URCA Hugo Harari-Kermadec

2008-2009 harari@ecogest.ens-cachan.fr

Économétrie 1 : le modèle linéaire

1 TD : Maintenance de véhicule

Une entreprise souhaite étudier comment varie le coût annuel de maintenance de ces véhicules utilitaires en fonction de leur âge. Nous disposons des données suivantes :

Coût annuel en euros	Âge en mois
480	15
430	8
770	36
890	41
500	16
400	8
560	21
620	21
1000	53
470	10
710	32
580	17
1020	58
350	6
600	20

☞ Q1 Proposez un modèle linéaire pour étudier ce problème. Écrire le modèle sous forme matricielle, en donnant les dimensions des vecteurs et des matrices. Que vaut l'estimateur des moindres carrés pour les paramètres ? Explicitez son expression pour le cas présent (une constante et seule variable explicative).

☞ Q2 Proposer un estimateur $\hat{Y}_i$ de Y_i qui est aussi un prédicteur de Y_i. Justifier votre réponse. Comment estimer la variance σ^2 des résidus ?

☞ Q3 Déterminer une prévision du coût de maintenance pour un véhicule utilitaire de 4 ans.

2 TP : Introduction à SAS et Proc GLM

2.1 Introduction

On peut trouver la documentation SAS en ligne à la page :
`http://support.sas.com/onlinedoc/913/docMainpage.jsp`

SAS est utilisé pour l'accès, la gestion, l'analyse et la présentation des données. Ces données peuvent être aussi bien des nombres que des chaînes de caractères alphanumériques. Ainsi, si vous souhaitez étudier les caractéristiques physiques d'un ensemble d'élèves, vous pourriez enregistrer pour chacun d'eux son nom, son sexe, son âge, sa taille et son poids. Le tableau suivant illustre ce fichier de données :

NOM	SEXE	AGE	Taille	Poids
Albert	M	14	155	52.5
Marc	M	13	145	50.0
Louis	M	15	132	35.5

Chacune des informations que vous avez enregistrées - sexe de Marc, poids de Louis, taille de Albert, ... - est une valeur. La valeur d'une donnée est une simple mesure : la taille d'un élève, le poids d'un élève, ... Les informations concernant chaque élève - nom, âge, sexe, taille, poids - forment une observation. Chaque ligne du tableau constitue une observation. Une observation est un ensemble de valeurs concernant un même individu. Les valeurs contenues dans une colonne du tableau constituent une variable. Une variable est un ensemble de valeurs concernant une même caractéristique, comme le poids d'une personne,... Les variables SAS peuvent être de type numérique ou alphanumérique. SAS identifie les variables par leur nom. Un nom de variable SAS peut comporter de un à huit caractères, et doit commencer par une lettre ou le caractère souligné (_). Les autres caractères doivent être des lettres, des chiffres ou le caractère souligné. Un nom de variable ne peut pas comporter de caractères blancs (espace). Il est conseillé d'utiliser des noms de variables rappelant leur contenu : NOM, AGE, POIDS,... plutôt que V1, V2, V3,... Une valeur manquante représente une valeur de donnée manquante ou non disponible. Elles sont représentées par des blancs ou des points, en fonction de la méthode de saisie et de lecture des données.

Ecriture de table SAS Un programme SAS comprend généralement une suite d'étapes, ces étapes pouvant être de type DATA (données) ou PROC (procédures). Chaque étape est elle-même constituée d'une suite d'instructions. Chaque instruction SAS doit obligatoirement se terminer par le caractère point-virgule (;). L'oubli de ce point-virgule constitue l'erreur la plus fréquente lorsque l'on commence à rédiger des programmes SAS. Les instructions SAS peuvent débuter à n'importe quelle colonne dans la ligne, et une instruction peut occuper une ou plusieurs lignes, pour autant que les noms de variables ou autres mots ne soient pas coupés. Un programme SAS est cependant plus lisible si une instruction occupe une ligne. Il est également autorisé d'insérer des lignes blanches, de façon à rendre le programme plus lisible. SAS n'est pas sensible à la mise en majuscules, de sorte qu'un programme SAS peut être introduit indifféremment en majuscules ou en minuscules.
On construit une table SAS avec les instructions suivantes :

```
/* exemple1.sas : lecture de données introduites dans le programme SAS */
DATA eleves;
INPUT nom $ sexe $ age taille poids;
CARDS;
Albert M 14 155 52.5
Marc M 13 145 50.0
Louis M 15 132 35.5
Valérie F 12 126 28.7
Mélanie F 14 138 35.2
;
RUN;
```

Chaque instruction de ce programme peut être décrite comme suit :

`/* exemple1.sas : lecture de données introduites dans le programme SAS */`

Cette instruction est un commentaire, qui sera ignoré par SAS. Les commentaires peuvent être introduits librement dans tout programme SAS, afin d'en améliorer la lisibilité et d'en permettre la documentation.

`DATA eleves;`
Une instruction DATA demande à SAS de lire des données et de les organiser en fichier SAS. Cette instruction comprend le mot-clé DATA et un nom-de-fichier choisi par l'utilisateur, dans ce cas-ci eleves.

`INPUT nom $ sexe $ age taille poids;`
L'instruction INPUT fournit l'information à SAS pour organiser les données en fichier SAS. Elle débute par le mot-clé INPUT suivi de la liste des noms de variables, dans ce cas-ci nom, sexe, age, taille et poids. Les variables nom et sexe sont suivies de $, pour indiquer que ces variables contiennent des valeurs alphabétiques. Les autres variables sont numériques. Cette forme de l'instruction INPUT en constitue la forme la plus simple, elle suppose que les valeurs des variables sont séparées par un blanc ou un espace minimum, et que les valeurs manquantes sont indiquées par un point. L'instruction CARDS indique que des lignes de données suivent, un point-virgule indique la fin des lignes de données. L'instruction RUN demande à SAS d'exécuter les instructions précédentes.

☞ Q1 Construire la table vehicule, issue de la partie TD.

Affichage d'une table SAS `PROC PRINT DATA=eleves;`
`RUN;`

L'instruction PROC PRINT demande à SAS d'imprimer les données. PRINT est une procédure SAS, une procédure étant un programme compris dans le logiciel SAS qui analyse et traite des données. Une instruction PROC comprend le mot-clé PROC, le nom de la procédure, dans ce cas-ci PRINT, et, si nécessaire, des options fournies par l'utilisateur, comme par exemple dans ce cas-ci, DATA=eleves. L'option DATA= spécifie le nom du fichier de données. Notez que, sans option DATA=, le système SAS lit automatiquement le dernier fichier SAS créé. L'option DATA= vous permet de changer cette option et de spécifier un fichier de données à votre choix. L'option DATA= permet aussi d'éviter les confusions lorsque vous travaillez avec deux ou plusieurs fichiers de données. L'instruction RUN indique que l'instruction PROC PRINT précédente est prête à être exécutée.

☞ Q2 Afficher la table vehicule.

2.2 Étapes DATA

On peut créer une nouvelle table en partant de la table eleve avec une étape data. La commande utile est alors SET :
DATA eleves2 ;
SET eleves ;
RUN ;

On peut créer de nouvelle variable simplement en en donnant la définition : pour créer la variable IMC on ajoute alors avant la commande RUN la ligne suivante :
IMC=Poids/(taille/100)**2 ;
On peut également effacer une colonne à l'aide de la commande DROP, ou effacer toutes les colonnes sauf une à l'aide de la commande KEEP. Pour travailler sur les lignes, on peut utiliser la commande IF. Pour ne garder que les lignes correspondant à des garçons, on exécute :
DATA garcons ;
SET eleves ;
IF sexe='M' ;
RUN ;

On peut également fusionner des tables. Si l'on veut coller deux tables l'une sous l'autre, il suffit d'utiliser SET avec 2 tables :
DATA elevesdouble ;
SET eleves eleves2 ;
RUN ;

Pour fusionner deux tables l'une à côté de l'autre, c'est plus compliqué. Par exemple si l'on dispose d'une table Coulyeux avec la couleur des yeux et le nom des élèves, on peut vouloir créer une grande table Grandetable avec toutes les informations. On utilise alors la commande MERGE :
DATA grandetable ;
MERGE eleves coulyeux ;
BY nom ;
RUN ;

Attention ! Il faut alors que les tables aient été préalablement triées par le nom. Pour cela il faut utiliser une procédure, la PROC SORT.

2.3 Étapes PROC

On a entrevu la PROC PRINT. Elles commencent toutes par PROC suivi du nom et terminent par RUN ; L'option principale est souvent le nom de la table à traiter.
PROC PRINT DATA=eleves ;
RUN ;

La PROC sort à la même syntaxe, avec une ligne supplémentaire permettant d'utiliser la commande BY, qui indique bien sûr pour quelle variable il faut trier.
PROC SORT DATA=eleves ;
BY nom ;
RUN ;

La PROC MEANS donnent de nombreuses statistiques descriptives sur la table considérée. La procédure décrit encore plus précisément une variable, spécifiée à l'aide de la commande VAR :
PROC UNIVARIATE DATA=eleves ;
VAR age ;
RUN ;

La procédure GPLOT La procédure GPLOT permet de dessiner des figures. Syntaxe :
PROC GPLOT < options > ;
PLOT (y1 y2)*(x1) / OVERLAY ;
RUN ;

L'option la plus courante est toujours de spécifier la table à utiliser : DATA=nomdetable. On donne ensuite les variables a mettre en ordonnées et en abscisses. Par defaut, SAS dessine un graphe par paire de variables, l'option OVERLAY permet de mettre les données dans le meme graphe. Pour dessiner plusieurs courbes, il faut préalablement utiliser la commande SYMBOL pour chaque courbe.
Exemple :
SYMBOL1 COLOR=RED
INTERPOL=JOIN
VALUE=DOT ;

previent SAS que la première courbe est à tracer en rouge, en joignant les points et en indiquant les observations par un point. La commande
SYMBOL2 INTERPOL=R ;

joint les points à l'aide de la régression linéaire. La procédure servant à faire une régression linéaire est GLM.
☞ Q3 Représentez graphiquement les données de la table vehicule.

2.4 La procédure GLM

La procédure GLM posséde des options innombrables, nous donnons ici les principales.
PROC GLM < options > ;
MODEL Y=var_explicatives < / options > ;
BY variables ;
OUTPUT < OUT=table_sortie > ;
RUN ;

La commande MODEL permet de donner le modèle, la variable expliquée à gauche et les explicatives à droite. L'option la plus courante est NOINT qui permet d'indiquer que l'on ne veut pas de constante. BY permet de faire une analyse pour différentes sous-population. nécessite que la table soit ordonnée (PROC SORT). La commande OUTPUT permet de sauvegarder certains resultats (et les données utilisées) dans une table sas.
Exemple
PROC GLM DATA=eleves ;
BY sexe ;
MODEL poids=taille age ;
OUTPUT OUT=sortie1 P=poidschap R=residu ;
RUN ;

La procédure GLM produit une grande quantité de résultats, affichés dans la fenêtre SORTIE, et éventuellement enregistrés à l'aide de la commande OUTPUT. SAS donne en particulier les p-value pour le modèle global et pour chaque paramètre, c'est-à-dire la probabilité que les paramètres soient nuls, tous ensemble et un par un.
On obtient alors la sortie suivante : The GLM Procedure
Dependent Variable : poids

Source	DF	Somme des carrés	Carré moyen	F	Pr > F
Model	2	279.2825014	139.6412507	7.68	0.0071
Error	12	218.3174986	18.1931249		
Corrected Total	14	497.6000000			

Le premier tableau donne

– 1ère colonne : le degré de liberté.
– 2ème colonne : la variance expliquée, la variance résiduelle et la variance totale.
– 3ème colonne le rapport des 2 précédentes.
– 4ème et 5ème colonnes : un test pour l'absence totale d'effet du modèle.

R-carré	Coeff Var	Racine MSE	poids Moyenne
0.561259	6.623199	4.265340	64.40000

Le deuxième tableau donne
– 1ère colonne : le R^2.
– 2ème colonne : un indicateur de la variance de Y.
– 3ème colonne : un estimateur de l'écart-type du bruit.
– 4ème colonne : la moyenne de Y.

Source	DF	Type I SS	Carré moyen	F	Pr > F
age	1	143.4243243	143.4243243	7.88	0.0158
taille	1	135.8581771	135.8581771	7.47	0.0182

Source	DF	Type III SS	Carré moyen	F	Pr > F
age	1	42.2541021	42.2541021	2.32	0.1534
taille	1	135.8581771	135.8581771	7.47	0.0182

Les troisième et quatrième tableaux testent les effets des variables explicatives, en les ajoutant dans l'ordre spécifié pour le type 1 et une par une pour le type 2.

Paramètre	Estimation	standard	Valeur du test t	Pr > |t|
Intercept	-71.39154324	37.87490351	-1.88	0.0839
age	1.05889043	0.69481668	1.52	0.1534
taille	0.64284833	0.23524439	2.73	0.0182

Le dernier tableau donne
– 1ère colonne : l'estimateur de theta.
– 2ème colonne : un estimateur de l'écart-type de cet estimateur.
– 3ème et 4ème colonnes : le test de Student de l'effet de chaque variable et la p-value associée.

☞ Q4 Appliquez la procédure GLM aux données vehicule. Interprétez les résultats.

☞ Q5 Utilisez la table de sortie pour afficher sur un même graphique les données (X_{i1}, Y_i), la droite de régression linéaire et les prévisions $(X_{i1}, \hat{Y}_i)$.