Projet d'économétrie

Mohamed Dhaoui, Matthieu Roussel

1. Régression

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

```
rawtable = readtable('mroz.txt', 'TreatAsEmpty', '.');
rawtable.Properties.VariableNames =
{'inlf','hours','kidslt6','kidsge6','age','educ','wage','repwage','hush
rs','husage','huseduc','huswage','faminc','mtr','motheduc','fatheduc','
unem','city','exper','nwifeinc','lwage','expersq'};
work_table = table2array(rawtable(rawtable.wage > 0,:));
disp('number of observations:')
disp(height(work_table))
```

```
number of observations:
   428
```

1.2. 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

```
% conversion en tableau pour plus de commodité
work_table = table2array(rawtable(rawtable.wage > 0,:));
```

- Sur l'ensemble des femmes de l'échantillon

```
% calcul des statistiques pour l'ensemble des 3 variables
results=[min(work_table(:,5:7))' max(work_table(:,5:7))' ...
    mean(work_table(:,5:7))' median(work_table(:,5:7))'  ...
    std(work_table(:,5:7))'];

% calcul des statistiques pour l'ensemble des variables
rowNames = {'age','educ','wage'};
colNames = {'Min','Max','Mean','Median','Std'};
sTable =
array2table(results,'RowNames',rowNames,'VariableNames',colNames)
```

	Min	Max	Mean	Median	Std
age	30	60	41.972	42	7.7211
educ	5	17	12.659	12	2.2854
wage	0.1282	25	4.1777	3.4819	3.3103

- Calcul du salaire médian du mari

```
h_med = median(work_table(:,12))
```

```
6.6831
```

- Statistiques pour les femmes dont le salaire du mari (huswage) est supérieur (ou égal) à 6.6831

```
% pour les femmes dont le salaire  est supérieur au salaire médian du
mari
F1 = (work_table(:,12)>h_med);
sum(F1)
wageF1=work_table(F1,:) ;
resultsF1=[min(wageF1(:,5:7))' max(wageF1(:,5:7))' ...
    mean(wageF1(:,5:7))' median(wageF1(:,5:7))' ...
    std(wageF1(:,5:7))']
sTableF1 =
array2table(resultsF1,'RowNames',rowNames,'VariableNames',colNames)
```

	Min	Max	Mean	Median	Std
age	30	59	42.276	43	7.3888
educ	5	17	13.243	12	2.359
wage	0.1616	25	4.8968	3.8464	4.0416

- Statistiques pour les femmes dont le salaire du mari (huswage) est inférieur à 6.6831

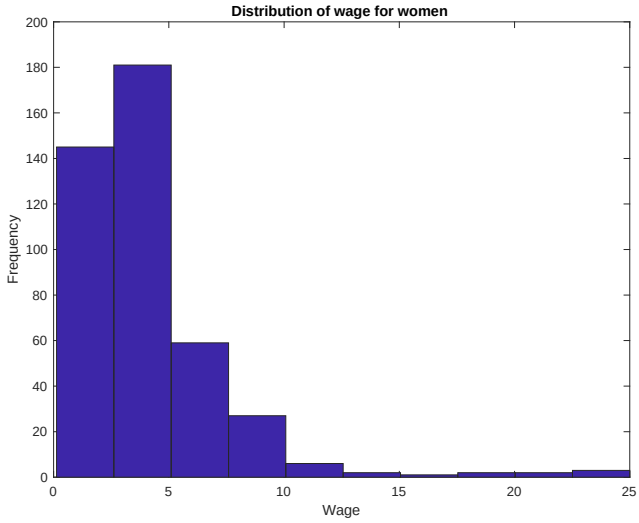
```
% pour les femmes dont le salaire  est inférieur ou égal au salaire
mÈdian du mari
F2 = (work_table(:,12)<=h_med);
sum(F2)
wageF2=work_table(F2,:) ;
resultsF2=[min(wageF2(:,5:7))' max(wageF2(:,5:7))' ...
    mean(wageF2(:,5:7))' median(wageF2(:,5:7))' ...
    std(wageF2(:,5:7))']
sTableF2 =
array2table(resultsF2,'RowNames',rowNames,'VariableNames',colNames)
```

	Min	Max	Mean	Median	Std
age	30	60	41.668	41	8.0455
educ	6	17	12.075	12	2.0542
wage	0.1282	18.267	3.4585	2.9718	2.1433

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

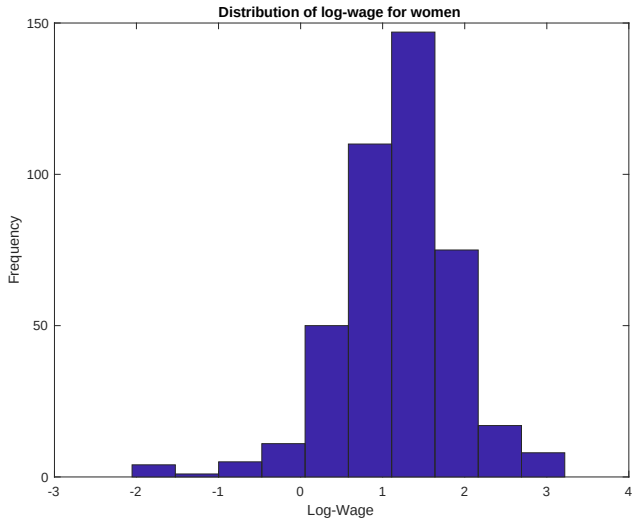
- Wage Histogram

```
% Wage histogram
f1=figure;
hist(work_table(:,7))
title('Distribution of wage for women')
xlabel('Wage')
ylabel('Frequency')
```



- Log-Wage Histogram

```
% log-wage
log_wage=log(work_table(:,7));
f2=figure;
hist(log_wage)
title('Distribution of log-wage for women')
xlabel('Log-Wage')
ylabel('Frequency')
```



Commentaire :
Le passage au logarithme permet de corriger l'asymétrie et ramener la distribution à une distribution plus « normale », cadre nécessaire à l'application des tests.

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

```
cor = corrcoef(work_table(:,15:16))
names = {'motheduc','fatheduc'}
sTableCor = array2table(cor,'RowNames',names,'VariableNames',names)
```

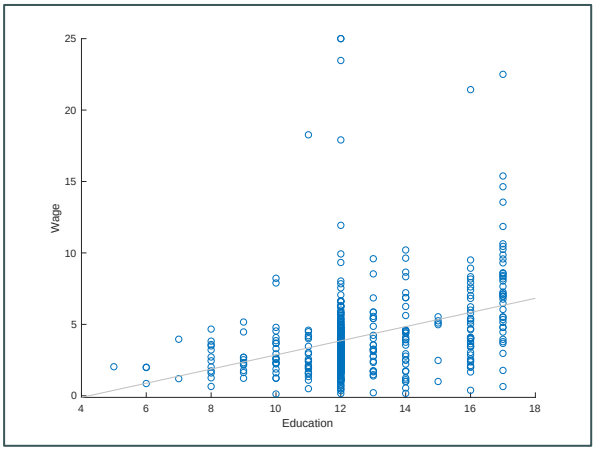
	motheduc	fatheduc
motheduc	1	0.55406
fatheduc	0.55406	1

Commentaire:
Le coefficient de corrélation de 0,55 n'est pas assez élevé pour montrer un réel problème de colinéarité entre les 2 variables.

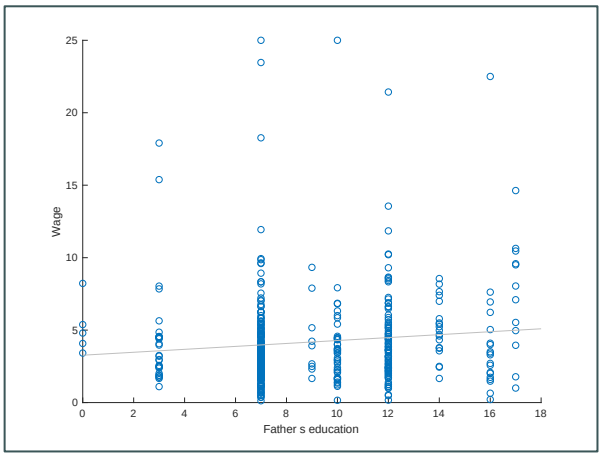
1.5. Faites un graphique en nuage de point entre wage et educ, wage et exper, wage et fatheduc. Commentez. S'agit-il d'un effet "toute chose étant égale par ailleurs ?"

```
% Variables
educ = work_table(:,6)
wage = work_table(:,7)
fatheduc = work_table(:,16)
exper = work_table(:,19)

% 1. Scatterplot between wage & educ
f1=figure;
scatter(wage,educ);
xlabel('Wage');
ylabel('Education');

% 2. Scatterplot between wage & exper
f2=figure;
scatter(wage,exper);
xlabel('Wage');
ylabel('Experience');

% 3. Scatterplot between wage & fatheduc
f3=figure;
scatter(wage,fatheduc);
xlabel('Wage');
```

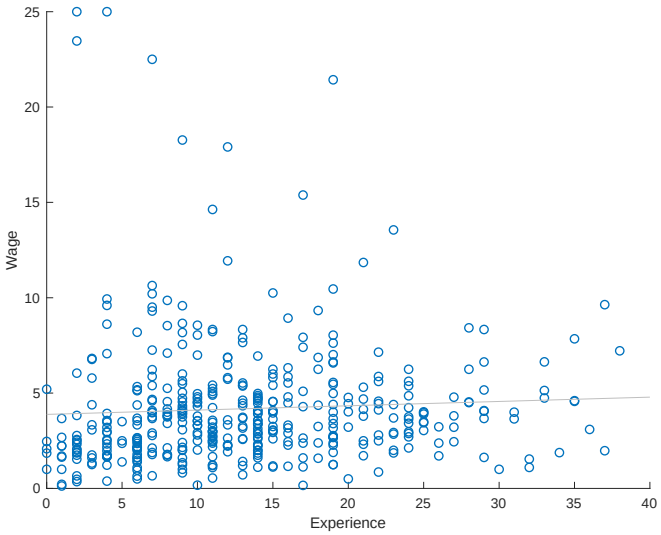
```
ylabel('Father s education');
```



Wage & Education



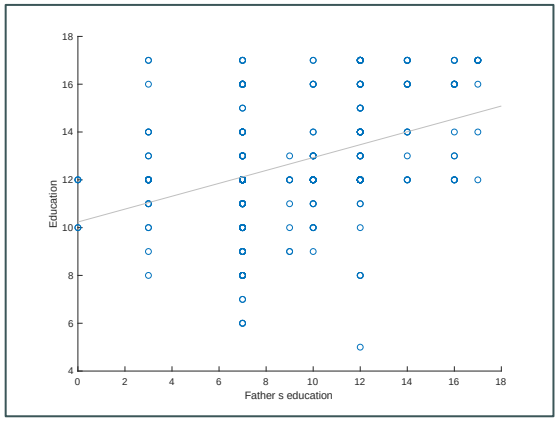
Wage & Father's education



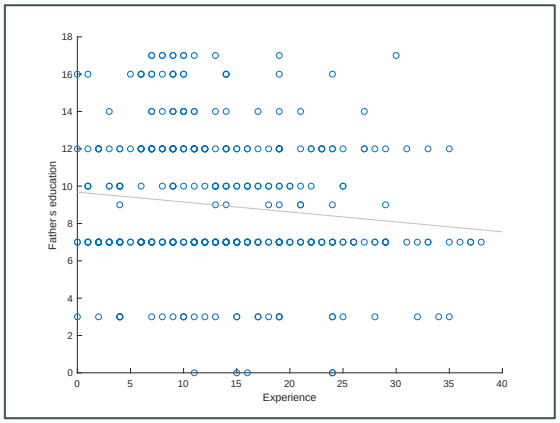
Wage & experience

Commentaires :

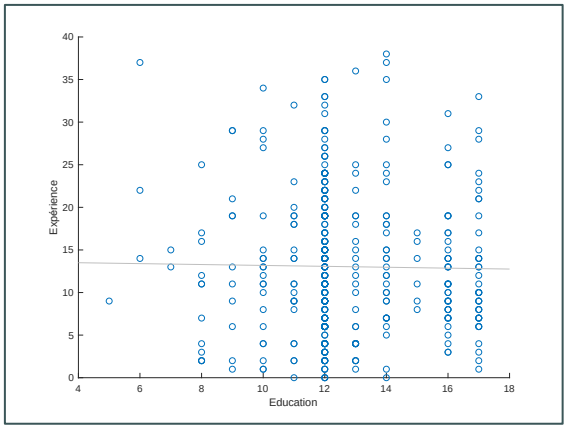
L'analyse visuelle permet d'observer une corrélation clairement positive entre le niveau d'éducation du sujet et son niveau de salaire. Dans une moindre mesure l'expérience entre en jeu, cependant perturbée cependant par quelques outliers avec un salaire élevé pour une faible expérience. Un passage au log pourrait s'avérer opportun pour mieux caractériser la relation. En revanche, le niveau d'éducation du père ne semble pas influer fortement sur le niveau de salaire du sujet. Nous ne pouvons caractériser un effet « toute chose égales par ailleurs » en nous contentant de croiser la variable à expliquer avec les variables explicatives. Par une analyse graphique des nuages de points entre les 3 variables explicatives, nous pouvons observer l'absence de corrélation entre l'expérience et l'éducation. Une corrélation positive (0,41) apparaît entre le niveau d'éducation et celui du père, avec une forte variabilité : le niveau d'éducation du père fixe un seuil minimal à l'éducation du sujet.



Education & Father's Education



Father's education & Experience



Experience & Education

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

Il faut que l'espérance des variables non observées soit nulle, de même que l'espérance de ces variables non observées conditionnellement aux variables explicatives soit nulles, ie qu'il n'y ait pas de multicolinarité entre les variables non observées et les variables explicatives.

Cas typique de biais par variable omise : il existe des variables non observables corrélées avec les variables explicatives et la variable à expliquer. Par exemple, si l'on tente d'expliquer le salaire par le niveau d'éducation, sans tenir compte de l'aptitude initiale du sujet, nous induisons un biais dans le modèle. En effet, l'aptitude initiale du sujet est corrélée au niveau de salaire, mais également au niveau d'éducation. E(x'u) n'est dans ce cas pas nulle.

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

```
% Variables
wage = work_table(:,7);
city = work_table(:,18);
educ = work_table(:,6);
exper = work_table(:,19);
nwifeinc = work_table(:,20);
```

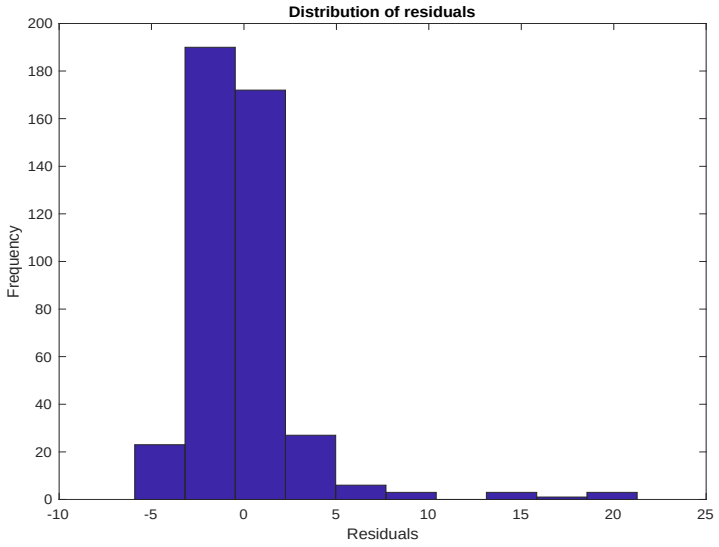
```
kidslt6 = work_table(:,3);
kidsgt6 = work_table(:,4);
% Model
y=wage;
[n,k]=size(wage);
X=[ones(n,1),city,educ,exper,nwifeinc,kidslt6,kidsgt6];
[n,k]=size(X);
% estimation of parameters
beta=inv(X'*X)*X'*y
% résiduals
u=y-X*beta;
% distribution of residuals
f=figure;
hist(u)
title('Distribution of residuals')
xlabel('Residuals')
ylabel('Frequency')
```

	beta	std	t	p_value
cste	-2.4035	0.9635	-2.4945	0.012995
city	0.36975	0.32657	1.1322	0.25818
educ	0.46005	0.070275	6.5464	1.7182e-10
exper	0.02382	0.02088	1.1408	0.2546
nwifeinc	0.015245	0.015499	0.98358	0.32589
kidslt6	0.036173	0.39697	0.091122	0.92744
kidsgt6	-0.061891	0.12538	-0.49361	0.62184

Commentaires :

Seul le coefficient correspondant à l'éducation est significatif. La nullité du coefficient correspondant à l'expérience n'est pas rejetée.

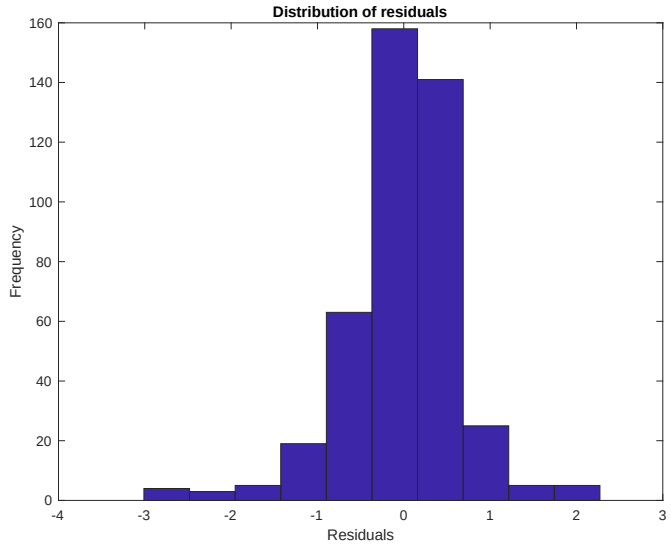
L'histogramme des résidus fait apparaître une asymétrie dans la distribution, due à des outliers. La distribution s'éloigne d'une hypothèse de normalité, cadre nécessaire pour l'application des tests.



1.8. Faire la régression de lwage sur une constante, city, educ, exper, nwifeinc, kidslt6, kidsgt6. Comparer l'histogramme obtenu à celui de la question 7.

```
% Variables
log_wage=log(work_table(:,7));
city = work_table(:,18);
educ = work_table(:,6);
exper = work_table(:,19);
nwifeinc = work_table(:,20);
kidslt6 = work_table(:,3);
kidsgt6 = work_table(:,4);
% Model
y=log_wage;
[n,k]=size(wage);
X=[ones(n,1),city,educ,exper,nwifeinc,kidslt6,kidsgt6];
[n,k]=size(X);
% estimation of parameters
beta=inv(X'*X)*X'*y
% residuals
u=y-X*beta;
% distribution of residuals
f=figure;
hist(u)
title('Distribution of residuals')
xlabel('Residuals')
ylabel('Frequency')
```

	beta	std	t	p_value
cste	-0.39898	0.20705	-1.9269	0.054659
city	0.035268	0.070178	0.50255	0.61555
educ	0.10225	0.015102	6.7706	4.3245e-11
exper	0.015488	0.004487	3.4517	0.00061537
nwifeinc	0.0048827	0.0033307	1.466	0.14341
kidslt6	-0.045303	0.085308	-0.53105	0.59566
kidsgt6	-0.011704	0.026944	-0.43436	0.66425



Commentaires :

Le passage au log permet de prendre en compte la vraie nature de la relation entre l'expérience et le salaire. La probabilité critique associée au coefficient de régression permet dorénavant de rejeter la nullité du coefficient. Dans ce cadre toutefois, seuls les coefficients correspondant à l'expérience et au niveau d'éducation sont significativement différents de 0.
La forme de la densité est plus normale après transformation logarithmique ; nous pouvons appliquer les tests de significativité plus sereinement dans ce cadre.

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

```
% ‡ 10%
C1 = tdis_inv(0.95,n-k);
% ‡ 5%
C2 = tdis_inv(0.975,n-k);
% ‡ 1%
C3 = tdis_inv(0.995,n-k);
seuils=[C1 C2 C3]
rowNames = {'quantile'};
colNames = {'pcts_10','pcts_5','pcts_1'};
sTableS =
array2table(seuils,'RowNames',rowNames,'VariableNames',colNames)
```

	beta	std	t	p_value
cste	-0.39898	0.20705	-1.9269	0.054659
city	0.035268	0.070178	0.50255	0.61555
educ	0.10225	0.015102	6.7706	4.3245e-11
exper	0.015488	0.004487	3.4517	0.00061337
nwifeinc	0.0048827	0.0033307	1.466	0.14341
kidslt6	-0.045303	0.085308	-0.53105	0.59566
kidsgt6	-0.011704	0.026944	-0.43436	0.66425

	pcts_10	pcts_5	pcts_1
quantile	1.6485	1.9656	2.5876

Commentaires :
Le test mené est :

- H0 : exper = 0
- H1 : exper != 0

Sous les hypothèses d'espérance nulle des résidus, de non multi-colinéarité entre les variables explicatives, d'homoscédasticité (non testé encore), la statistique de test t suit

une loi de Student à n-6-1 ddl. Avec une p-value de 0,6%, on rejettera donc H0 au risques alpha= 10%, 5%, ou même 1%

1.10. Tester l'hypothèse que le coefficient associé à educ est égal à 10% avec un seuil de significativité de 5% (test à alternatif des deux côtés)

```
t=(beta(3) - 0.1)/std(3)
p_val = tdis_prb(t,n-k)
results=[beta(3) std(3) t p_val]
rowNames = {'educ'};
colNames = {'beta','std','t','p_value'};
sTable =
array2table(results,'RowNames',rowNames,'VariableNames',colNames)
```

	beta	std	t	p_value
educ	0.10225	0.015102	0.14883	0.88176

Commentaires :

Le test mené est :

- H0 : educ = 0,1
- H1 : educ != 0,1

Avec une p-value de 88%, on ne rejettera pas H0 au risque alpha= 5%

1.11. Tester l'hypothèse jointe que le rendement de l'éducation est de 10% et que celui de l'expérience professionnelle est de 5%.

Le test d'hypothèses jointes à mener est :

- H0 : Beta_educ = 0,1 & Beta_exp = 0,05
- H1 : Beta_educ != 0,1 ou Beta_exp != 0,05

Sous H0, le modèle devient :

Lwage – 0,1 * educ – 0,05 * exp = Beta0 + Beta_city * city + Beta_nwifeinf * nwifeinf + Beta_kidslt6 * kidslt6 + Beta_kidsgt6 * kidsgt6

Nous allons calculer la statistique de de Fisher et la p-valeur associée pour le modèle non contraint et le modèle contraint (H0).

```
% H0 : Beta_educ = 0,1 ; Beta_exp = 0,05
% 1. Unconstrained model
log_wage=log(work_table(:,7));
city = work_table(:,18);
educ = work_table(:,6);
exper = work_table(:,19);
nwifeinc = work_table(:,20);
```

```
kidslt6 = work_table(:,3);
kidsgt6 = work_table(:,4);
% Unrestricted Model
y_ur=log_wage;
[n,k]=size(wage);
X_ur=[ones(n,1),city,educ,exper,nwifeinc,kidslt6,kidsgt6];
[n,k]=size(X_ur);
% estimation of parameters
beta_ur=inv(X_ur'*X_ur)*X_ur'*y_ur;
% residuals and ssr
u_ur=y_ur-X_ur*beta_ur;
ssr_ur = u_ur'*u_ur
% 2. Restricted Model
y_r=log_wage - 0.1.* educ - 0.05.* exper
[n, kb] = size(y_r);
X_r = [ones(n,1),X_ur(:,[2,5,6,7])];
[n, kb] = size(X_r)
q=k-kb
beta_r=inv(X_r'*X_r)*X_r'*y_r
u_r = y_r-X_r*beta_r;
ssr_r = u_r'*u_r;
% Comparer l SCE des 2 modËles (restreint et non restreint)
Ft = ((ssr_r-ssr_ur)/ssr_ur) * (n-k)/q;
p_val = fdis_prb(Ft,q,n-k);
% Mise en forme des resultats
rowNames = {'valeurs'};
colNames = {'F_stat','ddl1','ddl2','p_value'};
sTableS = array2table([Ft,q,n-
k,p_val],'RowNames',rowNames,'VariableNames',colNames)
```

	F_stat	ddl1	ddl2	p_value
valeurs	29.581	2	421	9.5355e-13

Conclusion :

La p-valeur associée à la statistique de Fisher est infiniment petite ; nous pouvons donc rejeter H0 avec une forte confiance.

Une des deux hypothèses est fausse. A priori, compte tenu des résultats du test précédent, vraisemblablement la décision de rejet est portée par l'assomption que le coefficient associé à l'expérience soit égal à 5%.

1.12. De combien augmente le salaire en pourcentage avec 10 années d'expérience ?

En pourcentage, le salaire augmente de 1,5488 * 10, soit 15,49% pour 10 années d'expérience.

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

Le test d'hypothèses jointes à mener est :

- H0 : Beta_kidslt6 = Beta_kidsgt6
- H1 : Beta_kidslt6 != Beta_kidsgt6

Le test est équivalent à :

- H0 : Beta_kidslt6 - Beta_kidsgt6 = 0
- H1 : Beta_kidslt6 - Beta_kidsgt6 != 0

Posons Theta = Beta_kidslt6 - Beta_kidsgt6 ⇔ Beta_kidslt6 = Theta + Beta_kidsgt6

Le modèle devient :

Lwage = Beta0 + Beta_city * city + Beta_educ * educ + Beta_exp * exp + Beta_nwifeinf * nwifeinf + Theta * kidslt6 + Beta_kidsgt6 *(kidslt6 + kidsgt6)

Le test est équivalent à :

- H0 : Theta = 0
- H1 : Theta != 0

Nous nous sommes ramenés à un simple test d'hypothèse de nullité d'un coefficient et allons calculer un simple test de Student sur le coefficient Theta.

```
% Variables
log_wage=log(work_table(:,7));
city = work_table(:,18);
educ = work_table(:,6);
exper = work_table(:,19);
nwifeinc = work_table(:,20);
kidslt6 = work_table(:,3);
kidsgt6 = work_table(:,4);
% Calcul de la somme des variables d'enfants
sum_k = kidslt6 + kidsgt6
% Model
y=log_wage;
[n,k]=size(wage);
X=[ones(n,1),city,educ,exper,nwifeinc,kidslt6,sum_k];
[n,k]=size(X);
% estimation of parameters
beta=inv(X'*X)*X'*y;
% résiduals and stddev
u=y-X*beta;
sig2=u'*u/(n-k);
std=sqrt(diag(sig2*inv(X'*X)));
% t-test
t=(beta(6))/std(6)
p_val = tdis_prb(t,n-k)
results=[beta(6) std(6) t p_val]
rowNames = {'kidslt6'};
colNames = {'beta','std','t','p_value'};
sTable =
array2table(results,'RowNames',rowNames,'VariableNames',colNames)
```

	beta	std	t	p_value
kidslt6	-0.033599	0.090382	-0.37175	0.71027

Conclusion :

Avec une p-valeur de 71%, on ne peut rejeter H0, à savoir l'égalité des coefficients kidslt6 et kidsgt6. Compte tenu de la non significativité de chacun des coefficients, les coefficients seraient conjointement égaux à 0.

1.14. En utilisant le modèle de la question 7, faire le test d'hétéroscédasticité de forme linéaire en donnant la p-valeur. Corriger le problème par rapport à la variable la plus importante en utilisant la méthode des MCG. Comparer les écarts-types des coefficients estimés avec ceux obtenus à la question 7. Commenter

- Test d'hétéroscédasticité

Nous allons tester dans un premier temps l'hypothèse H0 d'homoscédasticité du modèle dans le cadre de la question 7, c'est-à-dire avant passage au logarithme.

```
% Variables
wage = work_table(:,7);
city = work_table(:,18);
educ = work_table(:,6);
exper = work_table(:,19);
nwifeinc = work_table(:,20);
kidslt6 = work_table(:,3);
kidsgt6 = work_table(:,4);
% Test the homoskedasticity
% Model
y=wage;
[n,k]=size(wage);
X=[ones(n,1),city,educ,exper,nwifeinc,kidslt6,kidsgt6];
[n,k]=size(X);
% estimation of parameters
beta=inv(X'*X)*X'*y
% rÈsiduals
u=y-X*beta;
% H0 : _delta1 = _delta1_2 = _ _ _ = __delta1k = 0:
u2 = u.^2;
y = u2;
% Unrestricted model
beta=inv(X'*X)*X'*y
u=y-X*beta;
SSR0 = u'*u
% Restricted model
X = [ones(n, 1)];
```

```
[n, k0] = size(X);
beta=inv(X'*X)*X'*y
u=y-X*beta;
SSR1 = u'*u
% Test
q = k - k0
F = ((SSR1-SSR0)/SSR0) * (n-k)/q
fdis_prb(F,q,n-k)
```

Conclusion :

Le test à mener est le suivant:

- H0 : Var(u|x) = σ^2, le modèle est homoscédastique
- H1 : Var(u|x) n'est pas constante, le modèle est hétéroscédastique

La p-valeur obtenue est de 0.1477, ce qui conduirait à accepter H0. La méthode de correction n'est donc pas nécessaire.

Si l'on prend le modèle logarithmique de la question 8, la p-value est de 0,063. Là encore, l'homoscédasticité n'est pas rejetée à 5%, mais rejetée à 10%. Nous tenterons d'appliquer la correction par la méthode des MCG dans ce cadre uniquement (transformation logarithmique). La p-value devrait par conséquent augmenter, permettant d'accepter H0 avec une plus grande confiance.

Tentative de correction par la méthode des MCG par rapport à la variable la plus importante :

La variable educ est celle qui est la plus significative et qui pèse le plus sur l'explicabilité du salaire. Nous considérerons que l'hétéroscédasticité est de la forme Var (u|x) = σ^2h(x) avec h(x) = educ

Regardons l'impact du test sur la p-valeur.

```
% Variables
wage = work_table(:,7);
city = work_table(:,18);
educ = work_table(:,6);
exper = work_table(:,19);
nwifeinc = work_table(:,20);
kidslt6 = work_table(:,3);
kidsgt6 = work_table(:,4);
% Test the homoskedasticity
% Model
y=log(wage);
[n,k]=size(y);
X=[ones(n,1),city,educ,exper,nwifeinc,kidslt6,kidsgt6];
[n,k]=size(X);
y = y./sqrt(educ);
for i = 1:k
    X(:, i) = X(:, i)./sqrt(educ);
end
beta = inv(X'*X)*X'*y;
u = y - X * beta;
u2 = u.^2;
y = u2;
% Unrestricted model
```

```
beta=inv(X'*X)*X'*y;
u=y-X*beta;
SSR0 = u'*u;
% Restricted model
X = [ones(n, 1)./sqrt(educ)];
[n, k0] = size(X);
beta = inv(X'*X)*X'*y;
u = y - X * beta;
SSR1 = u'* u;
k
k0
q = k-k0;
F = ((SSR1-SSR0)/SSR0) * ((n-k)/q)
fdis_prb(F,q,n-k)
```

Le test amène toujours à accepter H0 avec une probabilité ctritique quasi-inchangée de 0,0642. La nature de l'éventuelle hétéroscédasticité n'a pas été capturée par notre choix de h(x).

.

1.15. Tester le changement de structure de la question 8 entre les femmes qui ont moins de 30 ans, entre 30 et 43 ans, plus de 43 ans (3 groupes mutuellement exclusifs). Donnez les pvaleurs.

Remarque : aucune femme n'a moins de 30 ans dans le jeu de données. Nous considérerons donc le cas où l'âge est inférieur ou égal à 30

Le test mené :

- H0 : similarité entre les groupes, ie égalité des coefficients de régression entre les 3 groupes
- H1 : H0 est fausse

Nous utiliserons le test de Chow pour 3 groupes, dont la statistique est la suivante :

$$F = 1 + \frac{[SSR - (SSR1 + SSR2 + SSR3)]}{SSR1 + SSR2 + SSR3} . \frac{[n - 3\,(k+1]}{k+1}$$

La statistique suit sous H0 une loi de Fisher à (6+1) et (n – 3 (6+1)) ddl

Le programme :

```
% Variables
log_wage=log(work_table(:,7));
city = work_table(:,18);
educ = work_table(:,6);
exper = work_table(:,19);
nwifeinc = work_table(:,20);
kidslt6 = work_table(:,3);
kidsgt6 = work_table(:,4);
% Calculate SSR for original model
y=log_wage;
[n0,k]=size(log_wage);
X=[ones(n0,1),city,educ,exper,nwifeinc,kidslt6,kidsgt6];
[n0,k]=size(X);
```

```
beta=inv(X'*X)*X'*y;
u=y-X*beta;
SSR0 = u'*u;
% Select women <= 30
work_table_sel = table2array(rawtable(rawtable.wage > 0 & rawtable.age
<=30,:));
% Variables
log_wage=log(work_table_sel(:,7));
city = work_table_sel(:,18);
educ = work_table_sel(:,6);
exper = work_table_sel(:,19);
nwifeinc = work_table_sel(:,20);
kidslt6 = work_table_sel(:,3);
kidsgt6 = work_table_sel(:,4);
% Calculate SSR for women aged <= 30
y=log_wage;
[n,k]=size(log_wage);
X=[ones(n,1),city,educ,exper,nwifeinc,kidslt6,kidsgt6];
[n,k]=size(X);
beta=inv(X'*X)*X'*y;
u=y-X*beta;
SSR_1 = u'*u
% Select women aged between 30 and 43
work_table_sel = table2array(rawtable(rawtable.wage > 0 & rawtable.age
> 30 & rawtable.age <= 43,:));
% Variables
log_wage=log(work_table_sel(:,7));
city = work_table_sel(:,18);
educ = work_table_sel(:,6);
exper = work_table_sel(:,19);
nwifeinc = work_table_sel(:,20);
kidslt6 = work_table_sel(:,3);
kidsgt6 = work_table_sel(:,4);
% Calculate SSR for women aged between 30 and 43
y=log_wage;
[n,k]=size(log_wage);
X=[ones(n,1),city,educ,exper,nwifeinc,kidslt6,kidsgt6];
[n,k]=size(X);
beta=inv(X'*X)*X'*y;
u=y-X*beta;
SSR_2 = u'*u
% Select women aged > 43
work_table_sel = table2array(rawtable(rawtable.wage > 0 & rawtable.age
> 43,:));
% Variables
log_wage=log(work_table_sel(:,7));
city = work_table_sel(:,18);
educ = work_table_sel(:,6);
exper = work_table_sel(:,19);
nwifeinc = work_table_sel(:,20);
kidslt6 = work_table_sel(:,3);
kidsgt6 = work_table_sel(:,4);
% Calculate SSR for women aged > 43
y=log_wage;
[n,k]=size(log_wage);
X=[ones(n,1),city,educ,exper,nwifeinc,kidslt6,kidsgt6];
[n,k]=size(X);
beta=inv(X'*X)*X'*y;
u=y-X*beta;
SSR_3 = u'*u
% Calculate F statistic
SSR_sum = SSR_1 + SSR_2 + SSR_3
```

```
F = ((SSR0 - SSR_sum) / SSR_sum) * ((n0-3*k)/k)
fdis_prb(F,k,n0-3*k)
```

La valeur obtenue pour la statistique est 1.1832 et la probabilité critique associée est de 0.3111. Avec un seuil de 5%, on ne rejettera pas la similarité entre les sous-populations par tranche d'âge.

1.16. A partir de la variable kidslt6, créer un ensemble de variables binaires pour le nombre d'enfants de moins de 6 ans. Refaire la question 8 avec ces variables et en utilisant comme référence les femmes qui ont des enfants de plus de 6 ans. Ces catégories sont-elles mutuellement exclusives ? Interprétez les paramètres associés aux variables binaires. Faire le test de non significativité de l'ensemble des variables binaires. Donnez les p-valeurs

Nous ne sommes pas certains d'avoir compris la question, ne comprenant pas la relation entre l'encodage en variables binaires de la variable kidslt6 et le fait de prendre comme référence une autre variable (enfant de plus de 6 ans) qui elle n'est pas binarisée…Voici donc le parti pris pour cette question : la variable kidslt6 sera binarisée et la modalité de référence sera l'absence d'enfant de moins de 6 ans.

La variable kidslt6 comprend trois modalités : 0 enfant, 1 enfant ou 2 enfants.
La référence sera de 0 enfants, permettant de mesurer l'influence du nombre d'enfant à partir de chaque coefficient associé à chacune des 2 variables binaires (1 ou 2 enfants).
Les variables sont mutuellement exclusives. Aucune interaction n'est donc à évaluer.

```
% Variables
log_wage=log(work_table(:,7));
city = work_table(:,18);
educ = work_table(:,6);
exper = work_table(:,19);
nwifeinc = work_table(:,20);
kidslt6 = work_table(:,3);
kidsgt6 = work_table(:,4);
y=log_wage;
[n,k]=size(y);
%hist(kidslt6)
% Creation of variables with dummy variables for kidslt6
kidslt6_1 = zeros(n,1);
kidslt6_1(kidslt6==1)=1;
kidslt6_2 = zeros(n,1);
kidslt6_2(kidslt6==2)=1;
% model
X=[ones(n,1),city,educ,exper,nwifeinc,kidslt6_1,kidslt6_2,kidsgt6];
[n,k]=size(X);
% estimation of parameters
beta=inv(X'*X)*X'*y
% ecarts-types
u=y-X*beta;
```

```
sig2=u'*u/(n-k);
std=sqrt(diag(sig2*inv(X'*X)));
t=beta./std
p_val = tdis_prb(t,n-k)
results=[beta std t p_val]
rowNames =
{'cste','city','educ','exper','nwifeinc','kidslt6_1','kidslt6_2','kidsg
t6'};
colNames = {'beta','std','t','p_value'};
sTable =
array2table(results,'RowNames',rowNames,'VariableNames',colNames)
```

	beta	std	t	p_value
cste	-0.39755	0.20757	-1.9152	0.056143
city	0.034609	0.070438	0.49134	0.62344
educ	0.10222	0.015121	6.7601	4.6277e-11
exper	0.015437	0.0045089	3.4237	0.00067853
nwifeinc	0.0049033	0.0033383	1.4688	0.14264
kidslt6_1	-0.05403	0.10807	-0.49994	0.61738
kidslt6_2	-0.065028	0.25855	-0.25151	0.80154
kidsgt6	-0.011594	0.026989	-0.42958	0.66772

Commentaire :

- Pour 1 enfant de moins de 6 ans, le salaire horaire baisserait de 5,4% ;
- Pour 2 enfants de moins de 6 ans, le salaire horaire baisserait de 6,5% ;
- Attention cependant, chaque coefficient associé n'est pas significativement différent de 0 donc l'interprétation ci-dessus ne tient pas.

Nous allons vérifier si l'ensemble des coefficients associés à ces variables sont conjointement nuls.

- H0 : kidslt6_1 = 0 et kidslt6_2 = 0
- H1 : H0 fausse

Sous H0, le modèle devient :

Lwage = Beta0 + Beta_city * city + Beta_educ * educ + Beta_exp * exp + Beta_nwifeinf * nwifeinf + Beta_kidsgt6 * kidsgt6

```
% 1. Unrestricted Model
ssr_ur = u'*u
% 2. Restricted Model
X_r = [ones(n,1),city,educ,exper,nwifeinc,kidsgt6];
[n, k_r] = size(X_r)
q=k-k_r
beta_r=inv(X_r'*X_r)*X_r'*y
u_r = y-X_r*beta_r;
ssr_r = u_r'*u_r;
% Comparer l SCE des 2 modÀles (restreint et non restreint)
Ft = ((ssr_r-ssr_ur)/ssr_ur) * (n-k)/q;
p_val = fdis_prb(Ft,q,n-k);
% Mise en forme des resultats
rowNames = {'valeurs'};
colNames = {'F_stat','ddl1','ddl2','p_value'};
```

```
sTableS = array2table([Ft,q,n-
k,p_val],'RowNames',rowNames,'VariableNames',colNames)
```

	F_stat	ddl1	ddl2	p_value
valeurs	0.14936	2	420	0.8613

On acceptera donc H0 acceptant la nullité jointe des coefficients.

1.17. A partir de l'échantillon global, faire une régression de inf sur une constante, city, educ, age, kidslt6, kidsgt6. Interprétez les coefficients estimés.

Nous réintégrons ici au périmètre de l'analyse l'ensemble des observations du jeu de données, c'est-à-dire les femmes sur le marché de l'emploi et les femme hors marché de m'emploi.

```
% Variables
work_table2 = table2array(rawtable);
infl=work_table2(:,1);
city = work_table2(:,18);
educ = work_table2(:,6);
age = work_table2(:,5);
kidslt6 = work_table2(:,3);
kidsgt6 = work_table2(:,4);
% Model
y=infl;
[n,k]=size(y);
X=[ones(n, 1),city,educ,age,kidslt6,kidsgt6];
[n,k]=size(X);
reso = ols(y,X)
stats = reso.tstat;
beta = reso.beta;
p_val = tdis_prb(stats,n-k)
results=[beta stats p_val]
rowNames = {'cste','city','educ','age','kidslt6','kidsgt6'};
colNames = {'beta','t','p_val'};
sTable =
array2table(results,'RowNames',rowNames,'VariableNames',colNames)
```

	beta	t	p_value
cste	0.70758	4.3648	1.4518e-05
city	-0.034085	-0.94383	0.34556
educ	0.04341	5.6558	2.2095e-08
age	-0.013026	-5.0809	4.7496e-07
kidslt6	-0.30747	-8.498	0
kidsgt6	-0.017341	-1.2306	0.21887

La construction d'un simple modèle de régression avec comme variable cible une variable binaire permet d'estimer une pseudo-probabilité d'appartenance à chaque classe ; cependant, rien ne garantit que les valeurs prédites de y soient effectivement des probabilités comprises entre 0 et 1.

L'analyse des coefficients permet de faire apparaître certains facteurs explicatifs quant à la participation des femmes au marché du travail :

- Facteurs jouant négativement :
  - Le nombre d'enfants de moins de 6 ans : lorsque celui-ci augmente d'un, la probabilité de participation au marché du travail diminue de 0,0307
  - Dans une moindre mesure : l'âge (-0,013),
  - Le nombre d'enfant de plus de 6 ans considéré seul, sans interaction
- Facteurs jouant positivement sur l' « activité » : le niveau d'éducation. Lorsque celui-ci augmente d'un an, la probabilité de participation au marché du travail augmente de 0,0434

Il convient d'intégrer ces facteurs à une modélisation plus appropriée (logit/probit) permettant d'affecter des probabilités « bornées » d'appartenance à chaque classe (femme professionnellement active ou inactive) pour éventuellement de réaliser des prédictions appropriées.

1.18. Estimer le modèle probit de l'inf sur une constante, city, educ, age, kidslt6, kidsgt6. Faire le test de non significativité jointes des coefficients associés à kidslt6 et à kidsgt6. Comparez le résultat du test à celui de la question 18.

- Modèle probit

```
% Variables
work_table2 = table2array(rawtable);
infl=work_table2(:,1);
city = work_table2(:,18);
educ = work_table2(:,6);
age = work_table2(:,5);
kidslt6 = work_table2(:,3);
kidsgt6 = work_table2(:,4);
% Model
y=infl;
[n,k]=size(y);
X=[ones(n, 1),city,educ,age,kidslt6,kidsgt6];
[n,k]=size(X);
resp = probit(y, X);
xb = X * resp.beta;
for i = 1:k
    ep(i) = mean(norm_pdf(xb) * resp.beta(i));
end
stats = resp.tstat
beta = resp.beta
p_val = tdis_prb(stats,n-k)
results=[beta stats p_val ep']
rowNames = {'cste','city','educ','age','kidslt6','kidsgt6'};
colNames = {'beta','t','p_val','ep'};
sTable =
array2table(results,'RowNames',rowNames,'VariableNames',colNames)
```

	beta	t	p_val	ep
cste	0.60504	1.2967	0.19512	0.21314
city	-0.086317	-0.84214	0.39998	-0.030407
educ	0.1234	5.469	6.1757e-08	0.043471
age	-0.03754	-5.0083	6.8587e-07	-0.013225
kidslt6	-0.88464	-7.8825	1.1324e-14	-0.31164
kidsgt6	-0.054234	-1.3507	0.1772	-0.019105

- Test joint de significativité des coefficients associés aux variables kidslt6, kidsgt6

  - *H0 : Beta_ kidslt6 = 0 & Beta_ kidsgt6 = 0*
  - *H1 : Beta_ kidslt6 != 0 ou Beta_ kidsgt6 != 0*

Comme vu en cours, dans ce contexte, le test à utiliser est celui du rapport de vraisemblance :

```
LU = resp.lik;
X=[ones(n, 1),city,educ,age];
resp2 = probit(y, X);
LR = resp2.lik;
RATIO = 2 * (LU - LR)
p_value = 1 - chi2cdf(RATIO,2)
```

On obtient les valeurs suivantes :

RATIO = 69.8366
P-value = 6.6613e-16

Commentaires :

- Les résultats entre les deux régressions convergent en termes de significativité des coefficients.

- Le test joint de significativité sur les coefficients liés au nombre d'enfants rejette la nullité jointe des deux coefficients.

- « Comparez le résultat du test à celui de la question 18 » => nous sommes sur la question 18, nous ne comprenons pas la question.

1.19. Calculer les effets partiels pour l'ensemble des variables explicatives. Comparer vos résultats à ceux obtenus à la question 17. Commentez.

Les effets partiels calculés précédemment (colonne ep, question 18) montrent l'influence de chaque variable sur le calcul de la probabilité. Les valeurs associées sont très proches de celles des coefficients de la régression linéaire initiale (Q17) sur les données.

Le nombre d'enfants de moins de 6 ans pénalise fortement la participation des femmes au marché de l'emploi. Au-delà de cette variable écrasant toutes les autres, le niveau d'éducation joue positivement (+4% par année). Plus les sujets sont âgés, plus la probabilité de travailler diminue (-4% par année).

1.20. Faire le test de non significativité jointes des coefficients associés à kidslt6 et à kidsgt6 en utilisant la méthode du rapport de vraisemblance. Comparez aux résultats de la question 18.

Test déjà mené en Q18.