Valentin Barthel

5 avril 2024

Table des matières

1 Question 1 : Importer le fichier Panel95light.csv pour en faire une table SAS

```
/*-----------------------------------QUESTION 1-----------------------------------*/

    /*Importons le fichier Panel95light.csv pour en faire une table SAS*/
PROC IMPORT DATAFILE='/home/u63824485/sasuser.v94/Panel95light.csv'
            OUT=PanelLight
            DBMS=csv
            REPLACE;
    GETNAMES=YES;
    DELIMITER=',';
    GUESSINGROWS=max;
RUN;
```

2 Question 2 : Représenter graphiquement et commenter la distribution des variables

```
/*-----------------------------------QUESTION 2-----------------------------------*/

/*Donnons et commentons la distribution des variables lw (logarithme du salaire horaire), e
potentielle, en mois), mois (nombre de mois depuis le debut du panel, en janvier 95 */
        /*Statistiques descriptives des moyennes et frequences*/
proc means data=PanelLight n mean min max std;
    var lw exper mois;
run;
proc freq data=PanelLight;
    tables etudes sexe;
run;
        /*Representations graphiques des distributions*/
/*lw*/
title "Distribution du Salaire Horaire";
proc sgplot data=PanelLight;
    histogram lw / binwidth=0.1;
```

```
    xaxis label="Logarithme du Salaire Horaire";
    yaxis label="Frequence";
run;
/*etudes*/
proc sql;
    create table etudes_percent as
    select etudes,
           count(*) as Frequency,
           (count(*) / (select count(*) from PanelLight))*100 as Percent
    from PanelLight
    group by etudes;
quit;
title "Nombre d'individus par niveau d'etude";
proc sgplot data=etudes_percent noautolegend;
    vbar etudes / response=Percent group=etudes datalabel;
    xaxis label="Niveau d'etudes";
    yaxis label="Nombre d'Observations";
run;
/*Diag. circulaire fréquence des sexes*/
title "Frequence des sexes";
proc gchart data=PanelLight;
        PIE sexe /percent= inside;
run;

/*exper*/
title "Distribution de l'experience";
proc sgplot data=PanelLight;
    histogram exper / binwidth=0.1;
    xaxis label="Nombre de mois d'experience";
    yaxis label="Frequence";
run;

title "Experience par niveaux d'etude";
proc sgplot data=PanelLight;
    vbox exper / category=etudes;
    xaxis label="Niveaux d'etudes";
    yaxis label="Experience (en mois)";
run;
```

```sql
proc sql;
    create table sexe_percent as
    select sexe,
           count(*) as Frequency,
           (count(*) / (select count(*) from PanelLight))*100 as Percent
    from PanelLight
    group by sexe;
quit;
title "Frequence des sexes";
proc sgplot data=sexe_percent noautolegend;
    styleattrs datacolors=(pink blue);
    vbar sexe / response=Percent group=sexe datalabel;
    xaxis label="Sexe";
    yaxis label="Pourcentage (%)";
run;
```

La procédure MEANS

Variable	N	Moyenne	Minimum	Maximum	Ec-type
lw	8856	3.8887611	1.7649612	5.5367284	0.4674402
exper	8124	193.4843673	0	565.0000000	123.6093377
mois	10548	18.5000000	13.0000000	24.0000000	3.4522162

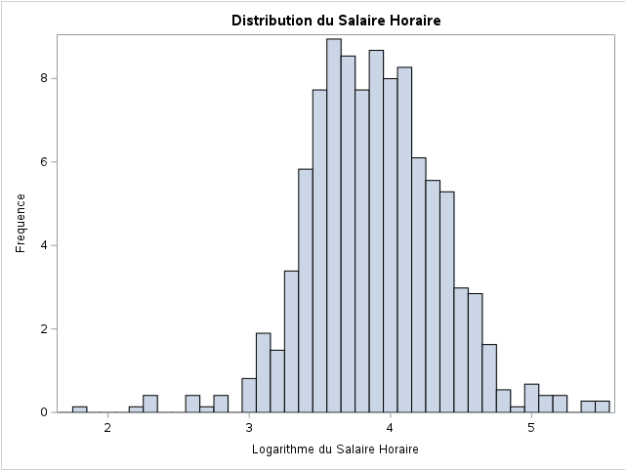
La procédure FREQ

etudes	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
deuxieme cycle	2460	23.32	2460	23.32
primaire	1320	12.51	3780	35.84
professionnel court	3288	31.17	7068	67.01
professionnel long	804	7.62	7872	74.63
secondaire	1656	15.70	9528	90.33
troisieme cycle	1020	9.67	10548	100.00

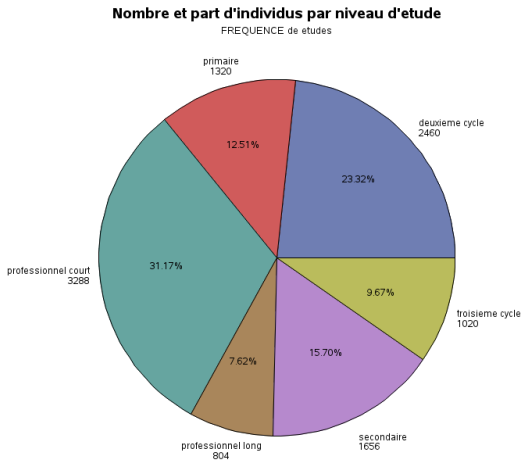
sexe	Fréquence	Pourcentage	Fréquence cumulée	Pourcentage cumulé
Femme	6408	60.75	6408	60.75
Homme	4140	39.25	10548	100.00

Ces tables nous présente la description des variables de notre jeu de données, contenant "lw", le logarithme du salaire horaire, "exper" déterminant le nombre de mois d'expérience, "mois" le nombre de mois depuis janvier 95 et également "etudes" et "sexe", présentant respectivement les différents niveau d'études et le sexe avec leur nombre d'individu par agrégat et la part qu'il occupe parmi l'ensemble des agrégats. Nous pouvons voir qu'au total les individus seront observés un an après 1995 sur une période d'une

année. S'agissant de données de panel, nous devrons prendre en compte dans notre analyse le fait qu'un individu soit observé douze fois, et par conséquent, nous avons un total de 10548 observations et 879 individus observés. Nous commenterons plus en détails ces statistiques descriptives avec la distribution des variables sous les figures suivantes.

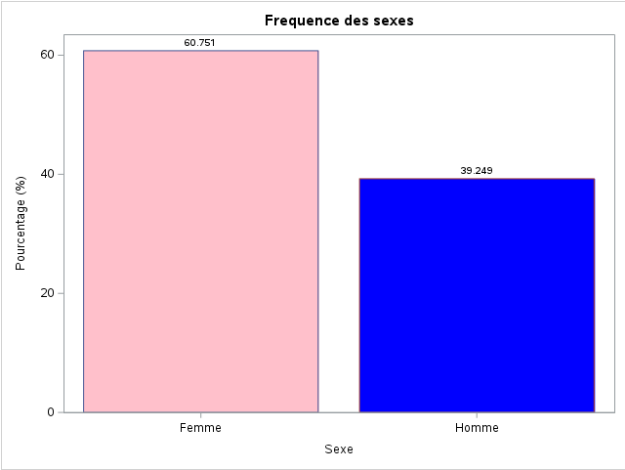


La distribution du log-salaire horaire semble suivre une loi normale légèrement asymétrique à droite avec une moyenne de 3,88 et un écart-type de 0,46. Nous observons des valeurs aberrantes pour des valeurs inférieures à 2 et supérieures à 5. La dispersion témoigne d'une variabilité des salaires au sein de la population observé, pouvant s'expliquer par un niveau d'expérience et d'étude différent ou encore le sexe de individu observé.

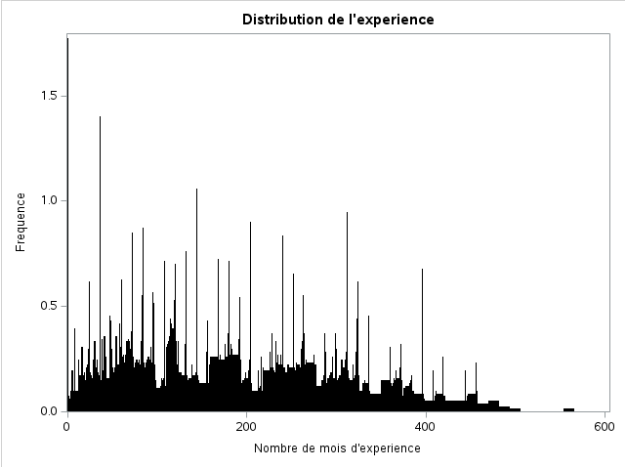


Ce diagramme en barre nous présente le nombre d'individus observé par niveau d'étude. Le niveau d'étude le plus représenté est le niveau "professionnel court" avec près de 3300 observations (soit 275 individus), suivi du niveau "deuxième cycle" atteignant quasiment les 2500 observations. Les individus issus de ces niveaux représentent la moitié de la population, l'autre moitié de la population se partageront

le reste des niveaux d'étude avec le niveau "professionnel court" et le "troisième cycle" atteignant respectivement 804 et 1020 observations, soit moins de 10% chacun, montrant une participation plus faible à ces niveaux d'éducation dans l'échantillon. Un tiers des individus présenteront les niveaux "primaire" et "secondaire" avec respectivement 1320 et 1656 observations.



Ce diagramme en barres présente la fréquence des sexes dans la population observée. Bien que la différence de fréquence des sexes ne soit pas importante, la population est composée majoritairement de femmes, soit à hauteur de 60%.



Cet histogramme nous présente la fréquence du nombre de mois d'expérience depuis 1995. Le nombre de moi d'expérience à légèrement tendance à se raréfier. En prenant en compte la dimension temporelle, il n'est pas étonnant d'observer que le nombre de mois d'expérience est quasiment identique sur l'ensemble de la période, s'expliquant par le fait que les individus prennent en expérience au fil des mois et cela est inclus dans nos observations. Par ailleurs, cela explique l'aspect périodique que l'on peut observer.

3 Question 3 : Générer des indicatrices pour chaque niveau d'études, ainsi que la variable numérique sexeN, qui vaut 1 pour les hommes, 2 pour les femmes, puis des indicatrices de sexe masculin et féminin

```
/*-----------------------------------QUESTION 3-----------------------------------*/

/*Generons une indicatrice pour chaque niveaux d'etude et pour les sexes*/
data PanelLight;
    set PanelLight;
    if etudes = 'professionnel court' then professionnel_court = 1;
    else professionnel_court = 0;
    if etudes = 'professionnel long' then professionnel_long = 1;
    else professionnel_long = 0;
    if etudes = 'primaire' then primaire = 1; else primaire = 0;
    if etudes = 'secondaire' then secondaire = 1; else secondaire = 0;
    if etudes = 'deuxieme cycle' then deuxieme_cycle = 1; else deuxieme_cycle = 0;
    if etudes = 'troisieme cycle' then troisieme_cycle = 1; else troisieme_cycle = 0;
    if sexe = 'Homme' then sexeN = 1; else sexeN = 2;
    if sexe = 'Homme' then sexeHomme = 1; else sexeHomme = 0;
    if sexe = 'Femme' then sexeFemme = 1; else sexeFemme = 0;
run;
```

4 Question 4 : Calculer le log-salaire moyen par niveau d'étude, ainsi que le log-salaire moyen dans l'échantillon

```
/*-----------------------------------QUESTION 4-----------------------------------*/

/* Calculer le log-salaire moyen de l'echantillon et le log-salaire par niveaux d'etude */

title "Moyenne du log-salaire global et par niveaux d'etude";
proc means data=PanelLight noprint;
    class etudes;
    var lw;
    output out=mean_salary(drop=_TYPE_ _FREQ_) mean(lw)=mean_lw;
```

```
run;
```

Figure 1. Moyenne du log-salaire global et par niveau d'étude

Obs.	etudes	mean_lw
1		3.8887611432
2	deuxieme cycle	4.0704466895
3	primaire	3.6030209247
4	professionnel court	3.7261213225
5	professionnel long	3.8576808902
6	secondaire	3.8352762183
7	troisieme cycle	4.2878874244

5 Question 5 :Estimer les différents modèles de régression du log du salaire suivants, en ajoutant à la table initiale le salaire prédit par chaque modèle :

5.1 Question 5a : tous les niveaux d'étude, "sans précaution"

```
        /* 5a. Tous les niveaux d'etude, "sans precaution"*/
/*Remarque:->Colinearite*/
proc reg data=PanelLight;
    model lw = professionnel_court professionnel_long primaire secondaire deuxieme_cycle t
    output out=PanelLight_Predicted p=predicted_lw;
run;
```

Figure 2. Modèle 5a

Nb d'observations lues	10548
Nb d'obs. utilisées	8856
Nombre d'observations avec valeurs manquantes	1692

Analyse de variance

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	5	375.12300	75.02460	425.70	<.0001
Erreur	8850	1559.69775	0.17624		
Total sommes corrigées	8855	1934.82075			

Root MSE	0.41981	R carré	0.1939
Moyenne dépendante	3.88876	R car. ajust.	0.1934
Coeff Var	10.79536		

Note: Model is not full rank. Least-squares solutions for the parameters are not unique. Some statistics will be misleading. A reported DF of 0 or B means that the estimate is biased.

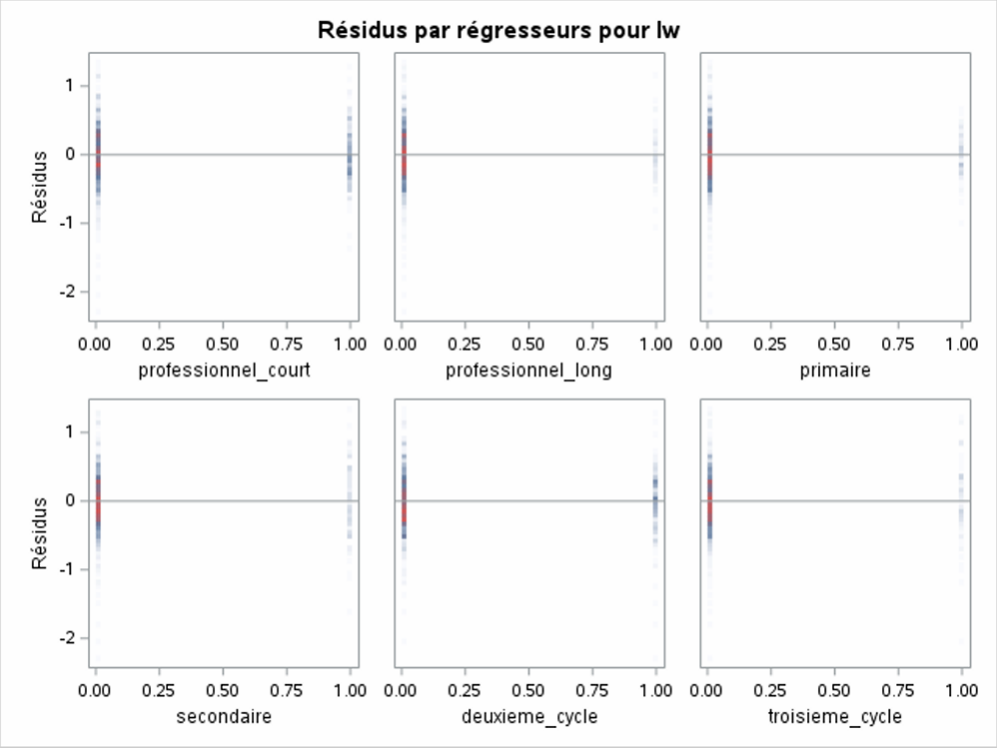
Note: The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

troisieme_cycle =	Intercept - professionnel_court - professionnel_long - primaire - secondaire - deuxieme_cycle

Paramètres estimés

Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	B	4.28789	0.01372	312.49	<.0001
professionnel_court	B	-0.56177	0.01595	-35.22	<.0001
professionnel_long	B	-0.43021	0.02071	-20.77	<.0001
primaire	B	-0.68487	0.01947	-35.18	<.0001
secondaire	B	-0.45261	0.01797	-25.18	<.0001
deuxieme_cycle	B	-0.21744	0.01628	-13.35	<.0001
troisieme_cycle	0	0	.	.	.

Figure 3. Résidus par régresseurs Modèle 5a



5.2 Question 5b : tous les niveaux d'étude, sans constante

```
        /* 5b. Tous les niveaux d'etude, sans constante*/
proc reg data=PanelLight;
    model lw = professionnel_court professionnel_long primaire secondaire deuxieme_cycle t
    output out=PanelLight_Predicted p=predicted_lw;
run;
```

Figure 4. Modèle 5b

Nb d'observations lues	10548
Nb d'obs. utilisées	8856
Nombre d'observations avec valeurs manquantes	1692

Note: No intercept in model. R-Square is redefined.

Analyse de variance

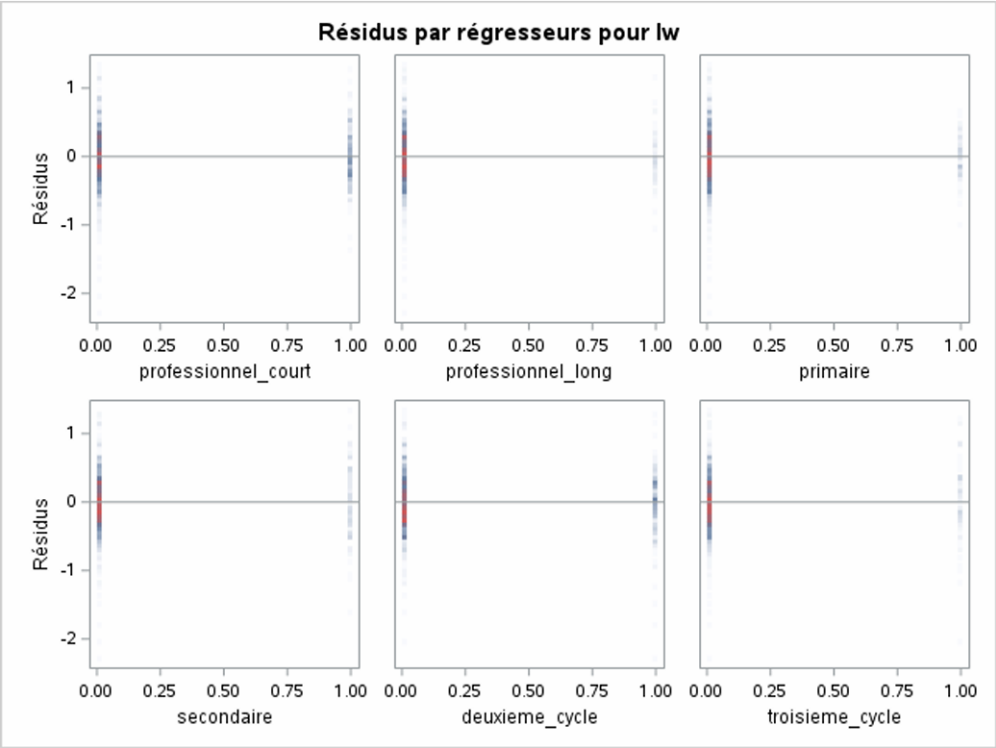
Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	6	134300	22383	127007	<.0001
Erreur	8850	1559.69775	0.17624		
Total sommes non corrigées	8856	135859			

Root MSE	0.41981	R carré	0.9885
Moyenne dépendante	3.88876	R car. ajust.	0.9885
Coeff Var	10.79536		

Paramètres estimés

Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
professionnel_court	1	3.72612	0.00813	458.12	<.0001
professionnel_long	1	3.85768	0.01552	248.62	<.0001
primaire	1	3.60302	0.01381	260.89	<.0001
secondaire	1	3.83528	0.01161	330.41	<.0001
deuxieme_cycle	1	4.07045	0.00877	464.20	<.0001
troisieme_cycle	1	4.28789	0.01372	312.49	<.0001

Figure 5. Résidus par régresseurs Modèle 5b



5.3 Question 5c : tous les niveaux d'étude sauf primaire (référence 1)

```
        /* 5c. Tous les niveaux d'etude sauf primaire (reference 1)*/
proc reg data=PanelLight;
    model lw = professionnel_court professionnel_long secondaire deuxieme_cycle troisieme_c
    output out=PanelLight_Predicted p=predicted_lw;
run;
```

Figure 6. Modèle 5c

Nb d'observations lues	10548
Nb d'obs. utilisées	8856
Nombre d'observations avec valeurs manquantes	1692

Analyse de variance

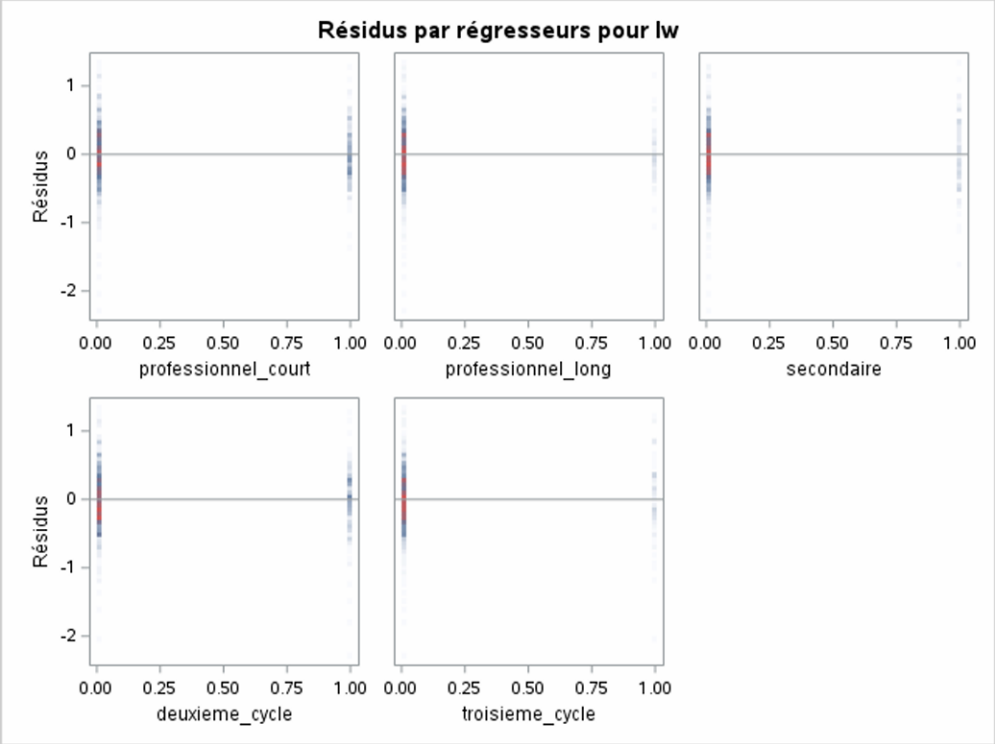
Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	5	375.12300	75.02460	425.70	<.0001
Erreur	8850	1559.69775	0.17624		
Total sommes corrigées	8855	1934.82075			

Root MSE	0.41981	R carré	0.1939
Moyenne dépendante	3.88876	R car. ajust.	0.1934
Coeff Var	10.79536		

Paramètres estimés

Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	1	3.60302	0.01381	260.89	<.0001
professionnel_court	1	0.12310	0.01603	7.68	<.0001
professionnel_long	1	0.25466	0.02077	12.26	<.0001
secondaire	1	0.23226	0.01804	12.87	<.0001
deuxieme_cycle	1	0.46743	0.01636	28.57	<.0001
troisieme_cycle	1	0.68487	0.01947	35.18	<.0001

Figure 7. Résidus par régresseurs Modèle 5c



5.4 Question 5d : tous les niveaux d'étude sauf cycle3 (référence 6)

```
        /* 5d. Tous les niveaux d'etude sauf cycle3 (reference 6)*/
proc reg data=PanelLight;
    model lw = professionnel_court professionnel_long primaire secondaire deuxieme_cycle;
    output out=PanelLight_Predicted p=predicted_lw;
run;
```

Figure 8. Modèle 5d

Nb d'observations lues	10548
Nb d'obs. utilisées	8856
Nombre d'observations avec valeurs manquantes	1692

Analyse de variance

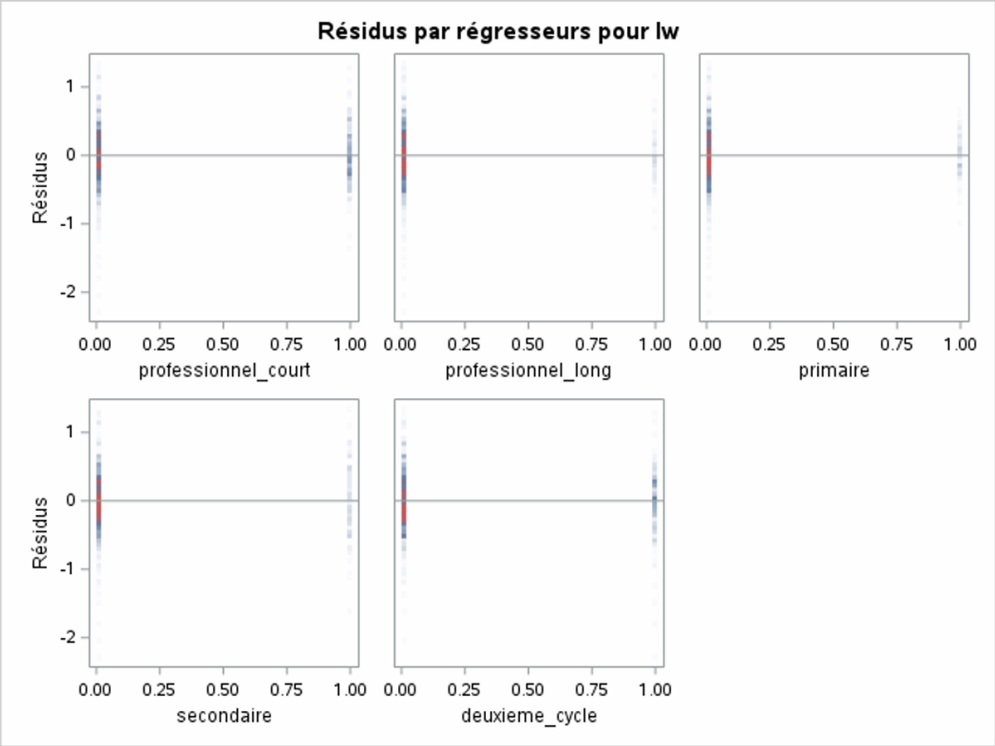
Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	5	375.12300	75.02460	425.70	<.0001
Erreur	8850	1559.69775	0.17624		
Total sommes corrigées	8855	1934.82075			

Root MSE	0.41981	R carré	0.1939
Moyenne dépendante	3.88876	R car. ajust.	0.1934
Coeff Var	10.79536		

Paramètres estimés

Variable	DDL	Valeur estimée des paramètres	Erreur type	Valeur du test t	Pr > |t|
Intercept	1	4.28789	0.01372	312.49	<.0001
professionnel_court	1	-0.56177	0.01595	-35.22	<.0001
professionnel_long	1	-0.43021	0.02071	-20.77	<.0001
primaire	1	-0.68487	0.01947	-35.18	<.0001
secondaire	1	-0.45261	0.01797	-25.18	<.0001
deuxieme_cycle	1	-0.21744	0.01628	-13.35	<.0001

Figure 9. Résidus par régresseurs Modèle 5d



5.5 Question 5e : tous les niveaux d'études, en imposant la nullité de la moyenne des coefficients des indicatrices, pondérée par les effectifs

La question n'a pas été comprise.

6 Question 6 : Comparer les résultats globaux des différents modèles, en particulier R2, R2 ajusté, Test de nullité globale de l'ensemble des coefficients et commenter

Les R2 et R2 ajustés et résultats de l'analyse de la variance (y compris les tests de nullité) sont identiques pour les modèles a, c et d (R2= 19,39%, R2 ajusté = 19,34%, Valeur F : 425,7, seuil <0,0001) indiquant un faible pouvoir explicatif du modèle. Cependant, le modèle b (sans intercept) montre un R2 et R2 ajusté bien plus élevé, de l'ordre de 98% et possède une F value à hauteur de 127000 Dans l'ensemble nous pouvons rejeter l'hypothèse nul à un seuil 0,01%, mettant en avant le fait que les niveaux d'études spécifiés ont une capacité significative à expliquer la variabilité de lw.

7 Question 7 : Comparer les prédictions des différents modèles

Les prédictions pour ces différents modèles seront identiques à la différence que la constante, si présente, sera différente. Les modèles auront également des coefficients différents selon leur spécification mais conduiront à une même prédiction. De ce fait, nous pouvons lire par exemple que le fait d'avoir un niveau d'étude de "troisième cycle" conduira à une hausse de revenu d'environ 68% (0,68*100(%)) par rapport à celui d'un niveau "primaire" (cf. Modèle c), ceteris paribus.

8 Question 8 : Exprimer les différents coefficients (betas, constante) des différents modèles en fonction des différents salaires moyens. En vous appuyant sur ces expressions, commenter les résultats des différents modèles, en particulier valeur et test de nullité de chaque coefficient

Ces modèles adoptent une spécification en semi-élasticité (log-lin), néanmois chacun de ces modèles auront une spécificité, et donc une lecture différente, bien qu'ils parviennent à obtenir les mêmes prédictions. Le modèle "a" intègre la variable "troisième cycle" comme valeur de référence par le biais de la constante pour les autres variables, ainsi un individu ayant un niveau "deuxième cycle" gagnera en moyenne 21% de moins qu'un individu ayant un niveau de troisième cycle, ceteris paribus.
A l'image du modèle a, les modèles c et d seront des modèles pour lesquels nous interpréterons la valeur de référence à partir de la constante (à l'exception que nous ne spécifions pas l'une des variables et que nous ne nous ferons pas corriger par SAS). Dans le modèle c, nous interpréterons les coefficients par rapport à la variable "primaire", le niveau d'étude qui gagnera le moins (un individu ayant un niveau d'étude "secondaire" gagnera en moyenne 23% de plus qu'un individu ayant un niveau primaire. Dans le cadre du modèle c, l'interprétation des coefficients se fera avec pour valeur de référence, la variable "troisième cycle", celle qui obtient un revenu le plus élevé et aura exactement la même interprétation que dans le modèle a.
Dans le deuxième modèle, n'ayant ni constante ni valeur de référence, nous devons soustraire la valeur des coefficients si nous voulons appliquer les niveau d'études ou appliquer une fonction exponentielle au coefficient appartenant à la variable observé pour linéariser l'opération afin de prédire le salaire d'un individu ayant un "x" niveau d'étude. Par exemple, un individu ayant un niveau troisième cycle pourra espérer gagner en moyenne exp(4,28) unité par heure, ceteris paribus.
Concernant le test de nullité des coefficients, pour l'ensemble des modèles (cf. modèles a, b, c et d) la t-value nous permet de rejeter H0 au seuil 0,01% pour l'ensemble des coefficients, et par conséquent,

conclure à la significativité statistique des coefficients.

9 Question 9 : Conclure sur les différences et similitudes, avantages et inconvénients des différents modèles. Quel est finalement le meilleur modèle ?

Parmi ces modèles, nous avons un modèle a et b qui seront affectés par de la multicolinéarité du fait qu'il y ai toutes les dummies. Il serait préférable d'enlever une des dummies, qui sera alors interprétable lorsque les autres dummies seront égales à 0, ce qui sera ensuite réalisé dans la construction des modèles c et d.
Le modèle b aura pour différence par rapport au modèle a, une suppression de l'intercept, ce qui pourrait être logique si l'on suppose que tout les individus observés ont un niveau d'étude qui existe dans notre jeu de donnée, ainsi, avec cette spécification, il y aura toujours un output non nul. Cependant, le R2 et R2 ajusté étant anormalement très élevé, s'expliquant par le fait que l'on suppose un passage de la droite des valeurs prédites par l'origine. Toutefois, nous retrouverons des outputs identique pour les modèles a et b, en effet, intercept + dummy observée dans le modèle a sera toujours égal à la dummy observée dans le modèle b
Les modèles c et d conduisent à un résultat identique au modèle a dans le cadre de l'analyse de la variance. Cependant les valeurs dans le tableau des paramètres estimés diffèrent. En effet, le fait de supposer une variable de référence lorsque les dummies sont nulles tels que l'individu observé aura un niveau d'étude "primaire" (cf. modèle c) ou un niveau d'étude "troisième cycle" (cf. modèle d) aura pour effet une modification de la magnitude des coefficients et de leur t-value, bien que cela n'affecte pas leur niveau de significativité.
Si nous observons le champ des outputs possible, nous nous rendons compte que pour l'ensemble de ces modèles (a, b, c, d), bien que la valeur des coefficients sera différente, l'output sera identique. En effet, le fait de supprimer l'intercept (cf. modèle b) ou choisir une variable de référence (cf. modèle c et d) n'aura aucune incidence sur l'output du fait que l'intercept prendra pour valeur estimée celle de la variable de référence en raison de la multicolinéarité parfaite dans ces modèles, justifiant par ailleurs un SSR identique pour l'ensemble des modèles. Ainsi, bien que l'analyse de variance affichent des résultats différents, nous ne pouvons pas affirmer qu'un modèle des modèles est meilleur qu'un autre en termes de performances. Cependant, la lecture des coefficients s'avère plus efficace pour le modèle C, du fait que celui-ci prenne comme variable de référence "primaire", cette variable ayant le coefficient le moins élevé parmi toutes les variables s'avère efficace comme outil de comparaison pour étudier le coefficient positifs des autres variables sur cette dernière et voir à quel point les individus gagnent plus en fonction de leur niveau. Le modèle d pourrait s'avérer tout aussi efficace si l'on souhaite avoir comme valeur de référence le niveau le

plus élevé, et par conséquent, à quel points les autres individus gagnent moins en fonction de leur niveau d'étude*Le résultat attendu pour le modèle 5 n'étant pas inclus dans notre analyse du fait que la question n'ait pas été comprise.*