SOCIO532.2
STATISTIQUES & INFORMATIQUE APPLIQUÉES AUX SCIENCES SOCIALES

El Hadj Touré, Ph D. Sociologie
Département de sociologie
Université Gaston Berger de Saint-Louis

Leçon 6

Régression linéaire multiple

1

Au programme

- Régression linéaire simple approfondie: quelques éléments de rappel
  - Diagramme de dispersion & équation de la droite de régression
  - Les coefficients de régression standardisés bêta
- Régression linéaire multiple
  - Diagramme de dispersion d'une relation entre 3 variables quant.
  - Prédiction à l'aide du plan de l'équation de régression
  - Force d'une relation et variance expliquée: bêta, r et r^2 multiples
  - Conditions d'application du modèle multivarié

2

Régression linéaire simple

Rappel

- Strictement parlant, l'analyse de régression linéaire est appropriée lorsque les deux variables sont quantitatives
  - Mais, il est possible d'inclure une VI qualitative dans le modèle de régression avec VD quantitative
  - But: prédire, mesurer la force d'une relation et la généraliser
- Le taux de fertilité varie-t-il selon le taux d'urbanisation dans les pays peuplés du monde (Fox, 1999)?

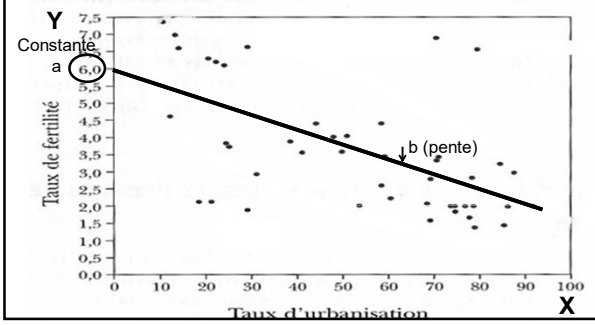
(n=50)

X Taux d'urbanisation → Taux de fertilité Y

20%, 81%... **Y = f (X)** 7.3, 1.6…enfants/femme

3

Régression linéaire simple



4

Régression linéaire simple

Les coefficients et leur signification statistique (SPSS)

- Le taux de fécondité selon le taux d'urbanisation (n=50)

Coefficients[a]

Modèle		Coefficients non standardisés B	Coefficients non standardisés Ecart standard	Coefficients standardisés Bêta	t	Sig.
1	(Constante)	5.720	.516		11.090	.000
	Taux d'urbanisation.	-.041	.009	-.555	-4.626	.000

a. Variable dépendante : Taux de fertilite.

Constante a et son erreur-type

Pente b et son erreur-type

- Équation de régression : $\hat{Y} = a + (bX)$
  - $\hat{Y} = 5{,}72 + (-0{,}041X)$
    - Fertilité prédite du Sénégal (taux d'urbanisation de 40%)
    - $\hat{Y} = 5{,}72 - 0{,}041(40) = 4{,}1$ enfants/femme

07:42 5

5

Régression linéaire simple

Deux relations bivariées

- Le taux de fertilité varie-t-il selon le taux d'urbanisation dans les pays peuplés du monde (Fox, 1999)?

X — Taux d'urbanisation → Taux de fertilité — Y

0% à 100% théoriquement — **Y = f (X)** — 7.3, 1.6…enfants/femme

- Le taux de fertilité varie-t-il selon l'indice de développement humain (IDH) dans les pays peuplés du monde (Touré)?

X — IDH → Taux de fertilité — Y

0 à 1 théoriquement — **Y = f (X)** — 7.3, 1.6…enfants/femme

07:42 6

6

Régression linéaire simple

Coefficient Bêta

- Taux de fertilité selon le taux d'urbanisation (0 à 100)

Modèle		Coefficients non standardisés B	Coefficients non standardisés Erreur standard	Coefficients standardisés Bêta	t	Sig.
1	(Constante)	5,720	,516		11,090	,000
	Taux d'urbanisation	,041	,009	,555	4,626	,000

a. Variable dépendante : Fertilite Taux de fertilité

- Taux de fertilité selon l'IDH (0 à 1)

Modèle		Coefficients non standardisés B	Coefficients non standardisés Erreur standard	Coefficients standardisés Bêta	t	Sig.
1	(Constante)	9,341	,761		12,280	,000
	Indice de développement humain	-8,181	1,048	-,748	-7,808	,000

a. Variable dépendante : Fertilite Taux de fertilité

07

7

Au programme

- Régression linéaire simple approfondie: quelques éléments de rappel
  - Diagramme de dispersion & équation de la droite de régression
  - Les coefficients de régression standardisés bêta
- Régression linéaire multiple
  - Diagramme de dispersion d'une relation entre 3 variables quant.
  - Prédiction à l'aide du plan de l'équation de régression
  - Force d'une relation et variance expliquée: bêta, r et r^2 multiples
  - Conditions d'application du modèle multivarié

8

8

Régression linéaire multiple

Problème de recherche

→ Toutes choses étant égales, le taux d'urbanisation et le nombre de radios/100hbts ont-ils un effet sur le taux de fertilité dans les pays peuplés du monde (Fox, 1999)?

→ On cherche à prédire le taux de fertilité par le taux d'urbanisation et le nombre de radios/100hbts…

X_1 Taux d'urbanisation → Taux de fertilité Y

X_2 Nombre de radios

$\mathbf{Y = f (X_1, X_2)}$

→ Le modèle de convergence (causalité convergente) est le modèle par excellence de la régression linéaire multiple

07:42

9

9

Régression linéaire multiple

Trois types majeurs de questions

- Investiguer si une VI prédit une VD quantitative, après avoir contrôlé l'effet d'autres prédicteurs
  - Équation de régression (constante et coefficients de régression)
- Investiguer de quelle manière un ensemble de VIs est associé à une VD, en augmentant la proportion expliquée
  - Force et variance expliquée par le modèle (r-deux, r-deux multiple, variation du r-deux) et VIs qui y contribuent de façon significative
- Investiguer laquelle des VIs est la plus importante, i.e. a un pouvoir prédictif plus important, toutes choses étant égales
  - Coefficients de régression bêta

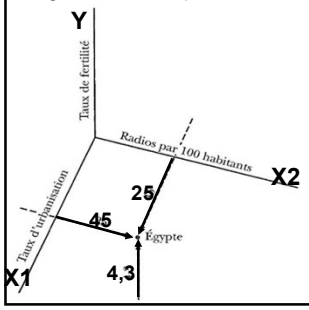
07:42 10

10

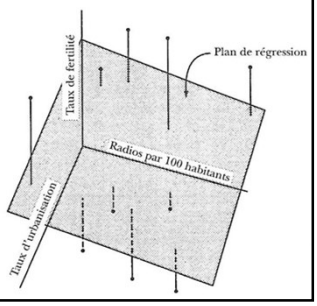
Régression linéaire multiple

Diagramme de dispersion & plan de régression (Fox:321-322)

Diagramme de dispersion à 3D



Plan de régression

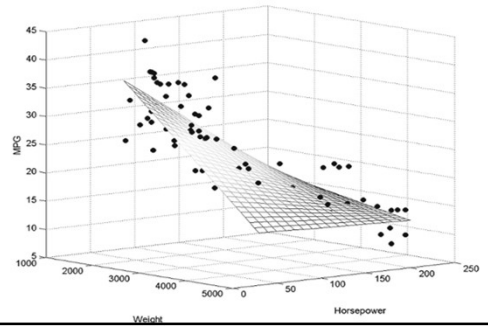


11

Régression linéaire multiple

Modèles de plans de régression

Plan de régression avec effets négatifs



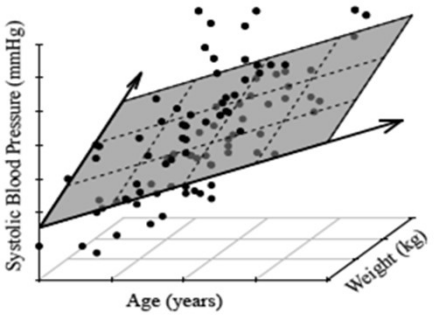
08:08

12

Régression linéaire multiple

Modèles de plans de régression

Plan de régression avec effets positifs



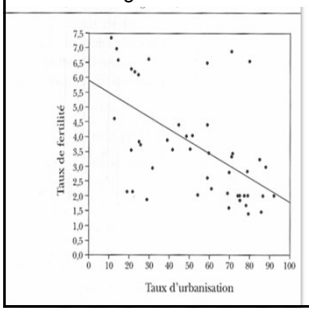
08:20

13

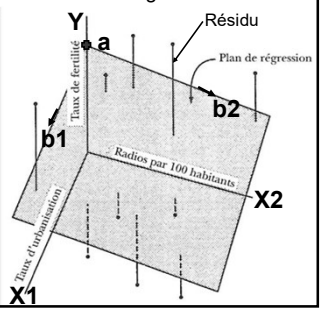
Régression linéaire multiple

Diagramme de dispersion & plan de régression (Fox:321-322)

Droite de régression



Plan de régression



14

Régressions linéaires simple & multiple

Équations linéaires

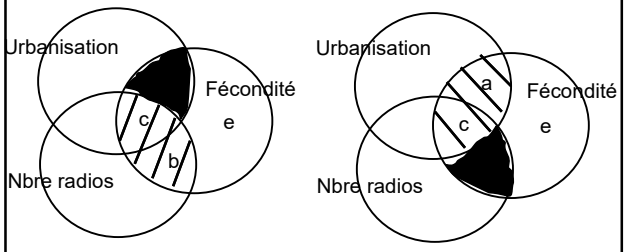
- Autant dans une relation bivariée, la droite de régression traverse le diagramme de dispersion à 2 D et permet de minimiser la SC des résidus ou erreurs de prédiction
  - La droite est représentée par l'équation : $Y = a + bX$
    - Fertilité = a + b(Urbanisation)
- De la même façon dans une relation multivariée, un plan de régression traverse l'espace tridimensionnel et permet de minimiser la SC des résidus ou erreurs de prédiction
  - Le plan est représenté par l'équation : $Y = a + b_1X_1 + b_2X_2$
    - Fertilité = a + b_1(Urbanisation) + b_2(Radios)

07:42 15

15

Équation de régression linéaire multiple

Estimation des paramètres: illustration



→ Chacun des coefficients de régression se calcule à partir de la relation entre un prédicteur et la VD, après avoir contrôlé, supprimé l'effet de l'autre prédicteur

16

Équation de régression linéaire multiple

Estimation des paramètres: coefficients

- Récapitulatif des estimés des paramètres de l'équation de régression sous forme tabulaire (n=49)

Taux de fertilité selon le taux d'urbanisation et le nbre de radios

Modèle	Coefficients non standardisés B	Coefficients standardisés Bêta
Constante (A)	5,59	
Taux d'urbanisation	-0,032	-0,428
Nbre de radios	-0,010	-0,207

Variable dépendante: Taux de fertilité

17

Équation de régression linéaire multiple

Estimation des paramètres & prédiction

- Strictement parlant, les coefficients de régression multiple (pentes) décrivent le changement dans la VD lorsque la VI augmente d'une unité, tout en contrôlant l'effet de l'autre VI
  - On les appelle ainsi des pentes partielles
- Voici l'équation de régression du taux de fertilité selon le taux d'urbanisation et le nombre de radios (Fox, p.324)

  Fertilité= *5,59* – *0,032*Urbanisation – *0,010*Radios
- Taux de fertilité prédit de l'Égypte, connaissant son taux d'urbanisation (45) et le nombre de radios (25)

  $\hat{Y}$ = 5,59 – 0,032(45) – 0,010(25) = 3,90 enfants/femme

18

Équation de régression linéaire multiple

Interprétation des paramètres estimés

- La constante est égale à 5,59, ce qui signifie que lorsque le taux d'urbanisation et le nombre de radios sont nuls, le taux de fertilité est de 5,59 enfants/femme
- La pente en X_1 est égale à -0,032. Une augmentation d'une unité (%) du taux d'urbanisation entraîne une diminution du taux de fertilité de 0,032 enfant/femme, en contrôlant l'effet du nombre de radios
- La pente en X_2 est égale à -0,010. Une augmentation d'une unité du nombre de radios entraîne une diminution du taux de fertilité de 0,010 enfant/femme, en contrôlant l'effet de l'urbanisation

19

Coefficients de régression standardisés (bêta)

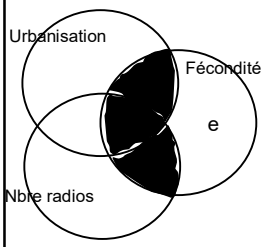
Interprétation statistique

- Le coefficient bêta de l'effet de l'urbanisation est β=-0.428. Il indique que lorsque le taux d'urbanisation augmente d'un écart-type (25.1), le taux de fertilité diminue en moyenne de 0,428 écart-type, en contrôlant l'effet du nbre de radios
  - 0,428 écart-type représente 0,80 enfant/femme (0,428*1,87)
- Le coefficient bêta pour le nbre de radios est β=-0.207. Il indique que lorsque le nbre de radios augmente d'un écart-type (38,99), le taux de fertilité décroît en moyenne de 0,207 écart-type, en contrôlant l'effet de l'urbanisation
  - 0,207 écart-type représente 0,39 enfant/femme (0,207*1,87)
- L'urbanisation affecte 2 fois+ la fertilité que le nb. de radios

20

Corrélation linéaire multiple

Illustration & calcul du r-deux, test de signification



- Le R^2 multiple

$$R^2_{Y.12} = \frac{\text{Variation expliquée}}{\text{Variation totale}} = \frac{56{,}527}{168{,}694} = 0{,}34$$

- La corrélation ® multiple

$$R_{Y.12} = racine\ carré(0{,}34) = 0{,}58$$

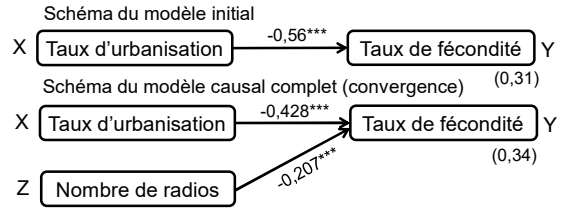
- Le R^2 multiple est la proportion de la variation expliquée par les 2 VIs X (a+b+c) par rapport à la variation totale dans Y (a+b+c+e)

21

Régression linéaire multiple

Présentation des résultats de l'analyse de convergence

- Figure 1. Schématisation du modèle de l'effet convergent du taux d'urbanisation et du nbre de radios sur le taux de fertilité



Note: Les chiffres sur les flèches sont des coefficients bêta. Les chiffres sous la variable dépendante correspondent au r^2 simple et au r^2 multiple ***p<0,001

22

Régression linéaire multiple

Interprétation statistique des résultats de l'analyse de convergence

- D'abord, le taux d'urbanisation a un effet négatif significatif sur le taux de fertilité, même après avoir contrôlé l'effet du nombre de radios/100hbts (β=-0,428; p<0.001).
- Ensuite, le nombre de radios/100hbts a un effet négatif significatif sur le taux de fertilité, même après avoir contrôlé l'effet du taux d'urbanisation (β=-0,207; p<0.001)
- Enfin, le taux d'urbanisation et le nombre de radios ont un effet convergent sur le taux de fertilité, chacun ayant un pouvoir explicatif et prédictif spécifique, mais l'effet de l'urbanisation s'avère plus important. Le modèle explique 34% de la variance de la VD (r^2=0,34; p<.001)

23

Régression linéaire multiple

Comment présenter les résultats d'une régression multiple?

- Présenter un tableau unique incluant… (voir labo SPSS)
  - les coefficients et leur erreur-type, la sig. le n, le r2
- Schématiser les modèles incluant les résultats
  - 1) Schéma du modèle initial et 2) schéma du modèle causal de convergence
  - Inscrire les résultats statistiques sur les deux schémas
    - On met les coefficients de régression bêta sur les flèches
    - On met des astérisques sur les bêtas s'ils sont significatifs (*p<0,05; **p<0,01; ***p<0,001)
    - Si le lien n'est pas sig., la flèche est barrée ou absente
    - On place le r-deux simple ou multiple sous la VD

24

Régression linéaire multiple

Postulats ou conditions d'application

- Trois postulats initiaux fondamentaux (comme en régression simple) à vérifier à l'aide du diagnostic des résidus
  - Linéarité
  - Normalité
  - Homogénéité

07:42 25

25

Régression linéaire multiple

Postulats ou conditions d'application

- + Absence de multi-colinéarité forte
  - Les VIs ne doivent pas être fortement corrélées entre elles: risque de redondance, d'inflation des erreurs-types
    - Examen de la matrice de corrélation entre les VIs, pour s'assurer que r < 0,70
    - Calcul de la tolérance et du "variance inflation factor" (VIF) en régressant chacun des prédicteurs selon les autres prédicteurs
      - Tolérance doit être proche de 1, et VIF pas trop excéder 1
      - Tolérance=1-R2
        - Minimale de 0.5: Tabachnick & Fidell, 2013)
      - VIF= 1/1-R2
        - VIF maximal de 2 (Tabachnick & Fidell, 2013)
  - Si colinéarité entre deux VIs, enlever une du modèle

08:32 26

26

Tout prochainement

- Prochaine séance
  - Régression logistique
- Au labo SPSS d'aujourd'hui
  - Régression linéaire simple: quelques éléments de rappel & calcul du coefficient bêta
  - Régression linéaire multiple
  - Diagnostic d'un modèle de régression linéaire multiple et vérification de la multicolinéarité

07:42 27

27