement négatif après deux périodes, avant de retrouver un effet quasi-nul. On peut possiblement interpréter la valeur négative de la seconde période comme un effet de compensation après une baisse trop importante de l'emploi à la période qui suit le choc initial.

L'inflation (graphique de gauche) suit une dynamique différente. Elle connaît un faible effet positif sur le chômage après une période, puis un faible effet négatif à la seconde période, avant de trouver son impact positif maximal en troisième période. On constate donc que l'économie prend un certain temps avant de transcrire dans l'activité réelle le changement de la variable nominale qu'est le niveau des prix.

On calcule finalement l'impact à long terme de l'inflation sur le chômage, qui n'est rien d'autre que la somme de tous les coefficients relatifs à l'inflation. On obtient :

The long-run impact of inflation on unemployment is 0.023475

Autrement dit, un choc unitaire d'inflation (une hausse permanente de 1 point de l'inflation) va entraîner, à terme, une hausse de 0.023 points du chômage. Cet effet est modeste. Il vient confirmer la neutralité à long terme de la politique monétaire, qui n'affecte pas l'activité réelle.