SOCIO 532.2
STATISTIQUES & INFORMATIQUE APPLIQUÉES AUX SCIENCES SOCIALES

El Hadj Touré, Ph D. Sociologie
Département de psychoéducation,
Université de Sherbrooke

Leçon 3

De l'analyse bivariée. Régression linéaire approfondie

17:00 1

1

Au programme

- Régression linéaire simple
  - Quelques éléments de rappel
  - Régression pour une relation entre deux variables quantitatives: quelques éléments de rappel
  - Valeurs prédites et diagnostic de régression
  - Les coefficients de régression standardisés (bêta): justification, calcul et interprétation
- Régression linéaire simple avec une VI qualitative
  - Modèle avec une VI dichotomique et parallèle avec l'ANOVA
  - Modèle avec une VI qualitative non dichotomique et parallèle avec l'ANOVA
  - Valeurs prédites et diagnostic de régression pour un modèle avec une VI

2

2

Rappel sur les statistiques bivariées

Considérations générales

- Répondent à des questions de recherche du genre:
  - Y a-t-il une <u>relation</u> d'association entre une variable X et une variable Y dans les données de l'échantillon ? Si oui, quelle en est l'<u>intensité</u>, et la <u>direction</u> éventuellement?
  - La relation entre X et Y existe-t-elle dans la population dont provient l'échantillon ?
- Pour ce faire, quelques éléments de pré-analyse:
  - Identifier la VD (Y) et la VI (X)
  - Savoir comment les variables sont mesurées afin de choisir les techniques statistiques appropriées
  - S'assurer que la relation est plausible sociologiquement

17:02 3

3

Rappel sur les statistiques bivariées

Du problème de recherche au choix statistique

- Tester une relation entre deux variables quali
- Tester une relation entre une variable indépendante quali dichotomique et une variable dépendante quanti
- Tester une relation entre une variable indépendante quali non dichotomique et une variable dépendante quanti
- Tester une relation entre deux variables quanti

17:01 4

4

Régression linéaire simple

Rappel

- Strictement parlant, l'analyse de régression linéaire est appropriée lorsque les deux variables sont quantitatives
  - Mesurer la force de la relation (signification réelle) : r, bêta, r-deux
  - Généraliser la relation (signification statistique) : test F
  - Procéder à la prédiction (équation de la droite) : Y = a + b(X)
- Le taux de fertilité varie-t-il selon le taux d'urbanisation dans les pays peuplés du monde (Fox, 1999)?

(n=50)

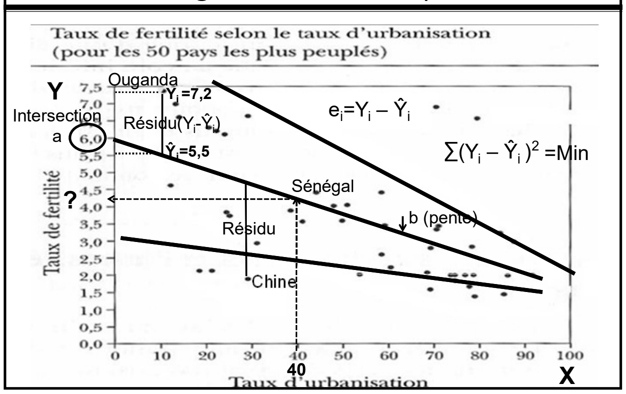
X Taux d'urbanisation → Taux de fertilité Y

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

17:00 5

5

Régression linéaire simple



6

Régression linéaire simple

Les coefficients et leur test (SPSS)

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

Coefficients[a]

Constante a et son erreur-type

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.

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

17:00 7

7

Régression linéaire simple

Scores prédits et résiduels (SPSS)

Fichier Edition Affichage Données Transformer Analyse Marketing direct Graphiques Utilitaires Extensions Fenêtre Aide

1 : Fertilite 6,70

Y_i $\hat{Y}_i$ $Y_i-\hat{Y}_i$ Scores stand. Z

	Pays	Tauxdurb	Fertilite	PRE_1	RES_1	ZPRE_1	ZRES_1
1	Arabie Saoudite	79	6,70	2,45	4,25	-1,08	2,72
2	Irak	70	6,86	2,82	4,04	-,72	2,59
3	Iran	57	6,40	3,36	3,04	-,20	1,95
4	Zaire **Sénégal**	40	6,70	4,06	2,63	,49	1,69
5	Pakistan	28	6,50	4,56	1,94	,97	1,24
6	Ouganda	11	7,15	5,26	1,88	1,65	1,21
7	Éthiopie	15	6,88	5,09	1,78	1,49	1,14
8	Soudan	23	6,19	4,76	1,42	1,17	,91
9	Tanzanie	21	6,25	4,85	1,40	1,25	,90
10	Kenya	25	6,06	4,68	1,37	1,09	,88
11	Nigéria	16	6,43	5,05	1,37	1,45	,88

8

Coefficient de régression standardisé (bêta)

Justification

- Entre le taux d'urbanisation (X) et l'indice du develop-pement humain (X), lequel des 2 prédicteurs a un effet plus important, sociologiquement, sur le taux de fertilité (Y)?
  - Les deux prédicteurs ne sont pas mesurés sur une même échelle

Statistiques descriptives

	N	Minimum	Maximum	Moyenne	Ecart type
Tauxdurb Taux d'urbanisation	50	8	92	52,14	24,910
IDH Indice de développement humain	50	,31	,94	,7062	,16966
Fertilite Taux de fertilité	50	1,37	7,15	3,5632	1,85572
N valide (liste)	50				

- Pour y répondre, on standardise les coefficients de régression en écarts-types pour obtenir des bêtas

17:00 9

9

Coefficient de régression standardisé (bêta)

Exemple illustratif

- 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

17 a. Variable dépendante : Fertilite Taux de fertilité

10

Coefficient de régression standardisé (bêta)

Formule et calcul

- Un coefficient standardisé bêta β se calcule en multipliant le coefficient de régression b par le rapport entre l'écart-type de X et l'écart-type de Y :

$$\beta = b^*(s_X / s_Y)$$

  - Calcul bêta de l'effet du taux d'urbanisation
    β = -0,041*(24,9/1,86) = -0,555
  - Calcul bêta de l'effet de l'indice du développement humain
    β = -8,181*(0,17/1,86) = -0,748
- Dans une relation bivariée, le coefficient de régression bêta est égal au coefficient de corrélation: $\beta = r$

17:00 11

11

Coefficient de régression standardisé (bêta)

Interprétation statistique

→ Les écarts-types du taux de fertilité, du taux d'urbanisation et de l'IDH sont respectivement de 1.86, 24.9 et 0.17

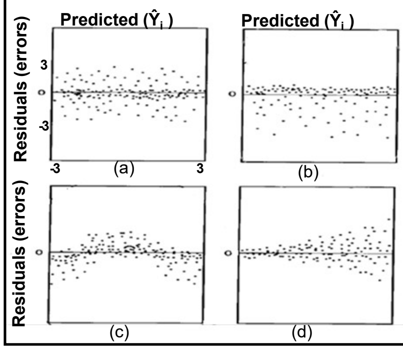
- Le coefficient bêta -0.56 indique que lorsque le taux d'urbanisation augmente d'un écart-type (24.9), le taux de fertilité diminue en moyenne de 0,56 écart-type
  - 0,56 écart-type représente **1,04 enfant/femme** (0,56*1,86)
- Le coefficient bêta -0.75 indique que lorsque l'IDH augmente d'un écart-type (0.17), le taux de fertilité décroît en moyenne de 0,75 écart-type
  - 0,75 écart-type représente **1,40 enfant/femme** (0,75*1,86)

17:00 12

12

Régression linéaire simple

Vérification des postulats: Diagnostic des résidus



a) Postulats respectés:
- Normalité (pas de cas déviants)
- Linéarité
- Homogénéité

b) Non normalité

c) Non linéarité

d) Non homogénéité des variances

13

13

Au programme

- Régression linéaire simple
  - Quelques éléments de rappel
  - Régression pour une relation entre deux variables quantitatives: quelques éléments de rappel
  - Valeurs prédites et diagnostic de régression
  - Les coefficients de régression standardisés (bêta): justification, calcul et interprétation
- Régression linéaire simple avec une VI qualitative
  - Modèle avec une VI dichotomique et parallèle avec l'ANOVA
  - Modèle avec une VI qualitative non dichotomique et parallèle avec l'ANOVA
  - Valeurs prédites et diagnostic de régression pour un modèle avec une VI

14

14

Régression avec VI qualitative

Généralités

- Théoriquement, dans un modèle de régression linéaire simple où la VD est quantitative (ex. taux de fertilité), la VI doit être aussi quantitative (ex. taux d'urbanisation)
- Cependant, en pratique, dans un modèle de régression, on peut y inclure une VI qualitative, après avoir procédé à quelques opérations de recodage ou transformation
  – Pour une VI qualitative dichotomique, elle doit être codée <u>0/1</u> pour être interprétée comme une échelle

    Région (0.Pauvre, 1.Riche) ⟶ Taux de fertilité

  – Pour une VI qualitative non dichotomique (k≥3), elle doit être transformée en variables factices dichotomiques <u>0/1</u>

    Région (0.Pauvre, 1.Emerg., 2.Riche) ⟶ Taux de fertilité

15

15

Régression avec VI qualitative dichotomique

Exemple illustratif

- Problème de recherche (n=50 pays)
  - VD: Taux de fertilité (échelle de 1 à 7 enfants/femme)
  - VI: Région (0= Pays pauvre; 1= Pays riche)
- Analyse de régression linéaire simple (SPSS)

Coefficients[a]

Modèle		Coefficients non standardisés B	Erreur standard	Coefficients standardisés Bêta	t	Sig.
1	(Constante)	4,428	,310		14,265	,000
	Region Riches vs Pauvres	-1,880	,458	-,510	-4,107	,000

a. Variable dépendante : Taux de fertilité

Récapitulatif des modèles

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	-,510[a]	,260	,245	1,61288

a. Prédicteurs : (Constante), Region Riches vs Pauvres

16

16

Régression avec VI qualitative dichotomique

Interprétation statistique & prédiction

- Équation de régression linéaire: $\hat{Y} = 4{,}43 - 1{,}88(X)$
- La constante (4,43) réfère au taux de fertilité prédit lorsque la région est pauvre (X=0), soit 4,43 enfants/femme
- Si l'on passe d'une région pauvre (0) à une région riche (1), la fertilité diminue de 1,88 enfant/femme (b=-1,88)
- Lorsque la région est riche (X=1), le taux de fertilité prédit est de 2,55 enfants/femme:

  $\hat{Y} = 4{,}43 - 1{,}88(X)$

  $\hat{Y} = 4{,}43 - 1{,}88(1) = 2{,}55$

17

17

Régression avec VI dichotomique

Scores prédits et résiduels

Fichie Editio Afficha Donné Transforr Analys Marketing d Graphiqu Utilitair Extensic Fenêtr Aide

1 : Fertilite | 4,40 | Visible : 15 variables sur 15

	Pays	Region	Fertilite (Y_i)	PRE_1 ($\hat{Y}_i$)	RES_1 ($Y_i - \hat{Y}_i$)
1	Afrique du Sud	Pauvre	4,40	4,43	-,03
2	Algérie	Pauvre	3,96	4,43	-,47
3	Allemagne	Riche	1,40	2,55	-1,15
4	Arabie Saoudite	Riche	6,70	2,55	4,15
5	Argentive	Riche	2,72	2,55	,17
6	Autralie	Riche	1,83	2,55	-,72
7	Bengladesh	Pauvre	4,55	4,43	,12
8	Brésil	Riche	2,49	2,55	-,06
9	Canada	Riche	1,84	2,55	-,71
10	Chine	Riche	1,85	2,55	-,70

18

Régression avec VI qualitative dichotomique

= ANOVA à un facteur

- Analyse de variance à un facteur (SPSS)

Taux de fertilité	Moyenne
0 Pauvres	4.428
1 Riches	2.550

			Somme des carrés	ddl	Carré moyen	F	Sig.
Taux de fertilité * Region Riches vs Pauvres	Inter-groupes	(Combinée)	43,875	1	43,875	16,866	,000
	Intra-groupes		124,867	48	2,601		
	Total		168,742	49			

	Eta	Eta carré
Taux de fertilité * Region Riches vs Pauvres	,510	,260

- La régression linéaire avec VI dichotomique et l'ANOVA à un facteur sont similaires: les résultats sont les mêmes
  - Valeurs prédites & moyennes = 4.428 (pauvre), = 2.55 (riche)
  - Corrélation R & Êta E = 0,51
  - R-deux & Êta-carré = 0,26

17:00 19

19

Régression avec VI qualitative non dichotomique

Création des variables factices dichotomiques

- Dans un modèle de régression, on peut y inclure une VI qualitative (k≥3) dont les catégories sont transformées en variables factices, i.e. dichotomisées en 0/1
- Exemple de RÉGION (k=3 catégories)
  0=Sous-développée
  1=Émergente
  2=Développée
  - On crée 2 variables factices, soit k-1 (3-1), à inclure dans le modèle, la 3[e] catégorie étant la référence (0) à exclure
    - DEVEL: 1=Développée; 0=Autrement
    - EMERG: 1=Émergente; 0=Autrement

Cas	Région	Code	
1	S-devel	0	0
2	S-devel	0	0
3	S-devel	0	0
4	Emerg	0	1
5	Emerg	0	1
6	Emerg	0	1
7	Devel	1	0
8	Devel	1	0
9	Devel	1	0

17:00

20

Régression avec VI qualitative non dichotomique

Exemple illustratif

- Problème de recherche (n=50 pays)
  - VD: Taux de fertilité (échelle de 1 à 7 enfants/femme)
  - VI: Région (0= Sous-développée; 1=Émergente; 2= Développée)

  → À inclure dans le modèle: Région développée et Région émergente
- Analyse de régression linéaire simple (SPSS)

Modèle		Coefficients non standardisés B	Erreur standard	Coefficients standardisés Bêta	t	Sig.
1	(Constante)	4,57	,309		14,79	,000
	Region2 Développée	-2,88	,528	-,67	-5,45	,000
	Region1 En émergence	-1,06	,492	-,26	-2,16	,036

a. Variable dépendante : Tauxdefert Taux de fert.

Modèle	R	R-deux	R-deux ajusté	Erreur standard de l'estimation
1	,623[a]	,388	,362	1,48259

a. Prédicteurs : (Constante), Region1 En émergence, Region2 Développée

17:00 21

21

Régression avec VI qualitative non dichotomique

Interprétation statistique des coefficients

- Constante a=5,57
  - Lorsque la région est pauvre (X=0), le taux de fertilité prédit est de 5,57 enfants/femme
- Coefficient de régression de Région développée b=-2,88
  - Si l'on passe d'une région pauvre (0) à une région développé (1), la fertilité diminue de 2,88 enfants/femme
- Coefficient de régression de Région émergente b=-1,06
  - Si l'on passe d'une région pauvre (0) à une région émergente (1), la fertilité diminue de 1,06 enfant/femme

17:00 22

22

Régression avec VI qualitative non dichotomique

Prédiction

→ 2 prédictions à l'aide de l'équation : $\hat{Y} = a + b(X)$

- Effet de Région développée (X=1)
  - Lorsque la région est développée (X=1), le taux de fertilité prédit est de 1,69 enfant/femme

    $\hat{Y} = 4{,}57 - 2{,}88(X)$

    Fertilité prédite= 4,57- 2,88 (1) = 1,69 enfant/femme
- Effet de Région émergente (X=1)
  - Lorsque la région est émergente (X=1), le taux de fertilité prédit est de 3,51 enfants/femme
  - $\hat{Y} = 4{,}57 - 1{,}06(X)$
  - Fertilité prédite= 4,57- 1,06 (1) = 3,51 enfants/femme

17:00 23

23

Régression avec VI qualitative non dichotomique

Scores prédits et résiduels

Fichier Edition Affichage Données Transformer Analyse Marketing direc Graphiques Utilitaires Extensions Fenêtre Aide

1 : Fertilite 4,40 Visible : 19 variables sur 19

Y_i $\hat{Y}_i$ $Y_i - \hat{Y}_i$

	Pays	Region2	Region1	Fertilite	PRE_2	RES_2
1	Afrique du Sud	Autres	Pays en émergence	4,40	3,51	,89
2	Algérie	Autres	Autres	3,96	4,57	-,61
3	Allemagne	Pays développés	Autres	1,40	1,69	-,29
4	Arabie Saoudite	Autres	Pays en émergence	6,70	3,51	3,19
5	Argentive	Autres	Pays en émergence	2,72	3,51	-,79
6	Autralie	Pays développés	Autres	1,83	1,69	,14
7	Bengladesh	Autres	Autres	4,55	4,57	-,02
8	Brésil	Autres	Pays en émergence	2,49	3,51	-1,02
9	Canada	Pays développés	Autres	1,84	1,69	,15
10	Chine	Autres	Pays en émergence	1,85	3,51	-1,66
11	Colombie	Autres	Autres	2,54	4,57	-2,03

24

Régression avec VI qualitative non dichotomique

= ANOVA à un facteur

- Analyse de variance à un facteur (SPSS), soit la région (développée, émergente, sous-développée)

Taux de fertilité	Moyenne
0 Sous-développée	4.570
1 En émergence	3.510
2 Développée	1.690

		Somme des carrés	ddl	Carré moyen	F	Sig.
Taux de fertilité * Region3 développée, émergente, sous développée	Inter-groupes (Combinée)	65,409	2	32,705	14,569	,000
	Intra-groupes	103,262	46	2,245		
	Total	168,671	48			

	Eta	Eta carré
Taux de fertilité * Region3 développée, émergente, sous développée	,623	,388

- La régression linéaire avec VI qualitative et l'ANOVA à un facteur sont similaires: les résultats sont les mêmes
  - Valeurs prédites & moyennes =4.57 (région sous-développée), =3.51 (région émergente), =1.69 (région développée)
  - Corrélation R & Êta E = 0,623
  - R-deux & Êta-carré = 0,388

17:00 25

25

Régression avec VI qualitative

Quelques remarques

- En définitive, le modèle de régression est utilisé lorsque la VD est quantitative, la VI pouvant être quantitative ou qualitative dichotomique
- Question: Si la régression linéaire avec une VI qualitative est équivalente à l'ANOVA à un facteur, alors pourquoi utiliser la régression qui est plus avancée comme modèle?
- Réponse: contrairement à l'ANOVA, le modèle de régression est plus flexible, car pouvant inclure de multiples prédicteurs sans compliquer inutilement l'interprétation de leur effet sur la VD

17:00 26

26

Tout prochainement

- Prochaine leçon
  - Introduction à l'analyse causale
- Au labo SPSS d'aujourd'hui
  - Estimer un modèle de régression linéaire simple et calculer les coefficients standardisés bêtas
  - Estimer un modèle de régression incluant une VI dichotomique
  - Créer des variables factices dichotomiques à partir d'une variable qualitative
  - Estimer un modèle de régression incluant des variables factices dichotomiques
  - Présenter et interpréter les résultats

17:00 27

27