statistique descriptive : >>Indices de position centrale

Moyenne arithmétique : (**caractères quantitatifs. Slmnt)**

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{p} n_i x_i = \sum_{i=1}^{p} f_i x_i$$

Pour caractères quantitatifs continues, On commet une légère erreur en remplaçant chacune des valeurs modalités par son centre de classe (CC)

Médiane :

50 % des valeurs lui sont inférieures ; 50 % des valeurs lui sont supérieures

Valable sur caractères quantitatifs et qualitatifs Ordinaux

Les valeurs exceptionnelles ne l'affectent pas; est qualifiée d'estimateur robuste

Médiane d'une variable continue :

On trouve la classe médiane (ayant Fi absolu cumulé ~N/2)

une interpolation linéaire à l'intérieur de la classe médiane :

Q2 = Binf + [(N/2 – F) / fMe] * E

- Binf : est la borne inférieure de la classe médiane ;
- F : la somme des fréquences absolues de toutes les classes précédant la classe médiane
- fMe : la fréquence absolue de la classe médiane
- E : l'étendu de la classe médiane

Médiane d'une variable discrète (Cas de valeurs répétées, eg : âge employé) :

Tri des val=>calcul effectifs cumulés Fi(%)=> qui corresponde à Fi = 50%

Médiane d'une variable discrète (Cas de valeurs distincte, T° paris par mois) :

Classer les n données dans l'ordre croissant

Si N est impair alors Q2 = x(k) avec k = (N+1)/2

Si N est pair alors Q2 = (x(k) + x(k+1))/2 avec k = N/2

Q1 = la valeur en dessous de laquelle se trouvent 25% des observations inférieurs

Q3 = la valeur en dessous de laquelle se trouvent 75% des observationsinférieures

Mode : la valeur du caractère la plus fréquente

Variance : Sert à Caractériser de façon globale l'écart plus ou moins important de l'ensemble des valeurs de la distribution par rapport à la valeur moyenne.

$$\boxed{S^2 = \frac{1}{N}\sum_{i=1}^{k} f_i (x_i - \overline{x})^2 \qquad f_i = \frac{n_i}{N}}$$

$$\boxed{S = \left(\frac{1}{N}\sum_{i=1}^{k} f_i (x_i - \overline{x})^2\right)^{1/2}}$$

Dans le cas d'une variable continue groupée en classes on utilise les centres de classes à la place des xi

L'écart interquartile = comprend 50% des observations, celles qui sont les plus centrales, l'espace compris entre les quartiles 1 et 3

EQ = Q3 –Q1

Indice de variabilité (coeff de variation)

$$\boxed{CV = \frac{S}{\overline{X}}(100)}$$

CV proche de 0 => groupe homogène

Hypothèses de regression lineaire comme methode descriptives :

Hypothèse 1 : La relation entre X et Y doit être linéaire;

Hypothèse 2 : le nombre d'observations doit être supérieur au nombre de variables;

Hypothèse 3 : les variables exogènes doivent être linéairement indépendantes.

Hypothèse 4 : Normalité et indépendance des Yi

Hypothèse 5 : Homoscedasticité

Hypothèse 6 : Les résidus doivent être Normaux, indépendants, centrés et non corrélées avec les variables explicatives. ; les résidus standarisé sont dans l'intervalle [-3,3] => variabilité acceptable

Relation entre la consommation et chacune des variables ?

La matrice de corrélation montre que la consommation est fortement corrélé avec les variables indépendantes (coefficient de pearson ~1)

De plus, **les diagrammes de dispersions** laissent penser que l'hypothèse de linéarité semble acceptable

Est-ce que les variables X1,…X4 sont indépendantes ?

Le tableau des variables introduites montrent que toutes les variables ont été introduite, donc l'introduction de chaque variables n'a pas fait passer la tolérance des autres variables au-dessous du seuil de tolérance (0,3) => les variables exogènes sont faiblement colinéaires

Conclusion ?on peut effectuer une regression linéaire comme méthode descriptive/ajustement linéaire

Méthode utilisée : méthode des moindres carrés ordinaire

Qualité du modèle : Récapitulatif du modèle : le R^2 ajusté = 0,948 => il restitue 94,8% d'information de la variabilité initiale

Expression de la relation : voir tableau des coefficients

Qu'est ce qui influence le plus la consommation d'essence d'une voiture :

La puissance ; ayant le plus grande coeff de corrélation et le plus grand beta ;

Rq : Bêta : permet de comparer la contribution de chaque variable ; t :doit être >1.96 pour que le coeff d'une var dans l'équation soit significatif

En analysant les résidus, pensez-vous que le modèle ajuste bien les données :

Oui : Les Résidus suivent une distribution normale (graphique de répartition de résidu par une répartition normale (**PP gaussien**))

Non : **L'histogramme (conso,effectif)** n'est pas symétrique donc la distribution n'est pas stable => nn stablité de Y

Existe de l'information non expliqué ?

Non, **la représentation des prédictions standardisés en fct des résidus standarisé** ne fait apparaitre aucun modèle particulier ce qui confirme l'hypo de val constante de la variance du terme d'erreur (homoscédasticité) et d'indépendance des termes d'erreur.

Interpretation du coefficient de la variable indépendante : Une pente de a implique une augmentation d'une unité en X entrainera une augmentation moyenne de a unité en Y

$$KMO = \frac{\sum_i \sum_j r_{ij}^2}{\sum_i \sum_j r_{ij}^2 + \sum_i \sum_j a_{ij}^2}$$

0,9	Merveilleux	0,6	médiocre
0,8	méritoire	0,5	misérable
0,7	moyen	> 5	inacceptable

Toutes les données ont meme unité =>matrice de covariance

Sinon : centrage&reduction =>matrice de correlation

Analyse factorielle : Analyse en composante principales (ACP)

étape 1 : Examen de la matrice des corrélation

mettre en évidence les relations entre les variables; évaluation des propriétés du modèle factoriel; décider du traitement des valeurs manquantes.

étape 2 : Extraction des facteurs

déterminer le nombre defacteurs requis; choix de la méthode d'extraction des facteurs.

étape 3 Transformation par rotation des facteurs

rendre les facteurs plus interprétables.

étape 4 :Calcul des scores

calcul des coefficients associés à chaque facteur pour servir à d'autres analyses.

1) pertinence de l'analyse factorielle : (s'applique qu'à des variables quantitatives)

Etude de liens entre les liens

-Matrice de corrélation => Variables fortement corrélées=>on est sur de réduire notre espace de n à p<n => un facteur regroupant chaque couple fortement corrélés

-Matrice des corrélations partielle /matrice anti-image : les relations 2à2 sont intrinsèques ? ou bien influencé par les autres var ?

Coeff est bien faible =>bonne réduction=>inter-relation transitant par toutes les variables du modèle

-Cela est aussi par le KMO : proche de 1=>confirme que corrélation partielle faibles

-Test de sphéricité de Bartlett : tester l'hypothèse nulle selon laquelle la matrice des corrélations est égale à la matrice identité.

Sig=0 => on rejete l'hypothèse de d'indépendance=> existence de corrélation ; en effet : nuage indépendant => inscrit ds une sphère

-MSAI (diagonale de la matrice anti-image): Plus msai est élevé et proche de 1, plus la variable correspondante contribue fortement dans la construction des facteurs.

=>coeff partiel de cette var avec les autre vars se trouver faible

-Statistiques descriptives=>variable d'unités différentes=> nécessité de centrage et de réduction (standarisation)

2) Nombre de facteurs à retenir :

Variance totale expliquée : %cumulé est de ~90% = on a restitué presque la totalité de la variance initiale.

Qualité de représentation à p facteur : ns constatons que les vars sont quasiment parfaitement expliquées

Corrélation reproduite : 0% de residu => pas de nécessité de passer à une dimension supérieur

3) nécessité de faire une rotation varimax : Nous remarquons que tt les vars sont corrélés avec le 1èr facteur => ns sommes en présence d'un effet de taille

Matrice des composantes : se lit verticalement : Facteur1(Y1)=a*var2+b*var2+…

Matrice des coefficients des coordonnées des composantes : se lit horizontalement : var1=a*Facteur1+b*Facteur2+…

Interprétation des facteurs si on décide de ne pas faire une rotation : aucune interpretation, problème d'interpretation

4) Interprétation des groupes d'individus : la relation var1=a*Facteur1+b*Facteur2 (a >0) =>(Facteur1 augmente => var1 augmente) ie : variables varient dans le même sens que les axes factoriels ;

La matrice de corrélation (facteurs , var expliquée) : égalera à l'identité => facteur indépendant, non corrélé =>on peut déduire de quel facteur dépend cette var expliquée

La méthode VARIMAX s'applique lorsque la plupart des variables sont représentées sur un seul axe. Il s'agit d'une méthode de rotation orthogonale qui minimise le nombre de variables qui ont des corrélations importantes avec un facteur.

La méthode QUARTIMAX s'utilise lorsqu'une variable est fortement corrélée à plusieurs axes à la fois. C'est une méthode de rotation qui minimise le nombre de facteurs requis pour expliquer le nombre de facteurs.

La méthode EQUAMAX est une combinaison des deux méthodes précédentes. Il s'agit d'une méthode de rotation, qui minimise à la fois le nombre de variables qui pèsent fortement sur un facteur et le nombre de facteurs requis pour expliquer une variable