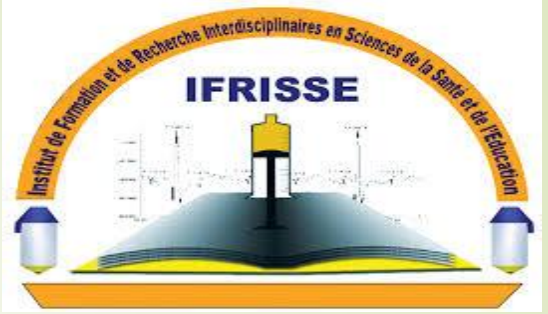
MODULE SPSS



Analyse multivariée

PRESENTATEURS DU COURS:
KABORE GERAUDE
OUEDRAOGO HOUDOU
DABRE ALI

Année académique_2021-2022

Analyse multivariée

Plan

- Introduction
- Régression linéaire simple
- Régression linéaire multiple
- Régression logistique
- Application sous SPSS

Introduction

Il est rare que la saisie d'une réalité sociale se fasse en mettant en relation deux phénomènes ou variables. Pour vérifier si la relation entre deux variables est stable, il faut effectuer des analyses mettant en jeu 3 variables ou plus.

Cette logique repose sur une observation d'Emile Durkheim selon laquelle lorsque 2 faits sociaux sont en relation et qu'on pense que l'un est la cause de l'autre, il faut se demander si cette relation ne serait pas due à quelque chose cachée.

Les différents types d'analyse multivariée

Il y a 3 critères qui déterminent la typologie d'analyse multivariée :

- Objectifs spécifiques de l'étude;
- Nature des variables ;
- Sujet d'analyse (variables ou individus).

Le premier critère conduit à la distinction entre méthodes descriptives et méthodes explicatives.

Les méthodes d'analyse multivariée se distinguent selon qu'elles s'appliquent aux variables métriques ou non métriques. La prise en compte de ces deux critères conduit au tableau suivant lorsque l'analyse porte sur les variables:

Objectif de l'étude	**Type de variables**	
	Quantitative	Qualitative
Descriptif	Analyse en facteurs communs et spécifiques Analyse en composantes principales (ACP)	Analyse factorielle des correspondances multiples (AFCM)
Explicatif	Régression linéaire multiple Analyse de la variance à plusieurs facteurs Analyse de classification multiple	Régression logistique binomiale Régression logistique multinomiale

Régression linéaire simple

Présentation du modèle

- Dans une régression linéaire simple on a une variable dépendante (expliquée) et une seule variable indépendante (explicative).
- La formulation mathématique du modèle est:

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t \,,\; t = 1.....n$$

Présentation du modèle

- Y :la variable endogène (dépendante, à expliquer) à la date t ;
- X : la variable exogène (indépendante, explicative) à la date t ;

β_0, β_1 : les paramètres inconnus du modèles ;

ε_t : l'erreur aléatoire du modèle ;

n: nombre d'observations.

Hypothèses du modèle

- H1: le modèle est linéaire en X par rapport au paramètres.
- H2 : $E(\varepsilon i)$=0 , (espérance de ε i =0) : En moyenne, les erreurs s'annulent c'est-à-dire que le modèle est bien spécifié
- H 3: La variance de l'erreur est constante et ne dépend pas de l'observation.
- H 4: l'erreur est indépendante de la variable exogène.
- H 5: les erreurs relatives à 2 observations sont indépendantes
- H 6 : les erreurs suivent une loi normale

Estimation et qualité du modèle

- L'estimation des paramètres β_0, β_1 est obtenue en minimisant la somme des carrés des erreurs (MCO)

- Le coefficient de détermination permet de juger de la qualité du modèle

- $0 \leq R \leq 1$, plus la valeur de R^2 est proche de 1, plus le modèle est plus significatif

Notion de contrôle de l'effet d'une variable

- La variable de contrôle est la variable qui est ajoutée par le chercheur dans une régression dans le but d'éviter un biais dans l'estimation du paramètre d'intérêt.

- Exemple: l'état palustre de l'enfant ne peut être expliqué par la seule variable « milieu de résidence ».

- Dans ce cas l'on aura beaucoup de chance d'avoir des résultats biaisés puisse qu'il y a d'autres variables qui expliquent l'état palustre de l'enfant, non inclus dans la régression.

- lorsqu'une variable capable d'expliquer la variable dépendante n'est pas spécifiée dans le modèle, celle-ci se retrouve de fait dans les résidus de la régression (toute l'information qui permet d'expliquer y qui n'est pas dans les x).

Notion de contrôle de l'effet d'une variable

- Si cette variable est corrélée aux variables explicatives spécifiées dans le modèle, l'hypothèse de non corrélation entre les variables explicatives et les résidus, hypothèse nécessaire à la bonne estimation du modèle par moindre carrés ordinaires, ne sera pas respectée.

- On aura alors un biais dans l'estimation des coefficients concernés.

- On voit donc l'importance de spécifier d'emblée toutes les variables x pertinentes pour expliquer une variable y, même si seul l'effet d'une variable x nous intéresse.

Notion de pondération

- La **pondération des données** consiste à accorder un **coefficient de pondération** (un poids) à chacun des individus d'un échantillon.
- il de corriger la représentativité de l'échantillon en fonction de certaines variables clés afin d'être en mesure d'extrapoler les résultats du sondage à la population.

- Soit une population qui compte 1000 individus dont la moitié est constituée d'hommes et l'autre moitié de femmes.

- Supposons que l'on fasse un sondage avec un échantillon de 100 individus et, qu'à cause de certains facteurs (ex. : taux de réponse, stratification), on obtient 80 femmes et 20 hommes.

Notion de pondération

- 80 femmes et 20 hommes.
- Dans ce cas, il y a un déséquilibre pour la variable «sexe», entre l'échantillon et la population : chaque homme de l'échantillon représente 25 hommes de la population (500/20) alors que chaque femme de l'échantillon représente 6,25 femmes de la population (500/80).
- Il faudra donc un coefficient de pondération qui aura pour effet de donner plus de poids aux réponses des 20 hommes (et moins à celles des femmes), et ce, afin de corriger le déséquilibre.
- Ainsi, le poids qu'on attribue à un individu de l'échantillon correspond au poids que cet individu représente dans la population.

Régression linéaire multiple

Présentation du modèle

- Les modèles de régression multiple
- Le modèle générale est une généralisation du modèle simple dans lequel figurent plusieurs variables explicatives :

$$Y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \ldots\ldots\ldots\ldots\ldots + \beta_k x_{kt} + \varepsilon_t \ldots\ldots\ldots, \qquad t = 1 \ldots\ldots\ldots n$$

Présentation du modèle

- Y_t = variable à expliquer a la date t
- X_{1t} = variable explicative 1 à la date t
- .
- .
- .
- X_{kt} = variable explicative k à la date t

Présentation du modèle

$\beta_0, \beta_1, \beta_2 \dots \dots \dots \beta_k$: les paramètres inconnus du modèles ;

ε_t : Erreur de spécification elle est inconnue et restera inconnue

n: nombre d'observations.

Hypothèses du modèle

- H1 : le modèle est linéaire
- H2 : xi pour tout i=1, …n est une variable certaine non aléatoire
- H3 : l'espérance mathématique des erreurs u est nulle $E(ut)=0$ pour tout t=1,…, T
- H4 : la variance des erreurs est constante (homoscédascité) $E(u^2t)=\sigma^2$ et les erreurs sont non corrélées $E(ut, ut')=0$ pour tout $t \neq t'$
- H5 : l'erreur est indépendante des variables explicatives $E(xit, ut)=0$
- H6 : les erreurs sont indépendamment et identiquement distribuées selon une loi normale.

Le modèle logistique

- La nature qualitative et binaire de la variable dépendante rend possible le recours à la régression logistique binaire.

- Cette méthode estime les risques de survenance d'un évènement en fonction de certaines variables indépendantes.

- La variable dépendante prend la valeur 1 quand l'évènement est réalisé et 0 dans le cas contraire.
- Ainsi, la régression logistique estime la probabilité pour un individu d'être dans un état donné.
- Une fois défini l'ensemble des événements auxquels on s'intéresse, on traduit par un nombre leurs « possibilités » de réalisation. Cela revient à affecter une mesure de « croyance » à chaque événement, c'est-à-dire un degré de certitude que l'on a que l'événement se produise ou non.
- Afin de correspondre à la notion intuitive, une probabilité sera un nombre associé à un événement, compris entre 0 et 1, pour pouvoir se convertir en pourcentage de « chances ». (Lecoutre, 2016).

- Il s'agit précisément d'estimer l'effet net des variables associées au risque d'être dans un état donné. Dans ce modèle, le logit de la probabilité (p) de réalisation de la variable à expliquer (Y) est exprimé en fonction d'un intercept (ordonnée à l'origine) , des variables explicatives rattachées à leurs coefficients et un terme d'erreur ε :

$$\ln\left(\frac{p}{1-p}\right) = logit(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots \beta_n X_n + \varepsilon$$

Remarque : si la variable dépendante est qualitative avec plus de deux modalités on utilise la régression logistique multinomiale.

- Après l'estimation des coefficients par la méthode du maximum de vraisemblance, on peut calculer les Odds Ratios (ORR) ou rapports de chances pour faciliter la lecture immédiate des résultats.

- Lorsque le OR est supérieur à 1, cela signifie que les individus de la modalité considérée ont OR fois ou 100*(OR-1)% plus de risque de subir le phénomène par rapport aux individus de la modalité de référence.

Un rapport de chance inférieur à 1 traduit que les individus de cette catégorie ont 100*(1-OR)% moins de risque de subir le phénomène par rapport à la catégorie de référence.

Application sous SPSS

Régression logistique

Identifier les facteurs explicatifs de la survenue du paludisme au Burkina Faso à partir de l'EIP 2017-2018 ?

Étape 1

Création des variables indépendantes sous forme dichotomique

Dichotomiser une variable revient à transformer chaque modalité de réponse en une nouvelle variable indicatrice de la présence de cette modalité.

Étape 2

Identifier la modalité de référence pour chaque variable

Étape 3

Avec la commande d'opération boîte de dialogue Analyse – Régression – Logistique binaire – Sélectionner les variables – Choisir la méthode –Ok

À travers la syntaxe:

logistic reg var = recours1 /method=enter sex1 coh1 coh2 instruis1 instruis2 milieu2.

Références bibliographiques

Cadario, R., Butori, R. & Parguel, B. (2017). Chapitre 2. Manipuler et mesurer les variables de l'expérimentation. Dans : , R. Cadario, R. Butori & B. Parguel (Dir), *Méthode expérimentale : analyses de modération et médiation* (pp. 35-48). Louvain-la-Neuve: De Boeck Supérieur.