Préparation à l'agrégation externe de Sciences Sociales

Statistique descriptive

2023-2024

Exercice 1

1. Si $Q = AL^\alpha K^\beta$, l'ajustement du type Cobb-Douglas consiste à trouver des valeurs de A, α, β les mieux adaptés àux donnés statistiques. En passant par ln, on obtient

$$Y = \ln Q = \ln(AL^\alpha K^\beta) = \ln A + \alpha X_1 + \beta X_2,$$

ce qui revient à faire l'ajustement linéaire de Y en fonction de X_1 et X_2.

2. (a) La matrice de variance-covariance M de (x_1, x_2) est une matrice 2×2. Pour $i, j \in \{1, 2\}$

$$M_{i,j} = \mathbf{Cov}(x_i, x_j) = \frac{1}{n}\sum_{k=1}^{n}(x_{i,k} - \overline{x_i})(x_{j,k} - \overline{x_j}) = \frac{1}{n}\sum_{k=1}^{n} x_{i,k}x_{j,k} = \frac{1}{n} x_i \cdot x_j,$$

avec $x_{i,k}$ la k-ème valeur du vecteur x_i et où on a utilisé le fait que x_i est centré.

Avec la calculette on a trouvé $M = \begin{pmatrix} 0.00630 & 0.01177 \\ 0.01177 & 0.02248 \end{pmatrix}$ On en déduit que $B = {}^t xx = nM = \begin{pmatrix} 0.0630 & 0.1177 \\ 0.1177 & 0.2248 \end{pmatrix}$.

(b) La calculatrice donne

$$B^{-1} = \begin{pmatrix} 727.2 & -380.8 \\ -380.8 & 203.8 \end{pmatrix}.$$

A part le fait que la calculatrice sait inverser la matrice B, il y a plusieurs justifications que B est inversible à priori.

i. Le déterminant de B est $0.063 \times 0.225 - 0.118^2 \neq 0$. Donc B est inversible.

ii. Le rang de B est 2 car les vecteurs colonnes $\begin{pmatrix} 0.063 \\ 0.118 \end{pmatrix}$ et $\begin{pmatrix} 0.118 \\ 0.225 \end{pmatrix}$ sont linéairement indépendants. Rappelons qu'une famille de deux vecteurs ne soit pas une famille linéairement indépendante si et seulement si les deux vecteurs sont colinéaires (l'un est multiple de l'autre). Une matrice carrée de taille k, k est inversible si et seulement si son rang est égal à k.

iii. Pour les matrices obtenues comme $B = {}^t XX$, le rang de B est égal au rang de x. Puisque le nombre d'individus est beaucoup plus grand que le nombre de variables explicatives, donc le nombre de ligne beaucoup plus grand que le nombre de colonnes de x. Le rang de x est facilement son rang nombre de colonnes. Donc B est de rang maximal.

3. (a) L'estimateur donné par la méthode des moindres carrés est $\hat{a} = B^{-1}\,{}^t xy$, ${}^t xy$ est un vecteur colonne de 2 coefficient, en haut $x_1 \cdot y = 0.132$ et en bas $x_2 \cdot y = 0.251$. On trouve que

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 0.832 \\ 0.679 \end{pmatrix}$$

La constante A est choisie de sorte que

$$\overline{Y} = \ln A + \alpha\overline{X_1} + \beta\overline{X_2},$$

donc

$$\ln A = \overline{Y} - \alpha\overline{X_1} - \beta\overline{X_2}.$$

Comme y est obtenu par Y en retranchant $\overline{Y}$, on peut obtenir facilement $\overline{Y}$ en regardant la différence entre une valeur de y et Y. On a $\overline{Y} = 11.9$, $\overline{X_1} = 9.14$ et $\overline{X_2} = 12.35$. Donc

$$\ln A = -4.090 \Rightarrow A = 0.0167.$$

On a ainsi l'ajustement du type Cobb-Douglas

$$Q = 0.0167L^{0.832}K^{0.679}.$$

(b) Les coefficients α et β correspondent à la répartition des revenus entre le travail et le capital.

(c) Les rendements d'échelle représentent l'accroisement d'efficience avec l'augmentation de facteurs de production.

La fonction de production du type Cobb-Douglas, $Q(K, L) = AL^{\alpha}K^{\beta}$, $Q(K, L)$ est $(\alpha + \beta)$-homogène, i.e.

$$Q(aK, aL) = A(aK)^{\beta}(aL)^{\alpha} = a^{\alpha+\beta}Q(K, L)$$

K et L étant des facteurs de production, typiquement capital et travail, et $a \geq 1$ étant le facteur d'échelle.

Donc elle possède des rendements d'échelle :

i. constants si $\alpha + \beta = 1$,

ii. croissants si $\alpha + \beta > 1$ (notre cas ici)

iii. décroissants si $\alpha + \beta < 1$.

Exercice 2

Expliquons d'abord la signification des coefficients de ces tableaux :

Nous disposons d'une statistique de 12 individus s_i, dont chaqu'un a cinq valeurs, respectivement $CA(i), AM(i), PTT(i), RES(i), CS(i)$.

— **Table 1 : Matrice des corrélations**

Les coefficients sont le coefficient de corrélation observée, qui est symétrique donc on ne donne que les valeurs du triangle inférieur. Par exemple pour la case AM, CA, le coefficient est calculé

$$\frac{\mathbf{Cov}(AM, CA)}{\sqrt{\mathbf{Var}(AM)\mathbf{Var}(CA)}} = \frac{\sum_{i=1}^{12}(AM_i - \overline{AM})(CA_i - \overline{CA})}{\sqrt{\sum_{i=1}^{12}(AM_i - \overline{AM})^2}\sqrt{\sum_{i=1}^{12}(CA_i - \overline{CA})^2}}.$$

Si la variance observée des quantités vaut 1, alors la matrice de corrélation coincide avec la matrice de variance-covariance. Pour faciliter l'interprétation et donner la même importance à chaque facteur, on renormalise les donnés de sorte que la variance soit 1, et considère la matrice de corrélation comme matrice de variance-covariance. Comme on étudie la dispersion de la distribution de ses individus, on peut aussi centrer (ou pas) les variables facteurs.

— **Table 2 : Valeurs propres**

Cela correspond au spectre de la matrice de corrélation classé dans l'ordre décroissant. Comme la matrice M est symétrique, elle est diagonalisable dans une base orthonormée. C'est à dire, on peut trouver une base $(C_1, \cdots, C_5)$ composé de 5 vecteurs propres, telles que :

$$C_i \cdot C_j = 1_{i=j} \text{ et } \mathbf{Cov}(C_i, C_j) = \lambda_j 1_{i=j}.$$

Les droites engendrés par C_i sont appelés des axes factoriels.

— **Table 3 : Composantes principales**

La ligne de s_1 dans ce tableau donne la décomposition de s_1 dans la base $(C_1, C_2, C_3, C_4, C_5)$.

$$s_1 = -0.648C_1 - 0.967C_2 + 0.090C_3 - 0.426C_4 + 0.434C_5.$$

Les coefficients peuvent être obtenus en utilisant le produit scalaire :

$$-0.648 = s_1 \cdot C_1.$$

— **Table 4 : Corrélations entre variables initiales et composantes principales.**

Par exemple

$$0.979 = \text{corrélation entre } CA \text{ et } C_1 = \frac{\mathbf{Cov}(CA, C_1)}{\sqrt{\mathbf{Var}(CA)\mathbf{Var}(C_1)}}$$

On peut vérifier que dans chaque ligne, la somme de carré des coefficients fait 1.

— **Table 5 : Norme au carré et Qualité de représentation sur le premier plan factoriel**

Le premier plan factoriel est le plan engendré par C_1 et C_2. La représentation en nuage de points est la figure à gauche, en prenant uniquement les deux premiers coefficients dans le tableau de composante principales. Les deux composantes principales correspondent aux deux premières directions orthogonales, où la dispersion est plus marquée. Et la représentation dans le premier plan factoriel correspondent à la projection orthogonale des points dans $\mathbb{R}^5$ dans le sous-espace vectoriel engendré par C_1 et C_2.
Pour s_1, la norme au carré est

$$\begin{aligned}\|s_1\|^2 &= \|-0.648C_1 - 0.967C_2 + 0.090C_3 - 0.426C_4 + 0.434C_5\|^2 \\ &= 0.648^2 + 0.967^2 + 0.090^2 + 0.426^2 + 0.434^2,\end{aligned}$$

car (C_i) est une base orthonormée. La qualité de représentation de s_1 est

$$\frac{0.648^2 + 0.967^2}{\|s_1\|^2} = 78.3\%.$$

— **Représentation des individus dans le premier plan factoriel** Le premier plan factoriel est le plan P engendré par C_1 et C_2. On représente les points s_i dans le premier plan factoriel, qui sont projection orthogonale des s_i dans P.
Rappelons que la projection orthogonale $\pi : \mathbb{R}^5 \to P$ prend un point $x \in \mathbb{R}^5$ envoie sur le point y qui vérifie les conditions équivalentes suivantes :

1. $y = \arg\min_{z \in P} \|x - z\|$. y est le point de P qui minimise la distance entre y et x.
2. $y \in P$ et $x - y \perp P$.

Donc $\pi(s_1) = -0.648C_1 - 0.967C_2$ on oublie tout simplement la partie $0.090C_3 - 0.426C_4 + 0.434C_5$. Car C_3, C_4, C_5 sont tous orthogonaux à C_1 et C_2, ainsi que leur combinaision linéaire.

— **Cercle de corrélation**
Le cercle de corrélation est obtenu en plaçant un point pour chaque facteur de coordonnés corrélations entre variables initiales et deux premières composantes principales, donc pour CA, on met un point de coordonnés $(0.979, 0.177)$.
Les règles de lecture du cercle des corrélations sont :
— On ne prend en compte que les variables proches du cercle des corrélations. Dans le cas contraire, la variable est non corrélée à la composante principale et est donc mal représentée.

— La liaison entre variables bien représentées s'analyse à travers la direction et le sens de leur vecteur :
  — si les vecteurs ont même direction et même sens, les variables sont corrélées positivement,
  — si les vecteurs ont même direction mais de sens contraire, les variables sont corrélées négativement,
  — si les vecteurs sont perpendiculaires, les variables sont non corréles.
— On synthétise chaque axe en précisant les variables qui contribuent le plus en positif ou en négatif.

Bien sûr ceci reste approximative puisque dans le premier plan factoriel ceci n'est qu'une représentation partielle.

Correction des questions :

1. L'inertie de l'axe de C_i de la statistique est sa valeur propre associée. L'inertie totale est la somme de l'inertie dans les cinq directions qui vaut 5, qui est aussi la trace de la matrice dont on a diagonalisé. Les pourcentages sont

1	2	3	4	5
55.9%	26.1%	15.4%	2.58%	0.05%

2. Le pourcentage de l'inertie expliqué est 82%.
3. La matrice des composantes principales donnes les coordonnés de s_i dans la décomposition dans la nouvelle base (C_j).
4. Voir explication.
5. C'est déjà calculé dans l'explication.

Exercice 3 (Calculatrice)

On considère les donées suivantes concernant la consommation d'eau chaude en litres et la température sur une période de 5 jours.

1. $\overline{X} = -0.6$, $\overline{Y} = 31.8$.
2. La droite de régression est : $Y = -1.48484848X + 30.90909090909091$