Valentin Barthel

6 avril 2024

Table des matières

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

```
/*-----------------------------------QUESTION 1-----------------------------------*/
                /* Définition de la librairie et appel de la base enregistrée dans l'exercice 1*/
libname maLib '/home/u63824485/sasuser.v94';

data panel95lightex1;
    set maLib.panel95lightex1;
run;

                /* Importation du fichier CSV dans SAS */
proc import datafile="/home/u63824485/sasuser.v94/PanelEuropeen95.csv"
        out = PanelEuropeen95_temp
        dbms = csv
        replace;
        getnames=yes;
run;

                /* Filtrage des colonnes spécifiques de l'ensemble de données importé */
data PanelEuropeen95;
        set PanelEuropeen95_temp(keep=mident mois actif actifp actifs agea);
run;

                /* Triage des ensembles de données par mident et mois */
proc sort data=PanelEuropeen95;
    by mident mois;
run;

proc sort data=panel95lightex1;
    by mident mois;
run;

                /* Fusion des ensembles de données triés */
proc sql;
    create table merged_panel as
    select a.*, b.*
    from panel95lightex1 as a
```

```
    inner join PanelEuropeen95 as b
    on a.mident=b.mident and a.mois=b.mois;
quit;
```

2 Question 2 : Régresser (MCO) actifs sur les différentes indicatrices d'éducation.

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

/*MCO actifs*/
proc reg data=merged_panel;
    model actifs = professionnel_court professionnel_long primaire secondaire deuxieme_cycle troisieme_cycle;
run;
```

La procédure REG
Modèle : MODEL1
Variable dépendante : actifs

Nb d'observations lues	10548
Nb d'obs. utilisées	10548

Analyse de variance

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	5	5.27005	1.05401	35.03	<.0001
Erreur	10542	317.21715	0.03009		
Total sommes corrigées	10547	322.48720			

Root MSE	0.17347	R carré	0.0163
Moyenne dépendante	0.03157	R car. ajust.	0.0159
Coeff Var	549.46839		

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	0.09020	0.00543	16.61	<.0001
professionnel_court	B	-0.07438	0.00622	-11.96	<.0001
professionnel_long	B	-0.07900	0.00818	-9.66	<.0001
primaire	B	-0.06217	0.00723	-8.60	<.0001
secondaire	B	-0.04128	0.00690	-5.98	<.0001
deuxieme_cycle	B	-0.06499	0.00646	-10.06	<.0001
troisieme_cycle	0	0	.	.	.

La procédure REG
Modèle : MODEL1
Variable dépendante : actifs

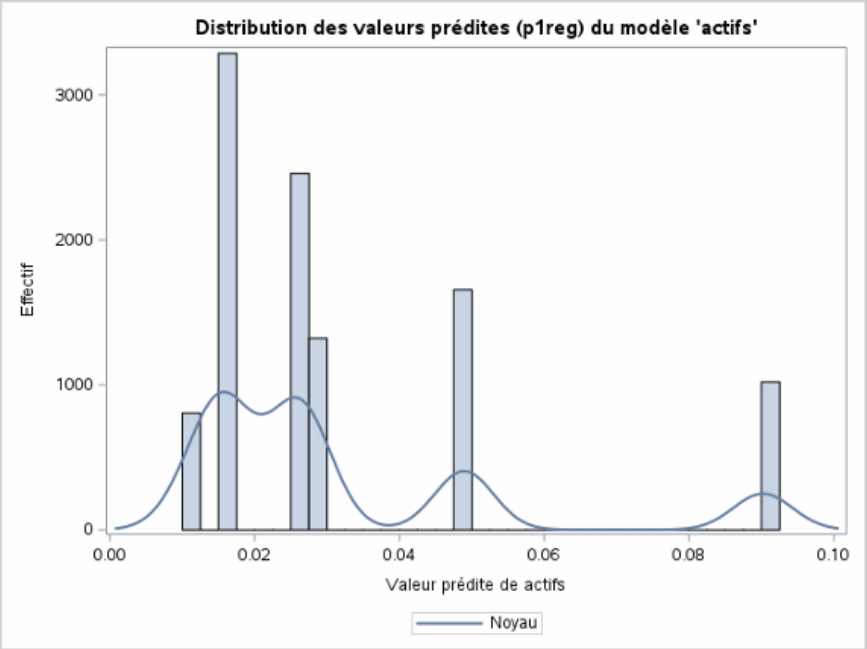
3 Question 3 : Ajouter à la table de travail la probabilité d'activité secondaire ou épisodique prédite par ce modèle, notée p1reg, ainsi qu'un estimateur de la variance du résidu, notée sigma1reg. Représenter graphiquement la distribution de p1reg et la commenter.

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

/*3- Ajout de p1reg et sigma1reg, et visualisation*/
proc reg data=merged_panel outest=reg_results;
    model actifs = professionnel_court professionnel_long primaire secondaire deuxieme_cycle troisieme_cycle;
    output out=predicted1 p=p1reg r=residuals;
run;

proc means data=predicted1;
    var residuals;
    output out=variance std=std_residuals;
run;
                /*Calcul de la variance des résidus*/
data variance;
    set variance;
    sigma1reg = std_residuals**2;
run;
                /*Ajout variance des résidus au df "merged_panel"*/
data merged_panel;
    if _N_ = 1 then set variance(keep=sigma1reg);
    set merged_panel;
    by mident mois;
    if _N_ = 1 then call symput('sigma1reg', sigma1reg);
run;
                /*Création d'un df "travail_final" à partir
                du merge de merged_panel et predicted1*/
data travail_final;
    merge merged_panel(in=a) predicted1(keep=mident mois p1reg);
    by mident mois;
    if a;
run;
```

```
/*Création d'un histogram pour représenter les valeurs prédites*/
proc sgplot data=travail_final;
    histogram p1reg / scale=count;
    density p1reg / type=kernel;
run;
```



Cet histogramme nous présente les valeurs prédites pour la variable "actifs". Les six valeurs différentes proviennent chacun des agrégats qui composent le niveau d'étude. L'histogramme présenté nous montre que la probabilité que actifs = 1, c'est-à-dire que la probabilité d'avoir une activité secondaire ou épisodique est faible. En effet, la concentration des barres se situe principalement autour de 20%. Si nous nous référons au tableau de résultat, le seul niveau conduisant à obtenir une activité secondaire ou épisodique serait le niveau "troisième cycle" (90%) à travers la constante suite à la correction de SAS, pour des raisons de multicolinéarité.

4 Question 4 : Régresser actifs sur les différentes indicatrices d'éducation par MCP et commenter les différences avec la régression précédente.

```
/*----------------------------------QUESTION 4----------------------------------*/
/*4- Regression MCP*/
```

```
proc pls data=merged_panel method=pls;
    model actifs = professionnel_court professionnel_long primaire secondaire deuxieme_cycle troisieme_cycle;
run;
```

La procédure PLS

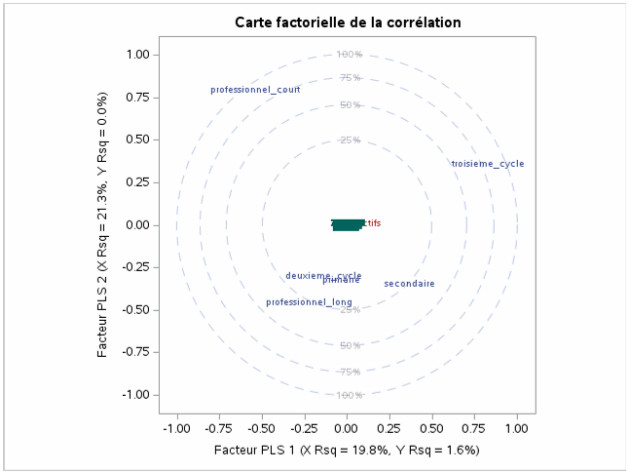
Table	WORK.MERGED_PANEL
Méthode d'extraction de facteurs	Moindres carrés partiels
Algorithme PLS	NIPALS
Nombre de variables de réponse	1
Nombre de paramètres du prédicteur	6
Gestion des valeurs manquantes	Exclude
Nombre de facteurs	6

Nombre d'observations lues	10548
Nombre d'observations utilisées	10548

La procédure PLS

Variation de pourcentage expliquée par Facteurs des moindres carrés partiels

	Effets du modèle		Variables dépendantes	
Nombre de facteurs extraits	En cours	Total	En cours	Total
1	19.7950	19.7950	1.6227	1.6227
2	21.3046	41.0996	0.0115	1.6341
3	20.8228	61.9223	0.0000	1.6342
4	18.7171	80.6395	0.0000	1.6342
5	19.3605	100.0000	0.0000	1.6342
6	0.0000	100.0000	0.0000	1.6342



5 Question 5 : Créer les variables agea2 = agea^2 $etagea3 = agea^3$

```
/*-----------------------------------QUESTION 5-----------------------------------*/
/*5- Créons les variables agea2 et agea3 en élevement respectivement la variable agea au carré et au cube*/
data merged_panel;
```

```
    set merged_panel;
    agea2 = agea**2;
    agea3 = agea**3;
run;
```

6 Question 6 : Régresser actifs sur agea, agea2, agea3 et les différentes indicatrices d'éducation, et l'indicatrice de sexe féminin.

```
/*-----------------------------------QUESTION 6-----------------------------------*/
/*Regressons les actifs sur agea, agea2, agea3 ainsi que les indicatrices d'éducation de sexe féminin*/
proc pls data=merged_panel method=pls;
    model actifs = agea agea2 agea3 professionnel_court professionnel_long primaire secondaire deuxieme_cycle tro
quit;
```

7 Ajouter à la table de travail la probabilité d'activité secondaire ou épisodique prédite par ce nouveau modèle, notée p2reg. Représenter graphiquement la distribution de p2reg et la commenter.

```
/*-----------------------------------QUESTION 7-----------------------------------*/
                /*Regression avec récupération de p2reg*/
proc reg data=merged_panel outest=reg_results;
    model actifs = agea agea2 agea3 sexeFemme professionnel_court professionnel_long primaire secondaire deuxieme
    output out=predicted2 p=p2reg r=residuals;
run;

                /* Visualisation de la distribution de p2reg */
proc sgplot data=predicted2;
    histogram p2reg / scale=count;
    density p2reg / type=kernel;
run;
```

La procédure REG
Modèle : MODEL1
Variable dépendante : actifs

Nb d'observations lues	10548
Nb d'obs. utilisées	10548

Analyse de variance

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	9	7.09627	0.78847	26.34	<.0001
Erreur	10538	315.39094	0.02993		
Total sommes corrigées	10547	322.48720			

Root MSE	0.17300	R carré	0.0220
Moyenne dépendante	0.03157	R car. ajust.	0.0212
Coeff Var	547.98844		

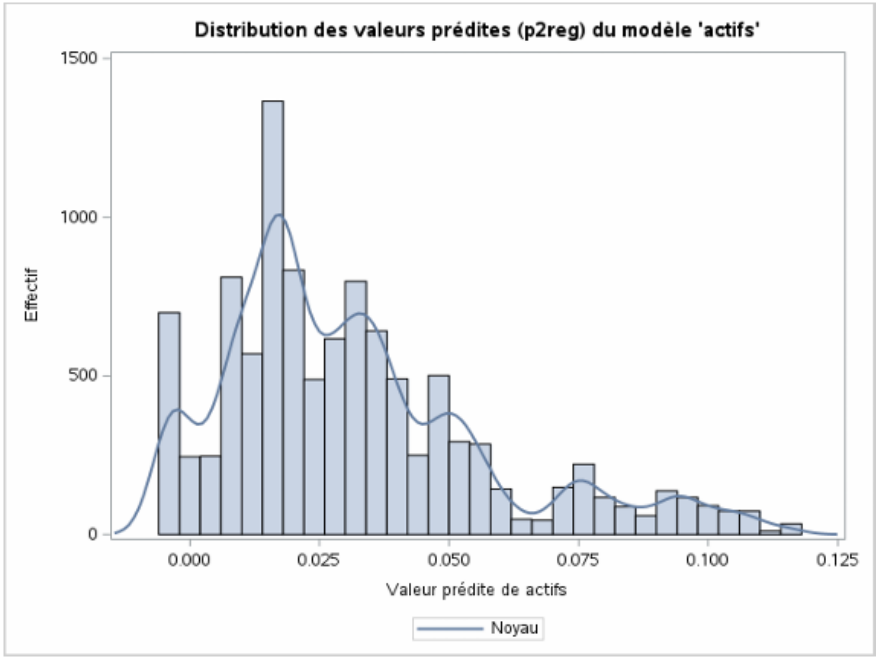
Note: Model is not full rank. Least-squares solutions for the parameters are not unique. Some statistics will be misleading. A reported DF of 0 or B means that the estimate is biased.

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

troisieme_cycle =	Intercept - 329E-14 * agea + 844E-16 * agea2 - 693E-18 * agea3 - 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	0.45086	0.10376	4.35	<.0001
agea	B	-0.02644	0.00825	-3.21	0.0014
agea2	B	0.00062803	0.00021059	2.98	0.0029
agea3	B	-0.00000470	0.00000173	-2.71	0.0068
sexeFemme	1	-0.02127	0.00349	-6.09	<.0001
professionnel_court	B	-0.07688	0.00621	-12.34	<.0001
professionnel_long	B	-0.07737	0.00817	-9.47	<.0001
primaire	B	-0.08586	0.00743	-8.84	<.0001
secondaire	B	-0.03952	0.00692	-5.71	<.0001
deuxieme_cycle	B	-0.06345	0.00645	-9.84	<.0001
troisieme_cycle	0	0	.	.	.



La distribution ci-dessus nous présente les prédictions dans le cadre d'un second modèle. Nous pouvons lire que les prédictions se situent autours également autour de 0,20, montrant une fois de plus la faible probabilité d'avoir une activité secondaire ou épisodique pour les individus observés bien que l'on ait ajouté la variable agea élevée au cube. Cependant, contrairement à la distribution des prédictions du premier modèle, cette distribution de p2reg nous dévoile les limites d'un modèle à probabilité linéaire. En effet, la distribution révèle des prédictions en dehors de l'intervalle [0,1], ce qui ne devrait logiquement pas être le cas du fait que notre variable "actifs" ne puisse prendre comme valeur 0 ou 1.

8 Question 8 : Peut-on appliquer la méthode des MCP ici ? Expliquer les différences par rapport au modèle 1.

Bien que le modèle à probabilité soit un modèle hétéroscédastique, nous ne pouvons pas appliquer de méthode MCP pour corriger l'hétéroscedasticité de notre modèle du fait que contrairement à notre premier modèle, nous obtenons des valeurs prédites en dehors de l'intervalle [0,1].

9 Question 9 : Comparer les résultats obtenus en remplaçant l'indicatrice de sexe féminin par l'indicatrice de sexe masculin, puis par la variable sexeN. Commenter précisément chacune des différences entre ces trois modèles.

9.1 a) Indicatrice

```
/*-----------------------------------QUESTION 9a-----------------------------------*/
        /*Indicatrice sexeHomme*/
                /*Regression avec récupération de p2reg1*/
proc reg data=merged_panel outest=reg_results;
    model actifs = agea agea2 agea3 sexeHomme professionnel_court professionnel_long primaire secondaire deuxieme
        output out=predicted21 p=p2reg1 r=residuals;
run;
```

La procédure REG
Modèle : MODEL1
Variable dépendante : actifs

Nb d'observations lues	10548
Nb d'obs. utilisées	10548

Analyse de variance

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	9	7.09627	0.78847	26.34	<.0001
Erreur	10538	315.39094	0.02993		
Total sommes corrigées	10547	322.48720			

Root MSE	0.17300	R carré	0.0220
Moyenne dépendante	0.03157	R car. ajust.	0.0212
Coeff Var	547.98844		

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 - 329E-14 * agea + 844E-16 * agea2 - 693E-18 * agea3 - 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	0.42959	0.10386	4.14	<.0001
agea	B	-0.02644	0.00825	-3.21	0.0014
agea2	B	0.00062803	0.00021059	2.98	0.0029
agea3	B	-0.00000470	0.00000173	-2.71	0.0068
sexeHomme	1	0.02127	0.00349	6.09	<.0001
professionnel_court	B	-0.07668	0.00621	-12.34	<.0001
professionnel_long	B	-0.07737	0.00817	-9.47	<.0001
primaire	B	-0.06566	0.00743	-8.84	<.0001
secondaire	B	-0.03952	0.00692	-5.71	<.0001
deuxieme_cycle	B	-0.06345	0.00645	-9.84	<.0001
troisieme_cycle	0	0	.	.	.

9.2 b) sexeN

```
/*-----------------------------------QUESTION 9b-----------------------------------*/
        /*Indicatrice sexeN*/

proc reg data=merged_panel outest=reg_results;
    model actifs = agea agea2 agea3 sexeN professionnel_court professionnel_long primaire secondaire deuxieme_cyc
        output out=predicted22 p=p2reg2 r=residuals;
run;
```

La procédure REG
Modèle : MODEL1
Variable dépendante : actifs

Nb d'observations lues	10548
Nb d'obs. utilisées	10548

Analyse de variance

Source	DDL	Somme des carrés	Moyenne quadratique	Valeur F	Pr > F
Modèle	9	7.09627	0.78847	26.34	<.0001
Erreur	10538	315.39094	0.02993		
Total sommes corrigées	10547	322.48720			

Root MSE	0.17300	R carré	0.0220
Moyenne dépendante	0.03157	R car. ajust.	0.0212
Coeff Var	547.98844		

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 - 329E-14 * agea + 844E-16 * agea2 - 693E-18 * agea3 - 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	0.47213	0.10377	4.55	<.0001
agea	B	-0.02644	0.00825	-3.21	0.0014
agea2	B	0.00062803	0.00021059	2.98	0.0029
agea3	B	-0.00000470	0.00000173	-2.71	0.0068
sexeN	1	-0.02127	0.00349	-6.09	<.0001
professionnel_court	B	-0.07668	0.00621	-12.34	<.0001
professionnel_long	B	-0.07737	0.00817	-9.47	<.0001
primaire	B	-0.06566	0.00743	-8.84	<.0001
secondaire	B	-0.03952	0.00692	-5.71	<.0001
deuxieme_cycle	B	-0.06345	0.00645	-9.84	<.0001
troisieme_cycle	0	0	.	.	.

9.3 c) Commentaire

Dans le cadre de ces modèles nous observons que le sexe a un impact sur la variable expliquée "actifs", en effet nous observons dans ces trois modèles que le fait d'être une femme réduit la probabibilité d'avoir une activité secondaire ou épisodique, ceteris paribus. Pour l'ensemble des modèles étudiés, l'âge a un impact significatif sur la probabilité d'être actif, indiquant que les changements dans l'âge (et ses effets non linéaires) ont un effet sur la détermination de l'activité, suggérant que l'activité est secondaire ou épisodique peut être différente selon l'âge de l'individu. Le signe des coefficients présente un effet de l'âge sur la probabilité d'avoir une activité secondaire ou épisodique commence par être négatif, puis s'atténue, avant de devenir fortement négatif à des âges plus avancés, ceteris paribus. Nous noterons également que les coefficients sont de même signe pour ces modèles, et que par conséquent, l'effet positif et négatif de ces derniers sur la variable "actifs" sera identique pour chacun de ces modèles. Ces modèles expliquent à hauteur de 22% les variations de nos variables dépendantes. Nous pouvons par ailleurs lire une significativité statistique pour l'ensemble des coefficients et un F-test nous suggérant une significativité globale de ces derniers. Ces modèles partagent donc les mêmes résultats bien que leur lecture et interprétation soit différente.