URCA Hugo Harari-Kermadec

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

Économétrie 1 : le modèle linéaire, suite

1 TD : Fonction de production

Nous disposons pour n entreprises, des valeurs du capital K_i, de la valeur ajoutée VA_i et de l'emploi L_i. Nous supposons que la fonction de production de ces entreprises est du type Cobb-Douglas :

$$VA_i = \alpha L_i^{\lambda} K_i^{\gamma}$$

☞ Q1 Proposez un modèle linéaire pour traiter ce problème. Écrivez ce modèle sous forme matricielle : $Y = X\theta + \mathcal{E}$, en précisant les dimensions des matrices et des vecteurs.

☞ Q2 Que vaut l'estimateur des moindres carrés ordinaires $\hat{\theta}$? Que vaut sa matrice de variance-covariance $Var(\hat{\theta})$? Proposez un estimateur $\widehat{\sigma^2}$ de la variance des résidus ? Proposez un estimateur de $Var(\hat{\theta})$

Pour 1658 entreprises, nous avons obtenu à l'aide de l'estimation MCO de θ :

$$\log VA_i = 3.136 + 0.738 \log L_i + 0.282 \log K_i$$

avec $R^2 = 0.945$, $SCR = \sum_i \hat{\varepsilon}_i^2 = 148.27$. Nous donnons également

$$(X'X)^{-1} = \begin{pmatrix} 0.0288 & 0.0012 & 0.0034 \\ 0.0012 & 0.0016 & 0.0010 \\ 0.0034 & 0.0010 & 0.0009 \end{pmatrix}$$

☞ Q3 Calculez une estimation non biaisée $\widehat{\sigma^2}$. En déduire une estimation de $Var(\hat{\theta})$. Rappelez la définition du R^2. En déduire la valeur de la variance empirique de la variable $\log VA$.

☞ Q4 Donnez un intervalle de confiance de niveau 95 % pour les paramètres α et λ. Pour un niveau de 95%, testez .

$$H_0 : \gamma = 0 \text{ contre } H_1 : \gamma > 0.$$

☞ Q5 Nous souhaitons tester l'hypothèse selon laquelle les rendements d'échelle sont constants. Nous rappelons qu'une fonction de production F est à rendements d'échelle constants si $\forall t > 0$, nous avons $F(tL, tK) = tF(L, K)$. Quelle contrainte vérifient les paramètres du modèle lorsque les rendements d'échelle sont constants ? Pour un niveau de 95%, réaliser le test de H_0 : "rendements d'éechelle constants" contre H_1 : "rendements d'échelle croissants".

☞ Q6 Nous nous intéressons à la fonction de production dans trois secteurs.
T07 : minerais et métaux ferreux, première transformation de l'acier,
T15B : fabrication de biens d'équipement ménager,
T18 : industries textiles et habillement.
Les trois secteurs sont-ils régis par la même fonction de production ? pour le savoir, nous avons estimé notre modèle par la méthode MCO sur chacun des trois secteurs, puis sur l'ensemble des entreprises :
T07 : : n = 40 et SCR = 2.57,
T15B : n = 24 et SCR = 1.17,
T18 : n = 238 et SCR = 17.8, Ensemble : n = 302 et SCR = 41.12 .
Pour un niveau de 95%, tester l'hypothèse selon laquelle la fonction de production est la même dans les trois secteurs.

2 TP : Proc REG

2.1 Syntaxe

La syntaxe et l'usage de la Proc REG sont très proches de ceux de la PROC GLM. La principale différence est la possibilité d'étudier plusieurs modèles en même temps.
PROC REG DATA=eleves ;
MODEL poids=age taille/ NOINT ;
MODEL poids=age taille / <options> ;
RESTRICT intercept-age ;
RUN ;

Dans le premier modèle, on considère deux variables explicatives (age et taille) et on impose l'absence de constante grâce à l'option NOINT.
La ligne RESTRICT permet de contraindre le second modèle. Dans cet exemple, les coefficients de la constante et de l'âge doivent être les mêmes.

2.2 Sortie

On obtient une sortie très proche de celle de la proc GLM, **pour chaque modèle**. On donne ici la sortie pour le premier modèle :
Analyse de variance

Source	DF	Somme des carrés	Carré moyen	Valeur F	Pr > F
Model	2	279.28250	139.64125	7.68	0.0071
Error	12	218.31750	18.19312		
Corrected Total	14	497.60000			

Le premier tableau est le même que pour la proc GLM.

Root MSE	4.26534	R-Square	0.5613
Dependent Mean	64.40000	Adj R-Sq	0.4881
Coeff Var	6.62320		

Dans le second tableau, on retient surtout le R^2 et **sa version ajustée** *Adj* R^2. Le troisième tableau est le même que pour la proc GLM.

2.3 Comparaison de modèles

On peut comparer des modèles en utilisant plusieurs fois la commande MODEL. Il faut ensuite choisir un critère de comparaison des modèles. On a vu en cours que le R^2 n'est pas un bon critère. Le R^2 ajusté est meilleur est vaut $ADJRSQ = 1 - \frac{(n-1)(1-R^2)}{n-p}$ où p est le nombre de variables explicatives. Pour que le R^2 ajusté soit calculé, il faut le spécifier en option de la commande MODEL. On spécifie en option qu'on veut le calculer et on créé une table stat contenant les statistiques par modèles, à l'aide de l'option globale OUTEST que l'on va afficher.
PROC REG DATA=eleves OUTEST=stat ;
MODEL poids=taille age / ADJRSQ ;
MODEL poids=taille / ADJRSQ ;
RUN ;

PROC PRINT DATA=stat ;
RUN ;

On obtient la sortie suivante :

obs	model	type	depvar	RMSE	intercept	taille	age	poids	R^2	ADJR^2
1	MODEL1	PARMS	poids	4.59083	-71.5712	0.78484	-0.19899	-1	0.49174	0.40703
2	MODEL2	PARMS	poids	4.47705	-72.2368	0.78227	.	-1	0.47634	0.43606

On se concentre sur les 2 dernières colonnes : le R^2 et le R^2 ajusté. On voit ici qu'en retirant l'âge, le R^2 diminue, puisque l'on passe à un modèle de dimension plus petite. Mais l'effet de l'âge étant peut significatif (le test de nullité à une p-value qui dépasse les 15%) on gagne en fait à retirer cette variable explicative. C'est ce qu'indique le R^2 ajusté.

3 TP : pollution

On veut étudier l'effet de la pollution sur la mortalité. On dispose d'une table pollution.txt avec, pour 60 villes américaines, les colonnes suivantes :

PREC	précipitation moyenne annuelle
JANT	température moyenne en janvier (en degré Fahrenheit)
JULT	température moyenne en juillet (en degré Fahrenheit)
OVR65	% de la population agé de plus de 65
POPN	taille moyenne des ménages
EDUC	niveau moyen d'éducation des adultes
HOUS	% des ménages équipés tout-confort
DENS	densité de population
NONW	% de non blancs
WWDRK	% d'employés dans un travail de bureau
POOR	% de familles pauvres
HC	pollution au monoxyde de carbone
NOX	pollution aux oxydes d'azote
SO	pollution au dioxyde de souffre
HUMID	Humidité moyenne sur l'année
MORT	Mortalité totale pour 100 000 personnes

☞ Q1 Enregistrez le ficher ronfle.dbf dans votre répertoire. Importez dans SAS le jeu de données Ronfle.dbf : menu "fichier" / "importer des données", spécifiez qu'il s'agit d'un fichier dBase et chercher le ficher. Commencez par faire une copie "save_ronfle" (pour être tranquille) à l'aide d'une étape DATA.

☞ Q2 On cherche à expliquer la mortalité Y à l'aide d'un modèle linéaire. Proposez un premier modèle avec du sens pour expliquer Y à l'aide des autres variables. Vous êtes fortement incités à vous renseigner sur le sujet avant de proposer un modèle. Expliquez votre choix.

☞ Q3 Utilisez SAS pour étudier ce modèle, et commentez les résultats.

☞ Q4 Essayez d'améliorer le modèle est ajoutant ou en retirant des variables, en utilisant les options éventuellement utiles.

Le tout ne doit pas faire plus de 5 pages. Ne mettez pas plus de 5-6 lignes de codes en tout. Vous pouvez rajouter un ou deux graphiques. C'est vos initiatives et votre analyse qui seront jugés, pas les sorties SAS ou le volume du projet.
Envoyez votre projet par groupe de 2 avant le lundi 26/01 à 23h59.