Econometrics Project

By: José Lucas Barretto and Lucas Celinga Agrizzi

Project 1

Married Women's Annual Labor Supply

It is assumed that labor force participation depends on other sources of income, including husband's earnings (nwifeinc , measured in thousands of dollars), years of education (educ),past years of labor market experience (exper), age, number of children less than six years old (kidslt6), and number of kids between 6 and 18 years of age (kidsge6). Using the data in MROZ.txt from Mroz (1987) 428 of the 753 women in the sample report being in the labor force at some point during 1975.

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

1. Lire le fichier mroz.txt. Ne sélectionner que les observations pour lesquelles la variable wage est strictement positive.

We excluded the non numbers from the wage and the values bigger than zero.

	inlf	hours	kidslt6	kidsge6	age	educ	wage	repwage	hushrs	husage	...	faminc	mtr	motheduc	fatheduc	unem	city	exper	nwifeinc	lwa
0	1	1610	1	0	32	12	3.3540	2.65	2708	34	...	16310	0.7215	12	7	5.0	0	14	10.910060	1.2101
1	1	1656	0	2	30	12	1.3889	2.65	2310	30	...	21800	0.6615	7	7	11.0	1	5	19.499980	0.3285
2	1	1980	1	3	35	12	4.5455	4.04	3072	40	...	21040	0.6915	12	7	5.0	0	15	12.039910	1.5141
3	1	456	0	3	34	12	1.0965	3.25	1920	53	...	7300	0.7815	7	7	5.0	0	6	6.799996	0.0921
4	1	1568	1	2	31	14	4.5918	3.60	2000	32	...	27300	0.6215	12	14	9.5	1	7	20.100060	1.5242

5 rows × 22 columns

2. Faire les statistiques descriptives du salaire, de l'age et de l'éducation pour l'ensemble des femmes puis, pour les femmes dont le salaire du mari est supérieure à la médiane de l'échantillon, puis pour les femmes dont le salaire du mari est inférieur à la médiane de l'échantillon

```
For women which the husband earns LESS
than the median husbands wage

WAGE :

nrows:     214
min_data:  0.1282
max_data:  18.267
mean:      3.4585406542056076
var:       4.572157433253777
median:    2.9718
*****
AGE :

nrows:     214
min_data:  30
max_data:  60
mean:      41.66822429906542
var:       64.42730806183947
median:    41.0
*****
EDUCATION :

nrows:     214
min_data:  6
max_data:  17
mean:      12.074766355140186
var:       4.200017468774566
median:    12.0
*****
```

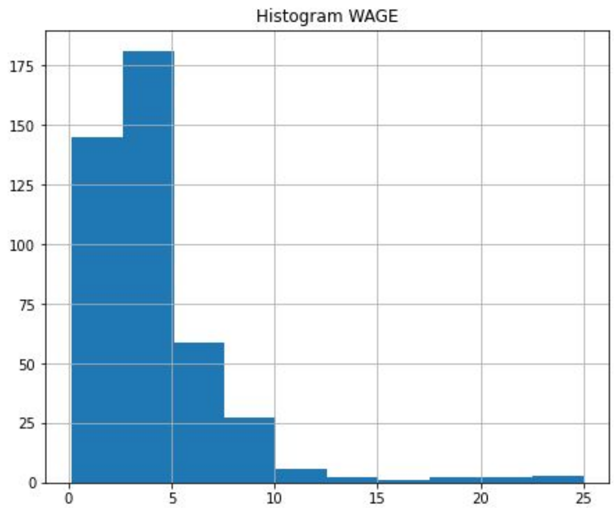
```
For women which the husband earns MORE
 than the median husbands wage

WAGE :

nrows:     214
min_data:  0.1616
max_data:  25.0
mean:      4.896822429906543
var:       16.258248838749235
median:    3.8464
*****
AGE :

nrows:     214
min_data:  30
max_data:  59
mean:      42.27570093457944
var:       54.33987684513931
median:    43.0
*****
EDUCATION :

nrows:     214
min_data:  5
max_data:  17
mean:      13.242990654205608
var:       5.5390863830902255
median:    12.0
*****
```

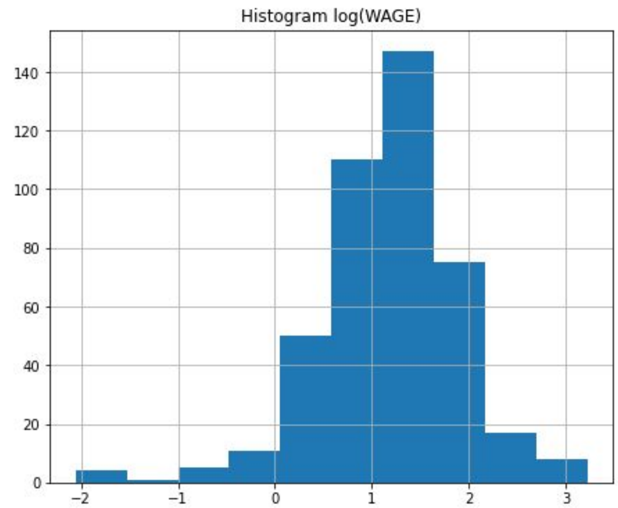
As we can notice, the woman with the husband who earns low usually earns less than the wifes who has a husband earning more, but the second one has more variance in the result.

We also saw that the education and age don't have much influence linked with the husband's salary.

All these affirmations are empirically.

3. Faire l'histogramme de la variable wage. Calculer le log de wage et faire l'histogramme. Comparez les deux histogrammes et commentez





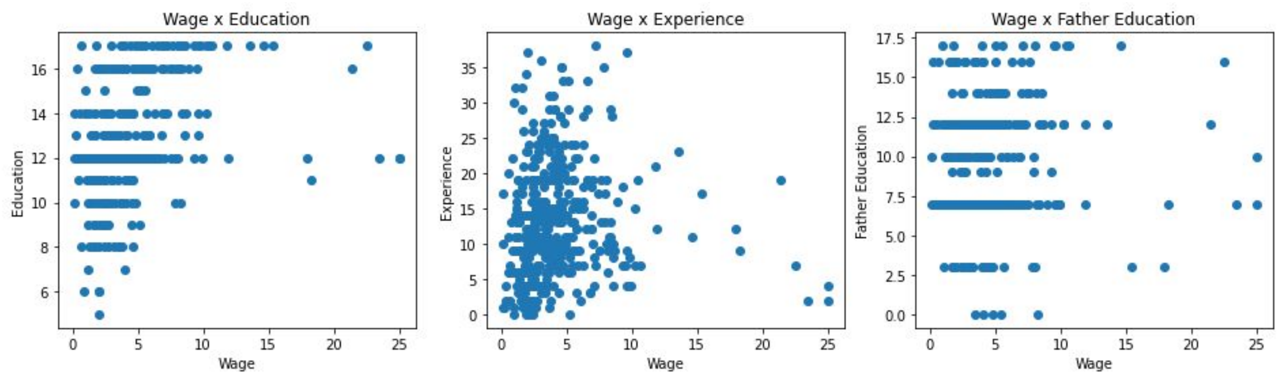
The first graph looks like the **wage** is near an exponential distribution, otherwize the histogram from **log(WAGE)** looks more like a normal distribution which is commonly better to work with because the statistiques from standard error and the theorems are symmetrical and has less variance.

4. Calculer les corrélations motheduc et fatheduc. Commentez. Il y a-t-il un problème de multicolinéarité si l'on utilise ces variables comme variables explicatives ?

- Correlation coef: 0.554063218431168

This value of correlation is significant and can change the regression because a part of one of these variables can be described as a linear combination of the other variables and it could affect the result of the OLS.

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



This graph can't show the data with the other variables constant, so we can note a simple relation between the wage and these other variables but there is a lot of interference from other variables in the wage value, who aren't associated with the other variable in the graph .

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

To grant that the estimators are unbiased, we have to assume that:

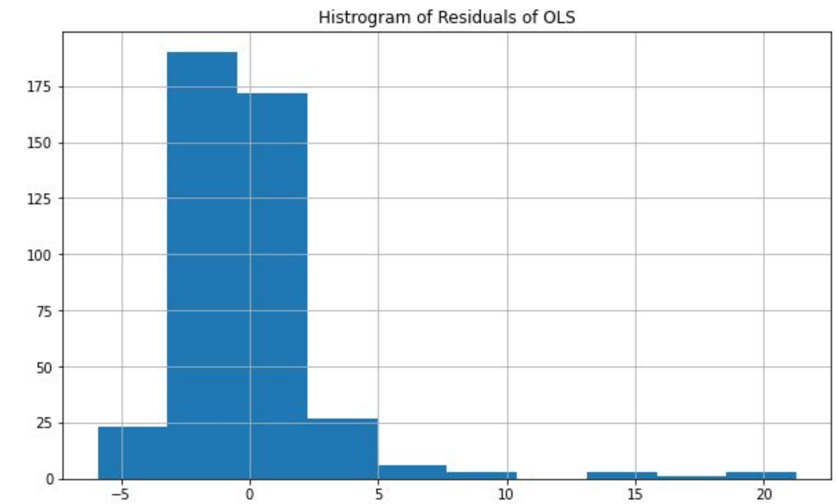
1. The model can be written as a linear combination of the variables
2. . The data has to be a random sample
3. 3. The data can not be in a perfect constant variable and a exact linear dependency between two independent variables
4. The error must have a null expected value

Omitted variable bias is when a variable truly belongs in a model but is not specified in the model, so your model excludes or under specifies an important variable of the regression. When you do this, the model will be unbiased if one of this conditions exists:

1. The slope of the true variable is null
2. The correlation with the true variable and all other variables in the model are null.

7. Faire la régression de wage en utilisant les variables explicatives un constante, city, educ, exper, nwifeinc, kidslt6, kidsgt6. Commentez l'histogramme des résidus.

- Residuals average: 3.98 e-15
- Residuals variance: 9.54



The OLS make a regression trying to approximate the residual average to zero. As we can see, the residuals mean is next to the null value and it looks like a normal distribution, but there are more values on the right of the curve making and their variance big. The variance is relationated with the goodness of a fit, when smaller is the variance, better is the fit.

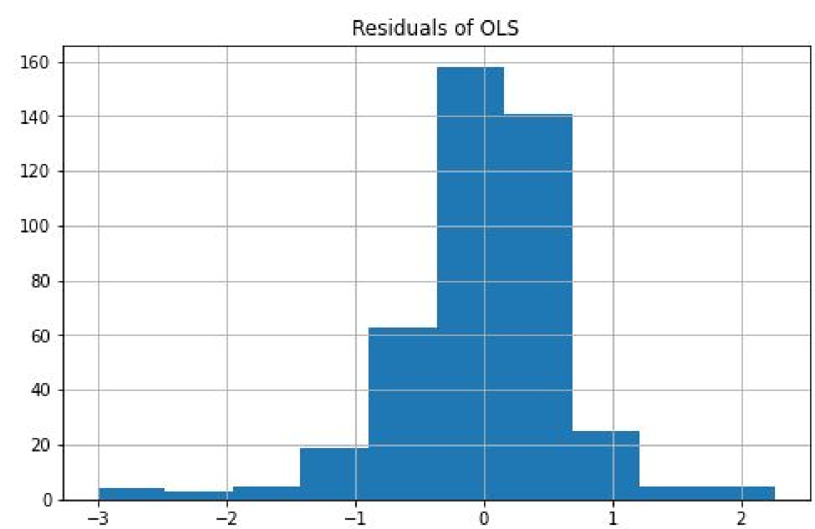
y = wage

X = const, x1 = city, x2 = educ, x3 = exper, x4 = nwifeinc, x5 = kidslt6, x6 = kidsge6

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                   wage   R-squared:                       0.127
Model:                            OLS   Adj. R-squared:                  0.115
Method:                 Least Squares   F-statistic:                     10.23
Date:                Sun, 22 Nov 2020   Prob (F-statistic):           1.41e-10
Time:                        13:18:07   Log-Likelihood:                -1090.0
No. Observations:                 428   AIC:                             2194.
Df Residuals:                     421   BIC:                             2222.
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -2.4035      0.963     -2.495      0.013      -4.297      -0.510
x1             0.3698      0.327      1.132      0.258      -0.272       1.012
x2             0.4600      0.070      6.546      0.000       0.322       0.598
x3             0.0238      0.021      1.141      0.255      -0.017       0.065
x4             0.0152      0.015      0.984      0.326      -0.015       0.046
x5             0.0362      0.397      0.091      0.927      -0.744       0.816
x6            -0.0619      0.125     -0.494      0.622      -0.308       0.185
==============================================================================
Omnibus:                      345.825   Durbin-Watson:                   2.056
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             6499.375
Skew:                           3.389   Prob(JB):                         0.00
Kurtosis:                      20.847   Cond. No.                         178.
==============================================================================
```

8. Faire la régrssion de lwage sur une constante, city, educ, exper, nwifeinc, kidslt6, kidsgt6. Comparer l’histogramme obtenu à celui de la question 7.

- Residuals average: 5.49 e-16
- Residuals variance: 0.44



The average of the residuals keeps next to zero as expected, but the data is better distributed around the zero making the variance being smaller than the other fitting, making this OLS better than the other.

y = wage

X = const, x1 = city, x2 = educ, x3 = exper, x4 = nwifeinc, x5 = kidslt6, x6 = 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:	Sun, 22 Nov 2020	Prob (F-statistic):	2.00e-13
Time:	13:25:51	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.

9. 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.

- Hipotesis Null   H0: X4 = 0
- P-value of nwifeinc: 0.1426

With this result the p-value indicates the level of significance to reject the hypothesis, so:

- We **can't reject** H0 with 1% of significance level
- We **can't reject** H0 with 5% of significance level
- We **can't reject** H0 with 10% of significance level

10. 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)

For do this we can find the *t-value* by:

$$t_{value} = (X_4 - 0.01)/std_{error}$$

t-value = *1.465*

- Hipotesis Null		H0: X4 = 0.01
- P-value of nwifeinc:	0.12444

With this result the p-value indicates the level of significance to reject the hypothesis, so:

- We **can't reject** H0 with 5% of significance level

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

For this question we have to make another OLS with the follow parameters:

y = lwage - 0.01*nwifeinc - 0.05*city
X = const, x1 = educ, x2 = exper, x3 = kidslt6, x4 = kidsge6

```
                            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:                Sun, 22 Nov 2020   Prob (F-statistic):           4.34e-12
Time:                        15:41:26   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.
==============================================================================
```

With this, we can compare the SSR from each one and the degrees of liberties and find the F statistique:

$$F = ((SSR_1 - SSR_0)/q) \; / \; (SSR_0/(n - k_1))$$

Obtaining:

- F-statistiques: 1.3434
- F-statistiques for 5% of significance: 0.95123

- P-value of the hypothesis 0.26206

So, for the hypothesis H_0 : $"nwifeinc" = 0.01 \; and \; "city" = 0.05$
So we **can't** reject the hypothesis H_0.

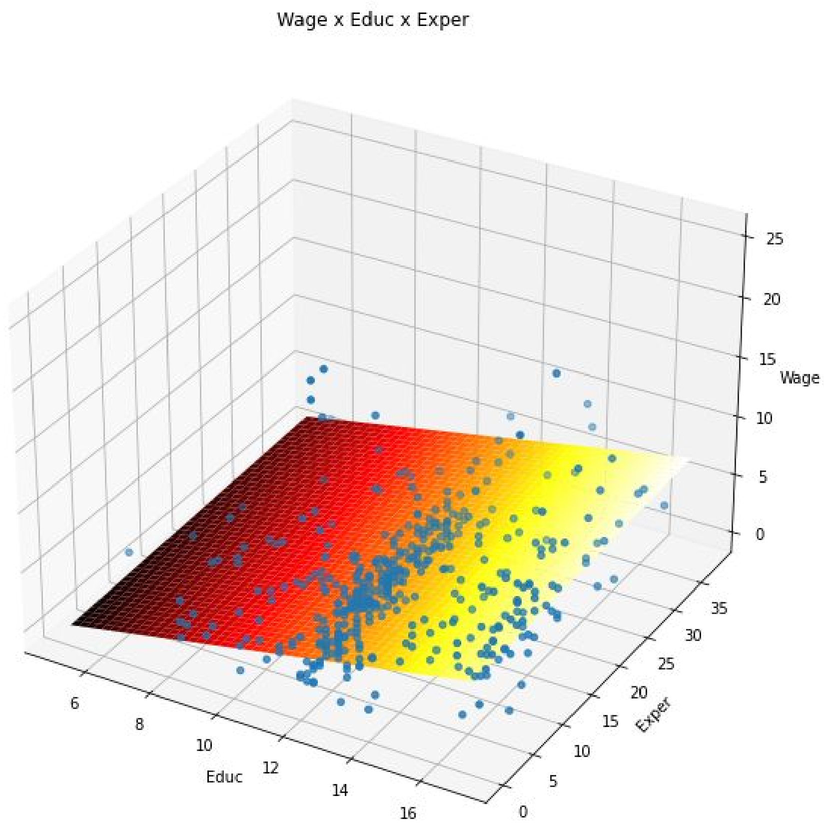
12. Faites une représentation graphique de la manière dont le salaire augmente avec l'éducation et l'expérience professionnelle. Commentez

To make this graph, we can make a OLS only with wage, educ and exper, and find the relationship between then.

```
                            OLS Regression Results                            
==============================================================================
Dep. Variable:                   wage   R-squared:                       0.121
Model:                            OLS   Adj. R-squared:                  0.116
Method:                 Least Squares   F-statistic:                     29.13
Date:                Sun, 22 Nov 2020   Prob (F-statistic):           1.39e-12
Time:                        16:47:09   Log-Likelihood:                -1091.6
No. Observations:                 428   AIC:                             2189.
Df Residuals:                     425   BIC:                             2201.
Df Model:                           2                                         
Covariance Type:            nonrobust                                         
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -2.4317      0.885     -2.746      0.006      -4.172      -0.691
x1             0.4966      0.066      7.537      0.000       0.367       0.626
x2             0.0247      0.019      1.323      0.186      -0.012       0.061
==============================================================================
Omnibus:                      348.306   Durbin-Watson:                   2.056
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             6624.423
Skew:                           3.421   Prob(JB):                         0.00
Kurtosis:                      21.018   Cond. No.                         113.
==============================================================================
```

So with this we find the relationship that is:
$wage = -2.432 + 0.497 \, educ + 0.025 \, exper$
These functions don't show all the information about wages, but can show the influence of the education and the experience in the wage. As we can notice, for the same experience, in each year of education the

wage increases 0.497 in average, and for the same education, each year of experience more, the wage increases in average 0.025.
And also show that for more years of education or more years of experience make you earn more in average.



13. Tester l'égalité des coefficients associés aux variables kidsgt6 et kidslt6. Interprétez.

As we saw in the classes and in the book basis. To evaluate if two constants are equal, as $x_1 = x_2$ in the linear system

$$y = x_1A + x_2B + x_3C + Cte + u.$$

We can create $K = x_1 - x_2$ to the equation be like

$$y = (K)A + x_2(A+B) + x_3C + Cte + u$$

and we try to make the hypothesis that $H_0 : K = 0$

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:	Sun, 22 Nov 2020	Prob (F-statistic):	2.00e-13
Time:	12:47:10	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.

Hipotesis Null $H_0 :\ K = 0$

P-value of K: 0.29769

The affirmation "We can reject H_0 with 5% of significance." is: False

So, we can’t reject the hypothesis $H_0 :\ 6 = 6$

14. En utilisant le modèle de la question 7, faire le test d'hétéroscédasticité de forme linéaire en donnant la p-valeur. Déterminer la ou les sources d’hétéroscédasticité et corriger avec les méthodes vues en cours. Comparer les écarts-types des coefficients estimés avec ceux obtenus à la question 7. Commenter.

For this question, we have to make the hypothesis of this regression follow the principle of homoscedasticity. And we have to reject or not this hypothesis. For this test we must calculate the residuals u from:

y = wage

X = const, x1 = city, x2 = educ, x3 = exper, x4 = nwifeinc, x5 = kidslt6, x6 = kidsge6

And after, we can make the follow OLS:

$y = u^2$

X = const, x1 = city, x2 = educ, x3 = exper, x4 = nwifeinc, x5 = kidslt6, x6 = kidsge6

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.022
Model:                            OLS   Adj. R-squared:                  0.008
Method:                 Least Squares   F-statistic:                     1.593
Date:                Sun, 22 Nov 2020   Prob (F-statistic):              0.148
Time:                        12:47:10   Log-Likelihood:                -2207.4
No. Observations:                 428   AIC:                             4429.
Df Residuals:                     421   BIC:                             4457.
Df Model:                           6
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.4856     13.111      0.113      0.910     -24.285      27.256
x1             5.9644      4.444      1.342      0.180      -2.770      14.699
x2             0.8077      0.956      0.845      0.399      -1.072       2.687
x3            -0.5341      0.284     -1.880      0.061      -1.093       0.024
x4             0.0435      0.211      0.206      0.837      -0.371       0.458
x5             4.9573      5.402      0.918      0.359      -5.661      15.575
x6            -0.4018      1.706     -0.236      0.814      -3.756       2.952
==============================================================================
Omnibus:                      638.793   Durbin-Watson:                   2.029
Prob(Omnibus):                  0.000   Jarque-Bera (JB):            96122.227
Skew:                           8.127   Prob(JB):                         0.00
Kurtosis:                      74.595   Cond. No.                         178.
==============================================================================
```

We achieve the follow statistiques:

- F: 1.5926
- P-value F stats: 0.1476

With this we **can't** reject the hypothesis of the homoscedasticity with 5% of significance level.

And avalianting the single p-values, we can't reject the hypothesis that each one of these variables individually is null with 5% of significance, and also can't conclude if there is heteroscedasticity in the data

The variable with more chance to reject this hypothesis is the exper variable which has 6.1% of significance level individually.

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. Refaire le test avec 3 groupes (mutuellement exclusifs) : les femmes de moins de 30 ans, entre 30 et 43 ans, plus de 43 ans. Donnez les p-valeurs

Our hypothesis is H_0 : $There\ isn't\ change\ in\ the\ structure\ when\ we\ split\ the\ data$

So we have to do the Chow test with the F statistics:

$F_{chow} = (SSR_0 - (sum(SSR_{splits})) * (n_0 - 2k) \;\; / \;\; ((sum(SSR_{splits}) * k)$

For the first group with

- Woman with: $age >= 43$
- Woman with: $age < 43$

We made the OLS and calculate the errors and find:

- F_chow: 1.1850
- P_value_chow: 0.30992

With this, we **can't** reject H_0 with 5%of significance level and can split the data.

****

For the first group with

- Woman with: $age >= 43$
- Woman with: $30 < age < 43$
- Woman with: $age <= 30$

We made the OLS and calculate the errors and find:

- F_chow: 1.5325
- P_value_chow: 0.15433

With this, we **can't** reject H_0 with 5%of significance level and can split the data.

16. Construire les variables binaires correspondant à l'âge des femmes de la question 15. Refaire la question 8 en ajoutant ces variables et en utilisant comme référence les femmes qui ont moins de 30 ans. Interprétez les paramètres associés aux variables binaires. Faire le test de non significativité de l'ensemble des variables binaires. Donnez les p-valeurs.

We separate the classes into:

Between 30 and 43 : $30 < age < 43$

More than 43 : $43 <= age$

Making two binary variables and the follow OLS:

y = lwage

X = const, x1 = city, x2 = educ, x3 = exper, x4 = nwifeinc, x5 = kidslt6, x6 = kidsge6, x7 = between30and43, x8 = more43

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  lwage   R-squared:                       0.161
Model:                            OLS   Adj. R-squared:                  0.145
Method:                 Least Squares   F-statistic:                     10.07
Date:                Sun, 22 Nov 2020   Prob (F-statistic):           7.08e-13
Time:                        15:22:51   Log-Likelihood:                -430.46
No. Observations:                 428   AIC:                             878.9
Df Residuals:                     419   BIC:                             915.5
Df Model:                           8
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.2210      0.248     -0.893      0.373      -0.708       0.266
x1             0.0475      0.071      0.672      0.502      -0.092       0.187
x2             0.1008      0.015      6.653      0.000       0.071       0.131
x3             0.0179      0.005      3.785      0.000       0.009       0.027
x4             0.0058      0.003      1.712      0.088      -0.001       0.012
x5            -0.0809      0.088     -0.920      0.358      -0.254       0.092
x6            -0.0183      0.029     -0.642      0.521      -0.074       0.038
x7            -0.1618      0.164     -0.986      0.325      -0.484       0.161
x8            -0.2558      0.169     -1.513      0.131      -0.588       0.077
==============================================================================
Omnibus:                       77.107   Durbin-Watson:                   1.996
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              282.164
Skew:                          -0.764   Prob(JB):                     5.36e-62
Kurtosis:                       6.673   Cond. No.                         254.
==============================================================================
```

Which means that people between 30 and 43 years old earns 16.18% less than people with less than 30 years old, and people with more than 43 years earns 25.5% less than people with less than 30 years old.

But we have also tested the hypothesis that these two new binary variables are null. So we made the comparison with the previous test:

y = lwage

X = const, x1 = city, x2 = educ, x3 = exper, x4 = nwifeinc, x5 = kidslt6, x6 = kidsge6

And compute the errors:

- SSR0 = 188.5899
- SSR5 = 187.3068

Finding a **F-statistiques: 1.44205** and **P-value of the hypothesis 0.2376.**

This means for the hypothesis $H_0 : ageBetween30and43 = age\_above43 = 0$

We **can't** reject the hypothesis H_0 with 5% of significance level.

Project 2

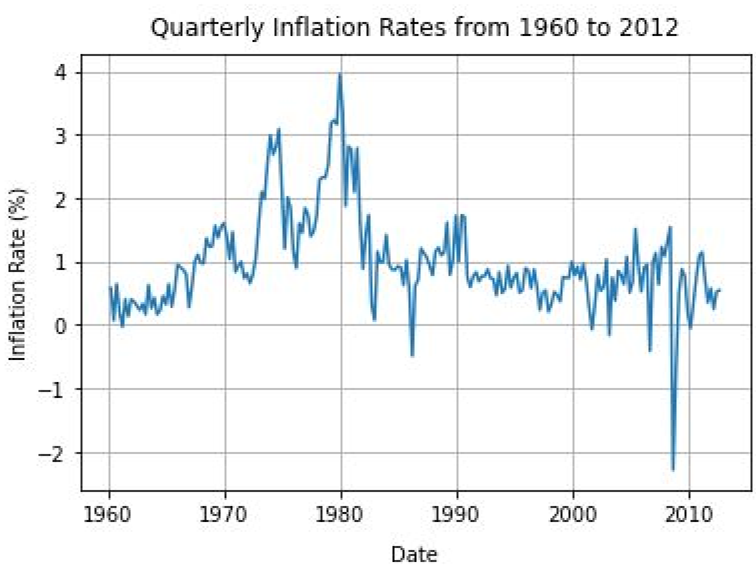
1. Importer les données du fichier quarterly.xls (corriger le problème éventuel d'observations manquantes).

We can import the dataset with the Pandas Framework, and do some pre-processing on the data. We found that there are no missing values or timestamps on the dataset.

	DATE	FFR	Tbill	Tb1yr	r5	r10	PPINSA	Finished	CPI	CPICORE	M1NSA	M2SA	M2NSA	Unemp	IndProd	RGDP	Potent	Deflator	Curr
0	1960-01-01	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	1960-04-01	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	1960-07-01	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	1960-10-01	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	1961-01-01	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

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

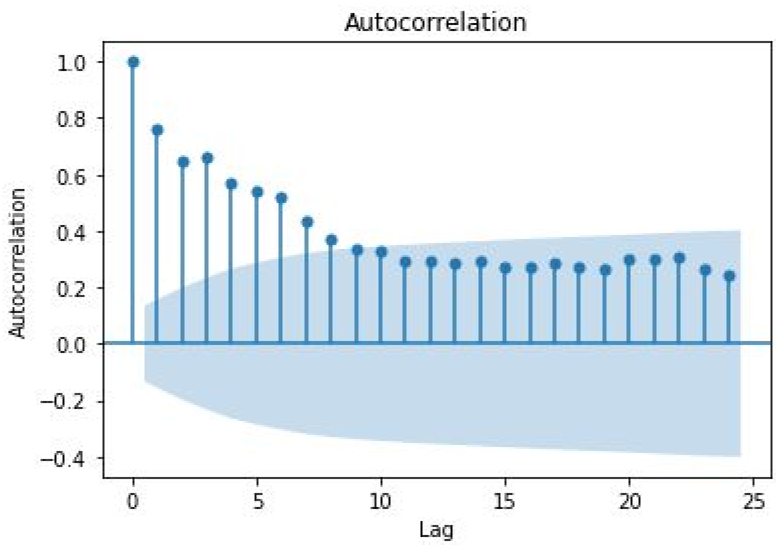
The inflation rate is the percentage change between two consecutive timestamps of the CPI variable. We can do this using Panda's *percent change* function.



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

First, we're going to plot the autocorrelogram for *inf*. The idea here is to calculate the correlation between a time series observation and its previous values, which is called the autocorrelation. The autocorrelogram, thus, is just a plot of the autocorrelation by lag. To this end, we can use statsmodels' plot_acf function, which will also plot the region of confidence (values outside the blue region have confidence over 95%).

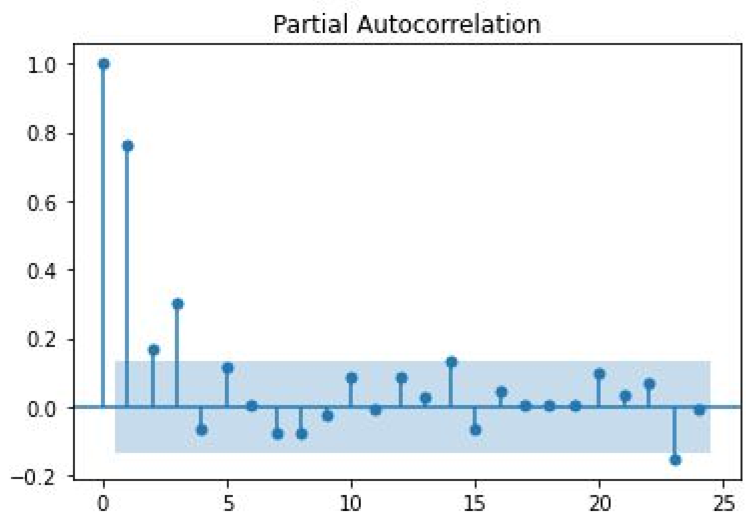
In our case, we want to see how the inflation rate of one quarter correlates with that of previous quarters.



We can see in the autocorrelogram above that the inflation rate of a quarter is highly correlated (> 0.5) with those of the 5 previous quarters. Also, since these autocorrelation values are outside the blue region, they have a high statistical confidence (over 95%).

Now, let's plot the partial autocorrelogram. The idea behind the partial autocorrelogram of a time series to obtain the conditional correlation between the observation at time t and the observation at time t-h (lag h), given that we observed what we observed in all timesteps between t and h. Therefore, the partial autocorrelation aims to remove the effects of the observations between the current observation and the observation at lag h, which also means that it removes indirect correlations that are included in the autocorrelogram.

Due to this property, for an AR model of order k, the partial autocorrelations are 0 for every lag beyond k. We can use this information to estimate the order of an AR model by counting the number of lags with non-zero partial autocorrelation.



From the partial autocorrelogram above, we can see that the partial autocorrelation is statistically significant for lags up to 3 (values outside the blue region, which means confidence over 95%) and that it oscillates around 0, which could suggest an AR model of order k=3 to predict the inflation.

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

In time series analysis, **stationarity** means that the joint distribution for random variables at times $(Y_{s+1}, Y_{s+2}, \ldots, Y_{s+T})$ does not depend on s. This means that the distribution of the processe's variables does not vary over time.

On the other hand, **ergodicity** means that the process does not depend on initial conditions, and that we can deduce statistical properties given sufficient random samples of a process.

If both of these conditions are satisfied, and the mean of variables is not infinite, then the temporal mean is equal to the spatial mean.

$$E(Y_t) = \frac{1}{T}\sum_{t=1}^{T} Y_t \rightarrow \mu$$

A **spurious regression** is a problem that happens when a regression shows evidence of a non-existing relationship between two variables. This means that the regression coefficient estimate should be zero (because the two variables are uncorrelated), but the regression returns a statistically significant value that is not zero, but has a high R^2 value (generalizes very badly). This can happen, if the time series are random walks, which are non-stationary.

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

We want to find the value of p that produces the best AR(p) model for the inflation values. We can do this by testing multiple AR(p) models, with varying values for p, and choosing the value that generates the model with lowest AIC value (which measures model quality while taking complexity into account). We find that p=3 generates the model with lowest AIC.

```
                            AutoReg Model Results
==============================================================================
Dep. Variable:                    CPI   No. Observations:                  211
Model:                     AutoReg(3)   Log Likelihood                -138.521
Method:               Conditional MLE   S.D. of innovations              0.471
Date:                Sat, 21 Nov 2020   AIC                             -1.458
Time:                        13:17:33   BIC                             -1.378
Sample:                             3   HQIC                            -1.425
                                  211
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept      0.1366      0.057      2.406      0.016       0.025       0.248
CPI.L1         0.5828      0.066      8.815      0.000       0.453       0.712
CPI.L2        -0.0184      0.077     -0.239      0.811      -0.170       0.133
CPI.L3         0.2979      0.066      4.515      0.000       0.169       0.427
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            1.0906           -0.0000j            1.0906           -0.0000
AR.2           -0.5144           -1.6774j            1.7545           -0.2974
AR.3           -0.5144           +1.6774j            1.7545            0.2974
-----------------------------------------------------------------------------
```

6. Estimer le modèle de la courbe de Phillips qui explique le taux de chômage (Unemp) en fonction du taux d’inflation courant et une constante.

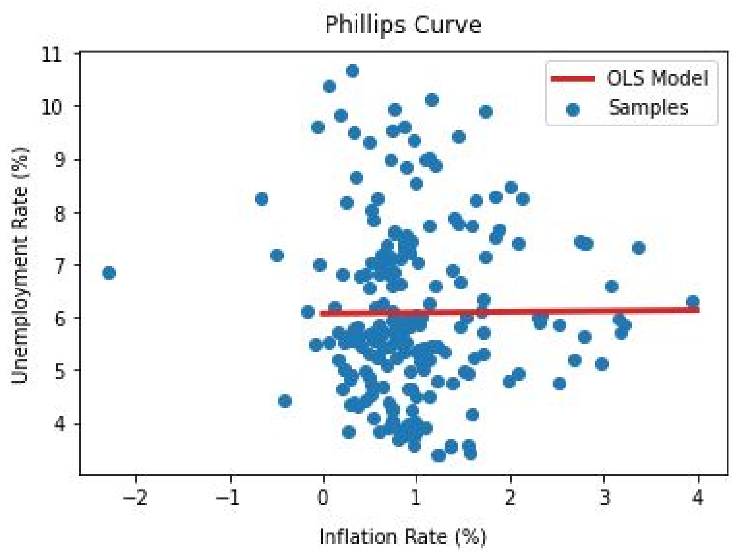
The Phillips Curve model is a simple static time-series model that expresses the unemployment rate at a time t in function of the inflation rate at the same time t:

$$(\textit{unemployment rate})_t = \beta_0 + \beta_1 (\textit{inflation})_t + u_t$$

We can build this using the OLS model from the statsmodels package.

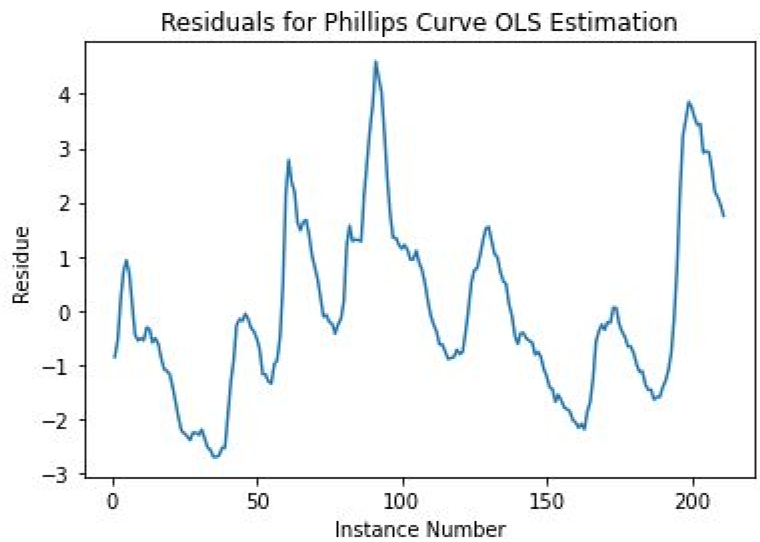
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Unemp   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                 -0.005
Method:                 Least Squares   F-statistic:                   0.01214
Date:                Sat, 21 Nov 2020   Prob (F-statistic):              0.912
Time:                        13:18:26   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
x1             0.0159      0.144      0.110      0.912      -0.269       0.301
==============================================================================
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.                         2.99
==============================================================================
```

We can plot the OLS predictions to see how it fit the data we used.



7. Tester l'autocorrélation des erreurs.

First, let's visualize the errors (or residuals) of the OLS model.



Now, we want to verify if errors are autocorrelated. To this end, we can test the hypothesis H_0 that the errors are serially uncorrelated. For an AR(1) model, $(u)_t = \rho\,(u)_{t-1}+e_t$, this hypothesis can be translated to:

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

Therefore, we can fit an AR(1) model to the residuals of the Phillips Curve OLS Estimator and check the t-value for ρ to test the null hypothesis.

```
                            AutoReg Model Results
==============================================================================
Dep. Variable:                      y   No. Observations:                  211
Model:                     AutoReg(1)   Log Likelihood                 -70.272
Method:               Conditional MLE   S.D. of innovations              0.338
Date:                Sat, 21 Nov 2020   AIC                             -2.140
Time:                        13:17:34   BIC                             -2.092
Sample:                             1   HQIC                            -2.121
                                  211
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept      0.0122      0.023      0.524      0.601      -0.034       0.058
y.L1           0.9800      0.014     67.714      0.000       0.952       1.008
                                    Roots
=============================================================================
                  Real          Imaginary           Modulus         Frequency
-----------------------------------------------------------------------------
AR.1            1.0204           +0.0000j            1.0204            0.0000
-----------------------------------------------------------------------------
```

We can see that the p-value for the coefficient ρ (appears as y.L1 in the model's summary) is approximately 0. Therefore, we reject the null hypothesis that the errors are serially uncorrelated at 5%, and conclude that the errors are autocorrelated.

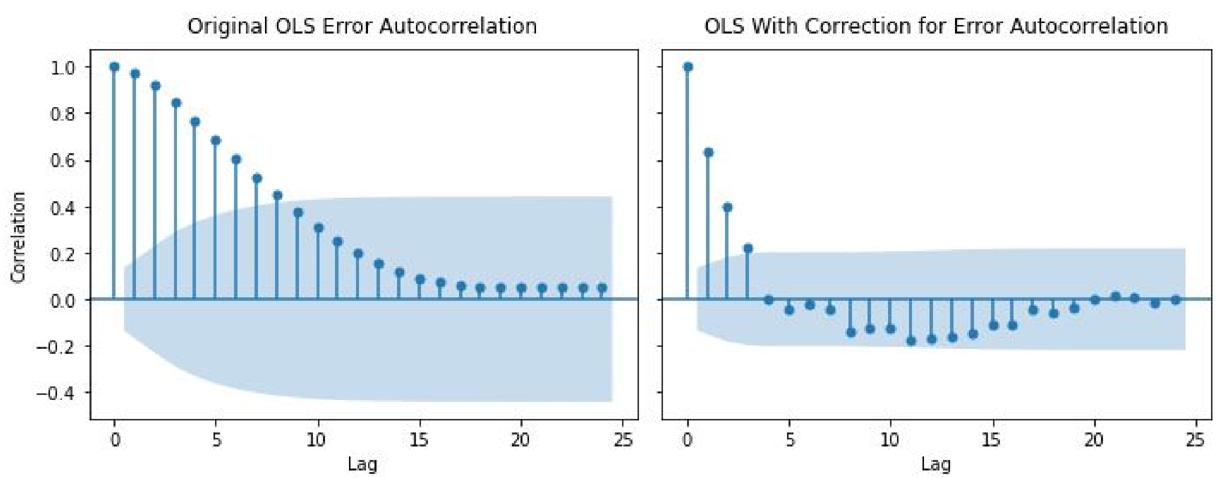
8. Corriger l'autocorrélation des erreurs par la méthode vue en cours.

To correct the issue of error autocorrelation, we can build the following regression model for the Phillips Curve:

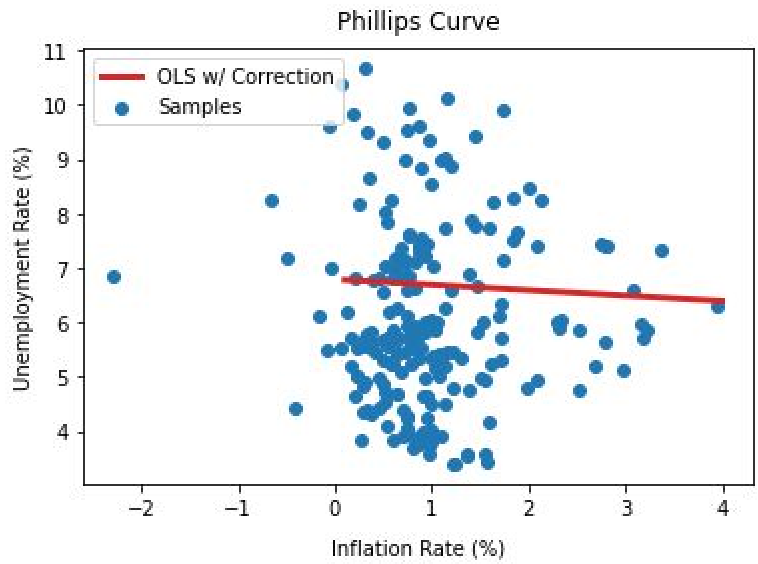
$$\widetilde{y}_t - \rho y_{t-1} = \beta_0(1-\rho) + \beta_1(\widetilde{x}_t) + e_t \quad \text{Where } \widetilde{y}_t = y_t - \rho y_{t-1} \text{ and } \widetilde{x}_t = x_t - \rho x_{t-1}.$$

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.024
Model:                            OLS   Adj. R-squared:                  0.020
Method:                 Least Squares   F-statistic:                     5.203
Date:                Sat, 21 Nov 2020   Prob (F-statistic):             0.0236
Time:                        16:31:57   Log-Likelihood:                -66.797
No. Observations:                 210   AIC:                             137.6
Df Residuals:                     208   BIC:                             144.3
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
x1             6.7936      1.152      5.899      0.000       4.523       9.064
x2            -0.0996      0.044     -2.281      0.024      -0.186      -0.014
==============================================================================
Omnibus:                       82.032   Durbin-Watson:                   0.725
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              283.478
Skew:                           1.600   Prob(JB):                     2.78e-62
Kurtosis:                       7.707   Cond. No.                         26.4
==============================================================================
```

Comparing the two models built previously, we can see that the residual autocorrelation is significantly reduced after we perform this correction in the OLS model:



Again, we can plot the OLS predictions to see how it fit the data we used, this time using the estimator with correction for autocorrelated errors.



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

First, we split the data into two, equal sized, subsamples, and then we perform a Chow Test to test the null hypothesis $\mathbf{H_0}$: there's no significant improvement in fit by splitting the data and fitting each subsample individually.

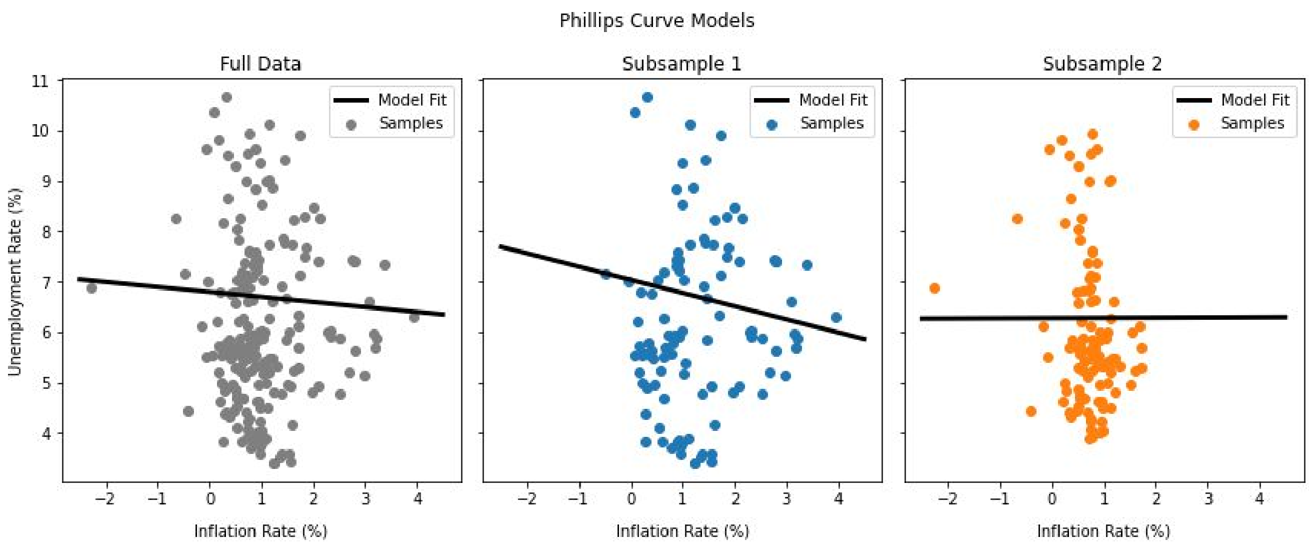
We can fit individual error corrected models for each subsample of the data and calculate the Chow Test F-statistic, which follows an F-distribution with k and $n{-}2k$ degrees of freedom. Therefore, we can calculate the critical f-value for this distribution as well as the measured statistic p-value and, thus, test hypothesis H_0.

SSR for full data model: 23.2287
SSR for subsample 1 model: 13.5797
SSR for subsample 2 model: 8.7855

Chow test F-value: 3.9768
p-value: 0.0202

We can see that the p-value is below 5%, and, therefore, we reject the null hypothesis that there's no improvement in fit by splitting the data and fitting each subsample individually. This indicates that the model for the full data is not stable throughout time.

Here's a visualization of each fitted model:



10. Estimer la courbe de Phillips 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.

To perform the Granger Causality Test, we will test the null hypothesis H_0: **inflation does not Granger cause unemployment**. To this end, we're going to fit two OLS models to the unemployment data: one containing the 4 previous lags of unemployment and inflation (*unrestricted model*, figure on the left), and the other containing only the 4 previous lags of unemployment itself (*restricted model*, figure on the right).

	coef	std err	t	P>|t|
const	0.1457	0.072	2.014	0.045
x1	1.5937	0.071	22.383	0.000
x2	-0.6472	0.134	-4.832	0.000
x3	0.0222	0.135	0.164	0.870
x4	-0.0080	0.070	-0.114	0.910
x5	0.0311	0.038	0.827	0.409
x6	-0.0236	0.041	-0.577	0.565
x7	0.0689	0.040	1.729	0.085
x8	0.0163	0.038	0.435	0.664

	coef	std err	t	P>|t|
const	0.2157	0.071	3.036	0.003
x1	1.6459	0.070	23.393	0.000
x2	-0.6975	0.135	-5.159	0.000
x3	0.0238	0.135	0.177	0.860
x4	-0.0078	0.070	-0.112	0.911

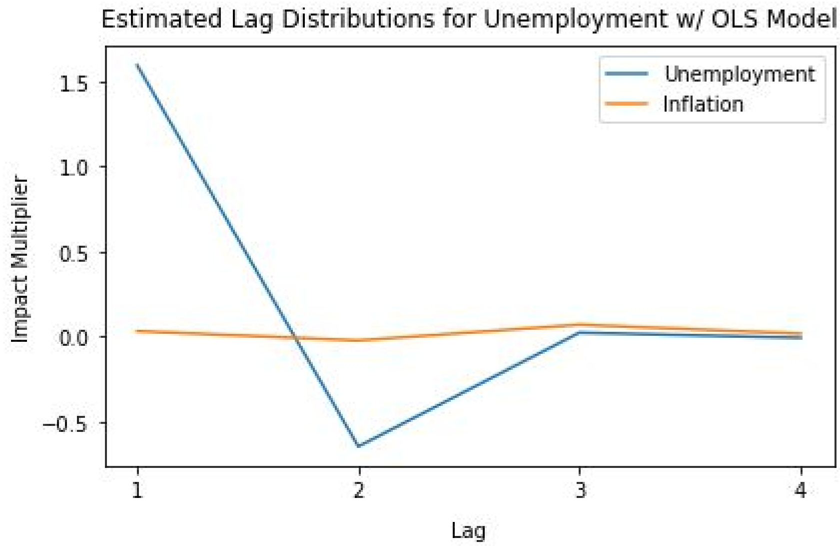
We can see that, in the unrestricted model, the coefficients for the previous 4 lags of inflation (x5 through x8) all pass the t-test (they all have p-values over 0.05). Now, we're going to run an F-test under the hypothesis that **adding the 4 previous lags of inflation to the model does not jointly provide a significantly better fit**:

F-statistics for Granger Causality Test 3.7967
p-value of measured F-statistic: 0.0054

The p-value indicates that we can reject this hypothesis at the 5% significance level. Therefore, we can conclude that adding the 4 previous lags of inflation to the model, in fact, does provide a significantly better fit. Thus, we satisfy conditions for the Granger Test, rejecting hypothesis H_0, and we can conclude that inflation Granger causes unemployment.

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

We can plot the OLS coefficient for each distributed lag in the previous model to obtain the estimated lag distribution for Unemployment:



With this plot, we can see the dynamic effect that a temporary increase in each variable (unemployment or inflation) has on unemployment. We can clearly see in the plot above that a temporary increase in the previous value of unemployment is a lot more impactful on unemployment itself than a temporary increase in the previous value of inflation.

We can calculate the inflation rate's long-run impact on the unemployment rate by adding up all of its coefficients.

Previous Inflation Rate long-run impact on Unemployment Rate: 0.0928

Therefore, we can estimate a 9.28% increase in the Unemployment due to a permanent one percent increase in the previous Inflation.