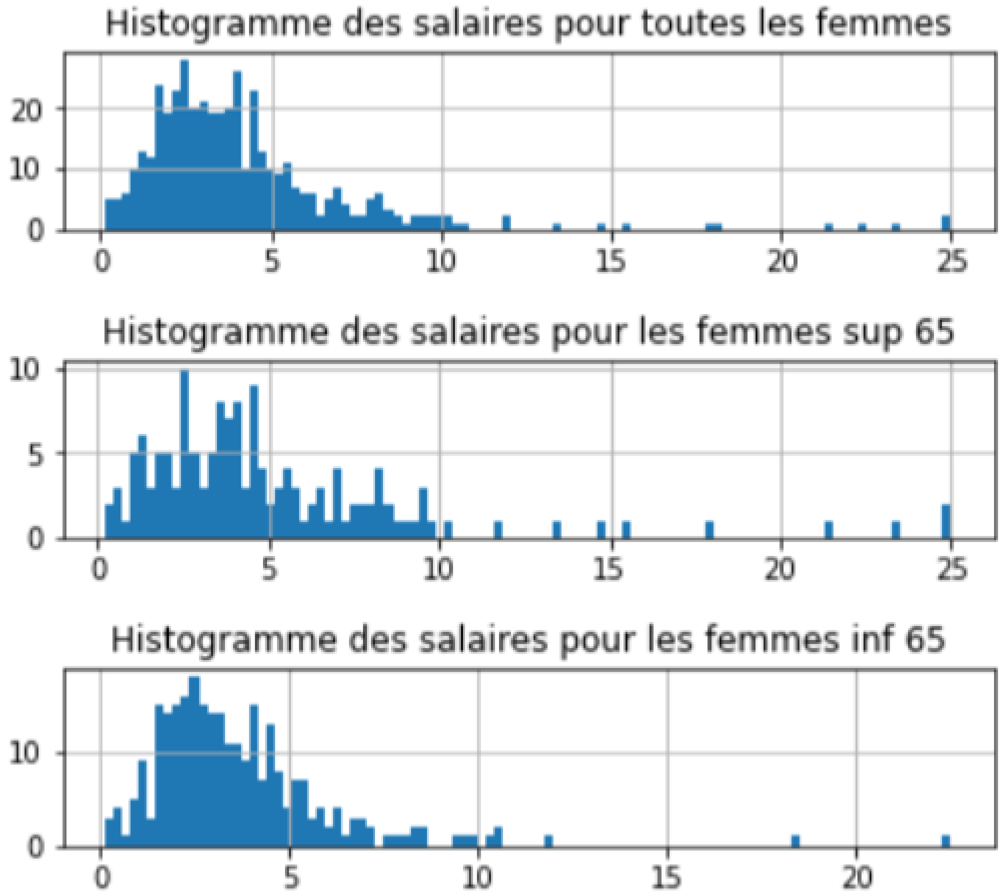
2. Faire les statistiques descriptives du salaire, de l'age et de l'éducation pour l'ensemble des femmes puis, pour les femmes dont le salaire du mari est supérieure au 65ème percentile de l'échantillon, puis pour les femmes dont le salaire du mari est inférieur au 65ème percentile de l'échantillon. Commenter

Sur le Salaire

Statistique élèmentaire:

	common Wage	Sup_65	Inf_65
count	428.000000	148.000000	276.000000
mean	4.177682	5.139315	3.653995
std	3.310282	4.351728	2.471311
min	0.128200	0.213700	0.128200
25%	2.262600	2.561925	2.139100
50%	3.481900	4.008050	3.169700
75%	4.970750	6.516300	4.508775
max	25.000000	25.000000	22.500000

Les histogrammes:

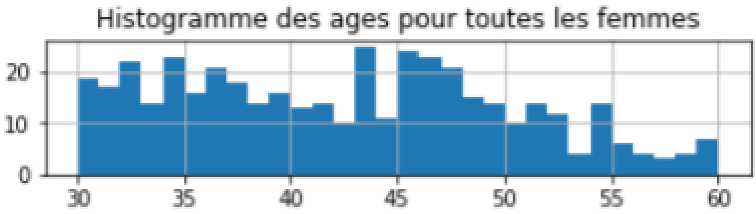


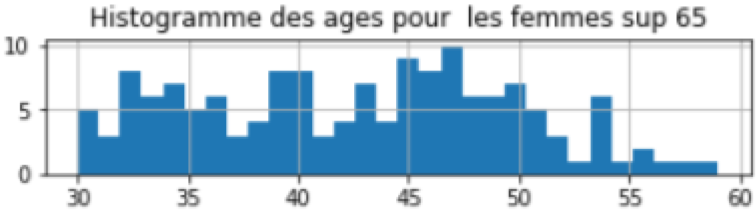
Les boxplots

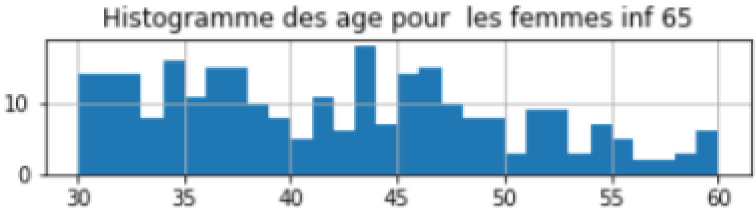
Pour la variable salaire, on remarque que les femmes dont le salaire du mari est supérieur au 65ème percentile de l'échantillon (groupe A), ont en général un salaire supérieur à la moyenne et un écart-type plus important, alors que les femmes dont le salaire du mari est inférieur au 65ème percentile de l'échantillon (groupe B) a en moyenne un salaire inférieur à la moyenne de notre echantillon et un écart-type plus faible. On peut donc faire l'hypothèse que les femmes du groupe B auront un profil généralement proche (un salaire moindre par rapport au commun des femmes) tandis que les femmes du groupe A auront des profils plus hétéroclites.

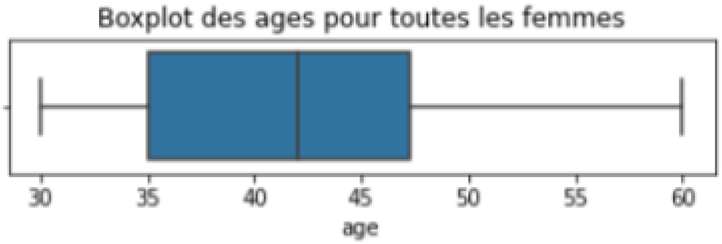
Sur l'age

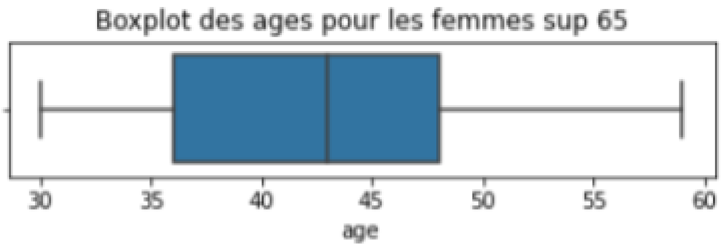
	common Wage	Sup_65	Inf_65
count	428.000000	148.00000	276.000000
mean	41.971963	42.52027	41.583333
std	7.721084	7.35168	7.910656
min	30.000000	30.00000	30.000000
25%	35.000000	36.00000	35.000000
50%	42.000000	43.00000	41.000000
75%	47.250000	48.00000	47.000000
max	60.000000	59.00000	60.000000

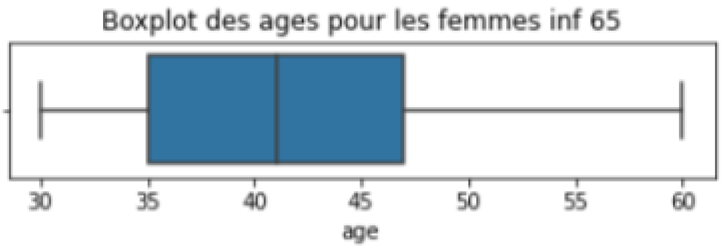








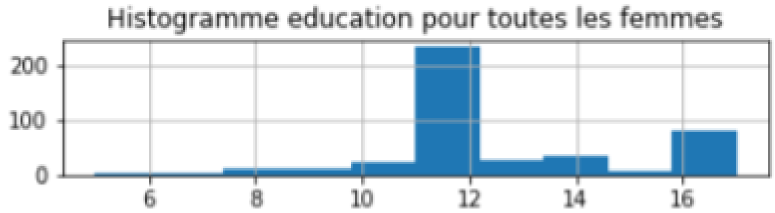


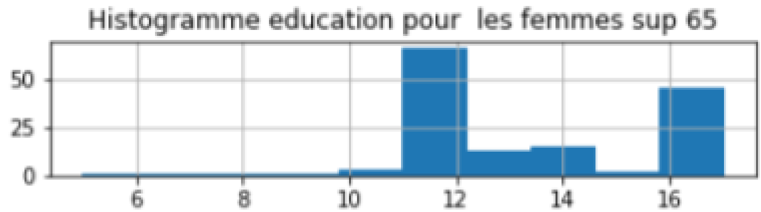


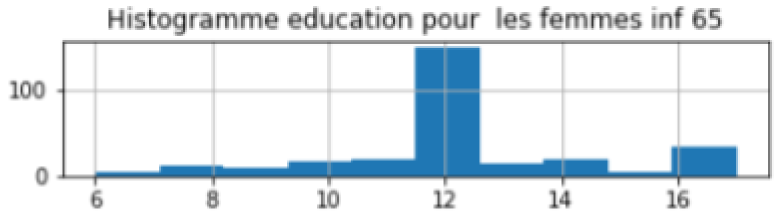
Au niveau de le variable âge, on remarque qu'en général les femmes faisant partie du groupe A sont plus âgées que la moyenne des femmes tandis que celle du groupe B est plus jeune, les écarts type entre chaque groupe quasiment similaire.

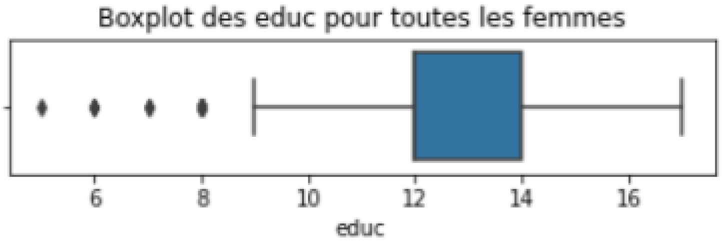
Sur l'education

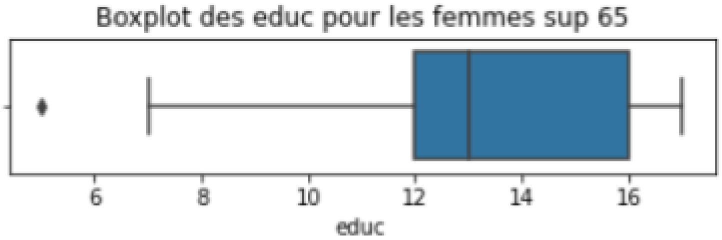
	common Wage	Sup_65	Inf_65
count	428.000000	148.000000	276.000000
mean	12.658879	13.520270	12.221014
std	2.285376	2.345845	2.126472
min	5.000000	5.000000	6.000000
25%	12.000000	12.000000	12.000000
50%	12.000000	13.000000	12.000000
75%	14.000000	16.000000	12.250000
max	17.000000	17.000000	17.000000

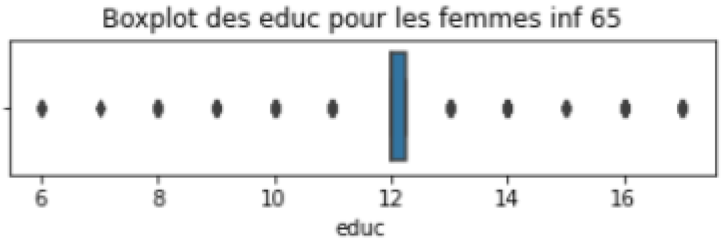






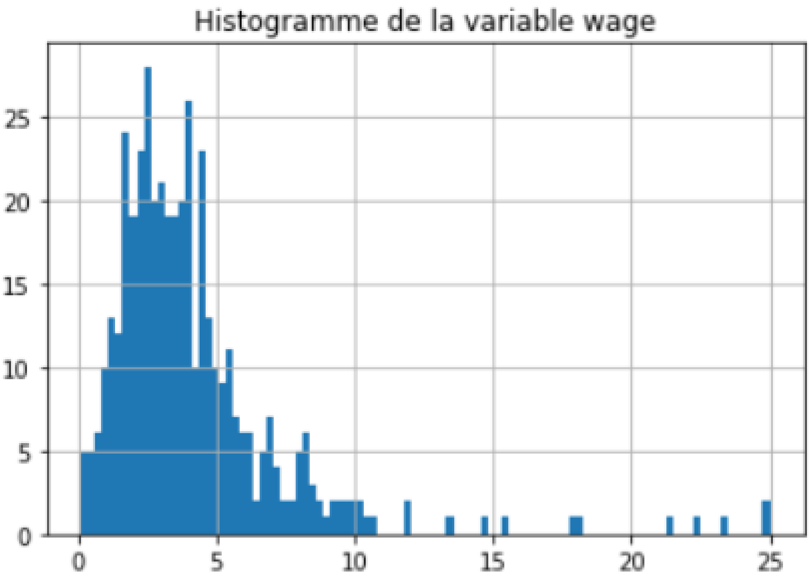


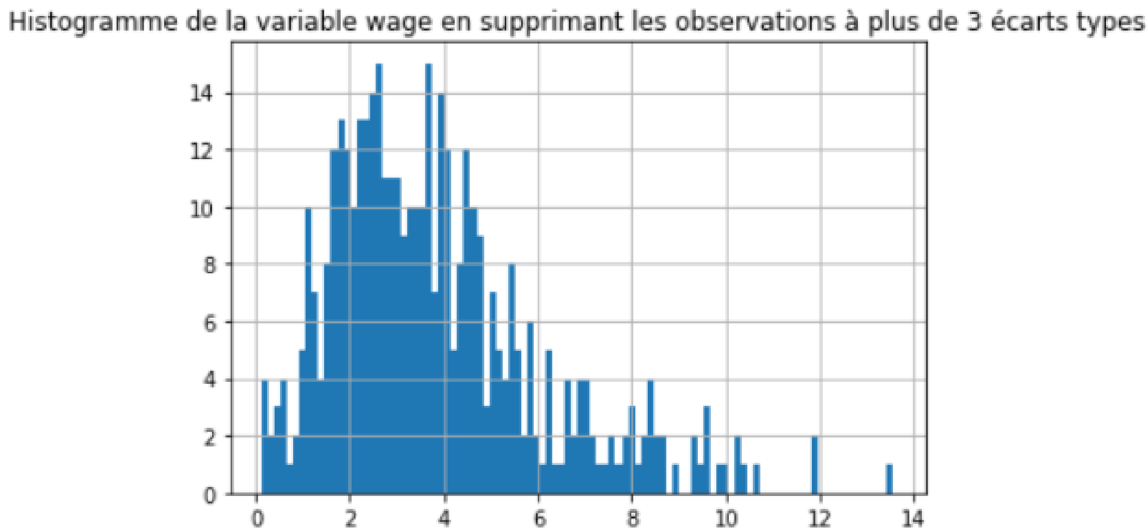




Pour la variable éducation, on constate que les femmes du groupe A feront en moyenne de plus longque étude par rapport à la moyenne de l'ensemble de l'Échantillon, tandis qu'à l'opposé les femmes du groupe B feront moins d'années d'études. On remarque également que les écarts-types de non trois group.
En conclusion les variables salaires, âge et duc sont corrélés avec la variable huswage.

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

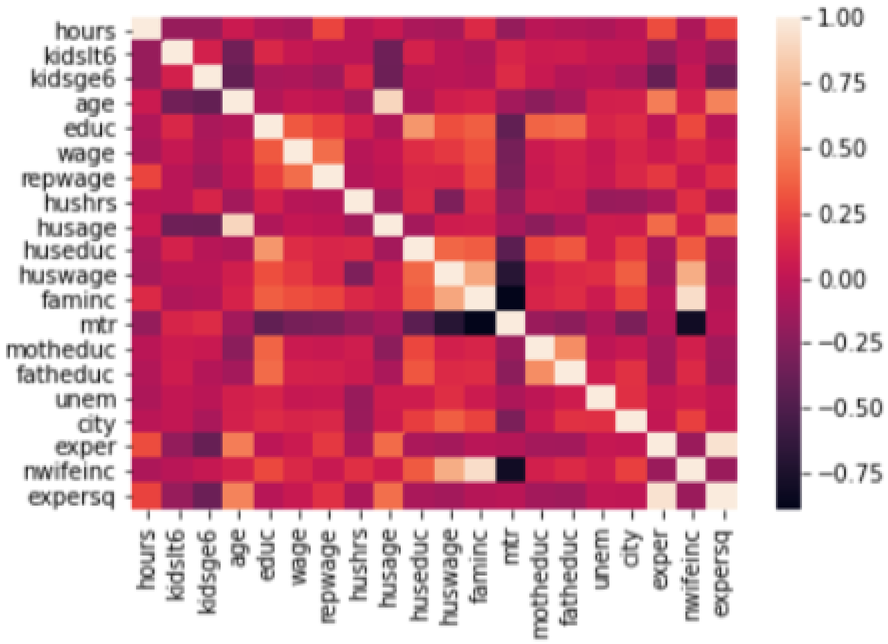




De manière générale, on observe une asymétrie positive. Le premier histogramme nous informe qu'il y a bon nombre d'outliers. Sur le second histogramme, même en supprimant les valeurs extrêmes, l'asymétrie positive persiste.

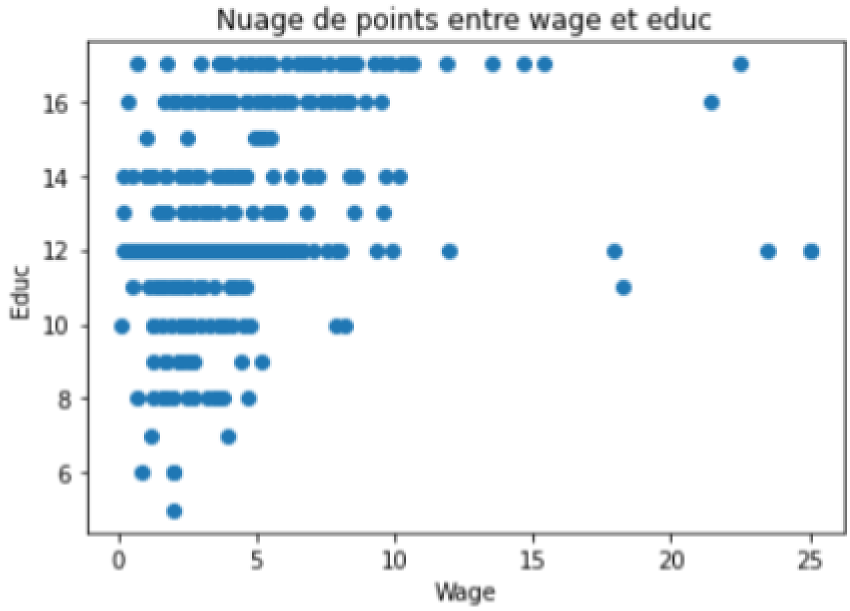
4. Calculer les corrélations motheduc et fatheduc. Expliquer le problème de multi-collinéarité. Commenter.

La corrélations entre motheduc et fatheduc est de 0.55. Les variables motheduc et fatheduc ont une corrélation positive assez élevé. Ceci pourrait expliquer que les niveaux d'éducation de la mère et du père sont en général corrélées. Ceci impliquerait donc la présence d'une colinéarité entre ces deux variables.
Dans une régression, la multicolinéarité est un problème qui survient lorsque certaines variables de prévision du modèle mesurent le même phénomène. Une multicolinéarité prononcée s’avère problématique, car elle peut augmenter la variance des coefficients de régression et les rendre instables et difficiles à interpréter.

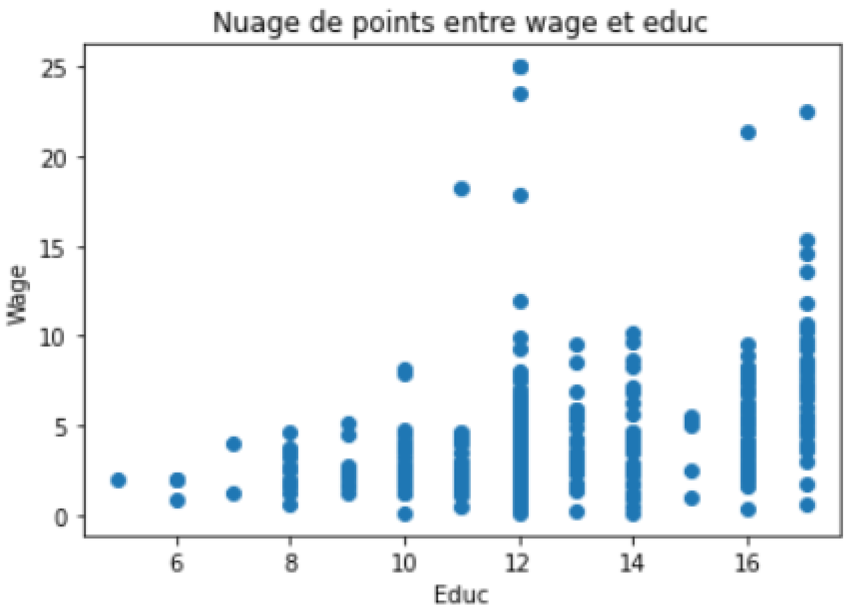


Du fait d'une forte corrélation positive entre les variables educ, huseduc, fatheduc et dans la moindre mesure motheduc, nous ne sommes pas vraiment dans le cas d'un problème à multicolinéarité. Cependant ces corrélations relativement élevées peuvent néanmoins poser des difficultés au modèle notamment dans la significativité des coefficients de régression.

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



Interversion des axes



On remarque une corrélation positive entre les variables wage et educ. Cependant, il ne s'agit pas d'un effet "toute chose étant égale par ailleurs". En effet, nous ne prenons pas en compte plusieurs variables (non observées) qui ont une forte influence sur le salaire, tel que la compétence physique ou intellectuelle d'un salarié, ou la pénibilité de la tache.

6. Quelle est l'hypothèse fondamentale qui garantit

des estimateurs non biaisés ? Expliquer le biais de variable omise.

L'hypothèse fondamentale qui garantit des estimateurs non biaisés est que le terme d'erreur μ ne soit pas corrélé avec les régresseurs (l'endogénéité) et que sa moyenne soit nulle ($E[\mu] = 0$). Ignorer cette hypothèse dans l'estimation viole l'hypothèse d'orthogonalité présente dans le théorème de Gauss-Markov. En effet , si une ou plusieurs variables explicatives sont corrélés avec le terme d'erreur, alors le coefficient estimé par l'estimateur des moindres carrés ordinaires (MCO) sera biaisé.

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.

On affiche le résumé de la régression du log de wage par rapport aux variables city, educ, exper, nwifeinc, kidslt6, kidsgt6

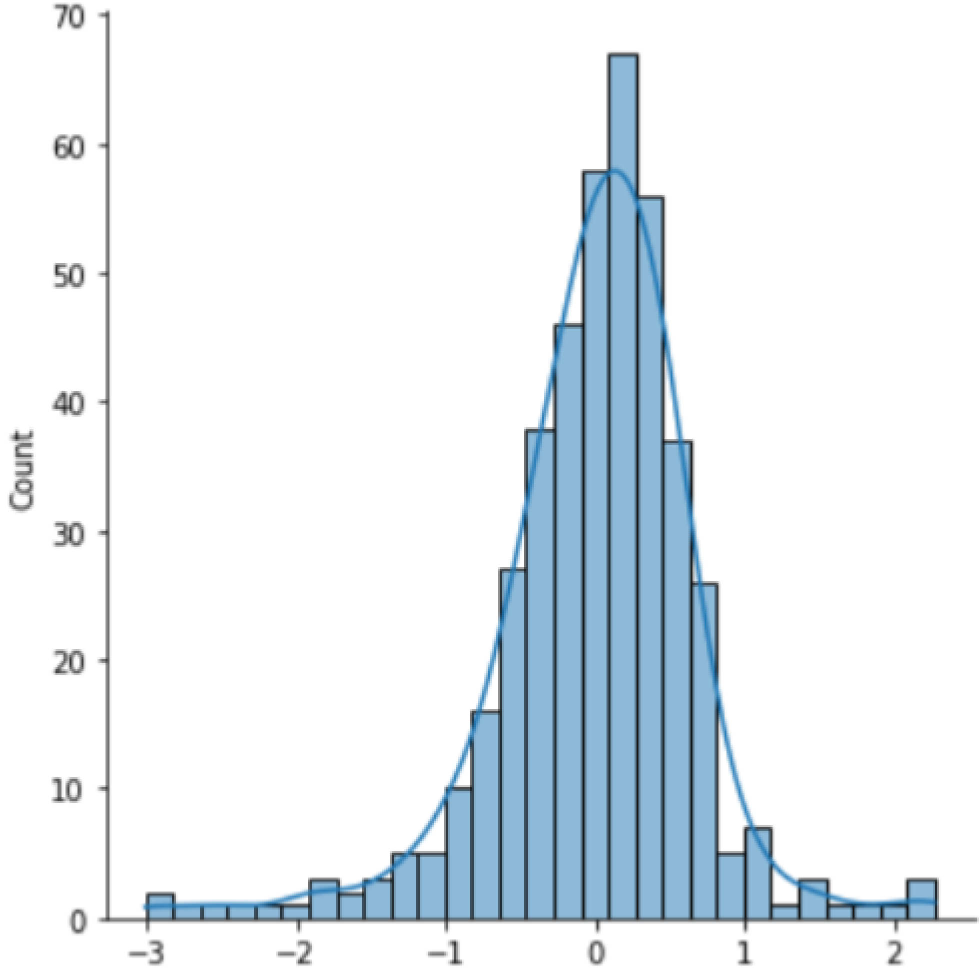
OLS Regression Results

Dep. Variable:	wage	**R-squared:**	0.156
Model:	OLS	**Adj. R-squared:**	0.144
Method:	Least Squares	**F-statistic:**	12.92
Date:	Sun, 16 May 2021	**Prob (F-statistic):**	2.00e-13
Time:	17:49:30	**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]**
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
const	-0.3990	0.207	-1.927	0.055	-0.806	0.008

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.

Histogrammes des résidus:



Du point de vue général, la forme de l'histogramme des résidus est proche d'une gaussienne centrée en 0. Néanmoins, la traine gauche de la courbe étant plus longue que celle de droite cela reflete donc une assymetire négative.

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.

On est sur un test de student du type bilatéral, c'est-à-dire pour déterminer si la variable nwifeinc est significative, il faudra tester l'hypothèse

$H_0 : a_{nwifeinc} = 0$
$H_1 : a_{nwifeinc} \neq 0$

On definit :

- $\hat{a}_{nwifeinc}$: estimateur du coefficient de nwifeinc dans le modèle etudié
- $a_{nwifeinc}$: coefficient de nwifeinc dans le modèle etudié
- $\sigma_{\hat{a}} = (Y - \hat{Y})^2 (X^T X)^{-1}$

Nous définissons la t-statistique $Z_{nwifeinc}$ telle que $Z_{nwifeinc} = \frac{\hat{a}_{nwifeinc} - a_{nwifeinc}}{\sigma_{\hat{a}}}$.

Sous l'hypothèse H_0 , nous avons $Z_{nwifeinc} = \frac{\hat{a}_{nwifeinc}}{\sigma_{\hat{a}}} \approx 1.47$

On pose α le seuil de risque de significativité tel que $\alpha = 0.01, 0.05, 0.1$

Si $|Z_{nwifeinc}| > 1 - \frac{alpha}{2}$ on rejette l'hypothèse nulle.

Ci dessous on affiche les différentes p-value pour les risque 1%, 5% et 10%

```
Pour le risque alpha =  0.01
On accepte l'hypothèse de non-significativité
p_value : 3.533096948479134e-07

Pour le risque alpha =  0.05
On accepte l'hypothèse de non-significativité
p_value : 9.878952107382961e-05

Pour le risque alpha =  0.1
On accepte l'hypothèse de non-significativité
p_value : 0.0010602435910751766
```

Les p-values sont inférieurs au seuils critiques pour tous les risques étudiés 0.01, 0.05 et 0.1. L'hypothèse H0 est rejetée et l'hypothèse H1 est acceptée pour tous les risques étudiés. Les coefficients de nwifeinc sont donc non nuls pour ce modèle.

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)

On est sur un test de student du type bilatéral, c'est-à-dire pour déterminer si la variable nwifeinc est significative, il faudra tester l'hypothèse

- $H_0 : a_{nwifeinc} = 0.01$
- $H_1 : a_{nwifeinc} \neq 0.01$

Ceci revient à étudier l'hypothèse suivante:

- $H_0 : a_{nwifeinc} - 0.01 = 0$
- $H_1 : a_{nwifeinc} - 0.01 \neq 0$

On definit :

- $\hat{a}_{nwifeinc}$: estimateur du coefficient de nwifeinc dans le modèle etudié
- $a_{nwifeinc}$: coefficient de nwifeinc dans le modèle etudié
- $\sigma_{\hat{a}} = (Y - \hat{Y})^2 (X^T X)^{-1}$

Nous définissons la t-statistique $Z_{nwifeinc}$ telle que $Z_{nwifeinc} = \frac{\hat{a}_{nwifeinc} - a_{nwifeinc}}{\sigma_{\hat{a}}}$.

Sous l'hypothèse H_0 , nous avons $Z_{nwifeinc} = \frac{\hat{a}_{nwifeinc}}{\sigma_{\hat{a}}} \approx 1.47$

On pose α le seuil de risque de significativité tel que $\alpha = 0.01, 0.05, 0.1$

Si $|Z_{nwifeinc}| > 1 - \frac{alpha}{2}$ on rejette l'hypothèse nulle.

Nous obtenons ce résultat:

```
Pour le risque alpha =  0.05
On accepte l'hypothèse de non-significativité
p_value = 0.050000000000006393
```

La p-value est légèrement supérieur au risque, donc H0 est acceptée, et le modèle contraint est validé.

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.

Le modèle est supposé satisfaire aux hypothèses des moindres carrés ordinaires (MCO), avec aléa normal (c'est-à-dire que les résidus ϵ (également appelé bruit) sont indépendants et de même loi normale N(0,$\sigma_{\hat{a}}$)), $\sigma_{\hat{a}}$ correspondant au coefficient en question.

Dans cette question l'hypothèse H_0 posée est jointe, car vise à étudier deux coefficients $a_{nwifeinc}$ et a_{city}. L'hypothèse H0 est la suivante:

- H0: $a_{nwifeinc}$ = 0.01 et a_{city} = 0.05
- H1: $a_{nwifeinc} \neq 0.01$ ou $a_{city} \neq 0.05$

Pour tester cette hypothèse, nous utiliserons le test de Fisher.

Sous H_0, le nombre des variables explicatives est diminué du nombre de conditions élémentaires (dans ce cas 2). On pose donc le modèle suivant :

- $log(wage) - 0.01nwifeinc - 0.05city = a_{educ}educ + a_{exper}exper + a_{kidslt6}kidslt6 + a_{kids}$

Tandis que sous H_1 nous avons le modèle standard de notre régression :

- $log(wage) = a_{city}city + a_{nwifeinc}nwifeinc + a_{educ}educ + a_{exper}exper + a_{kidslt6}kidslt6 +$

Ainsi sous l'hypothèse nulle, nous définissons F la statistique de Fisher, telle que:

$$F = \frac{\frac{SCR_0 - SCR_1}{dl0 - dl1}}{\frac{SCR1}{dl1}}$$

Où

- SCR_0 est la somme des carrés des résidus de la régression par les mco du modèle H_0. -SCR_1 est la somme des carrés des résidus de la régression par les mco du modèle H_1 - $dl0$ et $dl1$ sont respectivement les degrés de liberté des modèle H_0 et $H1$.

Si H_0 est vrai, la statistique F suit une loi de Fisher.

On obtient :

- la statistique F veut : 1.3338941768890733
- la P_Value est : 0.26456305438708805

Nous obetenons une P_Value $\approx$ 0.26 > 0.05, l'hypothèse H_0 est vrai.

11. Tester l'hypothèse joint que nwifeinc+city=0.1 et educ+exper=0.1

On rappelle que notre modèle principal est la suivante :

$\Rightarrow$

$log(wage) = a_{city}city + a_{nwifeinc}nwifeinc + a_{educ}educ + a_{exper}exper + a_{kidslt6}kidslt6 + a_{kids}$

Sous H_0 on pose: $a_{nwifeinc} + a_{city} = 0.1$ et $a_{educ} + a_{exper} = 0.1$, il est donc nécessaire de reformuler notre modèle sous l'hypothèse nulle :

\

$\Rightarrow log(wage) = a_{city}city + a_{nwifeinc}nwifeinc + a_{city}nwifeinc - a_{city}nwifeinc + a_{educ}educ +$

\

$\Rightarrow log(wage) = nwifeinc(a_{nwifeinc} + a_{city}) + a_{city}(city - nwifeinc) + exper(a_{exper} + a_{educ})$

\

$\Rightarrow log(wage) - 0.1nwifeinc - 0.1exper = a_{city}(city - nwifeinc) + a_{educ}(educ - exper) +$

\

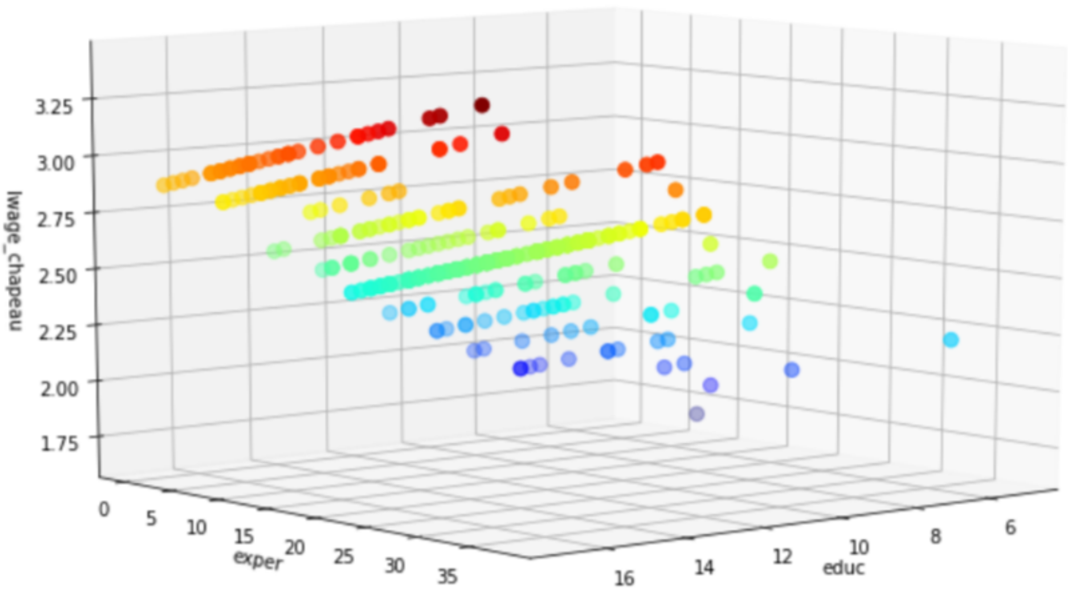
H_1 reste l'hypothése du modéle classique définit ci-dessus.

Nous obtenons :

- la statistique F veut : 0.920515430260816
- la P_Value est : 0.39911574764385327

Nous obetenons une P_Value $\approx$ 0.4 > 0.05, l'hypothèse H_0 est vrai

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



Le nuage de points observé ne forme pas un plan à proprement parlé, donc il n'y a pas de relation linéaire entre les variables wages, duc et expr, cependant on peut quand même constater que plus l'éducation et l'expérience sont importantes, plus le salaire sera élevé

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

L'hypothèse nulle est maintenant de la forme :
$H_0 : a_{kidsgt6} = a_{kidslt6}$. On réécrit le modèle standard

$$\Rightarrow log(wage) = a_{city}city + a_{nwifeinc}nwifeinc + a_{educ}educ + a_{exper}exper + a_{kidsgt6}kidslt6 + a_{kids$$

\

$$\Rightarrow log(wage) = a_{city}city + a_{nwifeinc}nwifeinc + a_{educ}educ + a_{exper}exper + a_{kidsgt6}(kidslt6 + kid$$

Notre résultat :

- la statistique F veut : 0.13786776145599483
- la P_Value est : 0.710597243130811

Nous obetenons une P_Value ≈ 0.7 > 0.05, l'hypothèse H0 est vrai. On en déduit que ce n'est pas l'âge des enfants qui influence un salaire, mais plutôt leur nombre.

14. 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.

Pour rappel hétéroscédasticité dans le contexte d'un modèle de regression, c'est lorsque la variance des residus n'est pas constante.

Afin de determiner si notre modéle est hétéroscédastique, nous allons realiser le "White test", avec un risque de %. Nous procéderons comme suit:

\

- 1) On entraine notre modèle de régression classique, $log(wage) = a_{city}city + a_{nwifeinc}nwifeinc + a_{educ}educ + a_{exper}exper + a_{kidsgt6}kidslt6 +$ et on récupère notre résidu r
- 2) on realise une seconde régression sur les résidus au carrés, sous la forme suivante : $r^2 = b_{city}city + b_{nwifeinc}nwifeinc + b_{educ}educ + b_{exper}exper + b_{kidsgt6}kidslt6 + b_{kidsgt6}kic$
- 3) On pose notre hypothèse nulle : $H_0 : b_{nwifeinc} = b_{educ} = \ldots = b_{kidsgt6} = 0$, Le modèle sous H_0 est donc => $r^2 = b_0 + \epsilon 2$. Si H0 est vrai, il existe une indépendance linéaire entre le carré des résidus et les variables étudiées, et donc l'hypothèse d'hétéroscédasticité (hypothèse nulle) est valide. Dans le cas contraire, il faudra déterminer les coefficient b_ non nuls, qui déterminent la dépendance linéaire des carrés des résidus par rapports à ces variables.

Nous obtenons :

- la statistique F veut : 2.003924882718778
- la P_Value est : 0.06398648165699261

Pour un risque α = 5% notre modèle est bien homoscédastique car le p-valu trouvé est environ égal à 0.06>0.05 (on est vraiment à la limite) mais pour α = 10% notre modèle est considéré comme heteroscdastique.

Pour corriger le problème d'hétéroscédasticité à 10% nous passons par l'estimateur WLS. Nous affichons les resultats obtenus :

p_value : 0.9853584944934253

L'estimateur WLS a rendu le modèle homoscédastique pour α=10%, en effet nous avons une p-value de 0.98.

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

Principe : Vérifier que la régression est la même dans les sous-parties de l'échantillon. Soit parce la relation est non-linéaire (point d'inflexion), soit parce qu'elles correspondent à des sous-populations (ce qui est notre cas) différentes. Nous allons donc realiser le test de chow. Pour ce faire,nous allons procéder comme suit :

- On crée une variable dont la modalité 1 sera attribuée aux femmes ayant 43 ans et plus, 0 sinon.

- On realise une regression non contrainte (notre modèle classique) :
$log(wage) = a_{city}city + a_{nwifeinc}nwifeinc + a_{educ}educ + a_{exper}exper + a_{kidsgt6}kidslt6 +$
. On récupère ensuite sa somme des résidus aux carrées RSS_R.

- Une seconde régression avec uniquement les femmes ayant 43 ans et plus :
$log(wage) = b_{city}city + b_{nwifeinc}nwifeinc + b_{educ}educ + b_{exper}exper + b_{kidsgt6}kidslt6 + b$
. On récupère ensuite sa somme des résidus aux carrées RSS_{+43}.

- Une dernière régression avec uniquement les femmes de moins de 43 ans:
$log(wage) = c_{city}city + c_{nwifeinc}nwifeinc + c_{educ}educ + c_{exper}exper + c_{kidsgt6}kidslt6 + c$
. On récupère ensuite sa somme des résidus aux carrées RSS_{-43}.

- Construction d'un test de Fisher : $F = \frac{\frac{RSS_R - RSS_{+43} - RSS_{-43}}{K}}{\frac{RSS_{+43} + RSS_{-43}}{T-2K}}$ avec K nombre de variable et T le nombre de modalité

Nous Obtenons :

p_value : 0.6856897347649467

Bien que le p-valu semble avoir une valeur complément aberrante (dans l'infiniment petite), il semble y à avoir un problème dans notre code que nous n'avons pas pu régler. Donc H_0 semble être rejeté par notre modèle (sans doute faux)

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

On découpe huseduc selon les différentes composant le quartile, pour obetenir 4 nouvelles variables : col*(3.999, 12.0], col*(12.0, 14.0], col_(14.0, 17.0]

On realise une regression linéaire du modèle :
$log(wage) = a_{city}city + a_{nwifeinc}nwifeinc + a_{educ}educ + a_{exper}exper + a_{kidsgt6}kidslt6 + a_{kids}$

ce modèle sera donc notre hypothèse H_1

L'hypothèse H_0 est donné par :
$H_0 : a_{q1} = a_{q2} = a_{q3} = 0$, Le modèle contraint qui en découle est donc :

$log(wage) = a_{city}city + a_{nwifeinc}nwifeinc + a_{educ}educ + a_{exper}exper + a_{kidsgt6}kidslt6 + a_{kids}$

Nous Obtenons :

- la statistique F veut : 1.684549426984688
- la P_Value est : 0.1696655025985131

Nous obetenons une P_Value ≈ 0.17 > 0.05, l'hypothèse H0 est vraie. On en déduisit le nombre d'années d'études dues marie n'est pas significatif sur le log(salaire) de la femme