SAS

Travaux Dirigés

III. La Statistique Multivariée avec SAS

Références Bibliographiques

- *A Handbook of Statistical Analyses using SAS*. G. Derr & B. S. Everitt, CRC Press, Chapitres 7, 8, 16, 17.

Pour toute question, commentaire ou suggestion, contacter : salim.lardjane@univ-ubs.fr

1) Lire le fichier …/derr/olympic.dat dans une table SAS nommée **decathlon** et la visualiser. Ce fichier fournit les performances des athlètes ayant participé à l'épreuve de décathlon des jeux olympiques de 1988.
   Les variables renseignées dans le fichier sont les suivantes (dans l'ordre) : nom de l'athlète, performance au 100m, performance au saut en longueur, performance au lancer du poids, performance au saut en hauteur, performance au 400m, performance au 110m haies, performance au lancer du disque, performance au saut à la perche, performance au lancer du javelot, performance au 1500m, score total.

   On souhaite obtenir une représentation graphique des performances des athlètes ainsi qu'une mesure statistique de performance globale qui puisse être comparée à leurs scores totaux, ces derniers étant calculés à l'aide de tables standards de conversion.

2) Le code utilisé est le suivant.

```
filename dir "d:/data/sasdata/derr";
data decathlon;
      infile dir(olympic.dat);
      input        nom $ 1-13
                   course_100m
                   saut_longueur
                   lancer_poids
                   saut_hauteur
                   course_400m
                   course_haies
                   lancer_disque
                   saut_perche
                   lancer_javelot
                   course_1500m
                   score
      ;
run;
```

```
proc print data = decathlon;
run;
```

On décide de représenter graphiquement les performances des athlètes à l'aide d'une Analyse en Composantes Principales. Avant de procéder à celle-ci, on doit toutefois traiter la question d'éventuelles valeurs aberrantes. A cet effet, on peut dans le cas présent s'intéresser à la distribution du score total.

Analyser celle-ci à l'aide de la **proc univariate** de SAS et la représenter à l'aide d'un diagramme en boîte. On utilisera le nom des athlètes comme label d'éventuelles valeurs aberrantes.

3) Le code utilisé est le suivant.

```
proc univariate data = decathlon;
       var score;
       id nom;
run;

proc sgplot data = decathlon;
       vbox score / datalabel = nom;
run;
```

Que remarque-t-on ?

4) Le score de l'athlète Kunwar est clairement une valeur aberrante ; ce dernier sera donc supprimé de l'analyse. Par ailleurs, afin de faciliter l'interprétation des résultats, il est souhaitable que les performances des athlètes dans les différentes épreuves soient évaluées dans la même direction, les valeurs faibles correspondant aux mauvaises performances et les valeurs hautes aux meilleurs. A cet effet, on décide de multiplier par -1 les résultats des athlètes aux épreuves de course. Rédiger un programme SAS permettant d'effectuer ces modifications.

5) Le code utilisé est le suivant.

```
data decathlon;
       set decathlon;
       if score > 6000;
       course_100m = course_100m*(-1);
       course_400m = course_400m*(-1);
       course_haies = course_haies*(-1);
       course_1500m = course_1500m*(-1);
run;
```

Nous sommes à présent en mesure d'effectuer une Analyse en Composantes Principales des performances des athlètes (à l'exclusion du score total). Utiliser la **proc princomp** de SAS pour la mener à bien. Combien d'axes factoriels retenir ? Obtenir une représentation graphique du nuage des variables avec le cercle des corrélations dans les trois premiers plans factoriels. Quelle interprétation donner des différents axes ?

6) Le code utilisé est le suivant.

```
proc princomp data = decathlon out = acpout
                          plots = pattern(ncomp = 3 vector) ;
       var course_100m--course_1500m;
run;
```

Les coefficients associés à la première composante sont tous positifs et cette dernière correspond clairement à une mesure de performance globale (on développera ce point plus loin). Cette composante explique 34% de la variation totale. La seconde composante oppose les performances aux épreuves de force (lancer du poids et du disque, notamment) à la performance à l'épreuve d'endurance (1500m). Prises ensembles, les deux premières composantes expliquent environ 60% de la variation totale. Seules ces deux composantes ont des valeurs propres associées supérieures à 1, ce qui suggère qu'elles fournissent une représentation adéquate et parcimonieuse des performances des athlètes.

Afin d'obtenir une représentation graphique de nos données, on décide de représenter le nuage des individus. Pour rendre celui-ci plus lisible, on peut étiqueter les points avec le classement final des athlètes à l'épreuve du décathlon. Obtenir ces classements à l'aide de la **proc rank** de SAS et représenter le nuage des points avec les rangs obtenus comme étiquettes.

7) Le code utilisé est le suivant.

```
proc rank data = acpout out = acpout descending;
      var score;
      ranks posn;
run;

proc sgplot data = acpout;
      scatter y = prin2 x = prin1 / markerchar = posn;
run;
```

On souhaite à présent comparer les scores totaux des athlètes à leurs scores sur les deux premières composantes principales et calculer les corrélations correspondantes. Utiliser la **proc sgscatter** et la **proc corr** de SAS pour mener à bien cette analyse.

8) Le code utilisé est le suivant.

```
proc sgscatter data = acpout;
      compare y = score x = (prin1 prin2);
run;

proc corr data = acpout;
      var score prin1 prin2;
run;
```

Qu'en conclure ?

9) Lire le fichier …/derr/usair2.dat dans une table SAS nommée usair et la visualiser. Les données correspondent aux valeurs de différentes variables liées à la pollution de l'air dans 41 villes des Etats-Unis. Ces variables sont (dans l'ordre) : nom de la ville, concentration de SO2 (dioxyde de soufre) dans l'air en microgrammes par mètre cube, température moyenne annuelle en Fahrenheit, Nombre d'entreprises industrielles employant 20 salariés ou plus, population en milliers d'habitants (recensement de 1970), vitesse moyenne annuelle du vent en miles par heure, précipitations moyennes annuelles en inches, nombre moyen de jours de pluie par an.

Convertir les unités de mesures des variables concernées des unités de mesures anglo-saxonnes aux unités de mesures internationales.

10) Le code utilisé est le suivant.

```
data usair;
	infile dir(usair2.dat) expandtabs;
	input ville $16. so2 temperature usines population vitesse_du_vent pluie
		nb_jours_de_pluie;
run;

proc print data = usair;
run;

data usair;
	set usair;
	vitesse_du_vent = vitesse_du_vent * 1.609344;
	pluie = pluie * 25.4;
	temperature = (temperature - 32)*(5/9);
run;
```

11) Etudier la distribution et la normalité des différentes variables à l'aide la **proc univariate** de SAS. Tester également le caractère exponentiel des différentes distributions.

12) Le code utilisé est le suivant.

```
proc univariate data = usair;
	var so2--nb_jours_de_pluie;
	id ville;
	histogram / normal kernel exponential;
run;
```

Ecrire un macro-programme permettant d'obtenir automatiquement les diagrammes en boîte correspondant aux différentes variables. On identifiera d'éventuelles valeurs aberrantes par la ville correspondante. Qu'en conclure ?

13) Le code utilisé est le suivant.

```
%let  var = so2 temperature usines population vitesse_du_vent pluie
nb_jours_de_pluie;

%macro boxplots;
	%do i = 1 %to 7;
		%let variable = %scan(&var,&i);
		proc sgplot data = usair;
			title "Variable : &variable";
			vbox &variable / datalabel = ville;
		run;
	%end;
	title;
%mend boxplots;


%boxplots;
```

Les ville de Chicago, Philadelphie, Providence, Miami, Detroit, Cleveland, Albuquerque et Phoenix présentent clairement des valeurs aberrantes (extrêmes) pour l'une ou l'autre des

variables. On décide donc de les supprimer de l'analyse. Ecrire un programme SAS permettant de faire cette modification.

14) Le code utilisé est le suivant.

```
data usair2;
       set usair;
       if ville not in ("Chicago" "Detroit" "Philadelphia" "Phoenix" "Miami"
                                        "Cleveland" "Albuquerque" "Providence");
run;
```

Afin d'étudier la relation entre la concentration de SO2 et les différentes autres variables, on décide de réaliser une Classification Ascendante Hierarchique (CAH) sur celles-ci par la méthode de Ward. Ecrire un programme SAS permettant de la mener à bien. Ne pas utiliser la variable SO2 dans la classification mais la sauvegarder dans la table obtenue pour usage ultérieur. Représenter graphiquement le dendrogramme (arbre de classification) obtenu à l'issue de l'analyse.

15) Le code utilisé est le suivant.

```
%let  var = temperature usines population vitesse_du_vent pluie nb_jours_de_pluie;

proc cluster data = usair2 method = ward ccc std outtree = ward;
       var &var;
       id ville;
       copy so2;
run;

goptions htext = 0.8;
axis1 label = (a = 90);

proc tree horizontal data = ward vaxis = axis1;
run;
```

Combien de classes retenir ?

16) On décide de retenir 4 classes. Etudier la distribution des différentes variables de la classification dans les différentes classes à l'aide de la **proc means** et de la **proc boxplot**. En déduire une première caractérisation des différentes classes.

17) Le code utilisé est le suivant.

```
proc tree data = ward out = clusters n = 4 noprint;
       copy ville so2 &var;
run;

proc sort data = clusters;
       by cluster;
run;

proc means data = clusters;
       var &var;
       by cluster;
run;

proc boxplot data = clusters;
```

```
      plot (&var)*cluster;
run;
```

On peut en déduire les premières caractérisations suivantes :
Classe 1 : températures moyennes à élevées, nombre d'usines faible à moyen, population faible à moyenne, vitesse du vent faible à moyenne, pluviosité moyenne à élevée, nombre de jours de pluie moyen à élevé (effectif : 17)

Classe 2 : températures moyennes, nombre d'usines faible à moyen, population faible à moyenne, vitesse du vent moyenne à élevée, pluviosité faible à moyenne, nombre de jours de pluie faible (effectif : 6)

Classe 3 : températures moyennes à élevées, nombre d'usines moyen à élevé, population moyenne à élevée, vitesse du vent moyenne, pluviosité moyenne à élevée, nombre de jours de pluie faible à moyen (effectif : 7)

Classe 4 : températures basses, nombre d'usines moyen à élevé, population moyenne, vitesse du vent élevée, pluviosité moyenne, nombre de jours de pluie moyen à élevé (effectif : 3)

18) Afin d'obtenir une caractérisation plus fine et pouvoir représenter graphiquement nos données de façon conjointe, on décide d'effectuer une ACP sur toutes les variables à l'exclusion de SO2. Mener cette analyse à bien à l'aide de la **proc princomp** de SAS. Représenter le nuage des variables avec le cercle des corrélations dans les trois premiers plans factoriels. Combien d'axes retenir ?

19) Le code utilisé est le suivant.

```
proc princomp data = clusters out = acpout
                        plots = (matrix pattern(ncomp = 3 vector));
      var &var;
run;
```

On décide de retenir les trois premiers axes factoriels. Quelle interprétation donner de chaque axe ?

20) La première composante s'interprète clairement comme un indicateur d'industrialisation, la deuxième comme un indicateur de température moyenne et la troisième comme un indicateur du caractère plus ou moins pluvieux du climat. Le plan factoriel (2,3) est donc de nature « climatique » en termes d'interprétation.

Représenter les différentes villes dans les trois premiers plans factoriels. Comparer les résultats obtenus avec une carte climatique des Etats-Unis. Est-ce cohérent ?

21) Le code utilisé est le suivant.

```
proc sgplot data = acpout;
      scatter y = prin2 x= prin1 / markerchar = ville;
run;

proc sgplot data = acpout;
      scatter y = prin3 x= prin1 / markerchar = ville;
run;

proc sgplot data = acpout;
      scatter y = prin3 x= prin2 / markerchar = ville;
run;
```

Représenter de nouveau le nuage des individus dans les trois premiers plans factoriels mais en distinguant les différentes classes obtenues à l'issue de la CAH. Quelle caractérisation des classes en déduire ?

22) Le code utilisé est le suivant.

```
proc sgplot data = acpout;
      scatter y = prin2 x= prin1 / group = cluster markerchar = ville;
run;

proc sgplot data = acpout;
      scatter y = prin3 x= prin1 / group = cluster markerchar = ville;
run;

proc sgplot data = acpout;
      scatter y = prin3 x= prin2 / group = cluster markerchar = ville;
run;
```

On en déduit la caractérisation suivante des 4 classes obtenues à l'issue de la CAH.

Classe 1 : villes à climat tempéré à industrialisation faible à moyenne (eff. 17)
Classe 2 : villes à climat aride ou semi-aride froid (eff. 6)
Classe 3 : villes à climat tempéré à industrialisation moyenne à forte (eff. 7)
Classe 4 : villes à climat humide et froid (eff. 3)

Comparer la répartition de la concentration de SO2 dans les différentes classes. Pour ce faire, on pourra utiliser des diagrammes en boîte. Qu'en conclure ?

23) Le code utilisé est le suivant.

```
proc sgplot data = clusters;
      vbox so2 / category = cluster datalabel = ville;
run;
```

Le niveau moyen de SO2 semble plus élevé dans les villes à climat tempéré que dans les villes à climat froid. Le degré d'industrialisation semble influer sur la *dispersion* des valeurs autour de la moyenne. Cela pourrait peut-être s'interpréter en disant que la concentration de SO2 dépend davantage du *type* d'industries implantées dans une ville que du nombre d'usines. Ceci étant dit, on peut souhaiter adopter une approche inférentielle et tester si les différences entre les moyennes et les variances de SO2 dans les différentes classes sont statistiquement significatives. A cet effet, on peut par exemple effectuer une analyse de variance. Mener celle-ci à bien à l'aide de la **proc glm** de SAS. Que peut-on en déduire ?

24) Le code utilisé est le suivant.

```
proc glm data = clusters;
       class cluster;
       model so2 = cluster;
       means cluster / hovtest;
run;
```

L'hypothèse d'égalité des moyennes n'est pas rejetée par le test de Fisher au seuil de 5%. On peut donc considérer que la CAH a permis de constituer des classes de niveaux moyens de SO2 comparables. De même, l'hypothèse d'égalité des variances n'est pas rejetée par le test de Levene au seuil de 5%. Ainsi, les différentes classes obtenues semblent être comparables en termes de répartition de la concentration en SO2 (moyennant l'hypothèse de normalité). Pouvait-on s'attendre à un tel résultat ?

25) Ce résultat peut être justifié en supposant que les industries les plus polluantes en SO2 s'installent naturellement dans les villes où le climat a pour effet d'atténuer l'effet de pollution, ce qui permet de maintenir un niveau en SO2 finalement comparable à celui des villes ayant un climat plus défavorable à ces industries. Rappelons toutefois qu'on a exclu certaines villes, présentant des valeurs extrêmes en l'une ou l'autre des variables, de l'analyse.

26) Lire les données …/derr/plasma.dat dans une table SAS nommée plasma. Celles-ci permettront d'étudier la relation entre la vitesse de sédimentation globulaire (VSG) du sang et la concentration de celui-ci en deux protéines, le fibrinogène (également appelée Facteur I) et la gamma-globuline, toutes deux mesurées en g/l. Pour un individu en bonne santé, la VSG doit être inférieure à 20 mm/h et comme sa valeur précise a relativement peu d'importance, la variable de réponse considérée est une variable dichotomique indiquant si ce seuil est franchi ou non. Une réponse égale à 0 signifie que l'individu est en bonne santé (VSG < 20) alors qu'une réponse égale à 1 signifie qu'il est en mauvaise santé (VSG $\geq$ 20). Les variables apparaissant dans le fichier de données sont (dans l'ordre) : concentration en fibrinogène, concentration en gamma-globuline, indicatrice de VSG.

27) Le code utilisé est le suivant.

```
ods graphics on;
options helpbrowser = sas; /* Win 64 */
filename dir "d:/data/sasdata/derr";

data plasma;
       infile dir(plasma.dat);
       input fibrinogene gamma vsg;
run;
```

Pour déterminer laquelle des deux protéines a l'influence la plus forte sur la VSG, on décide d'ajuster un modèle de régression logistique aux données, avec sélection de variables par la méthode « backward ». Ecrire un programme SAS permettant de mener à bien cette analyse.

On fera bien attention à spécifier que c'est la modalité 1 de la variable de réponse qui est ici la valeur dont on souhaite modéliser la probabilité (par défaut, SAS adopte pour cible la valeur la plus faible, ici 0).

28) Le code utilisé est le suivant.

```
proc logistic data = plasma desc;
	model vsg = fibrinogene gamma fibrinogene*gamma / selection = backward;
run;
```

On constate que seule la concentration en fibrinogène est retenue comme facteur explicatif par la méthode de sélection de variables. A l'aide de l'option **plots = effect** de la **proc logistic** représenter graphiquement la relation entre la probabilité que l'individu soit en mauvaise santé et la concentration de fibrinogène dans le sang. Que constate-t-on ?

29) Le code utilisé est le suivant.

```
proc logistic data = plasma desc plots = effect;
	model vsg = fibrinogene;
run;
```

Il y a clairement une relation croissante entre la concentration de fibrinogène et la probabilité que l'individu soit en mauvaise santé.

30) Lire les données …/derr/ghq.dat dans une table SAS nommée ghq. Celles-ci ont été recueillies dans le cadre d'une étude où un questionnaire de santé (Questionnaire de Santé Global) a été utilisé. On s'intéresse à l'effet conjoint du score au QSG et du sexe sur la probabilité qu'un individu soit en mauvaise santé. Les variable apparaissant dans le fichier de données sont (dans l'ordre) : score au QSG, sexe, nombre d'individus en mauvaise santé, nombre d'individus en bonne santé.

31) Le code utilisé est le suivant.

```
ods graphics on;
options helpbrowser = sas; /* Win 64 */
filename dir "d:/data/sasdata/derr";

data ghq;
	infile dir(ghq.dat) expandtabs;
	input qsg sexe $ nb_ms nb_bs;
run;
```

Définir une nouvelle variable nb_total correspondant au nombre d'individus enquêtés pour chaque combinaison de score au QSG et de sexe. Définir une nouvelle variable p_ms correspondant, pour chaque combinaison de score au QSG et de sexe, à la probabilité qu'un individu soit en mauvaise santé. La représenter graphiquement en fonction du score au QSG.

32) Le code utilisé est le suivant.

```
data ghq;
	set ghq;
	nb_total = nb_ms + nb_bs;
```

```
	pr_ms = nb_ms/nb_total;
run;

proc sgplot data = ghq;
	scatter y = pr_ms x = qsg;
run;
```

On souhaite à présent comparer graphiquement l'ajustement d'un modèle de régression linéaire simple et d'un modèle de régression logistique aux données, en utilisant le score au QSG comme seule variable explicative de la probabilité d'être en mauvaise santé. A cet effet, on commencera par utiliser la **proc logistic** et on sauvegardera les prédictions du modèle dans une variable lpred. En ce qui concerne la régression, on pourra utiliser directement l'instruction **reg** de la **proc sgplot**.

33) Le code utilisé est le suivant.

```
proc logistic data = ghq;
	model nb_ms / nb_total = qsg;
	output out = logout p = lpred;
run;

proc sort data = logout;
	by qsg;
run;

proc sgplot data = logout;
	series y = lpred x = qsg / legendlabel = "Logistique";
	reg y = pr_ms x = qsg / legendlabel = "Linéaire"
		lineattrs = (pattern = dash);
run;
```

Quelle conclusion tirer du graphique obtenu ? Quel modèle retenir ?

34) On souhaite à présent étudier l'influence conjointe du score au QSG et du sexe sur la probabilité d'être en mauvaise santé. Utiliser la proc logistic pour mener à bien cette analyse. On fera bien attention à spécifier que sexe est une variable qualitative dans le programme SAS soumis.

35) Le code utilisé est le suivant.

```
proc logistic data = ghq;
	class sexe;
	model nb_ms / nb_total = sexe qsg;
run;
```

Les trois tests de nullité gobale des coefficients conduisent à rejeter cette hypothèse au seuil de 5%. De plus, les estimations des coefficients de sexe et qsg sont toutes deux significativement différentes de 0 au seuil de 5%.

Afin de pouvoir interpréter de façon pratique les coefficients du modèle, on considère les odds-ratios (OR) qui s'en séduisent : l'OR pour qsg est de exp(0.7791) = 2.180 avec un intervalle de confiance à 95% donné par [1.795, 2.646], ce qui signifie qu'un accroissement d'une unité du score au QSG accroît les chances d'être en mauvaise santé par rapport à celles d'être en bonne santé de 1.8 à 3 fois, conditionnellement au sexe.

En ce qui concerne l'OR pour la variable sexe, celui-ci est obtenu comme exp(**2***0.648) = 2.55 car la variable est codée par SAS en -1/+1 (voir les sorties de la proc logistic). L'intervalle de confiance à 95% pour l'OR est donné par [1.088, 5.974], ce qui signifie que, conditionnellement au score au QSG, le fait d'être de sexe féminin accroît les chances d'être en mauvaise santé par rapport à celle d'être en bonne santé de 1.1 à 6 fois.

Afin d'illustrer graphiquement ce résultat, réaliser séparément une régression logistique de la probabilité d'être en mauvaise santé sur le score au QSG, pour les deux sexes, et comparer les graphiques obtenus.

36) Le code utilisé est le suivant.

```
proc sort data = ghq;
      by sexe;
run;

proc logistic data = ghq plots = effect;
      model nb_ms / nb_total = qsg;
      by sexe;
run;
```

Les résultats obtenus corroborent-ils ceux de l'analyse précédente ?

37) Lire les données …/derr/uscrime.dat dans une table SAS nommée uscrimi. Ces données vont nous permettre d'étudier, à l'aide d'un modèle de régression multiple, la relation entre le taux de criminalité et divers facteurs explicatifs de celle-ci, aux Etats-Unis (années 1960). Elles fournissent le taux de criminalité dans 47 états des Etats-Unis ainsi que les valeurs de 13 variables explicatives pouvant être envisagées comme facteurs explicatifs de la criminalité. Les variables sont, dans l'ordre d'apparition en colonnes du fichier :

- le taux de criminalité, c'est-à-dire le nombre de délits connus de la police pour 1 000 000 d'habitants
- le nombre d'individus de sexe masculins pour 1000 habitants
- une variable indicatrice identifiant les états du sud des Etats-Unis
- le niveau éducatif, c'est-à-dire le nombre moyen d'années de scolarisation multiplié par 10 de la population âgée de 25 ans et plus
- les dépenses de police totales par habitant pour l'année 1960
- les dépenses de police totales par habitant pour l'année 1959
- le taux d'activité des jeunes de 14 à 24 ans de sexe masculin, civils et urbains
- le nombre d'individus de sexe masculin pour 1000 individus de sexe féminin

- la population de l'état en centaines de milliers
- une variable décrivant l'uniformité de la population (non retenue dans l'analyse)
- le taux de chômage pour 1000, des habitants urbains de sexe masculin âgés de 14 à 24 ans
- le taux de chômage pour 1000, des habitants urbains de sexe masculin âgés de 35 à 39 ans
- la richesse mesurée par la valeur médiane du patrimoine et revenu des familles (en dizaines de dollars)
- l'inégalité de revenu, mesurée par le nombre de familles pour 1000 gagnant moins de la moitié du revenu médian

38) Le code utilisé est le suivant.

```
ods graphics on;
options helpbrowser = sas; /* Win 64 */
filename dir "d:/data/sasdata/derr";

data uscrimi;
	infile dir(uscrime.dat) expandtabs;
	input taux_crimi age sud education depenses_police_1960 depenses_police_1959
		taux_activite pop_masculine population uniformite_pop
		taux_chomage_jeunes taux_chomage_interm richesse
		inegalites_de_revenu;
run;
```

Supprimer la variable uniformite_pop de la table SAS uscrimi et définir une nouvelle variable id correspondant au numéro de l'observation.

39) Le code utilisé est le suivant.

```
data uscrimi;
	set uscrimi;
	drop uniformite_pop;
	id = _n_;
run;
```

Définir une macro-variable &var contenant la liste de toutes les variables explicatives de la criminalits. Rédiger un macro-programme permettant d'obtenir de façon automatique les diagrammes en boîte des différentes variables (y compris le taux de criminalité).

40) Le code utilisé est le suivant.

```
%let var = age sud education
		taux_activite pop_masculine population depenses_police_1960
depenses_police_1959
		taux_chomage_jeunes taux_chomage_interm richesse
inegalites_de_revenu;

%macro boxplots;
	%do i = 1 %to 13;
		%let variable = %scan(taux_crimi &var,&i);
		proc sgplot data = uscrimi;
```

```
                    title "Variable : &variable";
                    vbox &variable /datalabel = id;
                run;
        %end;
        title;
%mend boxplots;


%boxplots;
```

Les observations 4 et 26 sont clairement des valeurs aberrantes en termes de taux de criminalité et requièrent donc un traitement spécifique. Les supprimer de la table SAS uscrimi.

41) Le code utilisé est le suivant.

```
data uscrimi;
        set uscrimi;
        if id in (4 26) then delete;
run;
```

A l'aide la **proc sgscatter** représenter les nuages de points correspondant aux croisements des différentes variables explicatives retenues, entre elles et avec le taux de criminalité.

42) Le code utilisé est le suivant.

```
proc sgscatter data = uscrimi;
        matrix taux_crimi &var;
run;
```

Le graphique obtenu met en évidence des dépendances linéaires entre certaines variables. Un problème de multicollinéarité pourra donc se poser lors de l'estimation du modèle de régression multiple. Afin de quantifier celui-ci, on utilisera les Facteurs d'Inflation de la Variance (VIF) des différentes variables.

Rédiger un programme SAS permettant d'estimer le modèle de régression linéaire multiple du taux de criminalité sur les différentes variables explicatives retenues. Afficher les facteurs d'inflation de la variance pour les différentes variables. Qu'en conclure ?

43) Le code utilisé est le suivant.

```
proc reg data = uscrimi;
        model taux_crimi = &var / vif;
run;
```

Il apparaît que les dépenses de police pour 1959 et 1960 ont des facteurs d'inflation de la variance très élevée. On décide donc de supprimer les dépenses pour 1960 de l'analyse.
Modifier la macro-variable &var de façon à prendre en compte cette modification puis estimer de nouveau le modèle de régression linéaire multiple du taux de criminalité sur les différentes variables explicatives retenues. Commenter les résultats obtenus.

44) Le code utilisé est le suivant.

```
%let var = age sud education
            taux_activite pop_masculine population depenses_police_1959
            taux_chomage_jeunes taux_chomage_interm richesse inegalites_de_revenu;

proc reg data = uscrimi;
      model taux_crimi = &var / vif;
run;
```

On constate que plusieurs facteurs ont des coefficients non significativement différents de 0 au seuil 5% même si le test de nullité globale des coefficients conduit à rejeter cette hypothèse au seuil 5%. Par conséquent, on va adopter une approche en termes de sélection de variables pour identifier celles qui ont le plus d'influence sur le taux de criminalité.

Ecrire un programme SAS permettant d'implémenter la procédure « stepwise » de sélection de variables avec des seuils d'acceptation et de rejet fixés tous deux égaux à 10%. Commenter les résultats obtenus.

45) Le code utilisé est le suivant.

```
proc reg data = uscrimi;
      model taux_crimi = &var / selection = stepwise sle = .10 sls = .10;
run;
```

La procédure de sélection conduit à retenir comme variables explicatives : les dépenses de police pour 1959, le niveau éducatif, l'inégalité de revenu, la richesse et la proportion d'hommes de 14 à 24 ans. Toutefois on remarque que la variance des résidus semble être une fonction croissante de la valeur prédite du taux de criminalité. On décide par conséquent de modéliser plutôt le logarithme du taux de criminalité. Mener à bien cette analyse à l'aide de SAS, toujours avec une procédure de sélection « stepwise », et commenter les résultats obtenus.

46) Le code utilisé est le suivant.

```
data uscrimi;
      set uscrimi;
      log_taux_crimi = log(taux_crimi);
run;

proc reg data = uscrimi;
      model log_taux_crimi = &var / selection = stepwise sle = .10
      sls = .10;
run;
```

On constate que le problème d'hétéroscédasticité des résidus apparaît être résolu. De plus, le taux de chômage des hommes urbains âgés de 35 à 39 ans est retenu comme variable explicative par la méthode de sélection. On obtient un meilleur ajustement du modèle tel que mesuré par le R^2 et l'hypothèse de normalité de résidus semble vérifiée au vu des graphiques produits.