22/10/2024

Econométrie spatiale
Partie 3

M2 ECAP

Muriel Travers
LEMNA, IAE Economie et Management
Université de Nantes

Nantes Université

5) Estimation de modèles "spatiaux"

Spatial Lag, Spatial Autoregressive (SAR) : $y=\rho Wy+X\beta+\varepsilon$

➔ La valeur de y dans une unité spatiale peut avoir un impact sur la valeur de y dans une unité spatiale voisine
ρ : paramètre spatial autorégressif indiquant l'ampleur de l'interaction existant entre les observations de y

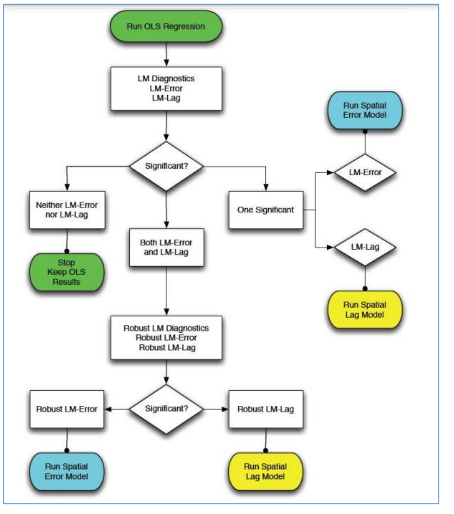
Spatial Error (SEM): $y= X\beta+u$, $u=\lambda Wu+\varepsilon$

➔ Les résidus ε dans une unité spatiale peuvent avoir un impact sur les résidus dans une unité spatiale voisine.
λ : paramètre représentant l'intensité de l'autocorrélation spatiale entre les résidus de la régression

A la différence du modèle précédent (SAR), on ne spécifie pas de relation particulière. La spatialisation est vu comme un « bruit» que l'on doit corriger : $y = X\beta + (I-\lambda W)^{-1} u$

➔ La détection de l'autocorrélation spatiale des erreurs : problème dans la spécification du modèle, telle que l'omission de variables significatives

<u>Choix entre les modèles LM-Error (SEM) et les modèles LM-Lag (SAR)</u> :
Approche « ascendante » (Le Gallo, 2002)



Si les deux sont significatifs, prendre celui qui a la p-value la plus faible

<u>Si les deux Robust LM tests ne sont pas significatifs alors que les 2 LM Diagnostics le sont</u> :

* <u>Test Robust LM-Lag non significatif</u> : l'hypothèse de la dépendance spatiale dans la variable dépendante n'est pas robuste après contrôle de l'autocorrélation spatiale des termes d'erreur.

* <u>Test Robust LM-Error non significatif</u> : l'hypothèse de l'autocorrélation spatiale dans les résidus n'est pas robuste après contrôle de la dépendance spatiale dans la variable dépendante

➔ Étant donné que les 2 Robust LM tests ne sont pas significatifs, cela implique qu'il n'y a pas de preuve solide en faveur d'un processus de retard spatial ou d'erreur spatiale exclusivement.

<u>Solutions</u> :

- Changer la spécification du modèle : variable omise ? Forme fonctionnelle à revoir
- Envisager d'autres types de modèles spatiaux : SDM, SLX

Attention : la lecture des effets dans le cas du modèle SAR n'est pas directe : il existe 2 types d'effets

$$\begin{aligned} Y &= (1-\rho W)^{-1} X\beta + (1-\rho W)^{-1}\varepsilon \\ &= \sum_{r=1}^{k} (1-\rho W)^{-1} \beta_r X_r + (1-\rho W)^{-1}\varepsilon \end{aligned}$$

La valeur prédite est donc : $\hat{y} = (1-\hat{\rho} W)^{-1} X\hat{\beta}$

Cette valeur prédite est donc différente de $X\hat{\beta}$ (MCO)

L'effet marginal pour une variable quantitative n'est plus β_r mais $(1\text{-}\rho W)^{-1}\beta_r$

- Termes diagonaux de $(1\text{-}\rho W)^{-1}\beta_r$: effets directs pour chaque zone d'une modification de la variable X_r dans la même zone.
- Autres termes de $(1\text{-}\rho W)^{-1}\beta_r$: effets indirects : impact de la modification de la variable X_r dans une zone sur une autre zone.

On peut donc calculer :

- Un effet direct moyen : moyenne des effets sur une zone d'une modification d'une unité de la variable X_r

- Un effet indirect moyen : moyenne des effets d'une modification d'une unité de la variable X_r dans une zone sur l'ensemble des zones

- Un effet total moyen = effet direct moyen + effet indirect moyen

Application sous R :

FVPTHH02 : Infractions contre les familles et les enfants (violence domestique) pour 1 000 ménages
Fonction de :
UNEMPP : % de chômeurs
SALESPC : Vente par habitant par an d'alcool, $

Estimation du modèle MCO

```
mco<-lm(FVPTHH02~UNEMPP+SALESPC, data=carte)
summary(mco)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	1.167254	0.404012	2.889	0.00423	**
UNEMPP	0.121642	0.065226	1.865	0.06346	.
SALESPC	- 0.002540	0.004064	-0.625	0.53259	

---
*Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

Residual standard error: 2.947 on 231 degrees of freedom
Multiple R-squared: 0.01483, Adjusted R-squared: 0.006304
F-statistic: 1.739 on 2 and 231 DF, p-value: 0.178

<u>Autocorrélation spatiale des résidus du MCO</u> : H0 : absence d'autocorrélation spatiale des résidus

$$I = \frac{\hat{\varepsilon}' W \hat{\varepsilon}}{\hat{\varepsilon}'\hat{\varepsilon}}$$

$$E(I) = tr(MW)/(N-K)$$
$$M = I - X(X'X)^{-1}X'$$
$$Var(I) = \frac{tr(MWMW') + tr(MWMW) + [tr(MW)]^2}{(N-K)(N-K+2)} - (E(I))^2$$

$$Z = \frac{I - E(I)}{\sqrt{Var(I)}} \sim N(0,1)$$

Si p-value <0,05 ➔ H_0 refusé

moran.lm<-lm.morantest(mco, WQueen, alternative="two.sided")
print(moran.lm)

Global Moran I for regression residuals

data:
model: lm(formula = FVPTHH02 ~ UNEMPP + SALESPC, data = carte)
weights: WQueen

Moran I statistic standard deviate = 4.2134, p-value = 2.515e-05
alternative hypothesis: two.sided
sample estimates:

Observed Moran I	Expectation	Variance
0.187391543	-0.004816203	0.002081008

➔ Au seuil de risque de 1%, les résidus sont corrélés spatialement (selon la matrice de poids choisie) ➔ MCO non adapté

LM<-lm.LMtests(mco, WQueen, test=c("LMerr", "LMlag", "RLMerr", "RLMlag"))
print(LM)

les deux derniers sont les versions robustes

Rao's score (a.k.a Lagrange multiplier) diagnostics for spatial dependence

model: lm(formula = FVPTHH02 ~ UNEMPP + SALESPC, data = carte)
test weights: listw
RSerr = 16.492, df = 1, p-value = 4.885e-05

Rao's score (a.k.a Lagrange multiplier) diagnostics for spatial dependence

model: lm(formula = FVPTHH02 ~ UNEMPP + SALESPC, data = carte)
test weights: listw
RSlag = 18.773, df = 1, p-value = 1.473e-05

Rao's score (a.k.a Lagrange multiplier) diagnostics for spatial dependence

model: lm(formula = FVPTHH02 ~ UNEMPP + SALESPC, data = carte)
test weights: listw
adjRSerr = 10.307, df = 1, p-value = 0.001325

Rao's score (a.k.a Lagrange multiplier) diagnostics for spatial dependence

model: lm(formula = FVPTHH02 ~ UNEMPP + SALESPC, data = carte)
test weights: listw
adjRSlag = 12.587, df = 1, p-value = 0.0003884

library(spatialreg)
sar<-lagsarlm(FVPTHH02~UNEMPP+SALESPC, data=carte, WQueen)
summary(sar)

Type: lag
Coefficients: (asymptotic standard errors)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.6946964	0.4019265	1.7284	0.08391
UNEMPP	0.0934239	0.0618525	1.5104	0.13093
SALESPC	-0.0014521	0.0038529	-0.3769	0.70625

Rho: 0.32803, LR test value: 16.189, p-value: 5.7325e-05

positif et significatif: impact des voisins

Asymptotic standard error: 0.079746
z-value: 4.1135, p-value: 3.8979e-05
Wald statistic: 16.92, p-value: 3.8979e-05

Log likelihood: -575.3167 for lag model
ML residual variance (sigma squared): 7.8026, (sigma: 2.7933)
Number of observations: 234
Number of parameters estimated: 5
AIC: 1160.6, (AIC for lm: 1174.8)
LM test for residual autocorrelation
test value: 6.5777, p-value: 0.010326

il y a problème d'auto spatiale dans les résidus donc il faut tester d'autres modèles car on peut pas utiliser le Sem?

impacts.sar<-impacts(sar, listw=WQueen)
impacts.sar

imaginons qu'on avait eu 0.11 au test précédent, on devrait calculer ce modèle

Impact measures (lag, exact):

	Direct	*Indirect*	*Total*
UNEMPP	*0.095887905*	*0.0431424129*	*0.139030318*
SALESPC	*-0.001490429*	*-0.0006705822*	*-0.002161012*

sem<-errorsarlm(FVPTHH02~UNEMPP+SALESPC, data=carte, WQueen)
summary(sem)

Type error
Coefficients: (asymptotic standard errors)

	Estimate	*Std. Error*	*z value*	*Pr(>|z|)*
(Intercept)	*1.3498805*	*0.4443509*	*3.0379*	*0.002383*
UNEMPP	*0.0698486*	*0.0638664*	*1.0937*	*0.274101*
SALESPC	*-0.0015187*	*0.0038553*	*-0.3939*	*0.693639*

Lambda: 0.32528, LR test value: 15.077, p-value: 0.0001032
Asymptotic standard error: 0.0801
z-value: 4.0609, p-value: 4.8874e-05
Wald statistic: 16.491, p-value: 4.8874e-05

Log likelihood: -575.8726 for error model
ML residual variance (sigma squared): 7.8432, (sigma: 2.8006)
Number of observations: 234
Number of parameters estimated: 5
AIC: 1161.7, (AIC for lm: 1174.8)

<u>Application sur Géoda :</u>
Cliquer sur Regression, sélectionner la variable à expliquer ainsi que les variables explicatives
Sélectionner Classic
Cliquer sur Run

pas appliquer dans le dossier

```
SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION
Data set             :  NCVACO Variables
Dependent Variable   :    FVPTHH02  Number of Observations:  234
Mean dependent var   :     1.64369  Number of Variables   :    3
S.D. dependent var   :     2.94982  Degrees of Freedom    :  231

R-squared            :    0.014834  F-statistic           :     1.73907
Adjusted R-squared   :    0.006304  Prob(F-statistic)     :    0.177976
Sum squared residual:      2005.93  Log likelihood        :    -583.411
Sigma-square         :     8.68369  Akaike info criterion :     1172.82
S.E. of regression   :     2.94681  Schwarz criterion     :     1183.19
Sigma-square ML      :     8.57236
S.E of regression ML:      2.92786

-----------------------------------------------------------------------------
       Variable      Coefficient      Std.Error    t-Statistic   Probability
-----------------------------------------------------------------------------
          CONSTANT       1.16725       0.404012        2.88916       0.00423
            UNEMPP      0.121642      0.0652257        1.86494       0.06346
           SALESPC   -0.00254001     0.00406404      -0.624996       0.53259
-----------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
MULTICOLLINEARITY CONDITION NUMBER   4.316351
TEST ON NORMALITY OF ERRORS
TEST                      DF           VALUE             PROB
Jarque-Bera                2         1752.3606           0.00000

DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                      DF           VALUE             PROB
Breusch-Pagan test         2           10.5140           0.00521
Koenker-Bassett test       2            1.5147           0.46890
```


Regression, sélectionner la variable à expliquer ainsi que les variables explicatives
Cliquer sur Weights File
Sélectionner Wqueen
Cliquer sur Spatial Lag
Cliquer sur Run

```
SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set              : NCVACO Variables
Spatial Weight        : WQueen
Dependent Variable    :    FVPTHH02  Number of Observations:  234
Mean dependent var    :     1.64369  Number of Variables   :    4
S.D. dependent var    :     2.94982  Degrees of Freedom    :  230
Lag coeff.   (Rho)    :    0.328032

R-squared             :    0.103302  Log likelihood        :    -575.317
Sq. Correlation       : -            Akaike info criterion :     1158.63
Sigma-square          :     7.80256  Schwarz criterion     :     1172.45
S.E of regression     :     2.79331

-----------------------------------------------------------------------------
       Variable      Coefficient      Std.Error        z-value   Probability
-----------------------------------------------------------------------------
        W_FVPTHH02      0.328032      0.0797462        4.11345       0.00004
          CONSTANT      0.694696       0.401926        1.72842       0.08391
            UNEMPP     0.0934239      0.0618525        1.51043       0.13093
           SALESPC   -0.00145213     0.00385285      -0.376897       0.70625
-----------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                        DF      VALUE        PROB
Breusch-Pagan test                          2         5.7778     0.05564

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL LAG DEPENDENCE FOR WEIGHT MATRIX : WQueen
TEST                                        DF      VALUE        PROB
Likelihood Ratio Test                       1        16.1890     0.00006
```

Regression, sélectionner la variable à expliquer ainsi que les variables explicatives
Cliquer sur Weights File
Sélectionner Wqueen
Cliquer sur Spatial Error
Cliquer sur Run

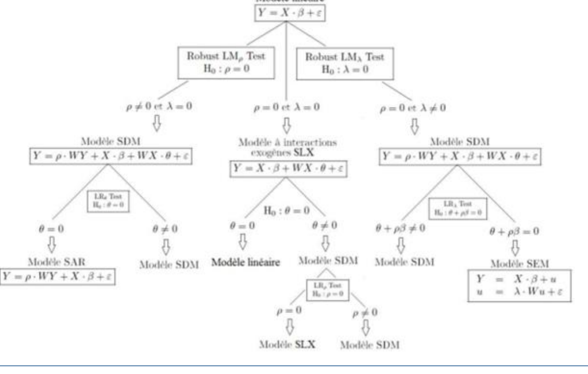
```
SUMMARY OF OUTPUT: SPATIAL ERROR MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : NCVACO Variables
Spatial Weight      : WQueen
Dependent Variable  :    FVPTHH02  Number of Observations:  234
Mean dependent var  :    1.643689  Number of Variables   :    3
S.D. dependent var  :    2.949819  Degrees of Freedom    :  231
Lag coeff. (Lambda) :    0.325284

R-squared           :    0.098635  R-squared (BUSE)      : -
Sq. Correlation     : -            Log likelihood        : -575.872550
Sigma-square        :     7.84317  Akaike info criterion :     1157.75
S.E of regression   :     2.80057  Schwarz criterion     :     1168.11

-----------------------------------------------------------------------------
       Variable       Coefficient     Std.Error       z-value    Probability
-----------------------------------------------------------------------------
          CONSTANT       1.34988      0.444351        3.03787      0.00238
            UNEMPP     0.0698486     0.0638664        1.09367      0.27410
           SALESPC   -0.00151868     0.0038553      -0.393921      0.69364
            LAMBDA      0.325284     0.0801005        4.06095      0.00005
-----------------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                    DF       VALUE        PROB
Breusch-Pagan test                      2         3.7522      0.15318

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL ERROR DEPENDENCE FOR WEIGHT MATRIX : WQueen
TEST                                    DF       VALUE        PROB
Likelihood Ratio Test                   1        15.0773      0.00010
```


<u>**Il existe d'autres types de modèles**</u> :

$y = \rho Wy + X\beta + WX\theta + \varepsilon$ Spatial Durbin Model : SDM

y = $X.\beta + WX.\theta + \varepsilon$: Spatial Lag of X model : SLX

➔ <u>Méthode de sélection</u> : Approche « mixte » : Elhorst (2010)



Etape 1 :
➔ Estimation par MCO

Etape 2 :
Tests de Lagrange (RLMerr, RLMlag)

- Si non significatifs ➔ Estimation du modèle SLX : $y = X.\beta + WX.\theta + \varepsilon$

Si θ=0 , on estime alors un modèle MCO

Etape 3 :
Si $\theta \neq 0$, on estime alors un modèle SDM et on le compare à un modèle SLX

Par conséquent, on teste
H_0 : ρ =0 (modèle SLX)
H_1 : $\rho \neq 0$ (modèle SDM)

LRstat= $-2(LL_{SLX}-LL_{SDM}) \sim \chi 2(1)$

Si LRStat > valeur théorique ➔ On retient le modèle SDM sinon on choisit le modèle SLX

Etape 2 :
Tests de Lagrange (RLMerr, RLMlag)

- Si un des 2 (ou les 2 sont) significatifs ➔ Estimation du modèle SDM

Etape 3 :
On teste tout d'abord
H_0 : θ=0 (modèle SAR)
H_1 : $\theta \neq 0$ (modèle SDM)

LRstat= $-2(LL_{SAR}-LL_{SDM}) \sim \chi 2(1)$
Si LRStat > valeur théorique ➔ On retient le modèle SDM sinon on choisit le modèle SAR

On teste ensuite :
H_0 : $\theta + \rho\beta = 0$ (modèle SEM)
H_1 : $\theta + \rho\beta \neq 0$ (modèle SDM)

LRstat= $-2(LL_{SEM}-LL_{SDM}) \sim \chi 2(1)$
Si LRStat > valeur théorique ➔ On retient le modèle SDM sinon on choisit le modèle SEM

Remarque :

Si on choisit le modèle SLX, l'effet d'une variable X_r est β_r. L'effet indirect est θ_r

Si on choisit le modèle SDM, du fait de la présence de variable décalée, l'effet marginal d'une variable X_r est : $(I-\rho W)^{-1}(I_n\beta_r + W\theta_r)$

➔ Il faut calculer des effets directs et indirects du même type que pour le modèle SAR

Application sous R :

```
#SLX non nécessaire ici puisque à l'étape 2, les tests sont significatifs
slx<-lmSLX(FVPTHH02~UNEMPP+SALESPC, data=carte, WQueen)
AIC(slx)
impacts(slx, listw=WQueen)

#estimation d'un modèle SDM
sdm<-lagsarlm(FVPTHH02~UNEMPP+SALESPC, data=carte, WQueen, type="mixed")
```


summary(sdm)

Type: mixed
Coefficients: (asymptotic standard errors)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.52011616	0.83529137	-1.8199	0.068780
UNEMPP	0.05232352	0.06346151	0.8245	0.409660
SALESPC	0.00069946	0.00386369	0.1810	0.856340
lag.UNEMPP	0.42110905	0.13674529	3.0795	0.002073
lag.SALESPC	0.00652359	0.00876196	0.7445	0.456553

le chômage des voisins (lag) a un impact sur le taux de violence du comté mais pas le chômage du comté même

$y = \rho W y + X\beta + WX\theta + \varepsilon$

FVPTHH02

Rho: 0.24765, LR test value: 8.4384, p-value: 0.0036738
Asymptotic standard error: 0.082887
z-value: 2.9878, p-value: 0.00281
Wald statistic: 8.9269, p-value: 0.00281

le rho est significatif

Log likelihood: -570.2191 for mixed model
ML residual variance (sigma squared): 7.5532, (sigma: 2.7483)
Number of observations: 234
Number of parameters estimated: 7
AIC: 1154.4, (AIC for lm: 1160.9)
LM test for residual autocorrelation
test value: 2.7638, p-value: 0.096422

au-dessus de 0.05, pas d'erreur au niveau des résidus et on peut garder ce modèle.

Etape 3 : Test du rapport de vraisemblance

#SDM : Modèle non contraint
#SEM : Modèle contraint
TestSDM_SEM<-LR.Sarlm(sdm,sem)
print(TestSDM_SEM)
Likelihood ratio for spatial linear models
Likelihood ratio = 11.307, df = 2, p-value = 0.003506
sample estimates:
Log likelihood of sdm Log likelihood of sem
-570.2191 -575.8726

#SDM: Modèle non contraint
#SAR : Modèle contraint
TestSDM_SAR<-LR.Sarlm(sdm,sar)
print(TestSDM_SAR)
Likelihood ratio for spatial linear models
Likelihood ratio = 10.195, df = 2, p-value = 0.006112
sample estimates:
Log likelihood of sdm Log likelihood of sar
-570.2191 -575.3167

➔ Il existe une différence significative entre SDM et SEM. On choisit le modèle ayant l'AIC le plus faible : SDM (1154 versus 1162)

➔ Il existe une différence significative entre SDM et SAR. On choisit le modèle ayant l'AIC le plus faible : SDM (1154 versus 1161)

impacts.sdm<-impacts(sdm, listw=WQueen)
impacts.sdm

Impact measures (mixed, exact):

	Direct	*Indirect*	*Total*
UNEMPP	*0.077468875*	*0.551801621*	*0.629270496*
SALESPC	*0.001087406*	*0.008513227*	*0.009600633*

➔ Faire le TD3 (cf. Madoc)