

Régression linéaire multiple

JM Galharret ¹

¹UFR de Psychologie
Université de Nantes

March 13, 2020

Régression
linéaire simple

JM Galharret

Introduction

Exemple

Matrice de corrélation

Modèle

Le test de Fisher

Les coefficients
standardisés β_j

1 Introduction

Exemple

Matrice de corrélation

2 Modèle

Le test de Fisher

Les coefficients standardisés β_j

Régression
linéaire simple

JM Galharret

Introduction

Exemple

Matrice de corrélation

Modèle

Le test de Fisher

Les coefficients
standardisés β_j

1 Introduction

Exemple

Matrice de corrélation

2 Modèle

Le test de Fisher

Les coefficients standardisés β_j

L'exemple du cours

Régression
linéaire simple

JM Galharret

Introduction

Exemple

Matrice de corrélation

Modèle

Le test de Fisher

Les coefficients
standardisés β_j

On a étudié le soutien apporté par un échantillon de 815 américains à la politique gouvernementale visant à atténuer la menace du changement climatique mondial :

- Pour cela on a posé 5 questions cotées de 1 (fortement opposé) à 7 (fortement en accord) et on a construit une variable govact en calculant le score moyen à ces 5 questions.
- On a également mesuré les réactions émotionnelles négatives (negemot) et positives (posemot) face au changement climatique. Ces variables ont été construites de façon qu'un score élevé reflète respectivement une forte réaction négative et une forte réaction positive.
- Les participants ont également donné une auto-évaluation de leur idéologie politique (ideology) sur une échelle de 1 à 7 (de très libéral à très conservateur).
- On connaît également l'âge des participants.

- On s'assure au préalable que la variable dépendante (Y) est liée aux prédicteurs (X_1, X_2, \dots, X_p) \rightsquigarrow **Matrice de corrélation**
- On écrit une équation de régression linéaire :

$$Y = b_0 + b_1X_1 + \dots + b_pX_p$$

(b_1, \dots, b_p) étant estimés par la **méthode des moindres carrés ordinaires**.

- On teste si l'équation de régression prédit significativement la variable Y .
 \rightsquigarrow **Table d'ANOVA du modèle**
- On teste chacune des variables X_1, \dots, X_p
 \rightsquigarrow **Test de Student sur un coefficient**
- On étudie le poids relatif de chacune des variables X_1, \dots, X_p dans l'équation de régression.
 \rightsquigarrow **Coefficients standardisés β** .

Table: Pearson Correlations

		govact	posemot	negemot	ideology	age
govact	Pearson's r	–				
	p-value	–				
posemot	Pearson's r	0.043	–			
	p-value	0.220	–			
negemot	Pearson's r	0.578	0.128	–		
	p-value	< .001	< .001	–		
ideology	Pearson's r	-0.418	-0.029	-0.349	–	
	p-value	< .001	0.402	< .001	–	
age	Pearson's r	-0.097	0.042	-0.057	0.212	–
	p-value	0.006	0.227	0.105	< .001	–

Concernant le lien entre les prédicteurs X_1, \dots, X_p et la variable réponse Y :

- Lien positif et significatif entre govact et negemot ($r = .578, p < .001$, donc lien fort).
- Lien positif et non significatif entre govact et posemot ($r = .043, p = .220$, lien très faible).
- Lien négatif et significatif entre govact et ideology ($r = -.418, p < .001$, donc lien moyen) et govact et age ($r = -.097, p < .001$, donc lien très faible).

Concernant les liens entre les prédicteurs X_1, \dots, X_p et govact certains sont significatifs (par exemple age et ideology). \rightsquigarrow ceci peut poser problème au niveau de l'équation de régression (voir chapitre 3).

Régression
linéaire simple

JM Galharret

Introduction

Exemple

Matrice de corrélation

Modèle

Le test de Fisher

Les coefficients
standardisés β_j

1 Introduction

Exemple

Matrice de corrélation

2 Modèle

Le test de Fisher

Les coefficients standardisés β_j

Retour sur les moindres carrés ordinaires

- Pour chaque individu de l'échantillon on connaît la valeur Y_i observée et on peut calculer la valeur prédite par l'équation de régression est :

$$\hat{Y}_i = b_0 + b_1 X_{1,i} + \dots + b_p X_{p,i}$$

- Comme dans le cas de la régression simple on associe le résidu ε_i :

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

- Comme dans le cas d'un seul prédicteur les coefficients de régression (b_0, b_1, \dots, b_p) minimiseront la somme des carrés des résidus:

$$S = \sum_{i=1}^N \varepsilon_i^2, \quad \text{Moindres Carrés Ordinaires}$$

↪ On ne peut pas calculer facilement ces coefficients.

L'estimation des coefficients

Régression
linéaire simple

JM Galharret

Introduction

Exemple

Matrice de corrélation

Modèle

Le test de Fisher

Les coefficients
standardisés β_j

Table: Coefficients de la regression de govact sur les variables posemot, negemot, ideology, age

	Unstand.	Stand.E	Stand.	<i>t</i>	<i>p</i>
(Intercept)	4.062	0.205		19.844	< .001
posemot	-0.027	0.028	-0.027	-0.966	0.334
negemot	0.441	0.026	0.496	16.766	< .001
ideology	-0.219	0.027	-0.243	-8.104	< .001
age	-0.001	0.002	-0.016	-0.576	0.564

Equation de régression

$$\text{govact} = 4.062 - 0.027(\text{posemot}) + 0.441(\text{negemot}) - 0.219(\text{ideology}) - 0.001(\text{age})$$

Exemples de calculs

Régression linéaire simple

JM Galharret

Introduction

Exemple

Matrice de corrélation

Modèle

Le test de Fisher

 Les coefficients
standardisés β_j

sujet	govact	posemot	negemot	ideology	age	sex	partyid
1	3.6	3.67	4.67	6	61	0	2
2	5	2	2.33	2	55	0	1
3	6.6	2.33	3.67	1	85	1	1
4	1	5	5	1	59	0	1
5	4	2.33	1.67	4	22	1	1
6	7	1	6	3	34	0	2

Pour l'individu 2 on a :

$$\widehat{govact} = 4.062 - 0.027 \times 2 + 0.441 \times 2.33 - 0.219 \times 2 - 0.001 \times 55 = 4.54$$

Donc l'erreur de modélisation (le résidu) pour l'individu 2 est de

$$\varepsilon = govact - \widehat{govact} = 5 - 4.54 = -0.46$$

La table d'ANOVA est identique à celle avec un seul prédicteur mais on ne peut plus faire les calculs correspondants. Le test de Fisher est identique :

Le test de FISHER

Soit l'équation de régression $Y = b_0 + b_1X_1 + \dots + b_pX_p$ où (b_0, b_1, \dots, b_p) ont été estimé par MCO. On teste l'hypothèse

$$H_0 : b_1 = b_2 = \dots = b_p = 0.$$

c'est-à-dire que l'équation de régression ne prédit pas significativement Y .

Si l'hypothèse H_0 est vraie, alors la variable $F = \frac{s_{Reg}^2}{s_{Res}^2}$ suit une loi de Fisher $\mathcal{F}(p, n - p - 1)$.

Table: ANOVA

	Sum of Squares	df	Mean Square	F	p
Regression	585.0	4	146.250	128.5	< .001
Residual	921.5	810	1.138		
Total	1506.5	814			

L'équation de régression obtenue par MCO prédit significativement la variable govact en fonction des variables posemot, negemot, ideology, age, $F(4, 810) = 128.5, p < .001$.

$R^2 = \frac{SCE_{Reg}}{SCE_Y} = \frac{585}{1506.5} = .388$. Cette équation prédit 38.8% de la variance de la variable govact.

Le test de STUDENT

On veut tester l'un des prédicteurs X_j de l'équation de régression. On pose $H_0 : b_j = 0$. On utilise la statistique de test $t = \frac{b}{SE(b)}$ qui sous H_0 suit une loi de Student à $(n - p - 1)$ ddl.

Exemple : Interprétation du coefficient de negemot dans l'équation de régression

$$t = \frac{0.441}{0.026} = 16.77.$$

	Unstand.	Stand.E	Stand.	t	p
negemot	0.441	0.026	0.496	16.766	< .001



Interprétation des coefficients

Régression linéaire simple

JM Galharret

Introduction

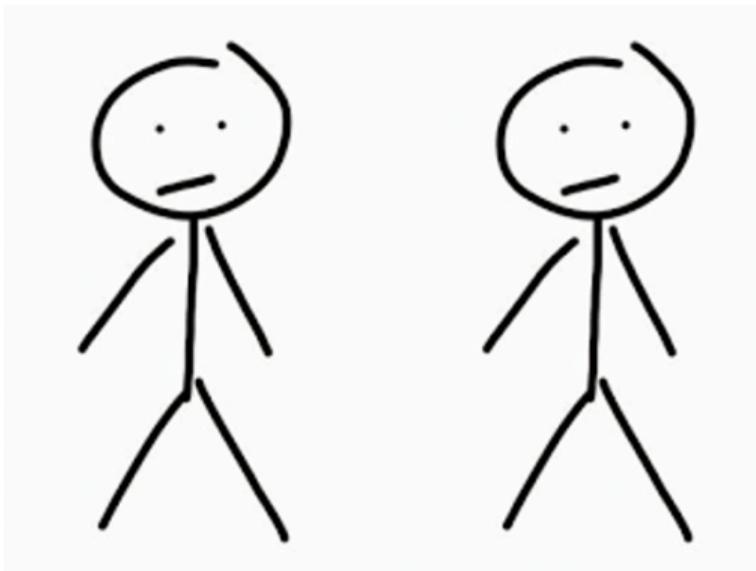
Exemple

Matrice de corrélation

Modèle

Le test de Fisher

Les coefficients
standardisés β_j



On considère deux individus ayant :

- même niveau d'émotion positive,
- même auto-évaluation de leur idéologie politique,
- même âge.

Si l'un des individus a une plus grande réaction négative par rapport au changement climatique alors le modèle prédit qu'il soutiendra significativement plus les actions gouvernementales ($b = 0.441$, $t(810) = 16.77$, $p < .001$).

Le coefficient standardisé β

- On veut comparer l'impact relatif de chacun des prédicteurs sur Y .
- On ne peut pas le faire directement à partir des estimations des coefficients b du modèle de régression. En effet, les prédicteurs X_j ne sont pas nécessairement sur la même échelle, ou n'ont pas les mêmes variabilités (ie variances).

La variable standardisée Z correspondante à la variable Y est définie par :

$$Z = \frac{Y - \bar{Y}}{s_Y}. \text{ Donc } \bar{Z} = 0, s_Z = 1.$$

On note Z_j les variables standardisées correspondantes aux variables X_j .

Définition

Les coefficients β_j sont définis par la relation :

$$Z = \beta_1 Z_1 + \beta_2 Z_2 + \dots + \beta_p Z_p$$

Un simple calcul montre que :

$$\beta_j = \frac{s_j}{s_Y} b_j$$

On en déduit que β_j et b_j sont de même signe et qu'ils ont la même significativité.

Table: Coefficients de la regression de govact sur les variables posemot, negemot, ideology, age

	Unstand.	Stand.E	Stand.	<i>t</i>	<i>p</i>
(Intercept)	4.062	0.205		19.844	< .001
posemot	-0.027	0.028	-0.027	-0.966	0.334
negemot	0.441	0.026	0.496	16.766	< .001
ideology	-0.219	0.027	-0.243	-8.104	< .001
age	-0.001	0.002	-0.016	-0.576	0.564