



MASTÈRE SPÉCIALISÉ BIG DATA GESTION ET ANALYSE DES DONNÉES MASSIVES

Projet SES 722 : Econométrie

Partie 1 : Régression

Partie 2 : Série temporelle

Télécom Paris

Julien LAIR, Hugo Michel

Année 2021 / 2022

Table des matières

Projet SES 722 : Econométrie	1
Télécom Paris	1
Année 2021 / 2022	1
Julien LAIR, Hugo Michel	1
Partie 1 : Regression	4
Pré-processing du dataframe mroz.raw	4
Question 1	7
Question 2	7
Question 3	10
Question 4	12
Question 5	13
Question 6	14
Question 7	15
Question 8	17
Question 9	20
Question 10	21
Question 11	25
Question 12	28
Question 13	29
Question 15	31
Question 15	49
Question 16	53
Partie 2 : Série temporelle	58
Question 1 : Importer les données du fichier quarterly.xls (corriger le problème éventuel d'observations manquantes)	58
Question 2	58
Question 3	58
Question 4	58
Question 5	58

Projet SES 722 : Econométrie

Question 6	58
Question 7	58
Question 8	58
Question 9	58
Question 10	59
Question 11	59
Question 12	59
Question 13	59
Question 14	59

1. Partie 1 : Regression

Notes :

Tous les tests d'hypothèses seront effectués à 5% sauf si un autre seuil de région de rejet est spécifié dans la question

1.1 Pré-processing du dataframe *mroz.raw*

Présentation du dataframe (avant pré-processing) :

Le nom des colonnes ont été renommées lors de la lecture du fichier *mroz.raw*

Le dataframe importé se présente comme suit :

	inlf	hours	kidslt6	kidsge6	age	educ	wage	repwage	hushrs	husage	huseduc	huswage	faminc	mtr	motheduc	fatheduc	unem	city	exp
0	1	1610	1	0	32	12	3.354	2.65	2708	34	12	4.0288	16310	0.7215	12	7	5.0	0	
1	1	1656	0	2	30	12	1.3889	2.65	2310	30	9	8.4416	21800	0.6615	7	7	11.0	1	
2	1	1980	1	3	35	12	4.5455	4.04	3072	40	12	3.5807	21040	0.6915	12	7	5.0	0	
3	1	456	0	3	34	12	1.0965	3.25	1920	53	10	3.5417	7300	0.7815	7	7	5.0	0	
4	1	1568	1	2	31	14	4.5918	3.60	2000	32	12	10.0000	27300	0.6215	12	14	9.5	1	
...	
748	0	0	0	2	40	13	.	0.00	3020	43	16	9.2715	28200	0.6215	10	10	9.5	1	
749	0	0	2	3	31	12	.	0.00	2056	33	12	4.8638	10000	0.7715	12	12	7.5	0	
750	0	0	0	0	43	12	.	0.00	2383	43	12	1.0898	9952	0.7515	10	3	7.5	0	
751	0	0	0	0	60	12	.	0.00	1705	55	8	12.4400	24984	0.6215	12	12	14.0	1	
752	0	0	0	3	39	9	.	0.00	3120	48	12	6.0897	28363	0.6915	7	7	11.0	1	

753 rows × 22 columns

Le dataframe initial (avant pré-processing) comporte **753 lignes** (i.e 753 observations) et **22 colonnes**

Nom des colonnes :

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

Les phases du pré-processing :

1. Remplissage des valeurs manquantes et outliers
2. Conversion des variables wage et lwage au format 'float'
3. Sélection des observations pour lesquelles wage > 0

Etape 1 : Remplissage des valeurs manquantes et outliers

Le type des colonnes du dataframe initial:

```

inlf      int64
hours     int64
kidslt6   int64
kidsge6   int64
age       int64
educ      int64
wage      object
repwage   float64
hushrs    int64
husage    int64
huseduc   int64
huswage   float64
faminc    int64
mtr       float64
motheduc  int64
fathedduc int64
unem      float64
city      int64
exper     int64
nwifeinc  float64
lwage     object
expersq   int64
dtype: object

```

Projet SES 722 : Econométrie

On remarque que les variables `wage` et `lwage` ne sont pas de type flottant. Il est nécessaire de convertir les variables `wage` et `lwage` en float.

Toutefois avant de pouvoir convertir ces variables en *float* il convient de gérer les valeurs manquantes ainsi que les outliers de la variable `wage`. Pour ce faire, on peut afficher toutes les valeurs prises par la variable `wage`.

```
1 df["wage"].value_counts()  
  
.  
2  
6.25  
2.5  
3.75  
...  
4.2347  
3.445  
7.0968  
8.19  
3.354  
Name: wage, Length: 374, dtype: int64
```

On remarque que les valeurs de la variable `wage` ne sont pas toujours un nombre. Par exemple, la variable `wage` contient 325 fois le caractère `"."`. A ce stade, il n'est pas possible de travailler avec cette variable. Pour pallier ce problème, l'idée est d'abord de remplacer les observations qui ne sont pas des nombres par des valeurs manquantes (`NaN` value) pour pouvoir ensuite réaliser la conversion en float des variables `wage` et `lwage` ($\log(wage)$). Enfin, on remplace les `NaN` par la valeur "0".

1.2 Question 1

Lire le fichier `mroz.raw`. Ne sélectionner que les observations pour lesquelles la variable `wage` est **strictement positive**.

```

1 # conversion de la colonne salaire en float avec une stratégie de force brute puis remplir NaN par 0
2 df["wage"] = pd.to_numeric(df["wage"], errors='coerce').fillna(0, downcast='infer')
3
4 # conversion de la colonne lwage en float avec une stratégie de force brute puis remplir NaN par 0
5 df["lwage"] = pd.to_numeric(df["lwage"], errors='coerce').fillna(0, downcast='infer')
6
7 # sélection des observations où le salaire > 0
8 df = df[df.wage > 0]
9 df

```

	inlf	hours	kidslt6	kidsge6	age	educ	wage	repwage	hushrs	hususage	huseduc	huswage	faminc	mtr	motheduc	fatheduc	unem	city	exp
0	1	1610	1	0	32	12	3.3540	2.65	2708	34	12	4.0288	16310	0.7215	12	7	5.0	0	
1	1	1656	0	2	30	12	1.3889	2.65	2310	30	9	8.4416	21800	0.6615	7	7	11.0	1	
2	1	1980	1	3	35	12	4.5455	4.04	3072	40	12	3.5807	21040	0.6915	12	7	5.0	0	
3	1	456	0	3	34	12	1.0965	3.25	1920	53	10	3.5417	7300	0.7815	7	7	5.0	0	
4	1	1568	1	2	31	14	4.5918	3.60	2000	32	12	10.0000	27300	0.6215	12	14	9.5	1	
...	
423	1	680	0	5	36	10	2.3118	0.00	3430	43	12	5.3061	19772	0.7215	7	7	7.5	0	
424	1	2450	0	1	40	12	5.3061	6.50	2008	40	8	7.2709	35641	0.6215	7	7	5.0	1	
425	1	2144	0	2	43	13	5.8675	0.00	2140	43	11	8.1776	34220	0.5815	7	7	7.5	1	
426	1	1760	0	1	33	12	3.4091	3.21	3380	34	12	7.1006	30000	0.5815	12	16	11.0	1	
427	1	490	0	1	30	12	4.0816	2.46	2430	33	11	6.5844	18000	0.6915	12	12	7.5	1	

428 rows × 22 columns

Après avoir gérer les outliers et les valeurs manquantes de la colonnes `wage` en sélectionnant les observations pour lesquelles la variable `wage > 0` on obtient un dataframe avec **428 observations** contre **753 initialement**.

1.3 Question 2

Faire les statistiques descriptives du salaire, de l'âge et de l'éducation pour :

- 1/ L'ensemble des femmes
- 2/ Pour les femmes dont le salaire du mari est supérieure au 65ème percentile de l'échantillon
- 3/ Pour les femmes dont le salaire du mari est inférieur au 65ème percentile de l'échantillon.

Commenter

Projet SES 722 : Econométrie

1/ Statistique descriptives du **salaire (wage)**, de l'**âge (age)** et l'**éducation (educ)** de l'ensemble des femmes

```
#### Statistique descriptives du salaire (wage) pour l'ensemble des femmes ####
```

```
count    428.000000
mean     4.177682
std      3.310282
min      0.128200
25%     2.262600
50%     3.481900
75%     4.970750
max     25.000000
Name: wage, dtype: float64
```

```
#### Statistique descriptives de l'âge (age) pour l'ensemble des femmes ####
```

```
count    428.000000
mean     41.971963
std      7.721084
min      30.000000
25%     35.000000
50%     42.000000
75%     47.250000
max     60.000000
Name: age, dtype: float64
```

```
#### Statistique descriptives de l'éducaton (educ) pour l'ensemble des femmes ####
```

```
count    428.000000
mean     12.658879
std      2.285376
min      5.000000
25%     12.000000
50%     12.000000
75%     14.000000
max     17.000000
Name: educ, dtype: float64
```

2/ Statistique descriptives du **salaire (wage)**, de l'**âge (age)** et l'**éducation (educ)** pour les femmes dont le salaire du mari est supérieure au 65ème percentile de l'échantillon

```
#### Statistique descriptives du salaire (wage) pour les femmes dont le salaire du mari est supérieure au 65ème percentile de l'échantillon ####
```

```
count    148.000000
mean     5.139315
std      4.351728
min      0.213700
25%     2.561925
50%     4.008050
75%     6.516300
max     25.000000
Name: wage, dtype: float64
```

```
#### Statistique descriptives de l'âge (age) pour les femmes dont le salaire du mari est supérieure au 65ème percentile de l'échantillon ####
```

```
count    148.000000
mean     42.52027
std      7.35168
min      30.00000
25%     36.00000
50%     43.00000
75%     48.00000
max     59.00000
Name: age, dtype: float64
```

Projet SES 722 : Econométrie

```
#### Statistique descriptives de l'éducaton (educ) pour les femmes dont le salaire du mari est supérieure au 65ème percentil  
e de l'échantillon ####
```

```
count    148.000000  
mean     13.520270  
std      2.345845  
min      5.000000  
25%     12.000000  
50%     13.000000  
75%     16.000000  
max      17.000000  
Name: educ, dtype: float64
```

3/ Statistique descriptives du *salaire (wage)*, de l'*âge (age)* et l'*éducation (educ)* pour les femmes dont le salaire du mari est inférieure au 65ème percentile de l'échantillon

```
#### Statistique descriptives du salaire (wage) pour les femmes dont le salaire du mari est inférieure au 65ème percentile d  
e l'échantillon ####
```

```
count    280.000000  
mean     3.669390  
std      2.458277  
min      0.128200  
25%     2.151600  
50%     3.203550  
75%     4.539500  
max      22.500000  
Name: wage, dtype: float64
```

```
#### Statistique descriptives de l'âge (age) pour les femmes dont le salaire du mari est inférieure au 65ème percentile de  
l'échantillon ####
```

```
count    280.000000  
mean     41.682143  
std      7.906875  
min      30.000000  
25%     35.000000  
50%     41.000000  
75%     47.000000  
max      60.000000  
Name: age, dtype: float64
```

```
#### Statistique descriptives de l'éducaton (educ) pour les femmes dont le salaire du mari est inférieure au 65ème percentil  
e de l'échantillon ####
```

```
count    280.000000  
mean     12.203571  
std      2.119542  
min      6.000000  
25%     12.000000  
50%     12.000000  
75%     12.000000  
max      17.000000  
Name: educ, dtype: float64
```

Commentaires

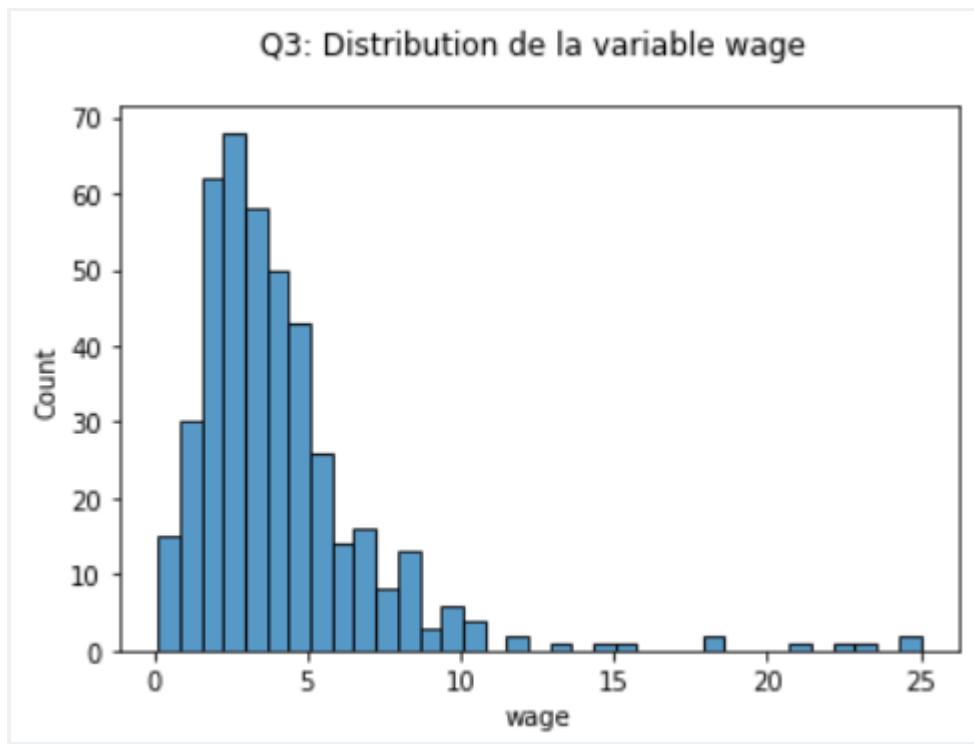
Au regard des statistiques descriptives, on peut remarquer que l'âge, le niveau d'éducation et le revenu de l'épouse sont corrélés à ceux du mari. Les femmes dont le revenu du mari est inférieur au 65ème percentile sont en moyenne plus jeunes et ont un niveau d'éducation et de revenu moindre que celui du groupe total. Les femmes dont le revenu du mari est supérieur au 65 ème percentile sont en moyenne plus âgées et ont un niveau d'éducation et de revenu supérieur à celui du groupe total. Cela semble logique car plus la personne est âgée plus ces années d'expériences sont élevées ce qui fait de facto augmenter le salaire de cette dernière.

Ce que l'on peut interpréter c'est que les personnes avec un profil similaire ont tendance à s'unir par le lien du mariage. Cela signifie que les époux ont quasiment le même âge, un revenu presque identique et un nombre d'année d'étude équivalent.

1.4 Question 3

Faire l'histogramme de la variable wage. Supprimer les observations qui sont à plus de 3 écart-types de la moyenne et refaire l'histogramme

Histogramme de la variable wage (distribution initiale)



On sélectionne uniquement les observations dont les valeurs de la variable wage sont strictement inférieure à 3 fois l'écart type de la moyenne. Cela revient finalement à supprimer les observations qui sont à plus de 3 écart-types de la moyenne

L'écart-type de la variable wage est: 3.31
La moyenne de la variable wage est: 4.178

```
1 criterion = moyenne_wage + (3 * ecart_type_wage)
2 print(criterion)
3 df_temp = df[df.wage < criterion]
```

14.108

Après avoir appliqué le critère de sélection, on obtient à DataFrame avec 419 lignes

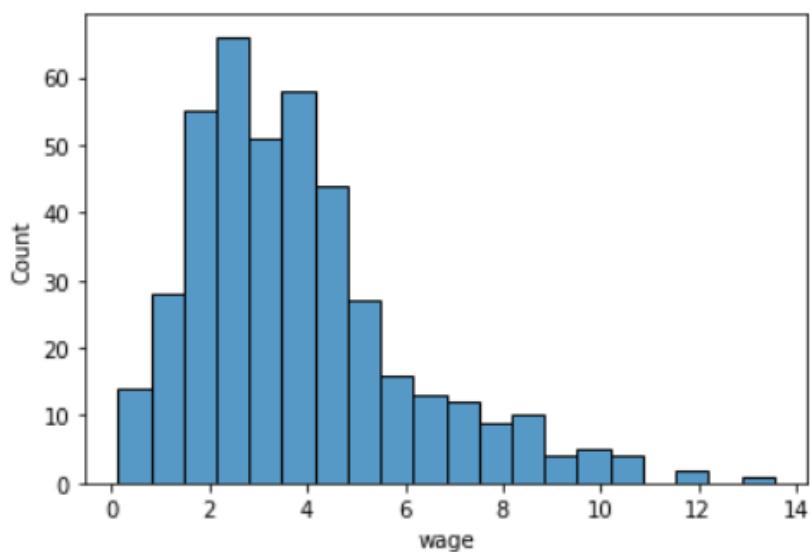
```
1 df_temp
```

	inlf	hours	kidslt6	kidsge6	age	educ	wage	repwage	hushrs	husage	huseduc	huswage	faminc	mtr	motheduc	fatheduc	unem	city	exp
0	1	1610	1	0	32	12	3.3540	2.65	2708	34	12	4.0288	16310	0.7215	12	7	5.0	0	
1	1	1656	0	2	30	12	1.3889	2.65	2310	30	9	8.4416	21800	0.6615	7	7	11.0	1	
2	1	1980	1	3	35	12	4.5455	4.04	3072	40	12	3.5807	21040	0.6915	12	7	5.0	0	
3	1	456	0	3	34	12	1.0965	3.25	1920	53	10	3.5417	7300	0.7815	7	7	5.0	0	
4	1	1568	1	2	31	14	4.5918	3.60	2000	32	12	10.0000	27300	0.6215	12	14	9.5	1	
...	
423	1	680	0	5	36	10	2.3118	0.00	3430	43	12	5.3061	19772	0.7215	7	7	7.5	0	
424	1	2450	0	1	40	12	5.3061	6.50	2008	40	8	7.2709	35641	0.6215	7	7	5.0	1	
425	1	2144	0	2	43	13	5.8675	0.00	2140	43	11	8.1776	34220	0.5815	7	7	7.5	1	
426	1	1760	0	1	33	12	3.4091	3.21	3380	34	12	7.1006	30000	0.5815	12	16	11.0	1	
427	1	490	0	1	30	12	4.0816	2.46	2430	33	11	6.5844	18000	0.6915	12	12	7.5	1	

419 rows × 22 columns

Le nouvel histogramme de la variable wage associé à ce dataframe est :

Q3: Distribution de la variable wage dont les observations sont inférieures à 3 écart-types de la moyenne



Commentaires

On observe qu'initiallement, la variable `wage` possède une distribution asymétrique négative puisque les valeurs ont tendance à davantage se concentrer vers des valeurs inférieures à la moyenne (4.178)

Après avoir appliqué le critère de sélection (à plus de 3 écarts-type de la moyenne) l'asymétrie négative persiste mais certaines valeurs extrêmes (salaire > 20) semblent disparaître alors qu'elles étaient présentes initialement.

1.5 Question 4

Calculer les corrélations `motheduc` et `fatheduc`. Expliquer le problème de multicollinearité. Commenter.

```
1 correlation = df['motheduc'].corr(df['fatheduc'])
2 print("Le coefficient de corrélation entre la variable mothereduc et f
3 print(correlation)
```

Le coefficient de corrélation entre la variable `mothereduc` et `fathereduc` est :

0.554063218431167

Commentaires

On en déduit que la corrélation entre l'éducation de la mère et celle du père est de 55%. Ainsi, dans notre cas, on peut conclure que les variables `fatheduc` et `motheduc` ne sont pas fortement corrélées.

Explication du problème de multi-collinearité

Le problème de multi-collinearité apparaît lorsque le coefficient de corrélation tend vers 1. Cela signifie alors qu'il existe une combinaison linéaire entre ces variables fortement corrélées. On dit alors que ces variables sont linéairement dépendantes. Cela pose problème car le déterminant de la matrice $X^T X$ devient alors nul. Par conséquent, la matrice $X^T X$ n'est pas inversible ce qui amène à estimer l'estimateur des moindres carrés ordinaires (MCO) avec une mauvaise précision. De ce fait, cela augmente la variance des coefficients de régression et les rends instables et difficiles à interpréter. Par ailleurs, cela se répercute également sur les intervalles de confiance qui risquent d'être beaucoup trop larges.

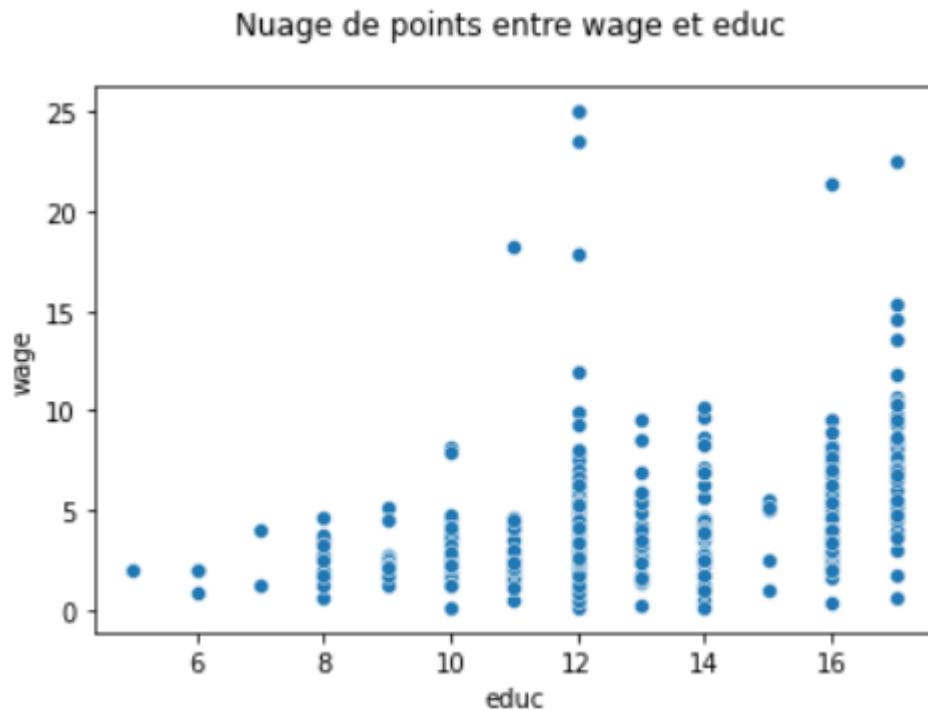
Les conséquences de coefficients instables sont les suivantes :

- Les coefficients peuvent sembler non significatifs
- Les coefficients de prédicteurs fortement corrélés vont évoluer en fonction de l'échantillon.

Toutefois, il est nécessaire de rappeler que la multi-collinearité n'a aucune influence négative sur le bon ajustement de la droite de régression et sur la qualité des prédictions. Cependant les coefficients individuels associés à chaque variable explicative ne peuvent pas être interprétés de façon fiable.

1.6 Question 5

Faites un graphique en nuage de point entre wage et educ. S'agit-il d'un effet "toute chose étant égale par ailleurs ?"



Commentaires

Au regard du nuage de points, le graphique semble mettre en évidence une corrélation entre les variables `wage` et `educ`.

S'agit-il d'un effet ceteris-paribus ?

Un effet ceteris paribus fait référence à l'effet direct que possède une variable X sur une autre variable Y tout en faisant l'hypothèse que le reste des variables du modèle restent constantes et inchangées au cours du temps. Autrement dit, une approche *ceteris paribus* permet de simplifier le travail d'analyse en supposant qu'une seule variable possède une influence sur une autre variable du modèle économique. Cette technique permet d'isoler les effets d'autres événements qui pourraient éventuellement se produire au même moment.

Sur la base de ce qui a été évoqué précédemment, L'effet observé dans notre cas ne semble pas être un effet *ceteris paribus* car d'autres variables différentes de celle de `educ` peuvent avoir une influence direct sur la variable `wage`. Par exemple, l'expérience professionnelle peut être une variable intéressante à prendre en compte pour le calcul du salaire car on sait intuitivement que les années d'expérience ont un influence importante sur le salaire d'un employé. Une personne disposant d'une expérience professionnelle de 20 ans dans un domaine d'activité précis possèdera un salaire plus élevé qu'une personne ne disposant que de 2 ans d'expertise dans ce même domaine.

1.7 Question 6

Quelle est l'hypothèse fondamentale qui garantit des estimateurs non biaisés ? Expliquer le biais de variable omise

Quelle est l'hypothèse fondamentale qui garantit des estimateurs non biaisés ?

L'hypothèse fondamentale qui garantit des estimateurs non biaisés fait référence au résidus (variable u non observée) du modèle qui doivent être de moyenne nulle et non corrélée avec les regresseurs X . On peut alors écrire la chose suivante : $E(u|X) = E(u) = 0$

Expliquer le biais de variable omise.

Si l'on omet une variable déterminante pour la prédiction du modèle, cela aboutit à un estimateur MCO biaisée. Autrement dit, si l'on oublie de prendre en compte une variable essentielle dans la qualité de la prédiction du modèle alors, cette variable va se retrouver dans le terme d'erreur. De plus, si cette variable explicative est corrélée avec d'autres variables explicatives du prisent en compte par le modèle alors $E(u|X) \neq 0$ car le résidu u intègrera une variable explicative. Par conséquent, l'estimateur des MCO se retrouvera biaisé.

NOTE (important)

A noter que pour les questions suivantes nous travaillerons exclusivement avec le dataframe dont les observations où le salaire > 0 (wage > 0)

La dataframe contient 428 lignes et 22 colonnes

	inlf	hours	kidslt6	kidsge6	age	educ	wage	repwage	hushrs	husage	huseduc	huswage	faminc	mtr	motheduc	fatheduc	unem	city	exp
0	1	1610	1	0	32	12	3.3540	2.65	2708	34	12	4.0288	16310	0.7215	12	7	5.0	0	
1	1	1656	0	2	30	12	1.3889	2.65	2310	30	9	8.4416	21800	0.6615	7	7	11.0	1	
2	1	1980	1	3	35	12	4.5455	4.04	3072	40	12	3.5807	21040	0.6915	12	7	5.0	0	
3	1	456	0	3	34	12	1.0965	3.25	1920	53	10	3.5417	7300	0.7815	7	7	5.0	0	
4	1	1568	1	2	31	14	4.5918	3.60	2000	32	12	10.0000	27300	0.6215	12	14	9.5	1	
...	
423	1	680	0	5	36	10	2.3118	0.00	3430	43	12	5.3061	19772	0.7215	7	7	7.5	0	
424	1	2450	0	1	40	12	5.3061	6.50	2008	40	8	7.2709	35641	0.6215	7	7	5.0	1	
425	1	2144	0	2	43	13	5.8675	0.00	2140	43	11	8.1776	34220	0.5815	7	7	7.5	1	
426	1	1760	0	1	33	12	3.4091	3.21	3380	34	12	7.1006	30000	0.5815	12	16	11.0	1	
427	1	490	0	1	30	12	4.0816	2.46	2430	33	11	6.5844	18000	0.6915	12	12	7.5	1	

428 rows × 22 columns

1.8 Question 7

Faire la régression du log de wage en utilisant comme variables explicatives une constante, city, educ, exper, nwifeinc, kidslt6, kidsgt6.

Commentez l'histogramme des résidus.

Calcul des estimateurs :

$$\beta = (X^T X)^{-1} X^T Y$$

Les estimateurs β de chaque variables explicatives du modèle sont :

```
{'constante': -0.39897522667501173,  
 'city': 0.03526789293230923,  
 'educ': 0.1022475478401905,  
 'exper': 0.015487872095843404,  
 'nwifeinc': 0.004882695681563886,  
 'kidslt6': -0.0453028703816894,  
 'kidsge6': -0.011703506704368392}
```

Calcul des résidus :

On sait que $Y = X\beta + u$

avec $u = \text{résidus}$

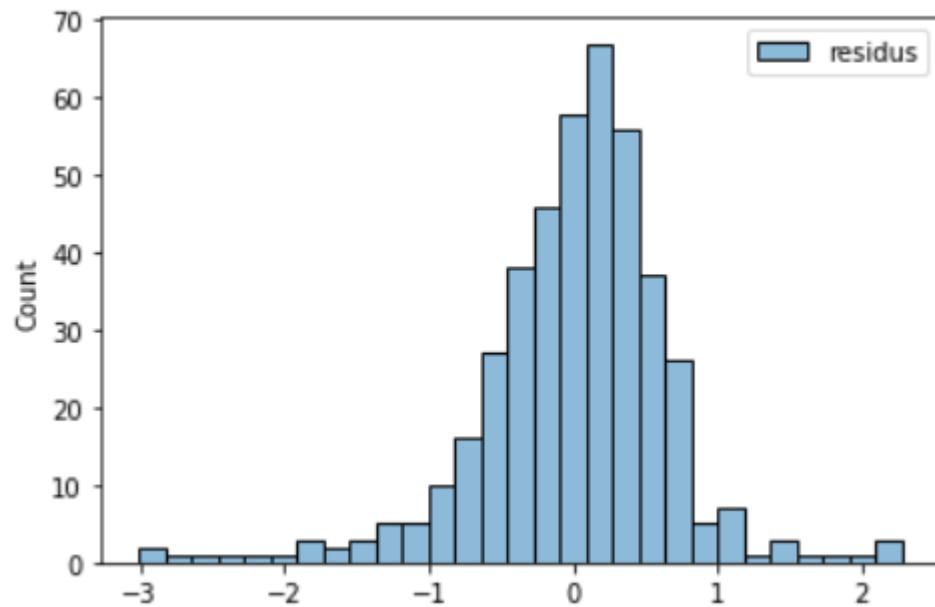
donc :

$$u = Y - X\beta$$

Les résidus u sont :

résidus	
0	0.157361
1	-0.683996
2	0.475451
3	-0.826891
4	0.318666
...	...
423	0.153203
424	0.381503
425	0.380930
426	0.040873
427	0.368391

Distribution des résidus



Commentaires

On remarque que l'histogramme des résidus semble suivre la distribution d'une loi normale car la moyenne des résidus est proche de 0. Cela signifie que la distribution est centrée en 0. A l'inverse, la distribution n'est pas symétrique et semble être légèrement asymétrique négativement. Par conséquent, des tests supplémentaires doivent être réalisés pour vérifier la pertinence de ces hypothèses.

1.9 Question 8

Tester l'hypothèse de non significativité de `nwifeinc` avec un seuil de significativité de 1%, 5% et 10% (test alternatif des deux côtés). Commentez les p-values.

Test d'hypothèse de non significativité

Tester l'hypothèse de non significativité de `nwifeinc` avec un seuil de significativité de 1%, 5% et 10% (test alternatif des deux côtés). Commentez les p-values.

$$X^T Y = (X^T X) \beta$$

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

$$Var(\hat{\beta}) = Var((X^T X)^{-1} X^T (X\beta + u)) = ((X^T X)^{-1} X^T Var(u) X(X^T X)^{-1}) = \sigma^2 (X^T X)^{-1}$$

$(X^T X)^{-1}$ est la matrice de covariance. On prendra la diagonale pour obtenir les coefficients par écart-type
 $std = diag(\sqrt{\sigma^2 (X^T X)^{-1}})$

On note : Tester l'hypothèse de non significativité de la variable `nwifeinc` revient à faire l'hypothèse :

$$H_0 : \beta_{nwifeinc} = 0$$

H_0 signifiant que la variable `nwifeinc` n'a pas une importance sur l'augmentation du salaire d'une personne.

On peut également poser $H_1 : \beta_{nwifeinc} \neq 0$

Calcul de la statistique de Student pour toutes les variables du modèle

Les statistiques de student associées à chaque variable pour le test d'hypothèse bilatérale sont :

```
{'constante': -1.9269466911402628,  
 'city': 0.5025468652951948,  
 'educ': 6.770561785365637,  
 'exper': 3.45171828081279,  
 'nwifeinc': 1.4659514850330864,  
 'kidslt6': -0.5310523188883772,  
 'kidsge6': -0.4343577482482585}
```

Calcul de la p-valeur pour la variable nwifeinc

La p-valeur pour la variable nwifeinc est (Test d'hypothèse bilatérale) :
0.14340798202513313

Calcul des region de rejets

- alpha = 1%
- alhpa = 5%
- alpha = 10%

Pour un risque alpha = 1% le seuil critique est : 2.5875575730543354

Pour un risque alpha = 5% le seuil critique est : 1.965614792008086

Pour un risque alpha = 10% le seuil critique est : 1.6484810571255268

Test de significativité pour la variable nwifeinc

Pour un risque alpha = 1%

On ne peut pas rejeter l'hypothèse H_0 car $0.143 > 0.01$

La variable nwifeinc n'est pas significative pour le salaire

Pour un risque alpha = 5%

On ne peut pas rejeter l'hypothèse H_0 car $0.143 > 0.05$

La variable nwifeinc n'est pas significative pour le salaire

Pour un risque alpha = 10%

On ne peut pas rejeter l'hypothèse H_0 car $0.143 > 0.1$

La variable nwifeinc n'est pas significative pour le salaire

Commentaires

Précédemment nous avons posé:

$$H_0 : \beta_{nwifeinc} = 0$$

$$H_1 : \beta_{nwifeinc} \neq 0$$

On a pu remarquer le $p - valeur_{nwifeinc} = 0.143 > \alpha = \{0.01, 0.05, 0.1\}$ donc on ne peut pas rejeter l'hypothèse nulle H_0 et on rejette l'hypothèse non nulle H_1 pour les seuils $\alpha = \{1\%, 5\%, 10\%\}$. Ne pas rejeter l'hypothèse H_0 et rejeter l'hypothèse H_1 signifie que la variable nwifeinc (non wife income) n'est pas significative pour le salaire d'une personne. Cela semble intuitif car le salaire d'une personne dépend fortement de l'année d'expérience et des nombre d'année d'étude.

On élargit l'étude des tests de significativité à toutes les variables du modèle

Les p-valeur associés à chaque variable pour le test d'hypothèse bilatérale sont :

```
{'constante': 1.9453410557297839,  
 'city': 0.6155456608558574,  
 'educ': 4.324526376139858e-11,  
 'exper': 0.0006133650790142258,  
 'nwifeinc': 0.14340798202513313,  
 'kidslt6': 1.4043374323234277,  
 'kidsge6': 1.3357487966211226}
```

Commentaires

En élargissant notre étude aux p-valeurs des autres variables, on remarque que les seules variables significatives pour le salaire sont respectivement les variables `educ` et `exper` car leur p-valeurs (resp. 6.10^{-4} et $4.32.10^{-11}$) < $\alpha = \{0.01, 0.05, 0.1\}$. Par conséquent pour ces variables on rejette l'hypothèse nulle H_0 et on ne peut pas rejeter l'hypothèse non nulle H_1 . Pour les autres variables, on peut tirer les mêmes conclusions de non significativité que ceux de la variable `nwifeinc`.

Cela confirme l'hypothèse que nous avions émis précédemment à savoir que le nombre des années d'étude et le nombre des années d'expérience sont des facteurs déterminants pour le salaire d'une personne.

1.10 Question 9

*Tester l'hypothèse que le coefficient associé à `nwifeinc` est égal à **0.01** avec un seuil de significativité de 5% (test à alternatif des deux côtés).*

Test d'égalité à une valeur

Pour ce test on pose :

$$H_0 : \beta_{nwifeinc} = 0.01$$

$$H_1 : \beta_{nwifeinc} \neq 0.01$$

La p-valeur pour la variable `nwifeinc` est (Test d'hypothèse bilatérale):
0.12519418591700024

Pour un risque alpha = 5%

Statistique de student = 1.536388985556214

p-valeur(`nwifeinc`) = 0.12519418591700024

Seuil critique = 1.965614792008086

On ne rejette pas l'hypothèse H_0 car $0.125 > 0.05$

La variable `nwifeinc` n'est pas significative pour le salaire

Commentaires

Pour ce nouveau test d'hypothèse d'égalité à une valeur, on a posé :

$$H_0 : \beta_{nwifeinc} = 0.01$$

$$H_1 : \beta_{nwifeinc} \neq 0.01$$

On remarque que $p - valeur_{nwifeinc} = 0.125 > \alpha = 0.05$. Donc on ne peut pas rejeter l'hypothèse $H_0 : \beta_{nwifeinc} = 0.01$ à un seuil de 5%.

1.11 Question 10

Tester l'hypothèse jointe que le coefficient de nwifeinc est égal à 0.01 et que celui de city est égal à 0.05.

Pour tester une hypothèse jointe, il est nécessaire de réaliser un Test de Fisher :

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

Pour ce faire on pose les hypothèses jointes suivantes :

$$H_0 : \beta_{nwifeinc} = 0.01 \text{ et } \beta_{city} = 0.05$$

$$H_1 : \beta_{nwifeinc} \neq 0.01 \text{ et } \beta_{city} \neq 0.05$$

On pose:

$$\begin{cases} \beta_{nwifeinc} = 0.01 \\ \beta_{city} = 0.05 \end{cases}$$

$$lwage = constante + \beta_{city}X_{city} + \beta_{educ}X_{educ} + \beta_{exper}X_{exper} + \beta_{nwifeinc}X_{nwifeinc} + \beta_{kidslt6}X_{kidslt6} + \beta_{kidsge6}X_{kidsge6}$$

$$lwage = constante + 0.05X_{city} + \beta_{educ}X_{educ} + \beta_{exper}X_{exper} + 0.01X_{nwifeinc} + \beta_{kidslt6}X_{kidslt6} + \beta_{kidsge6}X_{kidsge6}$$

$$lwage - 0.01X_{nwifeinc} - 0.05X_{city} = constante + \beta_{educ}X_{educ} + \beta_{exper}X_{exper} + \beta_{kidslt6}X_{kidslt6} + \beta_{kidsge6}X_{kidsge6}$$

Pour réaliser un test de fisher, il faut établir le modèle non contraint (i.e le modèle complet) et le modèle contraint.

Modèle non contraint (i.e modèle complet)*(constante, city, educ, exper, nwifeinc, kidslt6, kidsge6)*

OLS Regression Results						
Dep. Variable:	lwage	R-squared:	0.156			
Model:	OLS	Adj. R-squared:	0.144			
Method:	Least Squares	F-statistic:	12.92			
Date:	Tue, 03 May 2022	Prob (F-statistic):	2.00e-13			
Time:	16:41:27	Log-Likelihood:	-431.92			
No. Observations:	428	AIC:	877.8			
Df Residuals:	421	BIC:	906.3			
Df Model:	6					
Covariance Type:	nonrobust					
coef	std err	t	P> t	[0.025	0.975]	
const	-0.3990	0.207	-1.927	0.055	-0.806	0.008
x1	0.0353	0.070	0.503	0.616	-0.103	0.173
x2	0.1022	0.015	6.771	0.000	0.073	0.132
x3	0.0155	0.004	3.452	0.001	0.007	0.024
x4	0.0049	0.003	1.466	0.143	-0.002	0.011
x5	-0.0453	0.085	-0.531	0.596	-0.213	0.122
x6	-0.0117	0.027	-0.434	0.664	-0.065	0.041
Omnibus:	79.542	Durbin-Watson:	1.979			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	287.193			
Skew:	-0.795	Prob(JB):	4.33e-63			
Kurtosis:	6.685	Cond. No.	178.			

Warnings:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
SSRO = 188.5899801926394

Modèle constraint : $(constante, educ, exper, kidslt6, kidsge6)$

avec

$$X = constante + \beta_{educ} X_{educ} + \beta_{exper} X_{exper} + \beta_{kidslt6} X_{kidslt6} + \beta_{kidsge6} X_{kidsge6}$$

$$Y = lwage - 0.01 X_{nwifinc} - 0.05 X_{city}$$

OLS Regression Results						
Dep. Variable:	y	R-squared:	0.130			
Model:	OLS	Adj. R-squared:	0.122			
Method:	Least Squares	F-statistic:	15.84			
Date:	Tue, 03 May 2022	Prob (F-statistic):	4.34e-12			
Time:	16:41:27	Log-Likelihood:	-433.28			
No. Observations:	428	AIC:	876.6			
Df Residuals:	423	BIC:	896.9			
Df Model:	4					
Covariance Type:	nonrobust					
coef	std err	t	P> t	[0.025	0.975]	
const	-0.4287	0.206	-2.082	0.038	-0.833	-0.024
x1	0.0948	0.014	6.586	0.000	0.067	0.123
x2	0.0167	0.004	3.765	0.000	0.008	0.025
x3	-0.0316	0.085	-0.372	0.710	-0.199	0.135
x4	-0.0114	0.027	-0.422	0.673	-0.064	0.042
Omnibus:	76.581	Durbin-Watson:			1.976	
Prob(Omnibus):	0.000	Jarque-Bera (JB):			263.518	
Skew:	-0.779	Prob(JB):			6.00e-58	
Kurtosis:	6.514	Cond. No.			123.	

Warnings:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
SSR1 = 189.7878808521723

Calcul de Statistique de Fisher

```

1 F = ((SSR1 - SSRO)/2)/(SSRO/(n-k)) # on divise par car on a 2 contraintes
2 print("La statistique de Fisher de l'hypothèse jointe est : \n F-stat = ", F)

```

La statistique de Fisher de l'hypothèse jointe est :

F-stat = 1.3370704454929372

Calcul de la p-valeur

```

1 p_val = stats.f.sf(F,2,n-k)
2 print("La p-valeur de l'hypothèse jointe est \n p-valeur = ", p_val)

```

La p-valeur de l'hypothèse jointe est

p-valeur = 0.2637267136252519

Commentaires

On a posé les hypothèses jointes suivantes :

$H_0 : \beta_{nwifeinc} = 0.01$ et $\beta_{city} = 0.05$

$H_1 : \beta_{nwifeinc} \neq 0.01$ et $\beta_{city} \neq 0.05$

Pour ce test à un seuil $\alpha = 5$ on obtient :

- $stat - Fischer = 1.337$
- $p - valeur = 0.264$

On remarque que $p - valeur = 0.264 > \alpha = 0.05$. Donc on ne peut pas rejeter l'hypothèse jointe $H_0 : \beta_{nwifeinc} = 0.01$ et $\beta_{city} = 0.05$ à un seuil de 5%.

1.12 Question 11

Tester l'hypothèse jointe que

$$\beta_{nwifeinc} + \beta_{city} = 0.1 \text{ et } \beta_{educ} + \beta_{exper} = 0.1$$

Pour répondre à cette question on procèdera de la même manière que pour la question 10 à savoir, on réalisera un test de Fisher en établissant le modèle complet et le modèle contraint

Pour ce faire on pose les hypothèses jointes suivantes :

$$H_0 : \beta_{nwifeinc} + \beta_{city} = 0.1 \text{ et } \beta_{educ} + \beta_{exper} = 0.1$$

$$H_1 : \beta_{nwifeinc} + \beta_{city} \neq 0.1 \text{ et } \beta_{educ} + \beta_{exper} \neq 0.1$$

Pour ce test d'hypothèse jointe on pose:

$$\begin{cases} \beta_{nwifeinc} + \beta_{city} = 0.1 \\ \beta_{educ} + \beta_{exper} = 0.1 \end{cases}$$

\Leftrightarrow

$$\begin{cases} \beta_{nwifeinc} = 0.1 - \beta_{city} \\ \beta_{educ} = 0.1 - \beta_{exper} \end{cases}$$

$$lwage = constante + \beta_{city}X_{city} + \beta_{educ}X_{educ} + \beta_{exper}X_{exper} + \beta_{nwifeinc}X_{nwifeinc} + \beta_{kidslt6}X_{kidslt6} + \beta_{kidsge6}X_{kidsge6}$$

$$lwage = constante + \beta_{city}X_{city} + (0.1 - \beta_{exper})X_{educ} + \beta_{exper}X_{exper} + (0.1 - \beta_{city})X_{nwifeinc} + \beta_{kidslt6}X_{kidslt6} + \beta_{kidsge6}X_{kidsge6}$$

$$lwage = constante + \beta_{city}X_{city} + 0.1X_{educ} - \beta_{exper}X_{educ} + \beta_{exper}X_{exper} + 0.1X_{nwifeinc} - \beta_{city}X_{nwifeinc} + \beta_{kidslt6}X_{kidslt6} + \beta_{kidsge6}X_{kidsge6}$$

$$lwage - 0.1X_{educ} - 0.1X_{nwifeinc} = constante + \beta_{city}X_{city} - \beta_{exper}X_{educ} + \beta_{exper}X_{exper} - \beta_{city}X_{nwifeinc} + \beta_{kidslt6}X_{kidslt6} + \beta_{kidsge6}X_{kidsge6}$$

$$lwage - 0.1X_{educ} - 0.1X_{nwifeinc} = constante + \beta_{city}(X_{city} - X_{nwifeinc}) - \beta_{exper}(X_{exper} - X_{educ}) + \beta_{kidslt6}X_{kidslt6} + \beta_{kidsge6}X_{kidsge6}$$

Pour réaliser un test de fisher, il faut établir le modèle non contraint (i.e le modèle complet) et le modèle contraint.

Modèle non contraint (i.e modèle complet)

On réutilisera les valeurs calculés à la **question 10** car le modèle complet (non contraint) reste inchangé.

On rappel :

- $n = 428$
- $k = 7$
- SSRO = 188.5899801926394

(constante, city, educ, exper, nwifeinc, kidslt6, kidsge6)

Modèle contraint $(constante, kidslt6, kidsge6)$

avec

$$X = constante + \beta_{city}(X_{city} - X_{nwifeinc}) - \beta_{exper}(X_{exper} - X_{educ}) + \beta_{kidslt6}X_{kidslt6} + \beta_{kidsge6}X_{kidsge6}$$

$$Y = lwage - 0.1X_{educ} - 0.1X_{nwifeinc}$$

OLS Regression Results

Dep. Variable:	y	R-squared:	0.705			
Model:	OLS	Adj. R-squared:	0.702			
Method:	Least Squares	F-statistic:	252.9			
Date:	Tue, 03 May 2022	Prob (F-statistic):	9.73e-111			
Time:	16:41:27	Log-Likelihood:	-432.86			
No. Observations:	428	AIC:	875.7			
Df Residuals:	423	BIC:	896.0			
Df Model:	4					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	-0.2133	0.079	-2.700	0.007	-0.369	-0.058
x1	0.0948	0.003	29.720	0.000	0.089	0.101
x2	0.0141	0.004	3.275	0.001	0.006	0.023
x3	-0.0371	0.085	-0.437	0.662	-0.204	0.130
x4	-0.0159	0.026	-0.606	0.545	-0.068	0.036
Omnibus:	78.883	Durbin-Watson:			1.973	
Prob(Omnibus):	0.000	Jarque-Bera (JB):			272.220	
Skew:	-0.803	Prob(JB):			7.73e-60	
Kurtosis:	6.562	Cond. No.			59.4	

Warnings:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
SSR1 = 189.4166475528712

Calcul de Statistique de Fisher

```

1 F = ((SSR1 - SSRO)/2)/(SSRO/(n-k))
2 print("La statistique de Fisher de l'hypothèse jointe est : \n F-stat = ", F)

```

La statistique de Fisher de l'hypothèse jointe est :

F-stat = 0.9227079781812848

Calcul de la p-valeur

```

1 p_val = stats.f.sf(F,2,n-k)
2 print("La p-valeur de l'hypothèse jointe est \n p-valeur = ", p_val)

```

La p-valeur de l'hypothèse jointe est

p-valeur = 0.3982435347480464

Commentaires

On a posé les hypothèses jointes suivantes :

$H_0 : \beta_{nwifeinc} + \beta_{city} = 0.1$ et $\beta_{educ} + \beta_{exper} = 0.1$

$H_1 : \beta_{nwifeinc} + \beta_{city} \neq 0.1$ et $\beta_{educ} + \beta_{exper} \neq 0.1$

Pour ce test à un seuil $\alpha = 5\%$ on obtient :

- $stat - Fischer = 0.922$
- $p - valeur = 0.398$

On remarque que $p - valeur = 0.398 > \alpha = 0.05$. Donc on ne rejette pas l'hypothèse jointe $H_0 : \beta_{nwifeinc} + \beta_{city} = 0.1$ et $\beta_{educ} + \beta_{exper} = 0.1$ à un seuil de 5%.

1.13 Question 12

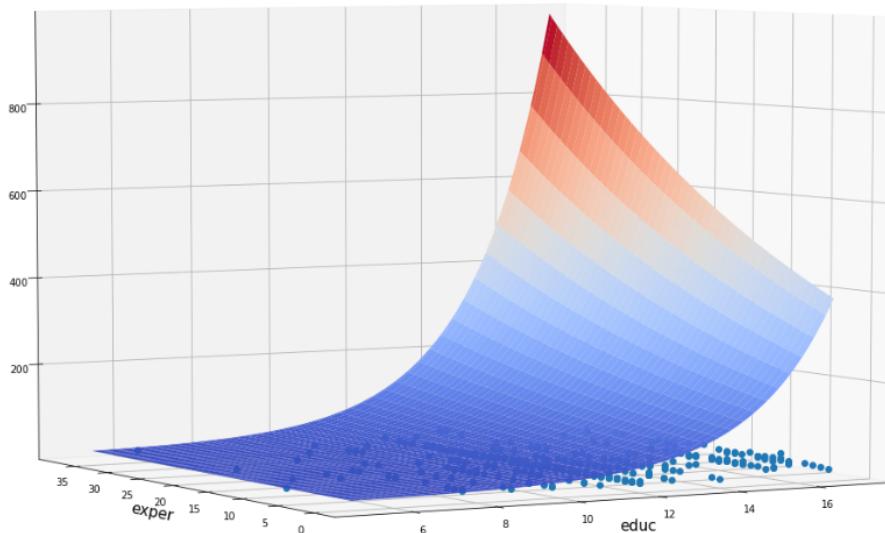
Faites une représentation graphique de la manière dont le salaire augmente avec l'éducation et l'expérience professionnelle. Commentez

Pour analyser la manière dont le salaire (`wage`) augmente avec le nombre d'années d'étude (`educ`) et le nombre d'années d'expérience professionnelle (`exper`), il convient de réaliser la régression linéaire suivante :

$$\exp(lwage) = \exp(constante + \beta_{educ}educ + \beta_{exper}exper)$$

De cette façon à l'aide d'un graphique en 3D, avec comme axe (x, y, z) = (`educ, exper, wage`), il sera possible de visualiser le plan de régression

Représentation graphique de la manière dont le salaire augmente (`wage`) avec l'éducation (`educ`) et l'expérience professionnelle (`exper`)



Commentaires

On remarque que nous obtenons un plan de régression non linéaire. Cela s'explique par le fait que nous avons estimé avec la variable `lwage` ($\log(wage)$). C'est pourquoi, le plan de régression semble adopter une forme convexe.

Le plan de régression montre bien que lorsque les variables `educ` et `exper` prennent des valeurs élevées, le salaire augmente de manière exponentielle. Cela signifie que plus le nombre des années d'études et des années d'expérience sont élevées, plus le salaire sera élevé.

1.14 Question 13

Tester l'égalité des coefficients associés aux variables `kidsgt6` et `kidsge6`. Interprétez.

Test d'égalité des coefficients

Tester l'égalité des coefficients associés aux variables `kidsgt6` et `kidsge6` revient à tester l'hypothèse
 $H_0 : \beta_{kidsgt6} = \beta_{kidsge6}$

Pour ce test d'hypothèse on pose:

$$\beta_{kidsgt6} = \beta_{kidsge6}$$

$$lwage = constante + \beta_{city}X_{city} + \beta_{educ}X_{educ} + \beta_{exper}X_{exper} + \beta_{nwifeinc}X_{nwifeinc} + \beta_{kidslt6}X_{kidslt6} + \beta_{kidsge6}X_{kidsge6}$$

$$lwage = constante + \beta_{city}X_{city} + \beta_{educ}X_{educ} + \beta_{exper}X_{exper} + \beta_{nwifeinc}X_{nwifeinc} + \beta_{kidslt6}(X_{kidslt6} + X_{kidsge6})$$

Pour cela il est nécessaire d'écrire le modèle en fonction du paramètre $\theta = \beta_{kidsgt6} - \beta_{kidsge6} = 0$

Ce qui revient à faire une régression de y sur : (*constante, city, educ, exper, nwifeinc, kidslt6, kidsge6 + kidslt6*)

Pour ce test d'hypothèse on pose alors :

- $H_0 : \beta_{kidsgt6} = \beta_{kidsge6}$
- $H_1 : \beta_{kidsge6} \neq \beta_{kidsgt6}$

Regressions du modèle :

$$X = (\text{constante}, \text{city}, \text{educ}, \text{exper}, \text{nwifeinc}, \text{kidslt6}, \text{kidsge6} + \text{kidslt6})$$

$$y = \text{lwage}$$

OLS Regression Results									
Dep. Variable:	lwage	R-squared:	0.156						
Model:	OLS	Adj. R-squared:	0.144						
Method:	Least Squares	F-statistic:	12.92						
Date:	Tue, 03 May 2022	Prob (F-statistic):	2.00e-13						
Time:	16:41:28	Log-Likelihood:	-431.92						
No. Observations:	428	AIC:	877.8						
Df Residuals:	421	BIC:	906.3						
Df Model:	6								
Covariance Type:	nonrobust								
	coef	std err	t	P> t	[0.025	0.975]			
const	-0.3990	0.207	-1.927	0.055	-0.806	0.008			
x1	0.0353	0.070	0.503	0.616	-0.103	0.173			
x2	0.1022	0.015	6.771	0.000	0.073	0.132			
x3	0.0155	0.004	3.452	0.001	0.007	0.024			
x4	0.0049	0.003	1.466	0.143	-0.002	0.011			
x5	-0.0336	0.090	-0.372	0.710	-0.211	0.144			
x6	-0.0117	0.027	-0.434	0.664	-0.065	0.041			
Omnibus:	79.542	Durbin-Watson:	1.979						
Prob(Omnibus):	0.000	Jarque-Bera (JB):	287.193						
Skew:	-0.795	Prob(JB):	4.33e-63						
Kurtosis:	6.685	Cond. No.	178.						

Warnings:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Commentaires

On a posé pour ce test d'hypothèse, les hypothèses suivantes:

- $H_0 : \beta_{kidsgt6} = \beta_{kidsgt6}$
- $H_1 : \beta_{kidsge6} \neq \beta_{kidsge6}$

Notre nouvelle variable θ en x_5 à comme $p - \text{valeur}_{x_5} = 0.071$

On remarque que $p - \text{valeur} = 0.071 > \alpha = 0.05$. Donc on ne rejette pas l'hypothèse $H_0 : \beta_{kidsgt6} = \beta_{kidsgt6}$ à un seuil de 5%. Donc, cela revient également à rejeter l'hypothèse $H_1 : \beta_{kidsge6} \neq \beta_{kidsge6}$

1.15 Question 14

Faire le test d'hétérosécédasticité de forme linéaire en donnant la p-valeur. Déterminer la ou les sources d'hétérosécédasticité et corriger avec les méthodes vues en cours. Comparer les écarts-types des coefficients estimés avec ceux obtenus à la question 7. Commenter.

Question 14

Faire le test d'hétérosécédasticité de forme linéaire en donnant la p-valeur. Déterminer la ou les sources d'hétérosécédasticité et corriger avec les méthodes vues en cours. Comparer les écarts-types des coefficients estimés avec ceux obtenus à la question 7. Commenter.

En régression linéaire, le fait que les résidus du modèle ne soient pas homoscédastiques a pour conséquence que les coefficients du modèle estimés par la méthode des MCO ne soient pas sans biais ce qui conduit à une estimation de leur variance qui n'est pas fiable.

Il convient donc, si l'on soupçonne que les variances ne sont pas homogènes (une simple représentation des résidus en fonction des variables explicatives peut révéler une hétérosécédasticité), d'effectuer un test d'hétérosécédasticité. Plusieurs tests ont été mis au point, avec pour hypothèses nulle et non nulle :

H_0 : Les résidus sont homoscédastiques

H_1 : Les résidus sont hétérosécédastiques

Pour tester l'hypothèse d'homoscédasticité du modèle, on peut utiliser la regression des résidus au carré (u^2) en fonction des variables du modèle.

Pour ce faire, on doit tester la significativité globale de la régression suivante :

$$u^2 = \text{constante} + \beta_{city} city + \beta_{educ} educ + \beta_{exper} exper + \beta_{nwifeinc} nwifeinc + \beta_{kidslt6} kidslt6 + \beta_{kidsge6} kidsge6$$

Ceci revient donc à tester :

$$H_0 : \beta_{city} = \beta_{educ} = \beta_{exper} = \beta_{nwifeinc} = \beta_{kidslt6} = \beta_{kidsge6} = \beta_{educ} = 0$$

Il convient alors de faire un test de Fischer

Regression du modèle :

$$X = (\text{constante}, \text{city}, \text{educ}, \text{exper}, \text{nwifeinc}, \text{kidslt6}, \text{kidsge6} + \text{kidslt6})$$

$$y = \text{lwage}$$

----- Rapport de Regression pour y = lwage -----

OLS Regression Results

Dep. Variable:	lwage	R-squared:	0.156			
Model:	OLS	Adj. R-squared:	0.144			
Method:	Least Squares	F-statistic:	12.92			
Date:	Tue, 03 May 2022	Prob (F-statistic):	2.00e-13			
Time:	16:41:28	Log-Likelihood:	-431.92			
No. Observations:	428	AIC:	877.8			
Df Residuals:	421	BIC:	906.3			
Df Model:	6					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	-0.3990	0.207	-1.927	0.055	-0.806	0.008
x1	0.0353	0.070	0.503	0.616	-0.103	0.173
x2	0.1022	0.015	6.771	0.000	0.073	0.132
x3	0.0155	0.004	3.452	0.001	0.007	0.024
x4	0.0049	0.003	1.466	0.143	-0.002	0.011
x5	-0.0453	0.085	-0.531	0.596	-0.213	0.122
x6	-0.0117	0.027	-0.434	0.664	-0.065	0.041
Omnibus:	79.542	Durbin-Watson:	1.979			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	287.193			
Skew:	-0.795	Prob(JB):	4.33e-63			
Kurtosis:	6.685	Cond. No.	178.			

----- Rapport de Regression pour $y = u^2$ -----

OLS Regression Results

Dep. Variable:	y	R-squared:	0.028			
Model:	OLS	Adj. R-squared:	0.014			
Method:	Least Squares	F-statistic:	2.009			
Date:	Tue, 03 May 2022	Prob (F-statistic):	0.0633			
Time:	16:41:28	Log-Likelihood:	-622.39			
No. Observations:	428	AIC:	1259.			
Df Residuals:	421	BIC:	1287.			
Df Model:	6					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]
const	0.4519	0.323	1.398	0.163	-0.183	1.087
x1	0.0967	0.110	0.883	0.378	-0.119	0.312
x2	0.0126	0.024	0.536	0.593	-0.034	0.059
x3	-0.0170	0.007	-2.423	0.016	-0.031	-0.003
x4	-0.0036	0.005	-0.683	0.495	-0.014	0.007
x5	0.1279	0.133	0.960	0.337	-0.134	0.390
x6	0.0278	0.042	0.661	0.509	-0.055	0.110
Omnibus:	454.545	Durbin-Watson:			1.642	
Prob(Omnibus):	0.000	Jarque-Bera (JB):			16789.400	
Skew:	4.875	Prob(JB):			0.00	
Kurtosis:	32.093	Cond. No.			178.	

Commentaires

On obtient les résultats suivant :

- statistique de Fisher (F-stat) = 2.009
- p-valeur = 0.0633

Pour le test d'hypothèse d'hétéroscléasticité de forme linéaire, on posé les hypothèses suivantes:

H_0 : Les résidus sont homoscédastiques

Autrement dit : $H_0 : \beta_{city} = \beta_{educ} = \beta_{exper} = \beta_{nwifeinc} = \beta_{kidslt6} = \beta_{kidsge6} = \beta_{educ} = 0$

H_1 : Les résidus sont hétéroscléastiques

Autrement dit : $H_1 : \beta_{city} = \beta_{educ} = \beta_{exper} = \beta_{nwifeinc} = \beta_{kidslt6} = \beta_{kidsge6} = \beta_{educ} \neq 0$

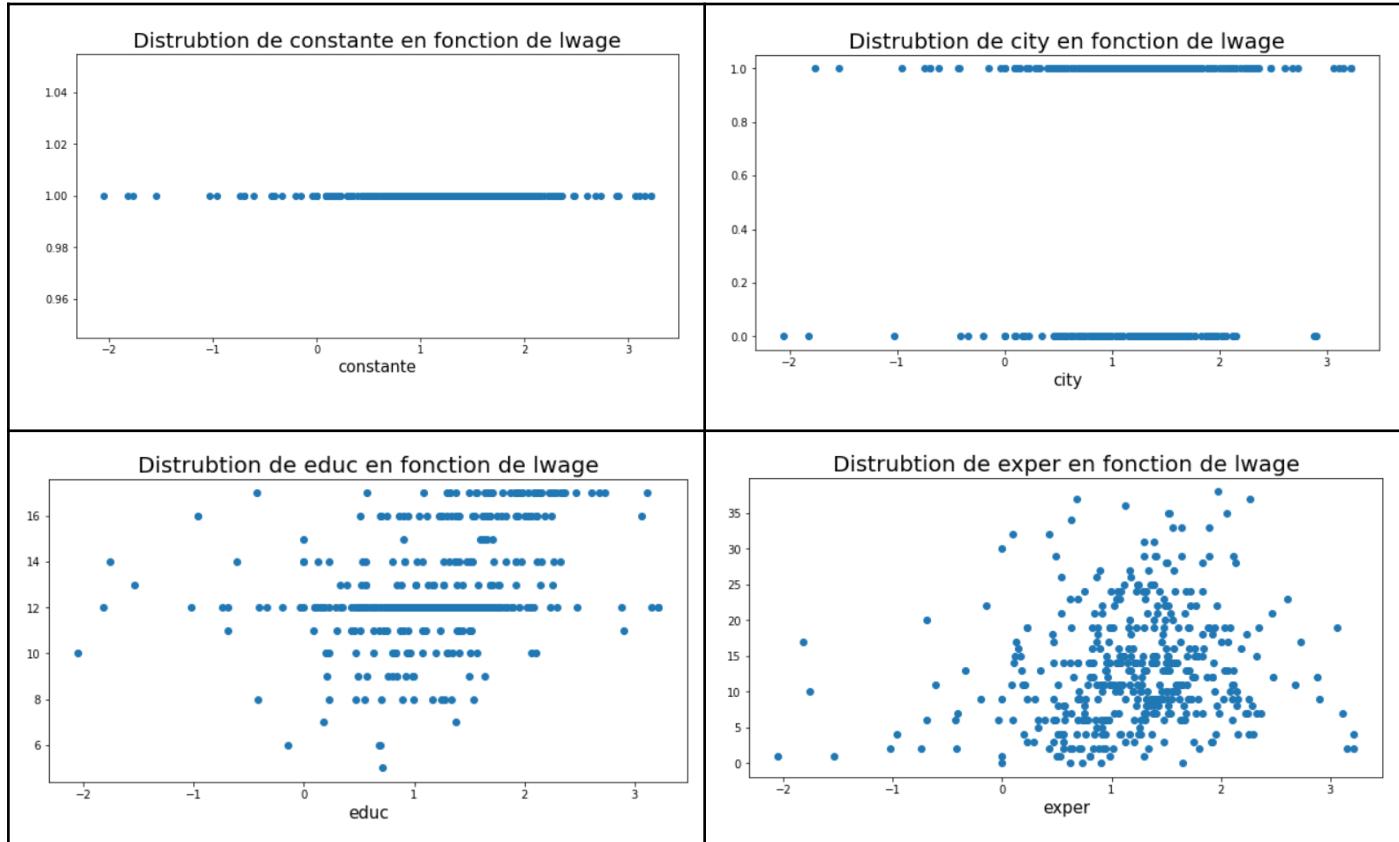
On remarque que $p - valeur = 0.633 > \alpha = 0.05$. Donc on ne rejette pas l'hypothèse H_0 : Les résidus sont homoscédastiques à un seuil de 5%. Par conséquent, cela revient également à rejeter l'hypothèse H_1 : Les résidus sont hétéroscléastiques à 5%.

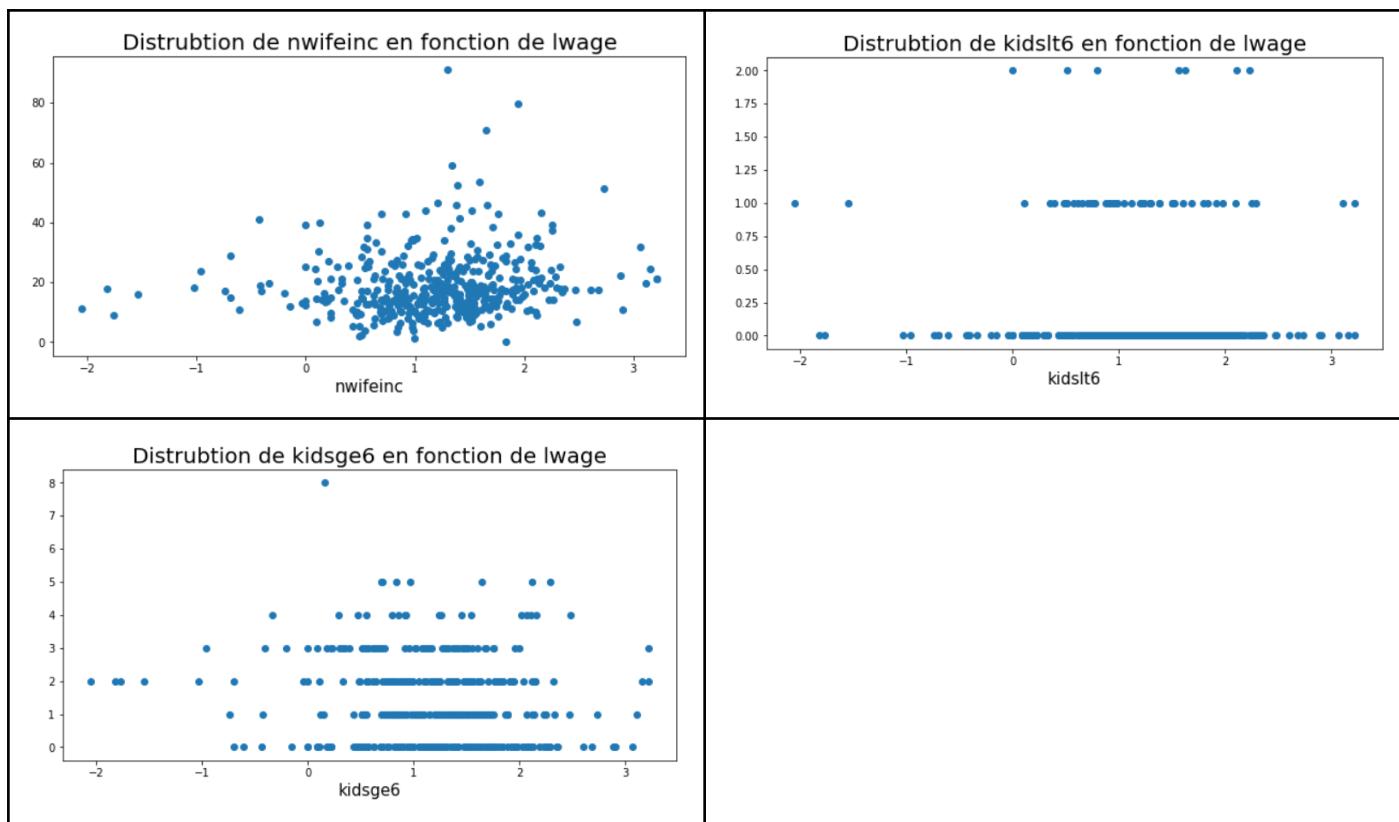
A l'inverse, on rejette l'hypothèse nulle à un seuil très proche de 5% à savoir le seuil de 7%. Ainsi, cela signifie que l'on ne peut pas clairement dire que le modèle est homoscédastique. On peut donc en déduire la présence d'hétéroscléasticité dans le modèle.

L'objectif sera alors de rendre ce modèle clairement homoscédastique en utilisant les méthodes adéquates de façon à faire augmenter la p-value et ainsi pouvoir ne pas rejeter largement l'hypothèse H_0 : modèle homoscédastique.

Analysons les sources potentielles d'hétéroscléasticité du modèle.

Pour accomplir cela, il convient d'afficher le nuage de points entre les variables explicatives (i.e `city` , `educ` , `exper` ,...) avec la variable cible du modèle `lwage`





Commentaires

Au regard des différents graphiques qui ont été tracés, on remarque que 2 phénomènes bien distincts se produisent. Par conséquent il est possible de ranger les graphiques associés à chaque variable explicative dans deux catégories:

- **Catégorie 1:** educ, exper et nwifeinc
- **Catégorie 2:** city, kidslt6, kidsge6

Concernant les variables de la **catégorie 1**, d'abord, on remarque que ces dernières sont des variables continues. De plus, on constate qu'ils semblent affecter non linéairement la variable cible `lwage`.

Par ailleurs, concernant les variables de la **catégorie 2**, on constate que ces dernières sont des variables discrètes car elles prennent uniquement un certains nombre de valeurs entières. Par exemple, la variable `city` est présente sous une forme binaire. Les variables `kidslt6` (`kids < 6 years`) et `kidsge6` (`kids 6-18`) sont présentes sous une forme continue mais au regard de la définition associée à ces 2 variables, elle seraient plus envisageable de les interpréter sous une forme binaire.

Ainsi on peut identifier 2 sources probables d'hétéroscédasticité.

Pour corriger l'hétéroscédasticité 2 choix s'offrent à nous :

1ère stratégie

La première méthode consiste à refaire le test d'hypothèse que nous avons réalisé précédemment tout en y effectuant des transformations sur les variables. La première transformation consiste à modifier les variables `educ`, `exper` et `nwifeinc` en log (un modèle log est généralement plus robuste au phénomène d'hétéroscédasticité). La seconde transformation consiste à transformer les variables `kidslt6` et `kidsge6` en un ensemble de variables binaires, la variable `city` étant déjà binaire.

2ème stratégie

La seconde méthode consiste à refaire le test d'hypothèse en utilisant un estimateur WLS. Cela revient à utiliser une variable pour pondérer les observations.

1ère stratégie

- Transformation de `educ`, `exper` et `nwifeinc` en log
- Transformation de `kidslt6` et `kidsge6` en un ensemble de variables binaires

Avant d'opérer la transformation des variables `kidslt6` et `kidsge6` en un ensemble de variables binaires il convient d'observer les valeurs prises par ces variables

Valeurs possibles prises par `kidslt6`

```
0    375  
1    46  
2     7  
Name: kidslt6, dtype: int64
```

Valeurs possibles prises par `kidsge6`

```
0    149  
1    99  
2    97  
3    58  
4    17  
5     7  
8     1  
Name: kidsge6, dtype: int64
```

Commentaires

`kidslt6` prend ses valeurs dans l'ensemble {0, 1, 2}

`kidsge6` prend ses valeurs dans l'intervalle {0, 1, 2, 3, 4, 5, 8}

Lors du passage en log des variables `educ`, `exper` et `nwifeinc` il est possible que des valeurs `-inf` ou `NaN` apparaissent

```
nombre de valeurs inf pour log(educ) 0  
nombre de valeurs -inf pour log(educ) 0  
nombre de valeurs NaN pour log(educ) 0
```

```
nombre de valeurs inf pour log(exper) : 0  
nombre de valeurs -inf pour log(exper) : 5  
nombre de valeurs NaN pour log(exper) 0
```

```
nombre de valeurs inf pour log(nwifeinc) : 0  
nombre de valeurs -inf pour log(nwifeinc) : 0  
nombre de valeurs NaN pour log(nwifeinc) 1
```

Commentaires

En passant la variable `exper` en log, on remarque que la variable prend **5** valeurs `-inf`. Avec ces valeurs la regression ne peut pas être faites correctement par python. Pour pallier ce problème, il convient de remplacer les valeurs `-inf` de la variable `exper` par 0

Par ailleurs on constate également que lors du passage de la variable `nwifeinc` en log, une valeur manquante `NaN` apparaît. Pour corriger ce problème, on remplace la valeur manquante par 0

Modèle de régression (Stratégie 1)

$X = (\text{constante}, \text{city}, \text{np. log(educ)}, \text{np. log(exper)}, \text{np. log(nwifeinc)}, \text{kidslt6}_0, \text{kidslt6}_1, \text{kidslt6}_2, \text{kidsge6}_0, \text{kidsge6}_1, \text{kidsge6}_2, \text{kidsge6}_3, \text{kidsge6}_4, \text{kidsge6}_5, \text{kidsge6}_8)$

$y = \text{l wage}$

OLS Regression Results											
Dep. Variable:	lwage	R-squared:	0.187								
Model:	OLS	Adj. R-squared:	0.164								
Method:	Least Squares	F-statistic:	7.974								
Date:	Wed, 04 May 2022	Prob (F-statistic):	1.78e-13								
Time:	10:56:53	Log-Likelihood:	-423.70								
No. Observations:	428	AIC:	873.4								
Df Residuals:	415	BIC:	926.2								
Df Model:	12										
Covariance Type:	nonrobust										
	coef	std err	t	P> t	[0.025	0.975]					
const	-1.8538	0.321	-5.772	0.000	-2.485	-1.222					
x1	0.0008	0.071	0.012	0.991	-0.139	0.140					
x2	1.1975	0.183	6.543	0.000	0.838	1.557					
x3	0.2000	0.046	4.358	0.000	0.110	0.290					
x4	0.1169	0.065	1.812	0.071	-0.010	0.244					
x5	-0.5811	0.123	-4.722	0.000	-0.823	-0.339					
x6	-0.5823	0.143	-4.078	0.000	-0.863	-0.302					
x7	-0.6903	0.222	-3.111	0.002	-1.126	-0.254					
x8	-0.1691	0.121	-1.402	0.162	-0.406	0.068					
x9	-0.1173	0.124	-0.944	0.346	-0.362	0.127					
x10	-0.2586	0.125	-2.063	0.040	-0.505	-0.012					
x11	-0.3159	0.129	-2.442	0.015	-0.570	-0.062					
x12	0.0592	0.171	0.346	0.730	-0.278	0.396					
x13	0.1935	0.240	0.805	0.421	-0.279	0.666					
x14	-1.2456	0.582	-2.141	0.033	-2.389	-0.102					
Omnibus:	72.924	Durbin-Watson:	2.016								
Prob(Omnibus):	0.000	Jarque-Bera (JB):	281.819								
Skew:	-0.700	Prob(JB):	6.37e-62								
Kurtosis:	6.720	Cond. No.	3.06e+16								

---- y = u² ----

OLS Regression Results

Dep. Variable:	y	R-squared:	0.047
Model:	OLS	Adj. R-squared:	0.019
Method:	Least Squares	F-statistic:	1.707
Date:	Wed, 04 May 2022	Prob (F-statistic):	0.0629
Time:	10:56:53	Log-Likelihood:	-603.01
No. Observations:	428	AIC:	1232.
Df Residuals:	415	BIC:	1285.
Df Model:	12		
Covariance Type:	nonrobust		
coef	std err	t	P> t [0.025 0.975]
const	0.3003	0.488	0.615 0.539 -0.660 1.260
x1	0.0905	0.108	0.839 0.402 -0.122 0.303
x2	0.1015	0.278	0.365 0.716 -0.446 0.648
x3	-0.2022	0.070	-2.898 0.004 -0.339 -0.065
x4	0.0407	0.098	0.415 0.679 -0.152 0.233
x5	0.0517	0.187	0.276 0.782 -0.316 0.420
x6	0.1021	0.217	0.470 0.638 -0.325 0.529
x7	0.1465	0.337	0.434 0.664 -0.517 0.810
x8	0.1249	0.183	0.681 0.496 -0.236 0.485
x9	-0.0539	0.189	-0.285 0.776 -0.425 0.317
x10	0.2722	0.191	1.428 0.154 -0.102 0.647
x11	0.0684	0.197	0.348 0.728 -0.318 0.455
x12	0.1965	0.261	0.754 0.451 -0.316 0.709
x13	-0.1377	0.365	-0.377 0.706 -0.856 0.580
x14	-0.1701	0.884	-0.192 0.848 -1.909 1.568
Omnibus:	449.718	Durbin-Watson:	1.620
Prob(Omnibus):	0.000	Jarque-Bera (JB):	16960.905
Skew:	4.775	Prob(JB):	0.00
Kurtosis:	32.324	Cond. No.	3.06e+16

Commentaires

On remarque que la méthode utilisée dans la 1ère stratégie ne s'est pas avérée fructueuse. En effet, en effectuant les transformations sur les variables en question, nous obtenons une $p - valeur_{strat1} = 0.0629$ quasi identique à celle du modèle initial qui était de $p - valeur_{initial} = 0.0633$. On ne peut toujours pas rejeter H_0 : Les résidus sont homoscédatiques. Par conséquent, les conclusions restent les mêmes que ceux émises pour le modèle initiale

2ème stratégie

La seconde méthode consiste à refaire le test d'hypothèse en utilisant cette fois-ci un estimateur WLS. Cela revient à utiliser un variable pour pondérer les observations.

A noter que l'on conservera les transformations de variable effectuées lors de la stratégie 1. Cela signifie que finalement, nous combinons la stratégie 1 (réutilisation du modèle utilisé pour la stratégie 1) et la stratégie 2 (réalisation d'un modèle WLS).

Pour ce faire, on peut utiliser la variable `exper` que l'on transformera en log pour pondérer les observations du modèle.

Modèle de régression WLS (Stratégie 2)

$X = (\text{constante}, \text{city}, \text{np.log(educ)}, \text{np.log(exper)}, \text{np.log(nwifelinc)}, \text{kidslt6}_0, \text{kidslt6}_1, \text{kidslt6}_2, \text{kidsge6}_0, \text{kidsge6}_1, \text{kidsge6}_2, \text{kidsge6}_3, \text{kidsge6}_4, \text{kidsge6}_5, \text{kidsge6}_8)$

$y = \text{lwage}$

Le poid choisi est : `weights = 1/(np.sqrt(np.log(educ))`

```
---- y = lwage ----
                    WLS Regression Results
=====
Dep. Variable:          lwage    R-squared:       0.186
Model:                 WLS    Adj. R-squared:   0.163
Method:                Least Squares    F-statistic:     7.922
Date:      Wed, 04 May 2022    Prob (F-statistic): 2.23e-13
Time:          10:56:53    Log-Likelihood:   -423.05
No. Observations:      428    AIC:             872.1
Df Residuals:         415    BIC:             924.9
Df Model:                  12
Covariance Type:        nonrobust
=====
            coef    std err          t      P>|t|      [0.025      0.975]
-----
const    -1.8041    0.315    -5.719      0.000    -2.424    -1.184
x1       0.0017    0.071     0.024      0.981    -0.137     0.141
x2       1.1659    0.179     6.507      0.000     0.814     1.518
x3       0.2006    0.046     4.395      0.000     0.111     0.290
x4       0.1190    0.064     1.849      0.065    -0.008     0.246
x5      -0.5666    0.122    -4.645      0.000    -0.806    -0.327
x6      -0.5713    0.142    -4.025      0.000    -0.850    -0.292
x7      -0.6663    0.222    -2.995      0.003    -1.104    -0.229
x8      -0.1631    0.119    -1.365      0.173    -0.398     0.072
x9      -0.1100    0.124    -0.891      0.374    -0.353     0.133
x10     -0.2511    0.125    -2.016      0.044    -0.496    -0.006
x11     -0.3041    0.128    -2.368      0.018    -0.557    -0.052
x12     0.0685    0.170     0.404      0.686    -0.265     0.402
x13     0.1948    0.237     0.821      0.412    -0.272     0.661
x14     -1.2392    0.579    -2.140      0.033    -2.378    -0.101
=====
Omnibus:           71.903    Durbin-Watson:      2.014
Prob(Omnibus):    0.000    Jarque-Bera (JB): 278.652
Skew:              -0.688    Prob(JB):        3.10e-61
Kurtosis:          6.705    Cond. No.        1.85e+16
=====
```

---- $y = u^2$ ----

OLS Regression Results

Dep. Variable:	y	R-squared:	0.047			
Model:	OLS	Adj. R-squared:	0.020			
Method:	Least Squares	F-statistic:	1.713			
Date:	Wed, 04 May 2022	Prob (F-statistic):	0.0615			
Time:	10:56:53	Log-Likelihood:	-602.53			
No. Observations:	428	AIC:	1231.			
Df Residuals:	415	BIC:	1284.			
Df Model:	12					
Covariance Type:	nonrobust					
coef	std err	t	P> t			
[0.025	0.975]					
const	0.2821	0.488	0.578	0.563	-0.677	1.241
x1	0.0896	0.108	0.832	0.406	-0.122	0.302
x2	0.1130	0.278	0.407	0.684	-0.433	0.659
x3	-0.2024	0.070	-2.904	0.004	-0.339	-0.065
x4	0.0400	0.098	0.408	0.683	-0.153	0.233
x5	0.0465	0.187	0.249	0.804	-0.321	0.414
x6	0.0974	0.217	0.449	0.653	-0.329	0.524
x7	0.1382	0.337	0.410	0.682	-0.524	0.801
x8	0.1225	0.183	0.669	0.504	-0.238	0.483
x9	-0.0564	0.189	-0.299	0.765	-0.427	0.315
x10	0.2690	0.190	1.414	0.158	-0.105	0.643
x11	0.0645	0.197	0.328	0.743	-0.322	0.451
x12	0.1927	0.260	0.740	0.459	-0.319	0.704
x13	-0.1372	0.365	-0.376	0.707	-0.855	0.580
x14	-0.1730	0.883	-0.196	0.845	-1.910	1.564
Omnibus:	449.611	Durbin-Watson:	1.618			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	16961.794			
Skew:	4.773	Prob(JB):	0.00			
Kurtosis:	32.326	Cond. No.	3.06e+16			

Commentaires

Avec cette seconde stratégie, nous obtenons une $p - valeur_{strat_2} = 0.0615$ pour une statistique de Fisher $F - stat_{strat_2} = 1.713$.

Etant donné que $p - valeur_{strat_2} = 0.0615 > \alpha = 0.05$. Ainsi on peut ne toujours pas rejeter l'hypothèse H_0 : Les résidus sont homoscédatiques à un seuil de 5%.

En conclusion, en combinant la stratégie 1 et la stratégie 2, les transformations effectuées, n'ont pas permises de corriger l'hétérosécédasticité du modèle. Par conséquent la situation n'a pas évolué par rapport à celle du modèle initial. Nous pouvons tirer les mêmes conclusions que ceux qui ont été évoqués pour le modèle initial.

Stratégie n°3

La 3ème stratégie est de réaliser un modèle GLS car les moindres carrés pondérés permettent de rendre plus probable l'homocédasticité.

Pour cette 3ème stratégie nous n'allons pas transformer les variables `educ`, `exper`, `nwifeinc` en *log* et les variables `kidslt6`, `kidsge6` en variables binaires. L'idée est ici de réaliser un modèle GLS sur les variables initiales du modèle.

Modèle de régression GLS (Stratégie 3)

$$X = (\text{constante}, \text{city}, \text{educ}, \text{exper}, \text{nwifeinc}, \text{kidslt6}, \text{kidsge6})$$

$$y = \text{l wage}$$

----- Rapport de Regression pour $y = \text{l wage}$ -----

OLS Regression Results

Dep. Variable:	l wage	R-squared:	0.156
Model:	OLS	Adj. R-squared:	0.144
Method:	Least Squares	F-statistic:	12.92
Date:	Wed, 04 May 2022	Prob (F-statistic):	2.00e-13
Time:	10:56:53	Log-Likelihood:	-431.92
No. Observations:	428	AIC:	877.8
Df Residuals:	421	BIC:	906.3
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t
const	-0.3990	0.207	-1.927
x1	0.0353	0.070	0.503
x2	0.1022	0.015	6.771
x3	0.0155	0.004	3.452
x4	0.0049	0.003	1.466
x5	-0.0453	0.085	-0.531
x6	-0.0117	0.027	-0.434

Omnibus:	79.542	Durbin-Watson:	1.979
Prob(Omnibus):	0.000	Jarque-Bera (JB):	287.193
Skew:	-0.795	Prob(JB):	4.33e-63
Kurtosis:	6.685	Cond. No.	178.

----- Rapport de Regression pour $y = u^2$ -----

OLS Regression Results									
Dep. Variable:	y	R-squared:	0.028						
Model:	OLS	Adj. R-squared:	0.014						
Method:	Least Squares	F-statistic:	2.009						
Date:	Wed, 04 May 2022	Prob (F-statistic):	0.0633						
Time:	10:56:53	Log-Likelihood:	-622.39						
No. Observations:	428	AIC:	1259.						
Df Residuals:	421	BIC:	1287.						
Df Model:	6								
Covariance Type:	nonrobust								
	coef	std err	t	P> t	[0.025	0.975]			

const	0.4519	0.323	1.398	0.163	-0.183	1.087			
x1	0.0967	0.110	0.883	0.378	-0.119	0.312			
x2	0.0126	0.024	0.536	0.593	-0.034	0.059			
x3	-0.0170	0.007	-2.423	0.016	-0.031	-0.003			
x4	-0.0036	0.005	-0.683	0.495	-0.014	0.007			
x5	0.1279	0.133	0.960	0.337	-0.134	0.390			
x6	0.0278	0.042	0.661	0.509	-0.055	0.110			

Omnibus:	454.545	Durbin-Watson:	1.642						
Prob(Omnibus):	0.000	Jarque-Bera (JB):	16789.400						
Skew:	4.875	Prob(JB):	0.00						
Kurtosis:	32.093	Cond. No.	178.						

----- NOUVEAU MODELE -----

----- Rapport de Regression GLS pour y = lwage -----

GLS Regression Results

Dep. Variable:	lwage	R-squared:	0.137
Model:	GLS	Adj. R-squared:	0.125
Method:	Least Squares	F-statistic:	11.18
Date:	Wed, 04 May 2022	Prob (F-statistic):	1.38e-11
Time:	10:56:53	Log-Likelihood:	-566.96
No. Observations:	428	AIC:	1148.
Df Residuals:	421	BIC:	1176.
Df Model:	6		
Covariance Type:	nonrobust		
coef	std err	t	P> t
const	-0.5814	0.247	-2.357
x1	0.0289	0.085	0.340
x2	0.1054	0.018	5.866
x3	0.0282	0.007	4.061
x4	0.0033	0.004	0.819
x5	-0.0334	0.070	-0.475
x6	0.0188	0.029	0.639
Omnibus:	108.367	Durbin-Watson:	1.997
Prob(Omnibus):	0.000	Jarque-Bera (JB):	1285.209
Skew:	-0.702	Prob(JB):	8.33e-280
Kurtosis:	11.373	Cond. No.	175.

----- Rapport de Regression pour $y = u^2$ -----

OLS Regression Results									
Dep. Variable:	y	R-squared:	0.020						
Model:	OLS	Adj. R-squared:	0.006						
Method:	Least Squares	F-statistic:	1.401						
Date:	Wed, 04 May 2022	Prob (F-statistic):	0.213						
Time:	10:56:53	Log-Likelihood:	-630.74						
No. Observations:	428	AIC:	1275.						
Df Residuals:	421	BIC:	1304.						
Df Model:	6								
Covariance Type:	nonrobust								
	coef	std err	t	P> t	[0.025	0.975]			

const	0.4122	0.329	1.251	0.212	-0.235	1.060			
x1	0.0774	0.112	0.693	0.488	-0.142	0.297			
x2	0.0137	0.024	0.569	0.570	-0.034	0.061			
x3	-0.0133	0.007	-1.863	0.063	-0.027	0.001			
x4	-0.0037	0.005	-0.692	0.489	-0.014	0.007			
x5	0.1181	0.136	0.870	0.385	-0.149	0.385			
x6	0.0316	0.043	0.738	0.461	-0.053	0.116			
=====									
Omnibus:	458.329	Durbin-Watson:	1.626						
Prob(Omnibus):	0.000	Jarque-Bera (JB):	17816.869						
Skew:	4.918	Prob(JB):	0.00						
Kurtosis:	33.039	Cond. No.	178.						
=====									

Commentaires

Pour ce test au seuil $\alpha = 5\%$ on obtient :

- $stat - Fischer_{strat_3} = 1.401$
- $p - valeur_{strat_3} = 0.213$

On constate que la p-valeur a bien augmenté ce qui signifie que l'hétéroscédasticité a encore diminué. Ainsi l'hétéroscédasticité du modèle a été corrigée.

Par conséquent étant donné que $p - valeur_{strat_3} = 0.213 > \alpha = 0.05$. Donc on ne peut encore moins rejeter l'hypothèse jointe $H_0 : \beta_{city} = \beta_{educ} = \beta_{exper} = \beta_{nwifeinc} = \beta_{kidslt6} = \beta_{kidsge6} = \beta_{educ} = 0$ à un seuil de 5%.

Comparaison des écarts-types des coefficients estimés avec ceux obtenus à la question 7

Coefficient des écart-types de chaque variable du modèle initial (question 7)

std err

0.207
0.070
0.015
0.004
0.003
0.085
0.027

Coefficient des écart-types de chaque variable du modèle transformé (hétéroscédasticité corrigé)

std err

0.247
0.085
0.018
0.007
0.004
0.070
0.029

Commentaires

On remarque gloagement que l'écart-type associé à chaque variable du modèle transformé (GLS) sont plus élevé que ceux du modèle intial (question 7) à l'exception de la variable `kidslt6`. Par ailleurs en regardant plus finement, on s'aperçoit que l'évolution de l'écart type des variables `educ`, `exper` et `nwifeinc` reste constant entre le modèle initial et le modèle GLS.

1.16 Question 15

Tester le changement de structure de la question 8 entre les femmes qui ont **plus de 43 ans** et les autres : test sur l'ensemble des coefficients. Donnez les p-valeurs

Regression sur le modèle original

$$X = (\text{constante}, \text{city}, \text{educ}, \text{exper}, \text{nwifeinc}, \text{kidslt6}, \text{kidsge6})$$

$$y = \text{lwage}$$

OLS Regression Results									
Dep. Variable:	lwage	R-squared:	0.156						
Model:	OLS	Adj. R-squared:	0.144						
Method:	Least Squares	F-statistic:	12.92						
Date:	Wed, 04 May 2022	Prob (F-statistic):	2.00e-13						
Time:	10:56:53	Log-Likelihood:	-431.92						
No. Observations:	428	AIC:	877.8						
Df Residuals:	421	BIC:	906.3						
Df Model:	6								
Covariance Type:	nonrobust								
	coef	std err	t	P> t	[0.025	0.975]			
---	---	---	---	---	---	---			
const	-0.3990	0.207	-1.927	0.055	-0.806	0.008			
city	0.0353	0.070	0.503	0.616	-0.103	0.173			
educ	0.1022	0.015	6.771	0.000	0.073	0.132			
exper	0.0155	0.004	3.452	0.001	0.007	0.024			
nwifeinc	0.0049	0.003	1.466	0.143	-0.002	0.011			
kidslt6	-0.0453	0.085	-0.531	0.596	-0.213	0.122			
kidsge6	-0.0117	0.027	-0.434	0.664	-0.065	0.041			
---	---	---	---	---	---	---			
Omnibus:	79.542	Durbin-Watson:	1.979						
Prob(Omnibus):	0.000	Jarque-Bera (JB):	287.193						
Skew:	-0.795	Prob(JB):	4.33e-63						
Kurtosis:	6.685	Cond. No.	178.						
---	---	---	---	---	---	---			

On affiche les p-valeurs associés à chaque variable

p-valeur associé à chaque variable du modèle original:

```
const      5.465894e-02
city       6.155457e-01
educ       4.324526e-11
exper      6.133651e-04
nwifeinc   1.434080e-01
kidslt6    5.956626e-01
kidsge6    6.642512e-01
dtype: float64
```

On sélectionne les observations pour la variable age

```
1 df_sup_43 = df[df.age > 43]
2 df_inf_43 = df[df.age <= 43]
```

On Effectue les test sur l'ensemble des coefficients pour le modèle dont les femmes qui ont plus de 43 ans

$$X_{sup_{43}} = (\text{constante}, \text{city}, \text{educ}, \text{exper}, \text{nwifeinc}, \text{kidslt6}, \text{kidsge6})$$

$$y_{sup_{43}} = \text{lwage}$$

```
OLS Regression Results
=====
Dep. Variable:          lwage    R-squared:       0.166
Model:                 OLS     Adj. R-squared:   0.138
Method:                Least Squares F-statistic:    5.925
Date:      Wed, 04 May 2022 Prob (F-statistic): 1.15e-05
Time:          10:56:53 Log-Likelihood:   -179.89
No. Observations:      186    AIC:             373.8
Df Residuals:          179    BIC:             396.4
Df Model:                  6
Covariance Type:        nonrobust
=====
            coef    std err        t      P>|t|      [0.025    0.975]
-----
const    -0.3232    0.276   -1.173     0.242     -0.867    0.221
city     -0.0188    0.110   -0.171     0.865     -0.236    0.199
educ      0.0853    0.020    4.228     0.000      0.045    0.125
exper      0.0178    0.006    3.047     0.003      0.006    0.029
nwifeinc   0.0093    0.005    1.957     0.052    -7.71e-05  0.019
kidslt6   -0.0091    0.332   -0.027     0.978     -0.664    0.646
kidsge6   -0.0249    0.051   -0.487     0.627     -0.126    0.076
=====
Omnibus:           37.611   Durbin-Watson:     2.352
Prob(Omnibus):    0.000    Jarque-Bera (JB): 114.134
Skew:              -0.787   Prob(JB):      1.64e-25
Kurtosis:          6.500    Cond. No.       212.
```

On affiche les p-valeurs associés à chaque variable

p-valeur associé à chaque variable du modèle (age > 43) :

```
const      0.242448
city       0.864526
educ      0.000038
exper     0.002665
nwifeinc   0.051898
kidslt6    0.978104
kidsge6    0.627090
dtype: float64
```

On Effectue les test sur l'ensemble des coefficients pour le modèle dont les femmes qui ont moins de 43 ans

$$X_{inf_{43}} = (\text{constante}, \text{city}, \text{educ}, \text{exper}, \text{nwifeinc}, \text{kidslt6}, \text{kidsge6})$$

$$y_{inf_{43}} = lwage$$

OLS Regression Results						
Dep. Variable:	lwage	R-squared:	0.168			
Model:	OLS	Adj. R-squared:	0.146			
Method:	Least Squares	F-statistic:	7.888			
Date:	Wed, 04 May 2022	Prob (F-statistic):	9.37e-08			
Time:	10:56:53	Log-Likelihood:	-248.69			
No. Observations:	242	AIC:	511.4			
Df Residuals:	235	BIC:	535.8			
Df Model:	6					
Covariance Type:	nonrobust					
	coef	std err	t	P> t	[0.025	0.975]

const	-0.5412	0.318	-1.702	0.090	-1.168	0.085
city	0.0850	0.092	0.920	0.358	-0.097	0.267
educ	0.1167	0.023	5.038	0.000	0.071	0.162
exper	0.0202	0.008	2.540	0.012	0.005	0.036
nwifeinc	0.0026	0.005	0.526	0.599	-0.007	0.012
kidslt6	-0.0933	0.094	-0.994	0.321	-0.278	0.092
kidsge6	-0.0233	0.036	-0.643	0.521	-0.095	0.048

Omnibus:	42.575	Durbin-Watson:	1.807			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	145.224			
Skew:	-0.684	Prob(JB):	2.92e-32			
Kurtosis:	6.540	Cond. No.	186.			

On affiche les p-valeurs associés à chaque variable

p-valeur associé à chaque variable du modèle (age <= 43) :

```
const      9.012661e-02
city       3.584252e-01
educ       9.375604e-07
exper      1.173559e-02
nwifeinc   5.991599e-01
kidslt6    3.212302e-01
kidsge6    5.211483e-01
dtype: float64
```

Calcul de la statistique de Fischer et la p-valeur

```
p-value : 0.5660364450328375
stat de test : 2.031702280317268
fisher : 0.8260374852756255
```

Commentaires

Par ailleurs, la $fischer_{sup_{43},inf_{43}} = 0.83$ est très inférieur à la valeur critique de **2.0317**. On peut alors en déduire de l'absence d'un changement structurel lié à l'âge.

La p-valeur semble confirmer cette absence de rupture car on a pu remarquer que la $p - valeur_{sup_{43},inf_{43}} = 0.566 > \alpha = \{0.01, 0.05, 0.1\}$. Par conséquent, on ne peut pas rejeter l'hypothèse nulle $H_0 : absence de rupture$. On peut donc conclure, qu'il n'y a pas de changement structurel lié à l'âge des femmes.

1.17 Question 16

Ajouter au modèle de la question 7 la variable `huseduc`. Faire ensuite la même régression en décomposant la variable `huseduc` en 4 variables binaires construites selon votre choix. Faire le test de non significativité de l'ensemble des variables binaires. Donnez les p-valeurs et commentez..

La variable `huseduc` correspond au nombre d'années d'études du mari

Modèle constraint = modèle initial :

$(constante, city, educ, exper, nwifeinc, kidslt6, kidsge6)$

$y = lwage$

Modèle complet :

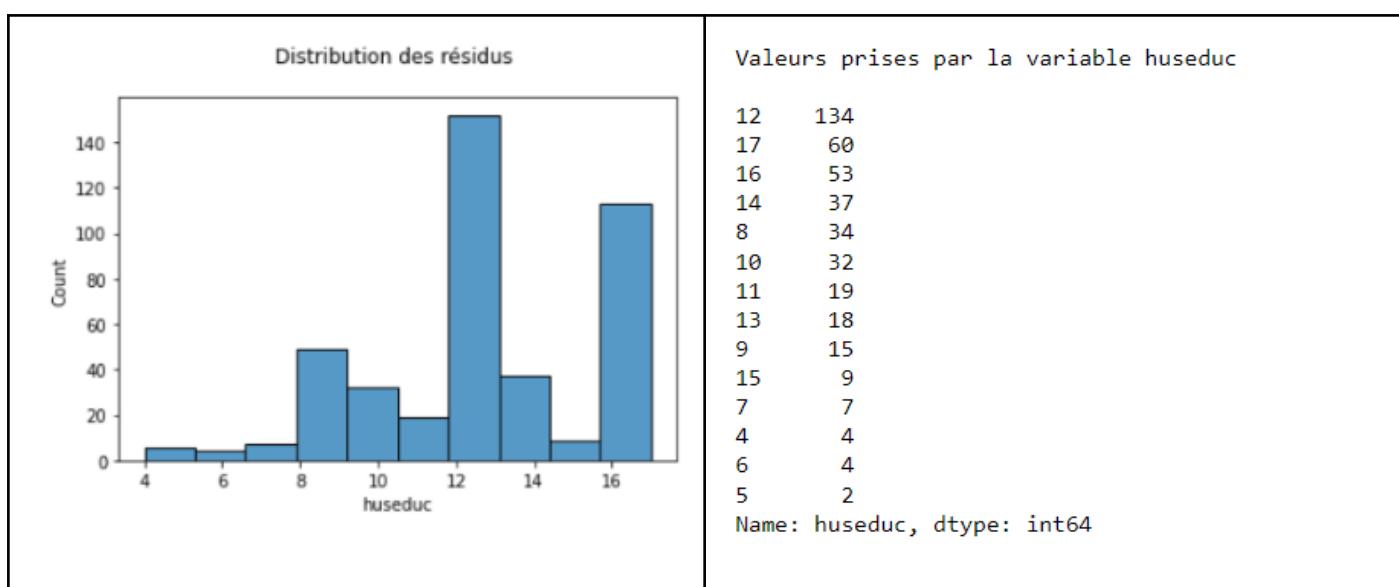
$(constante, educ, exper, kidslt6, kidsge6, huseduc_0, huseduc_1, \dots, huseduc_n)$

$y = lwage$

Le nombre de variable catégorielle de la variable `huseduc` dépendra de l'interprétation que l'on fera lors de la visualisation de la distribution de cette dernière.

Visualisation de la distribution de la variable `huseduc`

Pour transformer la variable `huseduc` en un ensemble de variables catégorielles, il convient d'abord d'observer la distribution des valeurs prises par cette dernière.



On remarque que la variable huseduc à pour maximum 17 et minimum 4.

De plus, au regard de l'histogramme il est possible de créer les 4 variables catégorielles de la manière suivante :

- 1 ère variable [4, 7] -> nombre année d'étude faible
- 2 ème variable [8, 11] -> nombre année d'étude moyen
- 3 ème variable [11, 14] -> nombre anné d'étude élevé
- 4 ème variable [14, 17] -> nombre année d'étude très élevé

Toutefois, il existe une fonction en python qui permet de transformer un variable quantitative en un ensemble de variable catégorielle. Il s'agit de la fonction `qcut` de la librairie `pandas`.

```

1 df_huseduc_cat = pd.DataFrame(pd.get_dummies(pd.qcut(df.huseduc, 4),prefix='col'))
2 df_new = pd.concat([df, df_huseduc_cat], axis=1)
3 df_new = df_new.rename(columns={"col_(3.999, 11.0)": "huseduc_faible",
4                               "col_(11.0, 12.0)": "huseduc_moyen",
5                               "col_(12.0, 16.0)": "huseduc_eleve",
6                               "col_(16.0, 17.0)": "huseduc_très_eleve"
7 })
8 df_new.head()

```

faminc	mtr	motheduc	fatheduc	unem	city	exper	nwifeinc	lwage	expersq	huseduc_faible	huseduc_moyen	huseduc_eleve	huseduc_très_eleve
16310	0.7215	12	7	5.0	0	14	10.910060	1.210154	196	0	1	0	0
21800	0.6615	7	7	11.0	1	5	19.499980	0.328512	25	1	0	0	0
21040	0.6915	12	7	5.0	0	15	12.039910	1.514138	225	0	1	0	0
7300	0.7815	7	7	5.0	0	6	6.799996	0.092123	36	1	0	0	0
27300	0.6215	12	14	9.5	1	7	20.100060	1.524272	49	0	1	0	0

Modèle constraint :

Ici le modèle constraint correspond au modèle initial

(*constante, city, educ, exper, nwifeinc, kidslt6, kidsge6*)

$$y = lwage$$

OLS Regression Results

Dep. Variable:	lwage	R-squared:	0.156			
Model:	OLS	Adj. R-squared:	0.144			
Method:	Least Squares	F-statistic:	12.92			
Date:	Wed, 04 May 2022	Prob (F-statistic):	2.00e-13			
Time:	10:56:54	Log-Likelihood:	-431.92			
No. Observations:	428	AIC:	877.8			
Df Residuals:	421	BIC:	906.3			
Df Model:	6					
Covariance Type:	nonrobust					
coef	std err	t	P> t			
[0.025	0.975]					
const	-0.3990	0.207	-1.927	0.055	-0.806	0.008
x1	0.0353	0.070	0.503	0.616	-0.103	0.173
x2	0.1022	0.015	6.771	0.000	0.073	0.132
x3	0.0155	0.004	3.452	0.001	0.007	0.024
x4	0.0049	0.003	1.466	0.143	-0.002	0.011
x5	-0.0453	0.085	-0.531	0.596	-0.213	0.122
x6	-0.0117	0.027	-0.434	0.664	-0.065	0.041
Omnibus:	79.542	Durbin-Watson:	1.979			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	287.193			
Skew:	-0.795	Prob(JB):	4.33e-63			
Kurtosis:	6.685	Cond. No.	178.			

Warnings:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
SSRO = 188.5899801926394

Modèle complet (i.e non contraint) :

Le modèle contraint correspond au nouveau modèle

(constante, city, educ, exper, nwifeinc, kidslt6, kidsge6, huseducFaible, huseducMoyen, huseducEleve, huseducTresEleve)

$$y = lwage$$

```

n = 428
k = 11
                    OLS Regression Results
=====
Dep. Variable:          lwage   R-squared:       0.166
Model:                 OLS    Adj. R-squared:   0.148
Method:                Least Squares   F-statistic:     9.250
Date:      Wed, 04 May 2022   Prob (F-statistic): 7.82e-13
Time:      10:56:54         Log-Likelihood:   -429.23
No. Observations:      428    AIC:             878.5
Df Residuals:          418    BIC:             919.1
Df Model:               9
Covariance Type:       nonrobust
=====
            coef    std err      t      P>|t|      [0.025      0.975]
-----
const    -0.5423    0.204    -2.664    0.008    -0.942    -0.142
x1        0.0511    0.070     0.726    0.468    -0.087    0.189
x2        0.1184    0.018     6.676    0.000     0.084    0.153
x3        0.0160    0.004     3.569    0.000     0.007    0.025
x4        0.0069    0.003     1.994    0.047    9.72e-05  0.014
x5       -0.0295    0.085    -0.345    0.730    -0.197    0.138
x6       -0.0094    0.027    -0.350    0.726    -0.062    0.043
x7       -0.0676    0.060    -1.120    0.263    -0.186    0.051
x8       -0.0299    0.063    -0.477    0.634    -0.153    0.093
x9       -0.2345    0.087    -2.687    0.008    -0.406    -0.063
x10      -0.2103    0.111    -1.902    0.058    -0.428    0.007
=====
Omnibus:           82.181   Durbin-Watson:    1.964
Prob(Omnibus):    0.000    Jarque-Bera (JB): 295.700
Skew:              -0.824   Prob(JB):       6.16e-65
Kurtosis:          6.723    Cond. No.       3.90e+16
=====
```

Warnings:

- [1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
 - [2] The smallest eigenvalue is 2.15e-28. This might indicate that there are strong multicollinearity problems or that the design matrix is singular.
- SSR1 = 186.23501394299845

Calcule de la statistique de Fischer F-stat

```

1 F = ((SSR0 - SSR1)/4)/(SSR1/(n-k)) # divise par 4 car on a 4 contraintes (i.e Les 4 variables binaires huseduc)
2 print("La statistique de Fisher de l'hypothèse jointe est : \n F-stat = ", F)
```

La statistique de Fisher de l'hypothèse jointe est :
F-stat = 1.3182549635924552

Calcule de la p-valeur

```
1 p_val = stats.f.sf(F,4,n-k)
2 print("La p-valeur de l'hypothèse jointe est \n p-valeur = ", p_val)
```

La p-valeur de l'hypothèse jointe est
p-valeur = 0.26235681163189795

Commentaires

En regardant les p-valeurs de chaque variable binaire on constate que l'ensemble des variables binaires `huseduc` ne sont pas significatives pour le salaire à un seuil de 5%. Toutefois, les variables binaires `husdeuc_faible` et `husdeuc_moyenne` sont significatifs à un seuil proche de 5% à savoir 7%. On pourrait interpréter la variable `huseduc_moyenne` comme la fin des études du lycée. Ainsi, on peut en déduire qu'il existe une différence de profil entre les femmes qui épousent un homme ayant fait des études supérieures et des hommes qui ont fait le choix de ne pas poursuivre d'étude après le lycée. Cela signifie que les femmes épousant un homme ayant fait des études supérieures ont tendance à appartenir à une classe sociale plus élevée, ce qui se répercute sur le salaire.

D'un point de vu plus globale, pour ce test au de seuil $\alpha = 5\%$ on obtient :

- $stat - Fischer = 1.318$
- $p - valeur = 0.262$

On remarque que $p - valeur = 0.262 > \alpha = 0.05$. Donc on ne peut pas rejeter l'hypothèse jointe $H_0 : \beta_{huseduc} = 0$ à un seuil de 5% ni même de 10%.

Par conséquent, on peut en déduire que la variable `huseduc` n'a pas d'effet significatif sur la variable `lwage`. On peut donc en conclure que le niveau d'éducation du mari sur le salaire de son épouse n'est pas significatif.

2. Partie 2 : Série temporelle

2.1 Question 1

Importer les données du fichier quarterly.xls (corriger le problème éventuel d'observations manquantes)

	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

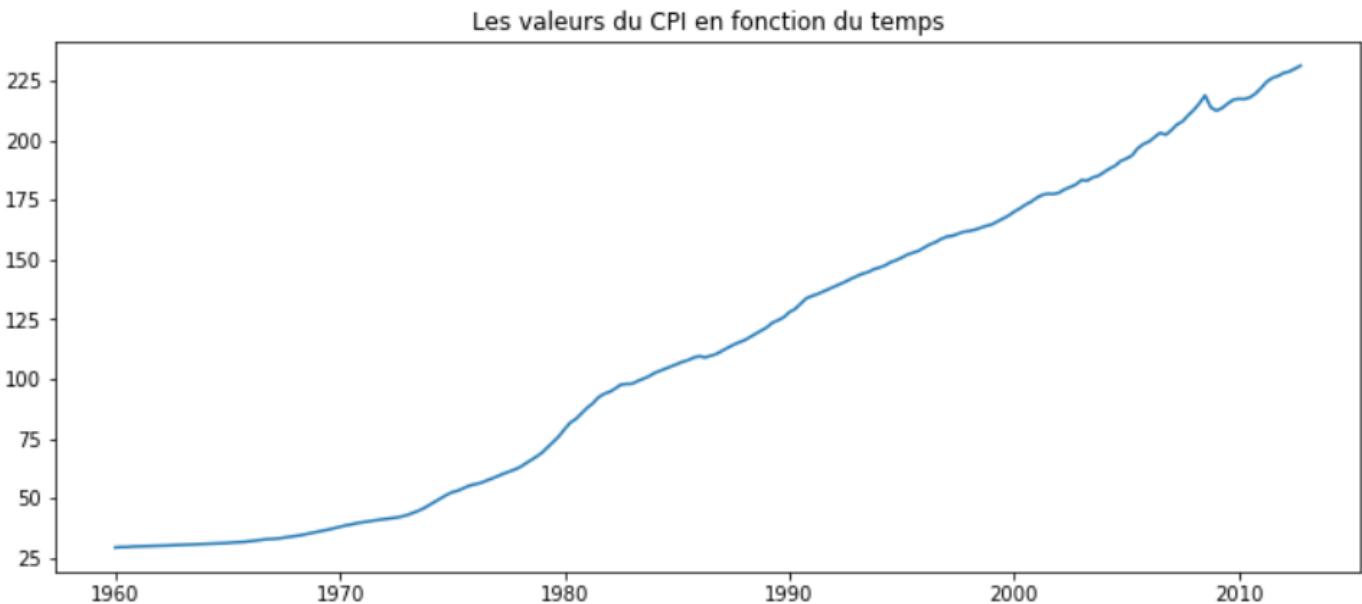
Les dimensions de notre table sont de 212 lignes pour 19 colonnes. Nous n'avons trouvé aucune donnée manquante. Nous n'avons donc pas besoin de faire de pré-traitement

2.2 Question 2

Stationnariser la série de CPI en utilisant la méthode de régression qui inclut un terme de tendance dont la forme fonctionnelle est à choisir (linéaire, quadratique, log, exponentielle, ...)

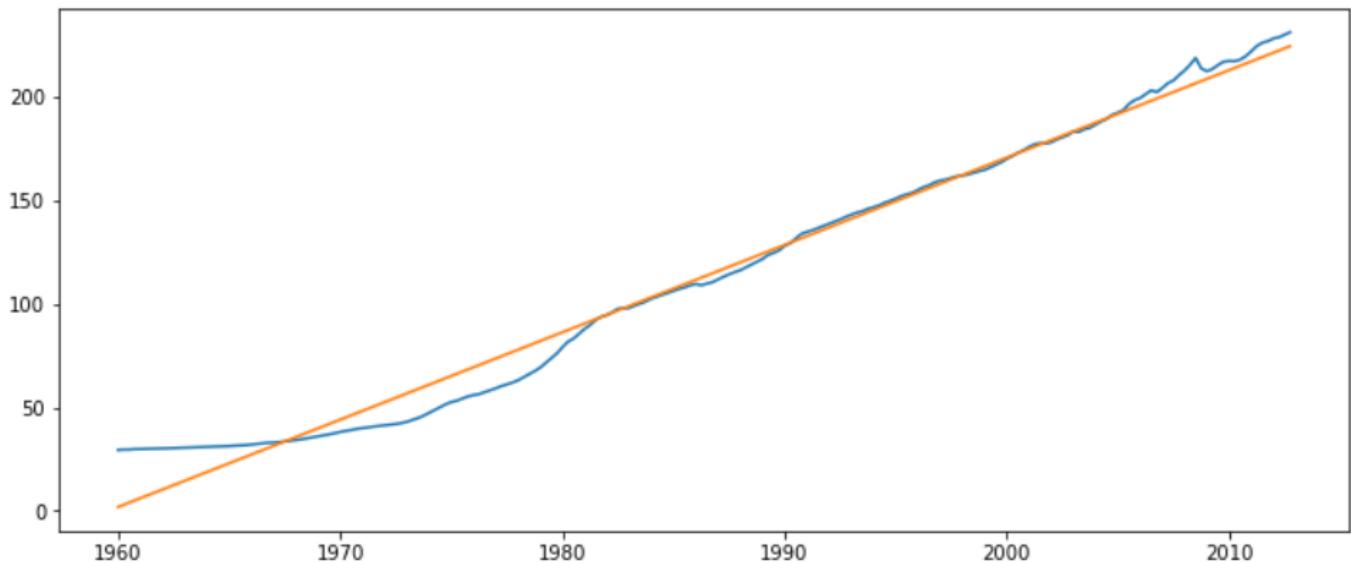
Le Consumer Price Index (CPI) est l'indice des prix à la consommation. Le CPI mesure le prix d'une panier de biens et de services dans toute l'économie. Il permet donc d'étudier les variations de prix d'un panier type de biens de consommation et de services achetés par les ménages.

Regardons le graph de la série CPI en fonction de la date



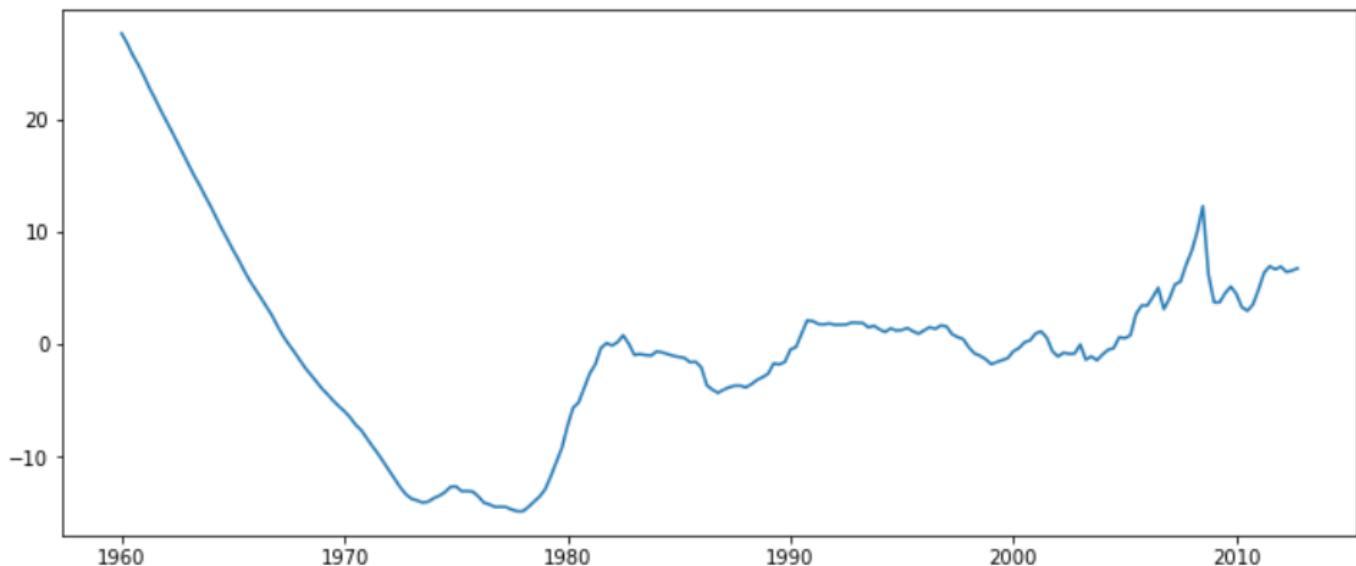
La tendance linéaire est celle qui s'approche le plus de nos données

Tendance de la série CPI



- La courbes bleues représentent la série de données CPI
- La courbe en orange représente la tendance

Notre série CPI stationnarisé

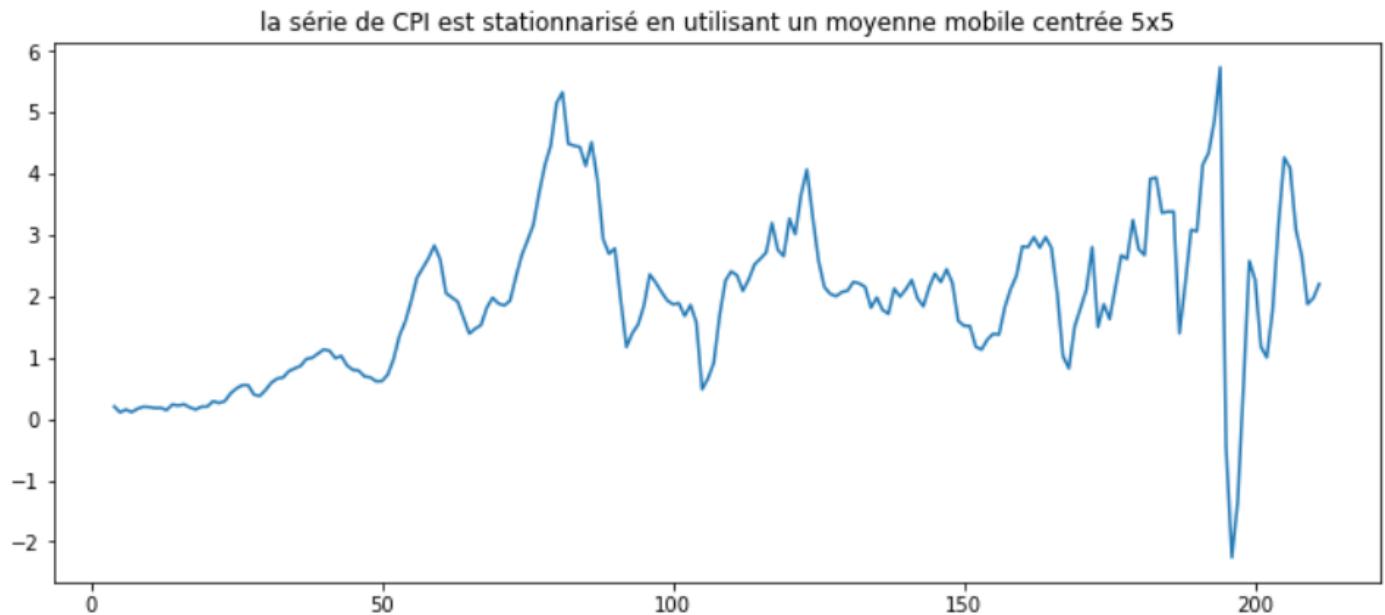


Cette courbe ci-dessus représente la série une fois la tendance enlevée. On peut remarquer que notre tendance linéaire nous renvoie de mauvais résultats sur la période allant de 1960 à 1980. Mais que celle-ci s'améliore après

2.3 Question 3

Stationnariser la série de CPI en utilisant un moyen mobile centrée 5x5.

Nous appliquons à notre série CPI une moyenne mobile centrée 5x5 pour pouvoir la stationnariser. Nous remarquons qu'avec ce processus, nous avons perdu des données : soit 4 observations à la fin et 4 observations au début. Nous obtenons alors :



Au regard de la question précédente, notre tendance que nous avions obtenue avec la moyenne mobile était proche de notre série d'origine. Notons l'impact de cette tendance après lissage sur notre série, obtenu après l'avoir stationnarité. La série que nous avons à l'origine n'a pas beaucoup de cycle au début de sa période, par ailleurs nous retrouvons peu de fluctuation sur cette période après avoir stationnarité notre série. Mais sur les années 2007 à 2012 nous retrouvons de fortes fluctuations ; ce qui représente une crise économique qui a eu lieu pendant cette période.

2.4 Question 4

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

Calculer le taux d'inflation à l'aide du CPI revient à passer notre variable à l'échelle logarithmique. Une fois cette transformation faite ; nous calculons un lag(1) avec le log(CPI) et nous faisons la différence première. Nous obtenons un taux d'inflation pour chaque trimestre ; nous pouvons multiplier ce taux par 4 pour avoir un taux annuel. Nous avons donc :



Nous pouvons observer que l'inflation a fortement varié au cours du temps et ces variations sont différentes selon les périodes. Ces fluctuations du taux d'inflation sont notamment dues aux événements économiques qui ont lieu au cours de différentes périodes : les chocs pétroliers de 1973 ou encore la crise des subprimes en 2008. Nous pouvons trouver facilement sur internet des listes de crises économiques qui ont eu lieu depuis les années 1970 :

lien : <https://cryptophilo.fr/2017/01/04/principales-etapes-de-la-crise-economique/>

15 août 1971

Nixon suspend la **convertibilité du dollar en or** : désormais, les États-Unis pourront rembourser leurs créanciers avec de la monnaie nouvellement émise, et non plus de l'or. Cela signifie que la masse monétaire (M3), créée par les banques via le crédit (sauf la monnaie centrale créée par la banque centrale, marginale) n'a plus de limite supérieure. Comme les banques ne gardent en réserve qu'une fraction de leurs prêts, faits avec intérêts, l'**exponentielle du crédit** commence.



24 janvier 1984

La France met fin à la séparation des activités bancaires de dépôt et d'investissement. Trois français (Jacques Delors à la Commission Européenne, Michel Camdessus au FMI et Henri Chavranski à l'OCDE) bâtent le nouvel environnement économique de globalisation financière : **déréglementation, décloisonnement, dérégulation**. Ces 3 "D" se retrouvent dans le cycle d'Uruguay de l'OMC et l'*Acte Unique Européen*, où la mobilité internationale des capitaux devient la norme. C'est le **consensus de Washington Paris**.



21 février 1995

Les banques américaines, toujours séparées entre banques de dépôt et banques d'affaires, entendent aussi profiter de cette mobilité des capitaux pour engranger d'énormes bénéfices via l'*effet de levier*. Dès 1995, le nouveau secrétaire d'Etat au Trésor Robert Rubin donne trois jours au président Clinton pour **abolir le Glass-Steagall Act**, ce qui sera fait en 1999. La même année, Robert Rubin intègre le Conseil d'Administration de Citigroup. C'est la constitution de **mégabanques "Too Big To Fail"** et le début de la **Bulle Internet**.



9 octobre 2002

Avec de retentissantes **fraudes comptables** (LTCM, Worldcom, ENRON...) et la remontée des taux d'intérêt à 10 ans, la **bulle Internet a explosé**. Le NASDAQ est au plus bas. Le 4 juin 2003, Le Président de la banque centrale américaine Alan Greenspan **abaisse les taux d'intérêt jusqu'à 1%** pour provoquer un nouveau cycle de crédit, donc une nouvelle bulle. Ce sera la **bulle immobilière**. En 2008, Greenspan plaidera une simple erreur devant le Congrès : "J'ai découvert une faille dans mon idéologie".



30 juin 2004

De nombreux **prêts hypothécaires à taux variable** sont accordés, souvent à des ménages peu ou pas solvables (tranche subprimes) convaincus que le marché ne pourrait baisser et à la merci de la moindre remontée des taux de la banque centrale. Ces prêts sont regroupés en tranches (selon leur qualité) dans des **titres financiers** (ABS, RMBS, CDO...etc) dont ils composent les **produits dérivés**. La bulle immobilière devient massive (**8400 milliards de dollars d'encours**) avec l'utilisation des **CDS**, un contrat d'assurance hors-bilan inventé par Blythe Masters (JP Morgan).

Alan Greenspan commence une série de 13 remontées successives (à 0,25%) des taux mais le capital issu de la bulle Internet alimente déjà la **bulle immobilière**.



3 mai 2005

Plusieurs fonds (Scion Capital, Paulson & Co, FrontPoint Partners, Cornwall Capital...) souscrivent des assurances (CDS) en cas de **perte de valeur des titres financiers hypothécaires**, pour plusieurs centaines de millions de dollars, auprès de la Deutsche Bank, Goldman Sachs et d'autres banques. **Les prix immobiliers baissent dès 2006**. En février 2007, HSBC passe d'importantes provisions (pertes potentielles) sur ces CDO.

Avec des mensualités en hausse, les **défaits de paiement des prêts hypothécaires** augmentent. Mais les CDO sont bien notés (AAA ou BB+) par les agences de notation, en **conflict d'intérêts** avec ceux qui les vendent.



15 septembre 2008

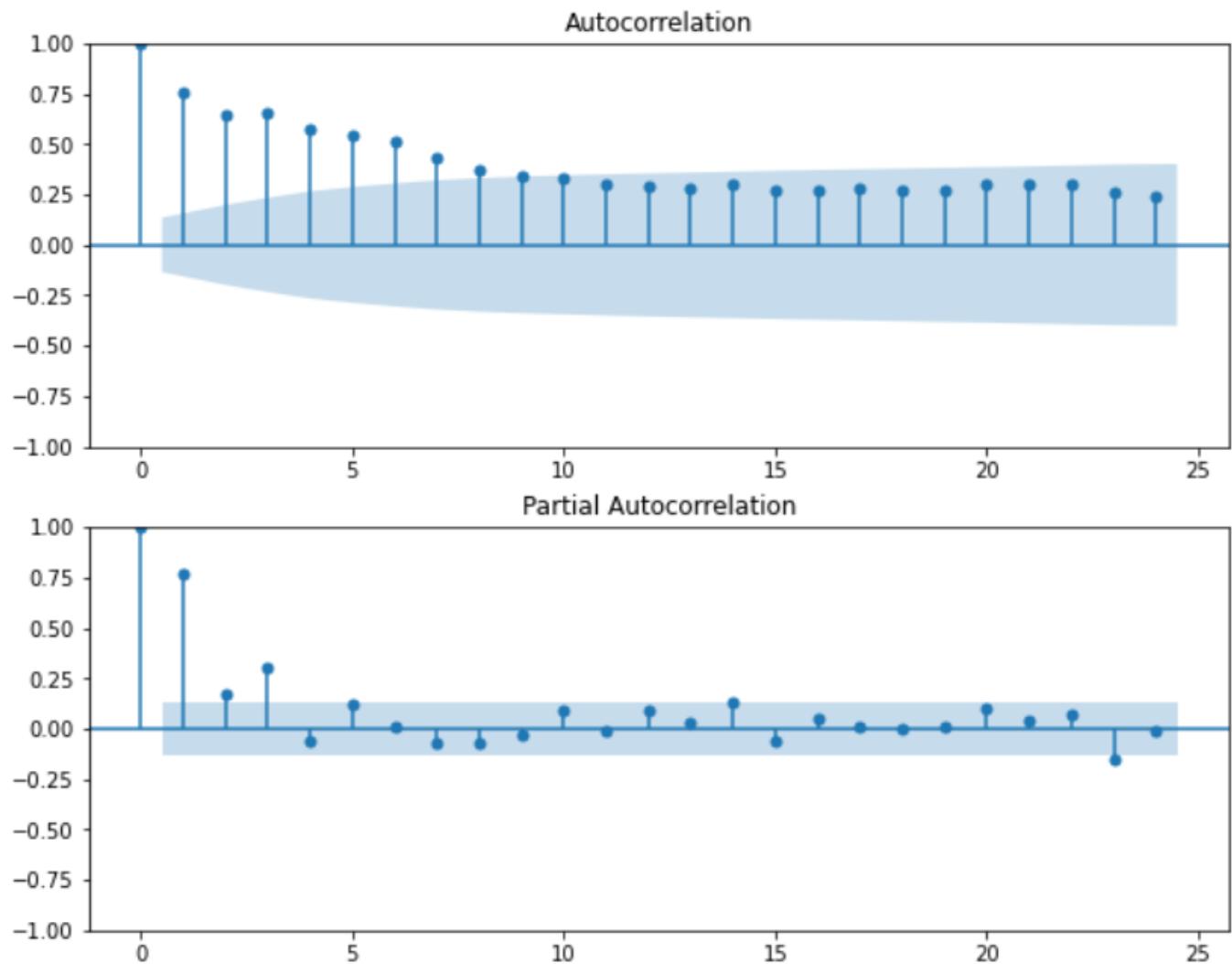
Les titres dérivés de l'immobilier perdent de la valeur et rendent les **banques réellement insolubles** (hors-bilan), même après la modification des normes comptables (abandon du mark-to-market). Elles ne prêtent plus ni aux entreprises, ni aux ménages ni aux autres banques (**gel de la liquidité**) et sont en concurrence pour **liquider en premier leurs actifs toxiques**. La **faillite de Lehman Brothers**, banque Too Big Too Fail, provoque un **krach boursier historique**.



2.5 Question 5

Interpréter l'autocorréogramme et l'autocorréogrammes partiels de inf. Quelle est la différence entre ces deux graphiques ?

Graphiques des fonctions d'autocorrélation et d'autocorrélation partielles de la série Inflation :



La fonction d'autocorrélation (ACF) est une mesure de la corrélation entre les observations d'une série chronologique qui sont séparées par des unités de temps k . En somme l'ACF nous indique la corrélation brute qui peut exister dans notre série entre le temps t et le temps $t+1$. Utilisons la fonction d'autocorrélation et les fonctions d'autocorrélation partielle pour identifier les modèles ARIMA. Examinons les pointes à chaque décalage pour déterminer si elles sont significatives. Un pic important dépassera les limites de signification, ce qui indique que la corrélation pour ce décalage n'est pas égale à zéro. Cette fonction joue un rôle important dans l'analyse des données visant à déterminer l'ampleur du lag dans un modèle autorégressif. L'utilisation de cette fonction a été introduite dans le cadre de l'approche Box-Jenkins de la modélisation des séries chronologiques, dans le cadre de laquelle le tracé des fonctions autocorrélatrices partielles permet de déterminer les décalages p appropriés dans un modèle AR (p) ou dans un modèle ARIMA étendu (p,d,q).

Dans l'analyse des times séries, la fonction d'autocorrélation partielle (PACF) donne la corrélation partielle d'une série chronologique stationnaire avec ses propres valeurs décalées, diminuée des valeurs de la série chronologique à tous les décalages plus courts. Il contraste avec la fonction d'autocorrélation, qui ne contrôle pas les autres décalages. La PACF va nous indiquer s'il existe dans une time série des corrélations entre le temps t et le temps $t+h$ quand nous avons supprimer les effet des corrélation de nos périodes $t+1, \dots, t+h-1$; nous obtenons les coefficients d'une régression linéaire multiple.

- L'autocorrélogramme indique que l'autocorrélation diminue avec le temps comme dans un processus ARMA(p,q)
- L'autocorrélogramme partiel oscille autour de 0 comme dans un processus de type Moving Average MA(1)

Cela signifie notamment qu'il existe une influence non négligeable du passé pour la détermination des valeurs présentes. On peut supposer que la série n'est pas stationnaire.

2.6 Question 6

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

La stationnarité fait référence aux distributions des variables aléatoires. Plus précisément, dans un processus stationnaire, toutes les variables aléatoires ont la même fonction de distribution, et plus généralement, pour chaque entier positif n et n moment t_1, \dots, t_n la distribution conjointe des n variables aléatoires $X(t_1), \dots, X(t_n)$ est la même que la distribution conjointe de $X(t_1 + \tau), \dots, X(t_n + \tau)$. C'est-à-dire, si on décale tous les instants de temps par τ , la description statistique du processus ne change pas du tout : le processus est stationnaire.

L'ergodicité, quant à elle, ne s'intéresse pas aux propriétés statistiques des variables aléatoires mais aux trajectoires de l'échantillon (sample path), c'est-à-dire à ce que l'on observe physiquement. Pour en revenir aux variables aléatoires, rappelons-nous que ces variables aléatoires sont des correspondances entre un échantillon et des nombres réels; ainsi différentes variables aléatoires mappent généralement un résultat donné à des nombres différents.

Par exemple : soit une expérience qui a abouti à un résultat ω dans l'espace de l'échantillon et ce résultat a été mappé sur des nombres réels (typiquement différent) par toutes les variables aléatoires dans le processus : Plus précisément, la variable aléatoire $X(t)$ a mappé ω avec un nombre réel que nous appellerons $x(t)$. Les nombres $x(t)$, considérés comme une forme d'onde, sont le chemin d'échantillon (sample path) correspondant à ω , et différents résultats nous donneront des chemins d'échantillon (sample path) différents. L'ergodicité traite ensuite des propriétés des chemins d'échantillonnage (sample path) et de la façon dont ces propriétés se rapportent aux propriétés des variables aléatoires composant le processus aléatoire.

un processus ergodique stationnaire est un processus stochastique qui présente à la fois la stationnarité et l'ergodicité. En substance, cela implique que le processus aléatoire ne changera pas ses propriétés statistiques avec le temps et que ses propriétés théorique statistiques (comme la moyenne théorique et la variance du processus) peuvent être déduites d'un échantillon unique et suffisamment long (réalisation) du processus.

Dans un modèle de régression linéaire, nous utilisons toujours le coefficient de détermination d'échantillon R^2 comme mesure de l'adéquation de l'équation de régression à la relation entre la variable explicative et la variation d'échantillon de la variable expliquée. Cependant, la corrélation d'échantillon entre les variables et la

corrélation globale sont deux concepts distinct. Bien que la relation entre les échantillons des variables économiques puisse expliquer dans une certaine mesure la relation entre les variables, il existe des exceptions, qui dépendent principalement de la distribution globale des variables économiques. Des études ont montré que lorsqu'un modèle de régression est construit avec deux séries chronologiques non stationnaires indépendantes, une équation de régression statistiquement significative est souvent obtenue. Nous l'appelons Régression parasite ou **spurious regression**.

Si les deux séries temporelles n'ont pas de relation de causalité au sens économique, la régression établie est une **spurious regression**. Par exemple, il peut y avoir un grand coefficient de corrélation entre le taux de croissance des arbres en bordure de route et le taux de croissance du PIB. Mais si on essaie de modéliser un tel modèle, on trouvera qu'il s'agit d'une pseudo-régression. En d'autres termes, il doit exister une relation causale entre des éléments liés, et un modèle capable de construire une régression.

2.7 Question 7

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

- Le test augmenté de Dickey-Fuller ou test ADF est un test statistique qui vise à savoir si une série temporelle est stationnaire c'est-à-dire si ses propriétés statistiques (espérance, variance, auto-corrélation) varient ou pas dans le temps.
- l'autocorrélation pour une série temporelle y , il s'agit de la corrélation entre y_t et y_{t-k} mesurée pour un délai (lag) k , donc la corrélation entre chaque mesure y et la mesure prise à k intervalles précédents.

Résultat de notre test test augmenté de Dickey-Fuller :

```
Pvalue : 0.04317651687154687  
lags inclus : 2
```

Notre Hypothèse nulle de départ H_0 : est que notre série a été générée par un processus qui présente une racine unitaire, et donc, que cette dernière n'est pas stationnaire. Ici nous rejettons cette hypothèse. Autre remarque : le nombre de lags à inclure est de 2

sachant : Plus la valeur de p valeur est petite, plus la probabilité de faire une erreur en rejetant l'hypothèse nulle est faible. Une valeur limite de 0,05 est souvent utilisée. Autrement dit, nous pouvons rejeter l'hypothèse nulle si la valeur de p est inférieure à 0,05.

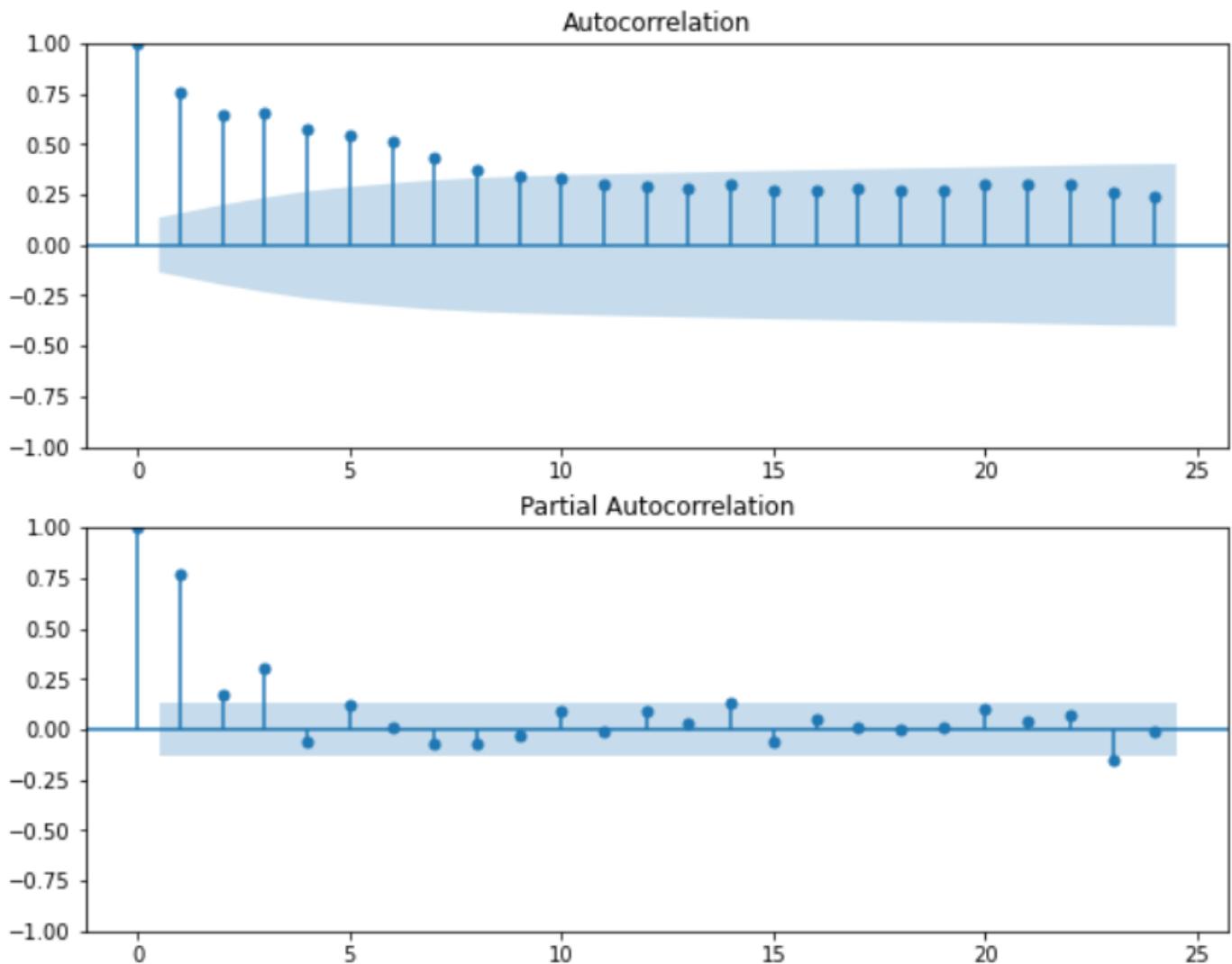
2.8 Question 8

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

Dans un premier temps déterminons le nombre de lags p que nous devons inclure dans notre modèle $AR(p)$. Quelles sont les méthodes que nous allons utilisées :

- Regardons les autocorrelations empiriques et comparons les aux autocorrelations théoriques des modèles $ARMA(p,q)$.
- Puis pour confirmer nos observations nous allons nous concentrer sur les critères AIC et BIC

Regardons les autocorrelations empiriques



La fonction d'autocorrélation décroît progressivement vers 0, ce qui signifie qu'on a un processus $MA(q)=0$. Nous pouvons sélectionner l'ordre p du modèle $AR(p)$ en fonction des pics significatifs du tracé $PACF$. La fonction d'autocorrélation partielle nous indique que les trois premiers lag sont significatifs. Donc après analyse des $PACF$ et ACF nous pouvons proposer un processus $AR(3)$.

Dans un deuxième temps nous allons utiliser les critères d'information AIC et BIC pour aider à sélectionné le *lag*. Nous calculons l' AIC et le BIC d'un $AR(p)$ pour chaque valeur de p

```

pour un p = 1
aic = 898.0754591499925
bic = 908.1310335504207

pour un p = 2
aic = 893.8233211715014
bic = 907.2307537054057

pour un p = 3
aic = 876.3508753957205
bic = 893.1101660631009

pour un p = 4
aic = 877.5942586320048
bic = 897.7054074328612

pour un p = 5
aic = 876.5824891289574
bic = 900.0454960632899

pour un p = 6
aic = 878.5733467681162
bic = 905.3882118359247

pour un p = 7
aic = 879.4074778227873
bic = 909.574201024072

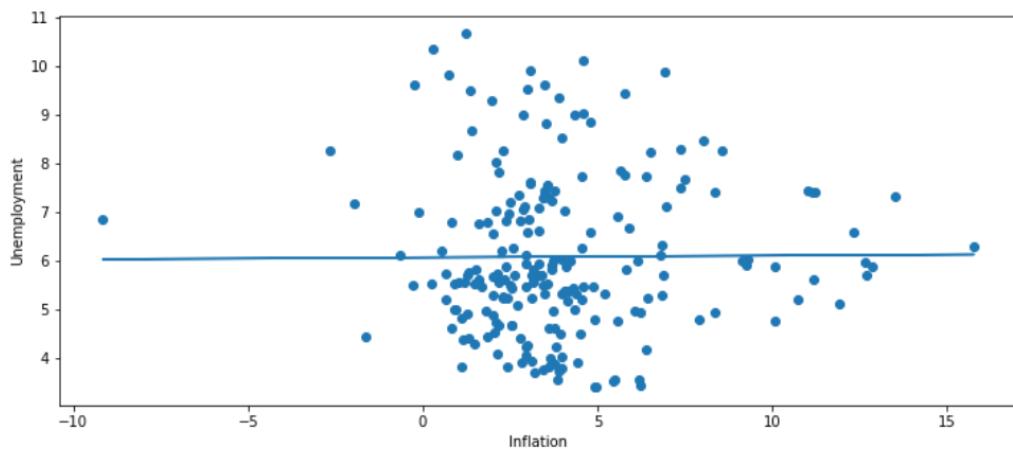
```

Les valeurs minimum pour nos critères d'informations *AIC* et *BIC* sont pour $p=3$. Ainsi au regard des résultats de nos critères on sélectionne un *lag*=3. Nos deux méthodes nous revoient un *lag*=3. On va donc proposer pour notre série *inflation* un modèle *AR(3)*.

2.9 Question 9

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.

La courbe de Philips :



Projet SES 722 : Econométrie

OLS Regression Results

Dep. Variable:	Unemp	R-squared:	0.000			
Model:	OLS	Adj. R-squared:	-0.005			
Method:	Least Squares	F-statistic:	0.01214			
Date:	Tue, 10 May 2022	Prob (F-statistic):	0.912			
Time:	08:57:41	Log-Likelihood:	-400.28			
No. Observations:	211	AIC:	804.6			
Df Residuals:	209	BIC:	811.3			
Df Model:	1					
Covariance Type:	nonrobust					
	coef	std err	t	P> t 	[0.025	0.975]
const	6.0708	0.181	33.576	0.000	5.714	6.427
inflation	0.0040	0.036	0.110	0.912	-0.067	0.075
	Omnibus:	13.872	Durbin-Watson:	0.044		
	Prob(Omnibus):	0.001	Jarque-Bera (JB):	15.356		
	Skew:	0.660	Prob(JB):	0.000463		
	Kurtosis:	2.937	Cond. No.	8.32		

On note que le model n'a pas les résultats attendu ; effectivement l'estimateur du coefficient directeur de notre de regression linéaire simple est positif (0.0040) alors qu'il est supposé être négatif. La courbe de Phillips est censé décrire une relation négative entre le taux de chômage et le taux d'inflation. Sur notre model l'inflation provoque une hausse du taux de chômage.

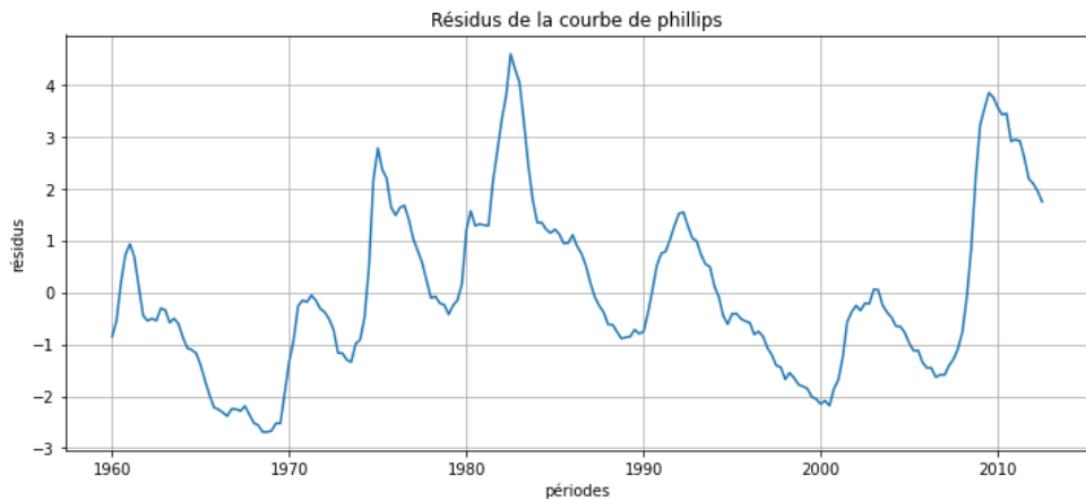
sachant : La valeur p valeur pour chaque terme teste l'hypothèse nulle que le coefficient est égal à zéro (aucun effet). Une faible valeur p (<0,05) nous indique que nous pouvons rejeter l'hypothèse nulle. En d'autres termes, un prédicteur qui a une faible valeur p est susceptible d'être un ajout significatif à notre modèle parce que les changements dans la valeur du prédicteur sont liés à des changements dans la variable de réponse. Inversement, une valeur p plus importante (non significative) suggère que les changements dans le prédicteur ne sont pas associés à des changements dans la réponse.

Au regard de la p-valeur du coefficient associé à l'inflation 0.912>0.05 ; en réalisé pour notre model il n'existe pas de relation entre le taux d'inflation et taux de chômage.

2.10 Question 10

Tester l'autocorrélation des erreurs.

graphique des résidus de la courbe de Phillips



Faisons un test d'autocorrélation avec la statistique de Durbin-Watson.

Le test de Durbin-Watson est un test statistique destiné à tester l'autocorrélation des résidus dans un modèle de régression linéaire.

Le test de Durbin-Watson cherche à vérifier la significativité du coefficient ρ dans la formule :

$$\epsilon_t = \rho \epsilon_{t-1} + u_t$$

où ϵ_t est le résidu estimé du modèle et u_t est un bruit blanc

L'hypothèse nulle (H_0) informe qu'il y a non auto-corrélation donc $\rho=0$. L'hypothèse alternative ($H1$) stipule qu'il y a auto-corrélation (Les résidus sont distribués selon un $AR(1)$) donc ρ est différent de 0 avec toujours $|\rho|<1$.

la statistique de test de Durbin-Watson vaut [0.04419413]

Pour tester l'autocorrélation positive à la signification α , la statistique de test d est

comparée aux valeurs critiques inférieure et supérieure ($d_{L,\alpha}$ et $d_{U,\alpha}$):

- Si $d < d_{L,\alpha}$, il existe une preuve statistique que les termes d'erreur sont positivement autocorrélés et on rejette l'hypothèse nulle.
- Si $d > d_{U,\alpha}$, il n'y a aucune preuve statistique que les termes d'erreur sont positivement autocorrélés et on ne rejette pas l'hypothèse nulle.
- Si $d_{L,\alpha} < d < d_{U,\alpha}$, le test n'est pas concluant.

Pour tester l'autocorrélation négative à la signification α , la statistique de test ($4-d$) est

comparée aux valeurs critiques inférieure et supérieure ($d_{L,\alpha}$ et $d_{U,\alpha}$):

- Si $(4 - d) < d_{L,\alpha}$, il existe une preuve statistique que les termes d'erreur sont autocorrélés négativement et on rejette l'hypothèse nulle.
- Si $(4 - d) > d_{U,\alpha}$, il n'y a aucune preuve statistique que les termes d'erreur sont négativement autocorrélés et on ne rejette pas l'hypothèse nulle.

- Si $d_{L,\alpha} < (4 - d) < d_{U,\alpha}$, le test n'est pas concluant.

TABLE de DURBIN-WATSON : Test unilatéral de $\rho = 0$ contre $\rho > 0$, au seuil de 5% (test bilatéral : seuil $\alpha = 10\%$)

n	k' = 1		k' = 2		k' = 3		k' = 4		k' = 5		k' = 6		k' = 7		k' = 8		k' = 9		k' = 10	
	d _L	d _u																		
15	1,08	1,36	0,95	1,54	0,82	1,75	0,69	1,97	0,56	2,21	0,45	2,47	0,34	2,73	0,25	2,98	0,17	3,22	0,11	3,44
16	1,10	1,37	0,98	1,54	0,86	1,73	0,74	1,93	0,62	2,15	0,50	2,40	0,40	2,62	0,30	2,86	0,22	3,09	0,15	3,30
17	1,13	1,38	1,02	1,54	0,90	1,71	0,78	1,90	0,67	2,10	0,55	2,32	0,45	2,54	0,36	2,76	0,27	2,97	0,20	3,20
18	1,16	1,39	1,05	1,53	0,93	1,69	0,82	1,87	0,71	2,06	0,60	2,26	0,50	2,46	0,41	2,67	0,32	2,87	0,24	3,07
19	1,18	1,40	1,08	1,53	0,97	1,68	0,86	1,85	0,75	2,02	0,65	2,21	0,46	2,40	0,46	2,59	0,37	2,78	0,29	2,97
20	1,20	1,41	1,10	1,54	1,00	1,68	0,90	1,83	0,79	1,99	0,69	2,16	0,60	2,34	0,50	2,52	0,42	2,70	0,34	2,88
21	1,22	1,42	1,13	1,54	1,03	1,67	0,93	1,81	0,83	1,96	0,73	2,12	0,64	2,29	0,55	2,46	0,46	2,63	0,38	2,81
22	1,24	1,43	1,15	1,54	1,05	1,66	0,96	1,80	0,86	1,94	0,77	2,09	0,68	2,25	0,59	2,41	0,50	2,57	0,42	2,73
23	1,26	1,44	1,17	1,54	1,08	1,66	0,99	1,79	0,90	1,92	0,80	2,06	0,71	2,21	0,63	2,36	0,54	2,51	0,46	2,67
24	1,27	1,45	1,19	1,55	1,10	1,66	1,01	1,78	0,93	1,90	0,84	2,03	0,75	2,17	0,67	2,32	0,58	2,46	0,51	2,61
25	1,29	1,45	1,21	1,55	1,12	1,66	1,04	1,77	0,95	1,89	0,87	2,01	0,78	2,14	0,70	2,28	0,62	2,42	0,54	2,56
26	1,30	1,46	1,22	1,55	1,14	1,65	1,06	1,76	0,98	1,88	0,90	1,99	0,82	2,12	0,73	2,25	0,66	2,38	0,58	2,51
27	1,32	1,47	1,24	1,56	1,16	1,65	1,08	1,76	1,01	1,86	0,92	1,97	0,84	2,09	0,77	2,22	0,69	2,34	0,62	2,47
28	1,33	1,48	1,26	1,56	1,18	1,65	1,10	1,75	1,03	1,85	0,95	1,96	0,87	2,07	0,80	2,19	0,72	2,31	0,65	2,43
29	1,34	1,48	1,27	1,56	1,20	1,65	1,12	1,74	1,05	1,84	0,97	1,94	0,90	2,05	0,83	2,16	0,75	2,28	0,68	2,40
30	1,35	1,49	1,28	1,57	1,21	1,65	1,14	1,74	1,07	1,83	1,00	1,93	0,93	2,03	0,85	2,14	0,78	2,25	0,71	2,36
31	1,36	1,50	1,30	1,57	1,23	1,65	1,16	1,74	1,09	1,83	1,02	1,92	0,95	2,02	0,88	2,12	0,81	2,23	0,74	2,33
32	1,37	1,50	1,31	1,57	1,24	1,65	1,18	1,73	1,11	1,82	1,04	1,91	0,97	2,00	0,90	2,10	0,84	2,20	0,77	2,31
33	1,38	1,51	1,32	1,58	1,26	1,65	1,19	1,73	1,13	1,81	1,06	1,90	0,99	1,99	0,93	2,08	0,86	2,18	0,79	2,28
34	1,39	1,51	1,33	1,58	1,27	1,65	1,21	1,73	1,15	1,81	1,08	1,89	1,01	1,98	0,95	2,07	0,88	2,16	0,82	2,26
35	1,40	1,52	1,34	1,58	1,28	1,65	1,22	1,73	1,16	1,80	1,10	1,88	1,03	1,97	0,97	2,05	0,91	2,14	0,84	2,24
36	1,41	1,52	1,35	1,59	1,29	1,65	1,24	1,73	1,18	1,80	1,11	1,88	1,05	1,96	0,99	2,04	0,93	2,13	0,87	2,22
37	1,42	1,53	1,36	1,59	1,31	1,66	1,25	1,72	1,19	1,80	1,13	1,87	1,07	1,95	1,01	2,03	0,95	2,11	0,89	2,20
38	1,43	1,54	1,37	1,59	1,32	1,66	1,26	1,72	1,21	1,79	1,15	1,86	1,09	1,94	1,03	2,02	0,97	2,10	0,91	2,18
39	1,43	1,54	1,38	1,60	1,33	1,66	1,27	1,72	1,22	1,79	1,16	1,86	1,10	1,93	1,05	2,01	0,99	2,08	0,93	2,16
40	1,44	1,54	1,39	1,60	1,34	1,66	1,29	1,72	1,23	1,79	1,17	1,85	1,12	1,92	1,06	2,00	1,01	2,07	0,95	2,14
45	1,48	1,57	1,43	1,62	1,38	1,67	1,34	1,72	1,29	1,78	1,24	1,84	1,19	1,90	1,14	1,96	1,09	2,00	1,04	2,09
50	1,50	1,59	1,46	1,63	1,42	1,67	1,38	1,72	1,34	1,77	1,29	1,82	1,25	1,87	1,20	1,93	1,16	1,99	1,11	2,04
55	1,53	1,60	1,49	1,64	1,45	1,68	1,41	1,72	1,38	1,77	1,33	1,81	1,29	1,86	1,25	1,91	1,21	1,96	1,17	2,01
60	1,55	1,62	1,51	1,65	1,48	1,69	1,44	1,73	1,41	1,77	1,37	1,81	1,33	1,85	1,30	1,89	1,26	1,94	1,22	1,98
65	1,57	1,63	1,54	1,66	1,50	1,70	1,47	1,73	1,44	1,77	1,40	1,80	1,37	1,84	1,34	1,88	1,30	1,92	1,27	1,96
70	1,58	1,64	1,55	1,67	1,52	1,70	1,49	1,74	1,46	1,77	1,43	1,80	1,40	1,84	1,37	1,87	1,34	1,91	1,30	1,95
75	1,60	1,65	1,57	1,68	1,54	1,71	1,51	1,74	1,49	1,77	1,46	1,80	1,43	1,83	1,40	1,87	1,37	1,90	1,34	1,94
80	1,61	1,66	1,59	1,69	1,56	1,72	1,53	1,74	1,51	1,77	1,48	1,80	1,45	1,83	1,42	1,86	1,40	1,89	1,37	1,92
85	1,62	1,67	1,60	1,70	1,57	1,72	1,55	1,75	1,52	1,77	1,50	1,80	1,47	1,83	1,45	1,86	1,42	1,89	1,40	1,92
90	1,63	1,68	1,61	1,70	1,59	1,73	1,57	1,75	1,54	1,78	1,52	1,80	1,49	1,83	1,47	1,85	1,44	1,88	1,42	1,91
95	1,64	1,69	1,62	1,71	1,60	1,73	1,58	1,75	1,56	1,78	1,54	1,80	1,51	1,83	1,49	1,85	1,46	1,88	1,44	1,90
100	1,65	1,69	1,63	1,72	1,61	1,74	1,59	1,76	1,57	1,78	1,55	1,80	1,53	1,83	1,51	1,85	1,48	1,87	1,46	1,90
150	1,72	1,75	1,71	1,76	1,69	1,77	1,68	1,79	1,66	1,80	1,65	1,82	1,64	1,83	1,62	1,85	1,60	1,86	1,59	1,88
200	1,73	1,78	1,75	1,79	1,73	1,80	1,73	1,81	1,72	1,82	1,71	1,83	1,70	1,84	1,69	1,85	1,68	1,86	1,66	1,87

Pour un seuil $\alpha=0.05$, un $k=2$ et un nombre d'observation $n \approx 200$ les valeurs critiques inférieure et supérieure sont :

$$d_{L,\alpha} = 1.75$$

$$d_{U,\alpha} = 1.79$$

Pour tester l'autocorrélation positive à la signification $\alpha=0.05$, la statistique de test $d=0.044$ est comparée aux valeurs critiques inférieure et supérieure $d_{L,\alpha} = 1.75$ et $d_{U,\alpha} = 1.79$:

nous obtenons :

$(4 - d) > d_{U,\alpha}$, il n'y a aucune preuve statistique que les termes d'erreur sont négativement autocorrélés.

$d < d_{L,\alpha}$ il existe donc une preuve statistique que les termes d'erreur sont positivement autocorrélés

2.11 Question 11

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

Pour corriger l'autocorrélation des erreurs nous utiliserons la méthode des moindres carrés généralisable faible aux erreurs AR(1). Nous obtenons $\hat{\epsilon}$ par MCO. Nous estimons $\hat{\epsilon}_i = \rho \hat{\epsilon}_{i-1} + u_i$. Avec comme estimateur de la covariance :

$$\sigma^2 \hat{\Psi} = \frac{\sigma_u^2}{1 - \hat{\rho}^2} \begin{pmatrix} 1 & \hat{\rho} & \hat{\rho}^2 & \cdots & \hat{\rho}^{T-1} \\ 1 & \hat{\rho} & \cdots & \hat{\rho}^{T-2} & \\ \ddots & & & & \vdots \\ & & 1 & \hat{\rho} & \\ & & & & 1 \end{pmatrix}$$

$$\hat{\Psi} = \begin{pmatrix} 1 & \hat{\rho} & \hat{\rho}^2 & \cdots & \hat{\rho}^{T-1} \\ 1 & \hat{\rho} & \cdots & \hat{\rho}^{T-2} & \\ \ddots & & & & \vdots \\ & & 1 & \hat{\rho} & \\ & & & & 1 \end{pmatrix}$$

L'estimation MGF est :

$$\hat{\beta}_{MCGF} = (X' \hat{\Psi}^{-1} X)^{-1} X' \hat{\Psi}^{-1} Y$$

Dep. Variable:	Unemp	R-squared:	0.024			
Model:	GLS	Adj. R-squared:	0.020			
Method:	Least Squares	F-statistic:	5.186			
Date:	Tue, 10 May 2022	Prob (F-statistic):	0.0238			
Time:	10:40:47	Log-Likelihood:	-68.505			
No. Observations:	211	AIC:	141.0			
Df Residuals:	209	BIC:	147.7			
Df Model:	1					
Covariance Type:	nonrobust					
<hr/>						
	coef	std err	t	P> t	[0.025	0.975]
const	6.3115	0.949	6.652	0.000	4.441	8.182
inflation	-0.0248	0.011	-2.277	0.024	-0.046	-0.003
<hr/>						
Omnibus:	64.399	Durbin-Watson:	0.723			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	200.061			
Skew:	-1.258	Prob(JB):	3.61e-44			
Kurtosis:	7.053	Cond. No.	87.1			

Première remarque : le coefficient de la variable *inflation* est de signe négatif – 0.0248 ; cette fois-ci cela correspond bien à la théorie de la courbe de Philips. Une seconde remarque : le coefficient *inflation* est devenu significatif avec une *p – valeur* = 0.024 < 0.05 ; on rejette donc l'hypothèse nulle (le coefficient est égal à zéro ; le coefficient n'a aucun effet) au seuil 5% .

2.12 Question 12

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

Le test de Chow nous permet de tester si les coefficients de régression de chaque ligne de régression sont égaux ou non.

Si le test détermine que les coefficients ne sont pas égaux entre les lignes de régression, cela signifie qu'il existe des preuves significatives qu'une rupture structurelle existe dans les données. En d'autres termes, le modèle dans les données est significativement différent avant et après ce point de rupture structurel.

Nous souhaitons tester la stabilité de notre relation chômage inflation sur deux sous périodes identique. Pour cela nous procédons à un test de chow, nous allons partager notre jeu de données en deux sous-échantillon de taille identique. Notre premier échantillon va de 1960-04-01 à 1986-04-01 et le deuxième échantillon va de 1960-07-01 à 1012-10-01. Regardons les régressions suivantes :

Soit le modèle initial :

$$y_t = a + b x_{1t} + \epsilon$$

Nous séparons en deux groupes notre modèles, on a:

$$\text{Model n°1 : } y_t = a_1 + b_1 x_{1t} + \epsilon$$

et

$$\text{Model n°2 : } y_t = a_2 + b_2 x_{1t} + \epsilon$$

Le test de Chow revient à tester l'hypothèse nulle : le modèle est stable $H_0 : a_1 = a_2, b_1 = b_2$. contre l'hypothèse alternative: le modèle est instable H_1 : l'un au moins des $a_i \neq b_i$ pour $i \in \{1, 2\}$

La p-valeur du test de Chow est de :

La p-valeur vaut 0.14467378426000327

La p-valeur = 0.14 > 0.05 ; nous ne rejetons pas l'hypothèse nulle d'égalité de nos coefficients sur les deux périodes. Sur deux sous-périodes de taille identique, il y a stabilité de la relation chômage inflation.

2.13 Question 13

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.

Nous calculons le modèle suivant maintenant :

$$\begin{aligned} unempi = & \text{ constante} + \beta_1 inf_{i-1} + \beta_2 inf_{i-2} + \beta_3 inf_{i-3} + \beta_4 inf_{i-4} + \gamma_1 unemp_{i-1} + \gamma_2 unemp_{i-2} \\ & + \gamma_3 nemp_{i-3} + \gamma_4 unemp_{i-4} + \epsilon_i \end{aligned}$$

Faisons le test de causalité de granger de l'inflation sur le chômage.

On test l'hypothèse $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$.

Contre l'hypothèse H_1 : l'un au moins des $\beta_i \neq 0$ avec $i \in \{1, 2, 3, 4\}$

Voici les résultats que nous obtenons ci-dessous :

```

Granger Causality
number of lags (no zero) 4
ssr based F test:      F=3.7967 , p=0.0054 , df_denom=198, df_num=4
ssr based chi2 test:   chi2=15.8771 , p=0.0032 , df=4
likelihood ratio test: chi2=15.2977 , p=0.0041 , df=4
parameter F test:      F=3.7967 , p=0.0054 , df_denom=198, df_num=4

{4: ({'ssr_ftest': (3.7966991982052365, 0.005351957807937859, 198.0, 4),
      'ssr_chi2test': (15.87710573794917, 0.0031885813965498565, 4),
      'lrtest': (15.297658304658057, 0.00412204780663406, 4),
      'params_ftest': (3.796699198205294, 0.005351957807937344, 198.0, 4.0)},
     [,
      <statsmodels.regression.linear_model.RegressionResultsWrapper at 0x20c75fbe370>,
      array([[0., 0., 0., 0., 1., 0., 0., 0., 0.],
             [0., 0., 0., 0., 1., 0., 0., 0., 0.],
             [0., 0., 0., 0., 0., 0., 1., 0., 0.],
             [0., 0., 0., 0., 0., 0., 1., 0., 0.]]])}

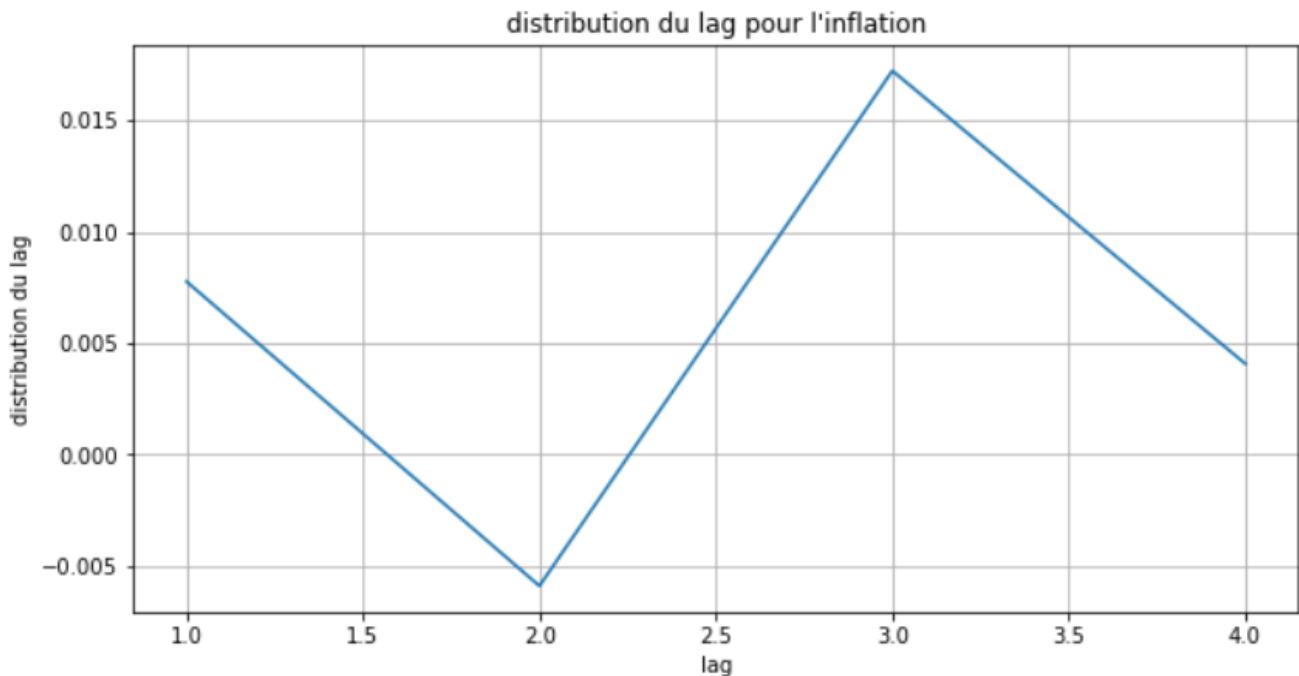
```

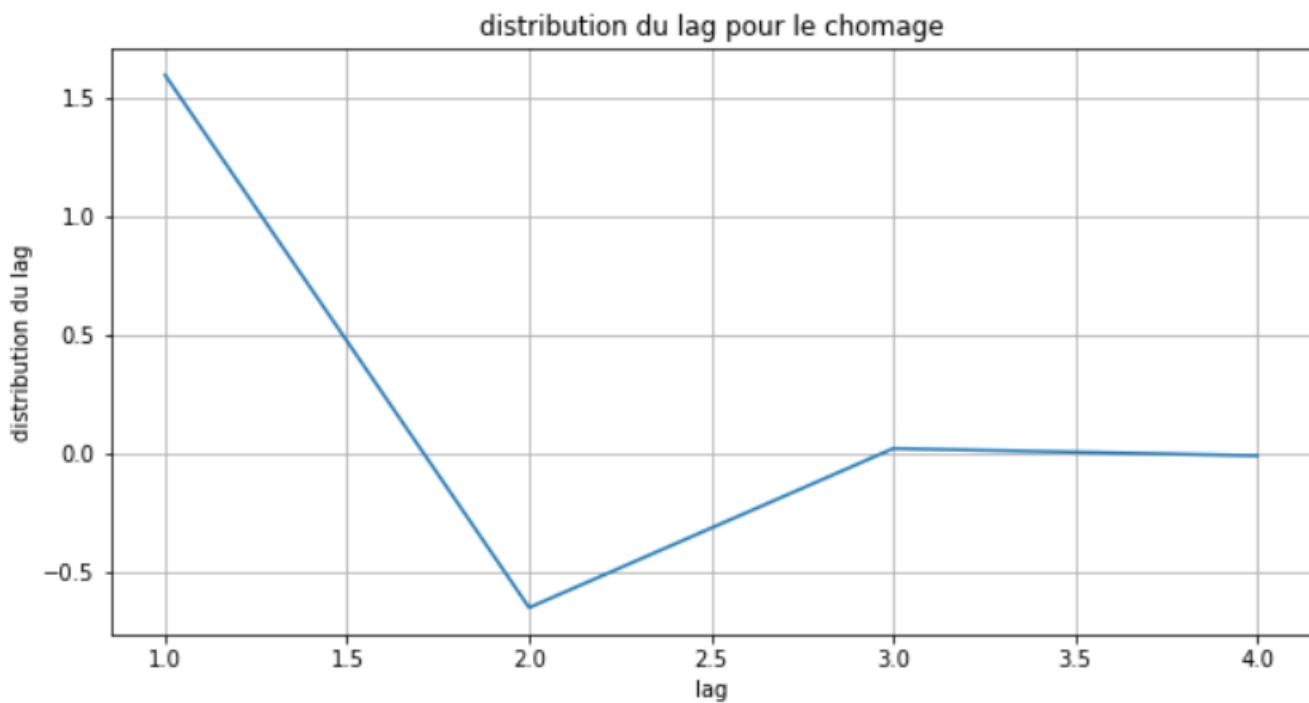
Au regard de la p-valeur qui est de $0.0054 < 0.05$ nous rejetons l'hypothèse nulle d'absence de causalité de Granger de l'inflation sur le chômage.

2.14 Question 14

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

Nous représentons ci-dessous les graphiques des délais distribués





Sur les résultats du graphique du retard distribué pour le chômage nous pouvons observer que le chômage peut s'expliquer par un passé très proche. Il y a une variabilité plus importante qui se produit après une crise ; puis nous observons des valeurs négatives sur la deuxième période et enfin les valeurs se stabilisent. Les valeurs fortement négatives dans la seconde période peuvent s'expliquer comme une compensation après une forte baisse du niveau d'emploi.

Sur les résultats du graphique du retard distribué pour l'inflation nous pouvons observer des résultats assez différents comparé au retard distribué pour le chômage. Il y a tout d'abord un faible effet positif sur l'inflation suivi d'un faible effet négatif ; la variation la plus importante se trouve sur la troisième période. Dans cette situation l'effet d'un choc met plus de temps à être retracé dans l'économie.

Nous pouvons calculer l'impact à long terme de l'inflation sur le chômage en faisant la somme des coefficients de l'inflation que nous avons calculer à l'aide de notre régression linéaire multiple :

Projet SES 722 : Econométrie

OLS Regression Results

Dep. Variable:	Unemp	R-squared:	0.979			
Model:	OLS	Adj. R-squared:	0.978			
Method:	Least Squares	F-statistic:	1145.			
Date:	Fri, 13 May 2022	Prob (F-statistic):	2.80e-161			
Time:	09:08:57	Log-Likelihood:	4.6497			
No. Observations:	207	AIC:	8.701			
Df Residuals:	198	BIC:	38.70			
Df Model:	8					
Covariance Type:	nonrobust					
	coef	std err	t	P> t 	[0.025	0.975]
const	0.1457	0.072	2.014	0.045	0.003	0.288
inf_1	0.0078	0.009	0.827	0.409	-0.011	0.026
inf_2	-0.0059	0.010	-0.577	0.565	-0.026	0.014
inf_3	0.0172	0.010	1.729	0.085	-0.002	0.037
inf_4	0.0041	0.009	0.435	0.664	-0.014	0.023
Unemp_1	1.5937	0.071	22.383	0.000	1.453	1.734
Unemp_2	-0.6472	0.134	-4.832	0.000	-0.911	-0.383
Unemp_3	0.0222	0.135	0.164	0.870	-0.245	0.289
Unemp_4	-0.0080	0.070	-0.114	0.910	-0.146	0.130
Omnibus:	29.127	Durbin-Watson:	1.997			
Prob(Omnibus):	0.000	Jarque-Bera (JB):	68.886			
Skew:	0.625	Prob(JB):	1.10e-15			
Kurtosis:	5.534	Cond. No.	170.			

L'impact à long terme de l'inflation sur le chômage est de 0,023. Nous pouvons interpréter ce résultat comme étant : chaque unité d'inflation va entraîner une hausse de 0,023 unité (points) de chômage en plus à long terme.