STATISTIQUES DESCRIPTIVES MULTIVARIÉES

SIMON COSTE ET PIERRE MONTAGNON

Exercice 1. Trois enfants de 7, 11 et 12 ans ont pour tailles respectives 117, 143 et 144 centimètres.

(1) Réaliser la régression linéaire de la variable `taille` par la variable `âge` sur l'échantillon dont on dispose.

(2) Calculer le R^2 de la régression et commenter.

Exercice 2. On cherche à étudier une éventuelle relation entre les bénéfices annuels B_i d'une entreprise i et son chiffre d'affaire C_i. On dispose pour cela des données pour $n = 68$ entreprises.

	Moy.	Méd.	Écart-type	Asym.	Erreur d'asym.	Min.	Max.
B_i	0,2236	0,2150	3, 07404	-6,726	0,291	-23,30	6,26
C_i	15,7096	10,4950	18,91607	2,619	0,291	0,48	102

Par ailleurs, on a $\frac{1}{n}\sum_{i=1}^{n} C_i B_i = 7,0908$.

(1) Calculer le coefficient de variation des variables B et C. Que peut-on en déduire sur la dispersion des valeurs de C, puis de B ?

(2) On cherche à étudier le modèle linéaire

$$B_i = \alpha C_i + \beta + \epsilon_i. \tag{1}$$

Écrire et démontrer les formules des estimations $\hat{\alpha}, \hat{\beta}$ des coefficients de régression linéaire, puis donner la valeur de ces estimateurs. En déduire la valeur du coefficient de corrélation linéaire entre ces variables.

(3) Le modèle linéaire est-il acceptable ? Effectuer un test au niveau de 5%.

Exercice 3. On souhaite mettre en évidence une corrélation entre le temps passé chaque jour devant la télévision (`time_tv`, en heures) et le taux de cholestérol (`cholesterol`, en mmol par litre de sang).

(1) Rappeler les hypothèses du modèle linéaire gaussien dans le cas d'une variable explicative et d'une variable expliquée.

(2) Énoncer les propriétés des estimateurs des coefficients du modèle linéaire gaussien par la méthode des moindres carrés.

(3) Commenter en détail les deux lignes inférieures du tableau de résultats suivant :

Date: 6 Avril 2021.

```
. regress cholesterol time_tv
```

Source	SS	df	MS		
Model	5.04902329	1	5.04902329	Number of obs	= 100
Residual	28.3220135	98	.289000137	F(1, 98)	= 17.47
				Prob > F	= 0.0001
Total	33.3710367	99	.337081179	R-squared	= 0.1513
				Adj R-squared	= 0.1426
				Root MSE	= .53759

cholesterol	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
time_tv	.0440691	.0105434	4.18	0.000	.0231461	.0649921
_cons	-2.134777	1.813099	-1.18	0.242	-5.732812	1.463259

(4) Proposez quelques pistes pour améliorer le R^2 de la régression.

Exercice 4. On dispose pour 55 États du monde des valeurs pour l'année 2003 des variables suivantes :

— Le budget de l'État `Betat`
— Le budget militaire `Bmilit`
— Le budget de la défense `DefBud`
— Les dépenses d'éducation `DepEduc`
— Le budget de la santé `Bsante`
— La population active `PA`
— Le PNB par habitant `pnbpop`

La planche ci-dessous recense les résultats d'une analyse en composantes principales et d'une régression linéaire menées sur ces données.

(1) Commenter le tableau 1 et faire le lien avec le R^2 de la régression linéaire représentée dans le graphique 4. Que devient le R^2 de la régression si l'on ajoute toutes les variables dont on dispose parmi les variables explicatives ? Pourquoi une telle manœuvre n'est-elle pas nécessairement habile ?

(2) Que sont les variables `prin1`, `prin2` et `prin3` données dans le tableau 3 ? À quoi correspondent les coefficients représentés en première ligne de chaque cellule ?

(3) Donner une interprétation des variables `prin1`, `prin2` et `prin3`.

ANNEXES : **Analyse en composantes principales**

(Suite 8)

I) Matrice des corrélations entre les variables initiales (TABLEAU 1)

	BEtat	DefBud	Dep. Educ	BSante	BMilit	PA	pnbpop
BEtat	1.0000	-.2680	0.0593	0.4216	-.1210	0.2386	0.4324
DefBud	-.2680	1.0000	-.0123	0.3131	-.0615	0.1695	0.4004
DepEduc	0.0593	-.0123	1.0000	-.0732	-.0529	0.1071	0.1219
BSante	0.4216	0.3131	-.0732	1.0000	-.1467	0.4029	0.7933
BMilit	-.1210	-.0615	-.0529	-.1467	1.0000	-.0865	-.1772
PA	0.2386	0.1695	0.1071	0.4029	-.0865	1.0000	0.4876
pnbpop	0.4324	0.4004	0.1219	0.7933	-.1772	0.4876	1.0000

II) Valeurs propres et inertie (TABLEAU 2)

	Eigenvalue	Difference	Proportion	Cumulative
1	2.58379066	1.30775018	0.3691	0.3691
2	1.27604048	0.23744373	0.1823	0.5514
3	1.03859675	0.09235791	0.1484	0.6998
4	0.94623884	0.28016800	0.1352	0.8350
5	0.66607084	0.34360652	0.0952	0.9301
6	0.32246432	0.15566621	0.0461	0.9762
7	0.16679811		0.0238	1.0000

III) Matrice des corrélations entre les variables initiales et les composantes principales : (TABLEAU 3)

	Prin1	Prin2	Prin3
BEtat	0.54188 <.0001 47	-0.71809 <.0001 47	-0.17920 0.2281 47
DefBud	0.39486 0.0060 47	0.83989 <.0001 47	0.10698 0.4741 47
DepEduc	0.10157 0.4969 47	-0.19516 0.1886 47	0.91727 <.0001 47
BSante	0.87322 <.0001 47	0.04999 0.7386 47	-0.23988 0.1044 47
BMilit	-0.28018 0.0565 47	0.09522 0.5244 47	-0.28109 0.0556 47
PA	0.65313 <.0001 47	-0.00998 0.9469 47	0.13055 0.3818 47
pnbpop	0.92539 <.0001 47	0.07224 0.6294 47	0.00771 0.9590 47

IV) - Régression linéaire

(GRAPHIQUE 4)