Rupture de Structure
Test de Chow

Régression Linéaire Multiple

Ricco RAKOTOMALALA

PLAN

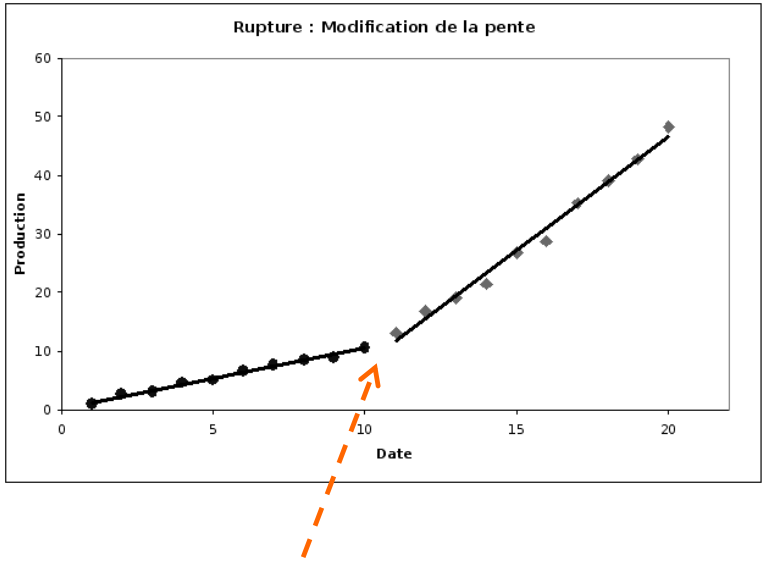
1. Changements structurels
2. Test de Chow – Principe
3. Stabilité de la constante
4. Stabilité de la pente

Changements structurels

Données longitudinales

Principe : La relation entre l'endogène et les exogènes est modifiée au cours du temps, à partir d'une date donnée. La régression sur l'ensemble de la période est différente de la régression sur les sous-périodes.

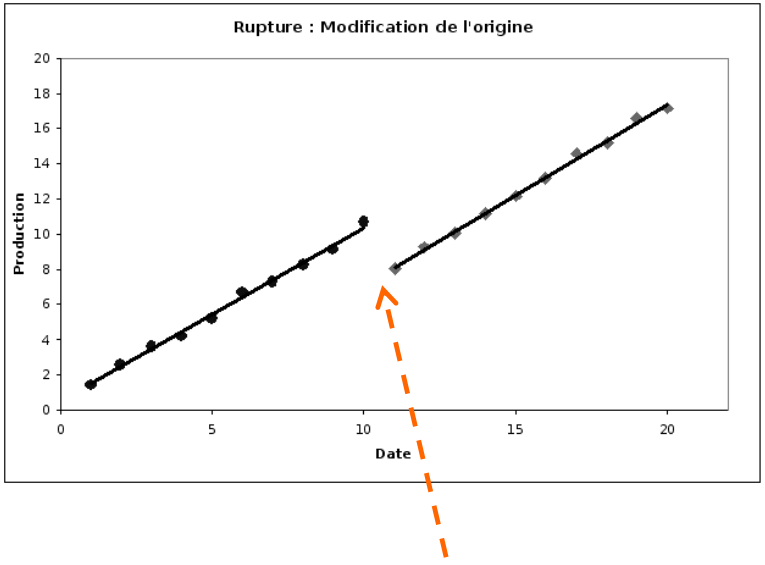
Ex. Niveau de production au cours du temps.



Mutation technologique

(Accélération de la production)

→ Modification de la pente



Crise, guerre, accidents…

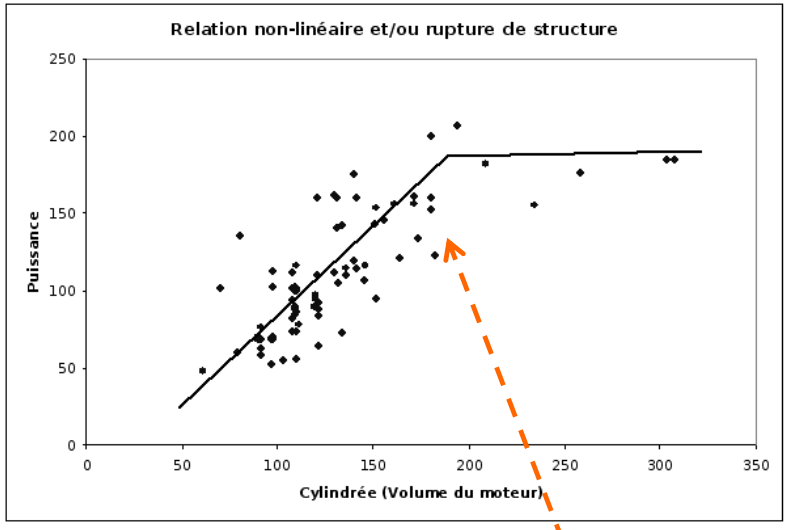
(Recul, puis poursuite avec la même pente)

→ Modification de l'origine

Changements structurels

Données transversales

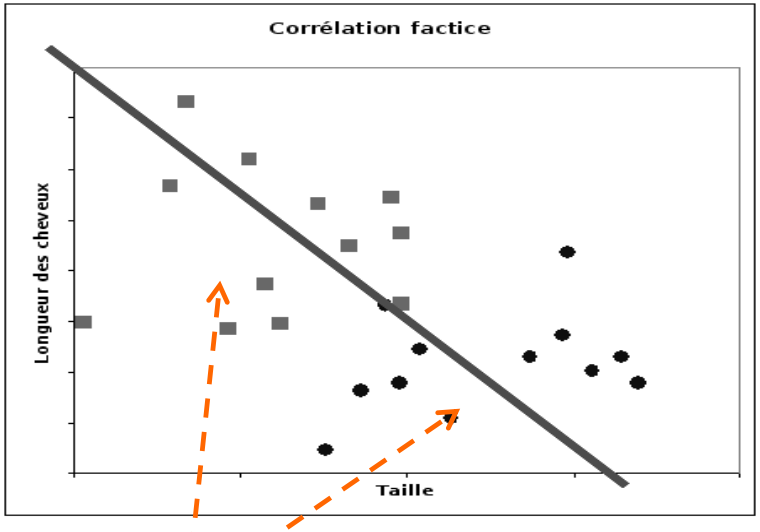
Principe : Vérifier que la régression est la même dans des sous-parties de l'échantillon. Soit parce la relation est non-linéaire (point d'inflexion), soit parce qu'elles correspondent à des sous-populations différentes.



Relation non-linéaire ou linéaire par morceaux.

Problème : comment deviner ces points d'inflexion dans une régression multiple ?

→ Souvent s'en tenir à des techniques graphiques simples (cf. les résidus partiels : résidus + aj x Xj [ordonnée] vs. Xj [abscisse])



Régression sur 2 populations.

Problème : comment détecter l'existence de sous populations ?

→ Connaissances experte

→ S'appuyer sur des variables disponibles (utilisées ou non dans la régression)

Tester les changements structurels – Test de Chow

Principe de la régression contrainte vs. Régressions non-contraintes

Principe : (1) Régression contrainte = régression sur l'ensemble de l'échantillon → SCR ; (2) Régressions non-contraintes = régressions sur les deux sous-échantillons définies sur les populations (ou sur les dates) → SCR1 et SCR2

Par construction : $SCR \geq SCR_1 + SCR_2$

- SCR = SCR1 + SCR2 si les régressions sont exactement les mêmes
- SCR >> SCR1 + SCR2 à mesure que les régressions diffèrent

Principe : Le test de Chow est aussi un test d'hypothèses

Les régressions à comparer

$$(1)\ y_i = a_0 + a_1 x_{i,1} + \cdots + a_p x_{i,p} + \varepsilon_i \quad (i = 1, \ldots, n)$$

$$(2)\begin{cases} y_i = a_{0,1} + a_{1,1} x_{i,1} + \cdots + a_{p,1} x_{i,p} + \varepsilon_{i,1} \quad (i = 1, \ldots, n_1) \\ y_i = a_{0,2} + a_{1,2} x_{i,1} + \cdots + a_{p,2} x_{i,p} + \varepsilon_{i,2} \quad (i = n_1 + 1, \ldots, n)[n_2 \text{ obs.}] \end{cases}$$

$$\rightarrow\ H_0 : \begin{pmatrix} a_0 \\ a_1 \\ \vdots \\ a_p \end{pmatrix} = \begin{pmatrix} a_{0,1} \\ a_{1,1} \\ \vdots \\ a_{p,1} \end{pmatrix} = \begin{pmatrix} a_{0,2} \\ a_{1,2} \\ \vdots \\ a_{p,2} \end{pmatrix}$$

Statistique du test :

$$F = \frac{\left[SCR - (SCR_1 + SCR_2)\right] / \text{ddl}_n}{(SCR_1 + SCR_2) / \text{ddl}_d} \equiv F(\text{ddl}_n, \text{ddl}_d)$$

Région critique (au risque α) : $F > F_{1-\alpha}(\text{ddl}_n, \text{ddl}_d)$

Comment calculer les d.d.l. ?

$$\text{ddl}_d = (n_1 - p - 1) + (n_2 - p - 1) = n - 2p - 2$$
$$= n - 2(p+1)$$
$$\text{ddl}_n = (n - p - 1) - \left[(n_1 - p - 1) + (n_2 - p - 1)\right]$$
$$= p + 1$$

Plus que les valeurs génériques, **c'est la démarche qu'il faut retenir**. On y reviendra pour tester des configurations particulières

Test de Chow

Un exemple

Tester si les régressions sont globalement identiques sur les 2 sous-périodes

Obs	Periode	Y	X
1	1	1	2
2	1	2	4
3	1	2	6
4	1	4	10
5	1	6	13
6	2	1	2
7	2	3	4
8	2	3	6
9	2	5	8
10	2	6	10
11	2	6	12
12	2	7	14
13	2	9	16
14	2	9	18
15	2	11	20

Régression globale

	X	const.
coef.	0.52	-0.07
	0.03	0.37
	0.95	0.71
	252.71	13
	127.44	**6.56**
		SCR

Régression période 1

	X	const.
coef.	0.44	-0.06
	0.05	0.43
	0.96	0.48
	66.82	3
	15.31	**0.69**
		SCR1

Régression période 2

	X	const.
coef.	0.51	0.40
	0.03	0.38
	0.97	0.56
	276.71	8
	85.53	**2.47**
		SCR2

ddl_n	2
ddl_d	11

SCR-(SCR1+SCR2)	3.40
SCR1+SCR2	3.16

F	5.91
p-value	0.0181

$$F = \frac{3.40/2}{3.16/11} = 5.91$$

À comparer avec 3.98 pour un test à 5%

→ Le résultat est cohérent avec la p-value

→ Au risque 5%, on conclut que les régressions (le lien entre Y et X) sont différentes sur les 2 sous-périodes.

Détecter la nature de la rupture

Tester la stabilité de la constante

Principe : Détecter si, sur les 2 sous-périodes (sous-populations), la constante de la régression est la même.

→ Rég. Contrainte : Régression « normale » sur tous les points (SCR)

→ Rég. Non-Contrainte : Régression avec « 2 constantes », une pour chaque sous-période (SCR3)

$$(1)\ y_i = a_0 + a_1 x_{i,1} + \cdots + a_p x_{i,p} + \varepsilon_i \quad (i = 1, \ldots, n)$$

$$(2)\ y_i = a_{0,1} d_{i,1} + a_{0,2} d_{i,2} + a_{1,1} x_{i,1} + \cdots + a_{p,1} x_{i,p} + \varepsilon_{i,1} \quad (i = 1, \ldots, n)$$

→ avec

$$d_{i,1} = \begin{cases} 1, i = 1, \ldots, n_1 \\ 0, i = n_1 + 1, \ldots, n \end{cases}$$

$$d_{i,2} = \begin{cases} 0, i = 1, \ldots, n_1 \\ 1, i = n_1 + 1, \ldots, n \end{cases}$$

Pourquoi pas 2 régressions distinctes pour (2) ?

Pourquoi ne pas mettre de constante dans la régression (2) ?

Un test de comparaison entre $a_{0,1}$ et $a_{0,2}$ dans (2) aurait fait l'affaire aussi.

Obs	Periode	Y	X	D1	D2
1	1	1	2	1	0
2	1	2	4	1	0
3	1	2	6	1	0
4	1	4	10	1	0
5	1	6	13	1	0
6	2	1	2	0	1
7	2	3	4	0	1
8	2	3	6	0	1
9	2	5	8	0	1
10	2	6	10	0	1
11	2	6	12	0	1
12	2	7	14	0	1
13	2	9	16	0	1
14	2	9	18	0	1
15	2	11	20	0	1

	D2	D1	X
coef.	0.55	-0.47	0.50
	0.34	0.30	0.03
	0.97	0.54	#N/A
	149.57	12	#N/A
	130.51	3.49	#N/A

SCR	6.56
SCR3	3.49
SCR-SCR3	3.07

ddl_n	1
ddl_d	12

F	10.5409
p-value	0.0070

Statistique F

$$F = \frac{(SCR - SCR_3)/ddl_n}{SCR_3/ddl_d} = 10.54$$

D.D.L.

$$ddl_n = (15-2) - 12 = 1$$

$$ddl_d = 15 - 3 = 12$$

La constante n'est pas la même dans les 2 régressions !

Détecter la nature de la rupture

Stabilité de la pente sur une variable : une formulation erronée

Principe : Détecter si, sur les 2 sous-périodes (sous-populations), le coefficient d'une des variables de la régression est le même.

→ Rég. Contrainte : Régression « normale » sur tous les points (SCR)

→ Rég. Non-Contrainte : Régression avec la variable scindée en 2 parties, une pour chaque sous-période c.-à-d.

$$z_{i,1} = \begin{cases} x_{i,1}, i = 1, \ldots, n_1 \\ 0, i = n_1 + 1, \ldots, n \end{cases}$$

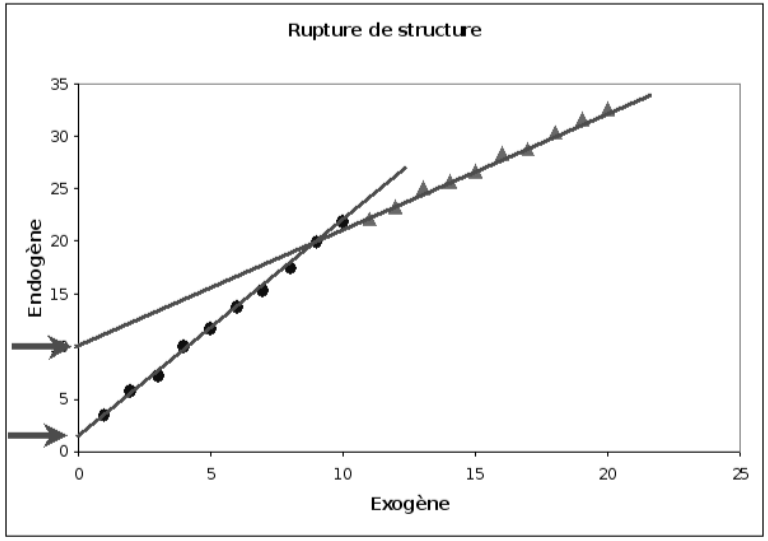
$$z_{i,2} = \begin{cases} 0, i = 1, \ldots, n_1 \\ x_{i,1}, i = n_1 + 1, \ldots, n \end{cases}$$

On opposerait ainsi…

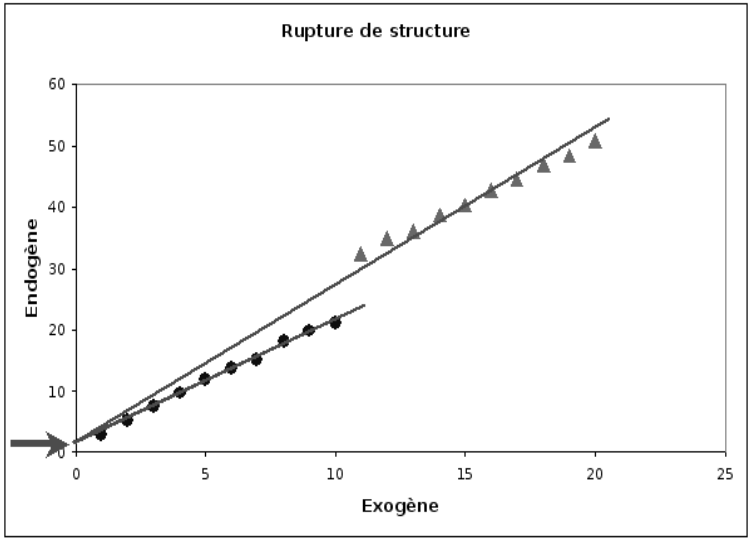
$$(1)\ y_i = a_0 + a_1 x_{i,1} + a_2 x_{i,2} + \cdots + a_p x_{i,p} + \varepsilon_i \quad (i = 1, \ldots, n)$$

$$(2)\ y_i = b_0 + b_{1,1} z_{i,1} + b_{1,2} z_{i,2} + b_2 x_{i,2} + \cdots + b_p x_{i,p} + \varepsilon_i \quad (i = 1, \ldots, n)$$

Mais ce n'est pas une bonne approche !



Un changement de pente entraîne forcément un changement d'origine



Forcer l'égalité de l'origine dans les deux sous-parties fausse l'appréciation de la pente

Détecter la nature de la rupture

Stabilité de la pente sur une variable : la formulation correcte

Principe : Détecter si, sur les 2 sous-périodes (sous-populations), le coefficient d'une des variables de la régression est le même. Il faut relâcher la contrainte d'égalité de constante dans le modèle de référence.

→ Rég. Contrainte : Régression avec constantes différentes sur les 2 sous-périodes (SCR3)

→ Rég. Non-Contrainte : En plus, la variable est scindée en 2 parties pour chaque sous-période (SCR4)

On oppose donc

$$(1)\ y_i = a_{0,1}d_{i,1} + a_{0,2}d_{i,2} + a_1x_{i,1} + a_2x_{i,2} + \cdots + a_px_{i,p} + \varepsilon_i \quad (i = 1,\ldots,n)$$

$$(2)\ y_i = b_{0,1}d_{i,1} + b_{0,2}d_{i,2} + b_{1,1}z_{i,1} + b_{1,2}z_{i,2} + b_2x_{i,2} + \cdots + b_px_{i,p} + \varepsilon_i \quad (i = 1,\ldots,n)$$

Obs	Periode	Y	X	D1	D2	Z1	Z2
1	1	1	2	1	0	2	0
2	1	2	4	1	0	4	0
3	1	2	6	1	0	6	0
4	1	4	10	1	0	10	0
5	1	6	13	1	0	13	0
6	2	1	2	0	1	0	2
7	2	3	4	0	1	0	4
8	2	3	6	0	1	0	6
9	2	5	8	0	1	0	8
10	2	6	10	0	1	0	10
11	2	6	12	0	1	0	12
12	2	7	14	0	1	0	14
13	2	9	16	0	1	0	16
14	2	9	18	0	1	0	18
15	2	11	20	0	1	0	20

	Z2	Z1	D2	D1
coef.	0.51	0.44	0.40	-0.06
	0.03	0.06	0.37	0.48
	0.98	0.54	#N/A	#N/A
	113.86	11	#N/A	#N/A
	130.84	3.16	#N/A	#N/A

SCR_3	3.49
SCR_4	3.16
SCR_3-SCR_4	0.33

ddl_n	1
ddl_d	11

F	1.15
p-value	0.3068

Statistique F

$$F = \frac{(SCR_3 - SCR_4)/ddl_n}{SCR_4/ddl_d} = 1.15$$

D.D.L.

$$ddl_n = (15-3) - 11 = 1$$

$$ddl_d = 15 - 4 = 11$$

La pente est la même dans les 2 régressions !

Conclusion

Détecter des ruptures de structures est primordial pour éviter les interprétations infondées.

Assez facile à mettre en œuvre pour les données longitudinales. Plus difficile à appréhender pour les données transversales.

Le principe du test repose sur une opposition entre régression contrainte (pas de rupture) et une régression non-contrainte (intégrant la rupture).

Attention à la tentation de la recherche systématique. En programmant des procédures testant toutes les éventualités par exemple. A force de chercher, on finit toujours par trouver quelque chose de significatif. Ici aussi, l'interprétation économique (l'expertise du domaine) est absolument nécessaire pour valider les résultats.

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.
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.