Traitement des variables exogènes qualitatives

Régression Linéaire Multiple

Ricco RAKOTOMALALA

PLAN

1. Introduction à l'analyse de variance
2. Codage disjonctif 0/1 : régression sans constante
3. Codage 0/1 « cornered effect »
4. Codage 0/1 « centered effect »

Remarque : si l'endogène est qualitative, on est dans un toute autre domaine (régression logistique et, plus généralement, dans le classement, la discrimination ou l'apprentissage supervisé)

ANOVA à 1 facteur
et
Régression Linéaire Multiple

Qu'est ce que l'ANOVA (analyse de variance),
quel rapport avec la régression ?

ANOVA à 1 facteur

Introduction et calculs

Loyer (Euro au m²)	Lieu Habitation
6.9	banlieue
6.3	banlieue
6.7	banlieue
6.5	banlieue
7.2	banlieue
5.6	campagne
4.9	campagne
5.3	campagne
7	centre
7.5	centre
8	centre
7.2	centre
8.4	centre
7.4	centre
8.3	centre

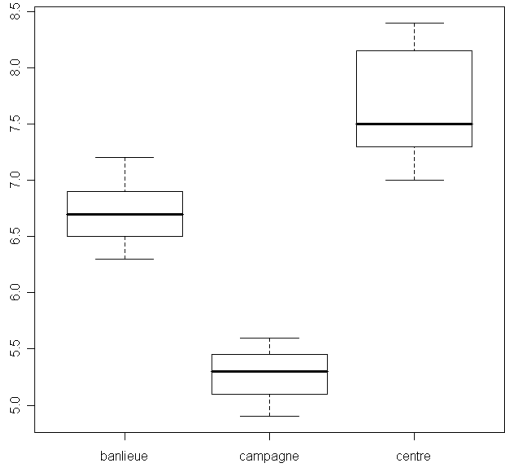
Exemple introductif : comparer les loyers selon la zone d'habitation

→ **Question** : le loyer est-il le même dans les différentes zones ?

Généralisation de la comparaison de moyenne à k populations

$$\begin{cases} H_0 : \mu_1 = \mu_2 = \cdots = \mu_k = \mu \\ H_1 : \exists j, \mu_j \neq \mu \end{cases}$$

Détection graphique : boîtes à moustaches conditionnelles



→ Décalage des points médians
→ Décalage des dispersions

Étude statistique : Analyse de variance (ANOVA)

Équation d'analyse de variance

$$SCT = SCE + SCR$$

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(y_{i,j}-\bar{y}\right)^2 = \sum_{j=1}^{k} n_j\left(\bar{y}_j-\bar{y}\right)^2 + \sum_{j=1}^{k}\sum_{i=1}^{n_j}\left(y_{i,j}-\bar{y}_j\right)^2$$

Tableau d'analyse de variance

Variation	Ddl	SC	CM
Expliqués (inter)	k-1	SCE	CME = SCE/(k-1)
Résiduels (intra)	n-k	SCR	CMR = SCR/(n-k)
Totaux	n-1	SCT	-

Statistique du test et région critique

$$F = \frac{CME}{CMR} \equiv F(k-1, n-k)$$

→ $R.C. : F > F_{1-\alpha}(k-1, n-k)$

ANOVA à 1 facteur (suite)

Relation avec la régression linéaire multiple

$$SCT = SCE + SCR$$

$$\sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{i,j}-\bar{y})^2 = \sum_{j=1}^{k} n_j(\bar{y}_j-\bar{y})^2 + \sum_{j=1}^{k}\sum_{i=1}^{n_j}(y_{i,j}-\bar{y}_j)^2$$

Dans notre exemple, il y a une différence significative entre les loyers moyens. On peut affiner l'analyse en essayant de détecter celle qui s'écarte le plus des autres, etc. (cf. cours ANOVA, analyse des contrastes, comparaisons multiples, etc.)

Lieu Habitation	Moyenne	n	n x Ecart moyenne²
banlieue	6.72	5	0.13
campagne	5.27	3	7.81
centre	7.69	7	4.54
Globale	6.88	15	

Tableau ANOVA			
Source	**ddl**	**SC**	**CM**
SCE	2	12.48	6.24
SCR	12	2.54	0.21
SCT	14	15.02	-

F	**29.4446**
p-value	**0.0000**

Loyer (Euro au m²)	Lieu Habitation
6.9	banlieue
6.3	banlieue
6.7	banlieue
6.5	banlieue
7.2	banlieue
5.6	campagne
4.9	campagne
5.3	campagne
7	centre
7.5	centre
8	centre
7.2	centre
8.4	centre
7.4	centre
8.3	centre

Quel lien avec la régression ?

Les valeurs de l'endogène peut se modéliser en plusieurs composantes

$$y_{i,j} = \mu + \alpha_j + \varepsilon_{i,j}$$

Où, ε est une composante d'erreur sur laquelle on peut faire des hypothèses

$$\varepsilon_{i,j} \equiv N(0,\sigma)$$

C'est ni plus ni moins qu'une régression

H0 de l'ANOVA devient H0 du test de significativité globale de la régression

$$\begin{cases} H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_k = 0 \\ H_1 : \exists j, \alpha_j \neq 0 \end{cases}$$

Dans la pratique, **tout dépend du type de codage de la variable exogène dans la régression**. Pour :
(1) Interpréter correctement les coefficients calculés
(2) Mettre en place les tests voulus (globalement pour une ANOVA, une modalité contre les autres, etc.)

Régression simple, à une seule variable qualitative exogène

Codage 0/1 et Régression sans constante

Codage disjonctif complet

Pour la régression à une seule exogène qualitative

Principe du codage 0/1

$$Z_{i,j} = \begin{cases} 1, X_i = j \\ 0, \text{sinon} \end{cases}$$

Écriture de la régression

$$loyer = a_1.z_{banlieue} + a_2.z_{campagne} + a_3.z_{c-ville} + \varepsilon$$

Pourquoi ne faut-il pas mettre la constante dans cette régression ?

Résultats Régression Linéaire

Loyer	Habitation	banlieue	campagne	centre
6.9	banlieue	1	0	0
6.3	banlieue	1	0	0
6.7	banlieue	1	0	0
6.5	banlieue	1	0	0
7.2	banlieue	1	0	0
5.6	campagne	0	1	0
4.9	campagne	0	1	0
5.3	campagne	0	1	0
7	centre	0	0	1
7.5	centre	0	0	1
8	centre	0	0	1
7.2	centre	0	0	1
8.4	centre	0	0	1
7.4	centre	0	0	1
8.3	centre	0	0	1

	centre	campagne	banlieue
coef.	7.69	5.27	6.72
std.dev	0.17	0.27	0.21
R^2	0.83	0.46	#N/A
	19.63	12	#N/A
	12.48	2.54	#N/A

Résultats ANOVA

Lieu Habitation	Moyenne	n	n x Ecart moyenne²
banlieue	6.72	5	0.13
campagne	5.27	3	7.81
centre	7.69	7	4.54
Globale	6.88	15	

Tableau ANOVA			
Source	**ddl**	**SC**	**CM**
SCE	2	12.48	6.24
SCR	12	2.54	0.21
SCT	14	15.02	-

F	**29.4446**
p-value	**0.0000**

- Les coefficients se lisent comme des moyennes conditionnelles
- SCE et SCR cohérents
- Le test global de significativité F = (12.48/2)/(2.54/12) = 29.4446

(Attention, les logiciels ne tiennent pas compte de la même manière des DDL dans la régression sans constante)

-> Très bonne solution, on retrouve nos résultats…

…MAIS, elle est inopérante pour une régression avec plusieurs exogènes qualitatives

Régression multiple, à une ou plusieurs variables qualitatives exogènes

Comment coder une exogène qualitative sans interférer avec les autres ?
Comment lire les coefficients estimés ?

Codage « cornered effect »

Principe

Omettre une des modalités (la dernière par ex.) de X puisqu'elle peut être déduite des autres (c'est la modalité de référence)

$$X_i = k \Rightarrow Z_{i,1} = Z_{i,2} = \cdots = Z_{i,k-1} = 0$$

Régression linéaire multiple

En omettant la modalité de référence

$$loyer = a_0 + a_1.z_{banlieue} + a_2.z_{campagne} + \varepsilon$$

Généralisable aux cas de plusieurs exogènes, dont certaines qualitatives recodées de la même manière

---

Résultats et interprétation des coefficients

Loyer	Habitation	banlieue	campagne
6.9	banlieue	1	0
6.3	banlieue	1	0
6.7	banlieue	1	0
6.5	banlieue	1	0
7.2	banlieue	1	0
5.6	campagne	0	1
4.9	campagne	0	1
5.3	campagne	0	1
7	centre	0	0
7.5	centre	0	0
8	centre	0	0
7.2	centre	0	0
8.4	centre	0	0
7.4	centre	0	0
8.3	centre	0	0

	campagne	banlieue	constante
coef.	-2.42	-0.97	7.69
	0.32	0.27	0.17
	0.83	0.46	#N/A
	29.44	12	#N/A
	12.48	2.54	#N/A

Moyenne conditionnelles

campagne	banlieue	centre
5.27	6.72	7.69

Test significativité globale

F	29.44
ddl1	2
ddl2	12
p-value	0.0000

Quel faut-il y lire ?

- Constante = Moyenne de la modalité de référence
- Coefficient = Écart de la moyenne de la modalité étudiée avec la moyenne de la modalité de référence
- Test de significativité d'un coefficient = Test d'écart significatif avec la modalité de référence
- Test global = Test si les différences de toutes les autres avec la référence est négligeable (Remarque : le F global est calculé correctement maintenant)

Codage « centered effect »

Principe

Omettre toujours une des modalités de X (la k-ème par ex.), le codage des variables indicatrices tient compte explicitement de cette modalité

$$Z_{i,j} = \begin{cases} +1, \text{si } X_i = j \\ -1, \text{si } X_i = k \\ 0 \end{cases}$$

Régression linéaire multiple

En omettant la modalité de référence

$$loyer = a_0 + a_1.z_{banlieue} + a_2.z_{campagne} + \varepsilon$$

Généralisable aux cas de plusieurs exogènes, dont certaines qualitatives recodées de la même manière

---

Résultats et interprétation des coefficients

Loyer	Habitation	banlieue	campagne
6.9	banlieue	1	0
6.3	banlieue	1	0
6.7	banlieue	1	0
6.5	banlieue	1	0
7.2	banlieue	1	0
5.6	campagne	0	1
4.9	campagne	0	1
5.3	campagne	0	1
7	centre	-1	-1
7.5	centre	-1	-1
8	centre	-1	-1
7.2	centre	-1	-1
8.4	centre	-1	-1
7.4	centre	-1	-1
8.3	centre	-1	-1

	campagne	banlieue	constante
coef.	-1.29	0.16	6.56
	0.20	0.17	0.13
	0.83	0.46	#N/A
	29.44	12	#N/A
	12.48	2.54	#N/A

Moyenne conditionnelles

campagne	banlieue	centre
5.27	6.72	7.69

Test significativité globale

F	29.44
ddl1	2
ddl2	12
p-value	0.0000

Quel faut-il y lire ?

- Constante = Valeur centrale / Référence : Moyenne non pondérée des moyennes conditionnelles (Rq. : ce n'est pas la moyenne globale sauf si effectif équilibrés)
- Coefficient = Écart de la moyenne conditionnelle de la variable avec cette valeur centrale
- Test de significativité d'un coefficient = Test d'écart significatif avec la valeur centrale
- Test global = Test si les différences de toutes les autres avec la valeur centrale est négligeable (Remarque : le F global est calculé correctement maintenant)

Conclusion

On peut utiliser la régression sur variables qualitatives pour réaliser une ANOVA (*comparaisons de moyennes en général*)

On ne peut pas utiliser directement les variables qualitatives, il faut les recoder.

Le type de codage adopté conditionne la lecture des résultats et l'interprétation des coefficients.

Attention aux pertes d'informations et à l'introduction d'une information fictive

- codage {1, 2, 3…} pour les variables nominales
- codage des variables ordinales

Bibliographie

En ligne

R. Rakotomalala, « Pratique de la Régression Linéaire Multiple – Diagnostic et sélection de variables ». Support de cours.
http://eric.univ-lyon2.fr/~ricco/cours/cours/La_regression_dans_la_pratique.pdf

R. Rakotomalala. Portail économétrie.
http://eric.univ-lyon2.fr/~ricco/cours/cours_econometrie.html

Wikipédia.
http://fr.wikipedia.org/wiki/Régression_linéaire_multiple

Ouvrages

M. Tenenhaus, « Statistique – Méthodes pour décrire, expliquer et prévoir », Dunod, 2006.

R. Bourbonnais, « Econométrie – Manuel et exercices corrigés », Dunod, 1998.

Y. Dodge, V. Rousson, « Analyse de régression appliquée », Dunod, 2004.