

MODULE SPSS



Année académique_2021-2022

KABORE T.A. Geraude

Démographe

Mail: kaboralimata@gmail.com

Tel: (+226) 73382412/ (+226) 66757694

OUEDRAOGO Houdou

Statisticien Economiste

Mail: houdbarack@gmail.com

Tel: (+226) 70129773 /

COULIBALI Moussa

Démographe

Mail: coulibalymoussa330@gmail.com

Tel: (+226) 72555573 / (+226) 74066271

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:

$$Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t, \quad t = 1, \dots, n$$

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 modèle ;

- 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 $\varepsilon_i = 0$) : 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

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 :

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

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 modèle ;

ε_t : Erreur de spécification elle est inconnue et restera inconnue

n: nombre d'observations.

Hypothèses du modèle

- H1 : le modèle est linéaire
- H2 : x_i pour tout $i=1, \dots, n$ est une variable certaine non aléatoire
- H3 : l'espérance mathématique des erreurs u est nulle $E(u_t)=0$ pour tout $t=1, \dots, T$
- H4 : la variance des erreurs est constante (homoscédasticité) $E(u^2_t)=\sigma^2$ et les erreurs sont non corrélées $E(u_t, u_{t'})=0$ pour tout $t \neq t'$
- H5 : l'erreur est indépendante des variables explicatives $E(x_{it}, u_t)=0$
- 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.

- 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 (X) rattachées à leurs coefficients et un terme d'erreur ε :

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

- 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 $100 \times (OR - 1)\%$ 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 $100 \times (1 - \text{OR})\%$ 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.
```