ESTADISTICA ESPACIAL

REGRESIÓN ESPACIAL

De vuelta…

- “Todos los lugares están relacionados pero los lugares más cercanos están más relacionados que los lugares lejanos”
- Los fenómenos sociales y físicos están frecuentemente agrupados en el espacio:

ej. Patrones de votación regional, segregación racial, cinturón de pobreza, cáncer de pulmón, valor de la vivienda, delincuencia, incendios forestales, hábitats de animales, especies de plantas, química del suelo

Análisis Espacial

- Frecuentemente estas relaciones espaciales son ignoradas

Debilidad de nuestra habilidad para generar inferencias significativas sobre los procesos que estudiamos

- Los modelos de regresión espacial incluyen relaciones entre variables y sus valores de vecindad
  - Incluye como variables explicativas el valor del error, valores X o Y en regiones cercanas

- Nos permite examinar el impacto que una observación tiene sobre observaciones próximas

¿Por qué preocuparse por las similitudes espaciales? (1)

- Nos dice algo más sobre lo que estamos estudiando
  - ✓ Hay un proceso que no se puede medir que afecta la salida sobre lo que estamos interesados?
  - ✓ Este proceso se manifiesta en el espacio?
  - ✓ Ej: procesos de interacción, difusión, legado étnico o histórico, efectos programáticos

¿Por qué preocuparse por las similitudes espaciales? (2)

- Violación de los supuestos de regresión
  - ✓ Los residuales no están correlacionados con cada uno
  - ✓ La varianza no es constante
- Si ignoramos las relaciones espaciales en nuestros datos:
  - ✓ Nuestros coeficientes de regresión estimados están sesgados/inconsistentes
  - ✓ El estadístico R2 es exagerado
  - ✓ Haremos hecho inferencias incorrectas
  - ✓ Nunca será publicado (no debería ser)
- Si los efectos espaciales están presentes y no se tienen en cuenta, el modelo no es exacto!!!

Si la autocorrelación espacial ocurre

- Puede haber Xs no medidas las cuales causan el fallo de independencia
    - ✓ Error no especificado
- Puede haber un proceso de “contagio” en el trabajo (Ys en una sola ubicación puede afectar las Ys en ubicaciones adyacentes)
- Los valores de Y pueden depender del valor de X en el mismo sitio así como en sitios cercanos
- Los errores estimados pueden estar correlacionados espacialmente entre unidades

Cómo tratar el componente espacial?

- Como un efecto sustantivo de interés
  - ✓ Incorporardentro de un modelo/explorar
  - ✓ Ej. Retardo espacial, regímenes espaciales
- Como un efecto de ruido debido a errores de especificación
  - ✓ Eliminar/Controlar
  - ✓ Ej. Error espacial

Pregunta

- “Un error entre la unidad espacial de observación y la extensión espacial de un fenómeno bajo consideración resultará en errores de medida espacial y autocorrelación espacial entre estos errores y ubicaciones contiguas?” Anselin & Bera, 1998
- ¿Por qué?

El Modelo de error espacial

- Examina la autocorrelación espacial entre los residuales de áreas adyacentes
- Trata la correlación espacial principalmente como un ruido
  - ✓ No tiene en cuenta la idea que la correlación espacial puede reflejar algún proceso significativo
- Errores espaciales positivos pueden reflejar un modelo no especificado (particularmente una variable omitida que es espacialmente agrupable)
- Si ignoramos los errores espaciales en los residuales:
  - ✓ Coeficientes imparciales
  - ✓ Los errores estándar están equivocados (p-valor equivocado)

Autocorrelación espacial en residuales
Modelo de error espacial

- Incorpora efectos espaciales a través del error

$$y = x\beta + \varepsilon$$

$$\varepsilon = \lambda W\varepsilon + \xi$$

Donde:

- ε Es el vector de términos de error, ponderados espacialmente por la matriz de pesos
- λ es el coeficiente de error espacial
- ξ es el vector de errores no correlacionados
- Si no hay correlación espacial entre los errores, entonces $\lambda = 0$

Modelo de retardo espacial

- Incorpora dependencia espacial adicionando un “rezago espacial” DV (y) en la parte derecha de la ecuación de regresión
  - ✓ Otros modelos más complejos tampien incluyen rezafos expaciales IV (x)

- Tratar la correlación espacial como procesos o efectos de interés
  - ✓ Los valores de Y en un área están directamente influenciados por los valores de Y que se encuentran en las áreas vecinas.
  - ✓ Depende en cómo se defina el vecindario

Modelo de retardo espacial

- Un retardo espacial positivo provee evidencia que las Ys en áreas adyacentes covarían
- Si ignoramos la influencia del retardo espacial:
  - ✓ Coeficientes estarán sesgados
    - ❖ Si hay un efecto positivo de la vecindad en Ys, usualmente los coeficientes estarán sesgados hacia arriba
    - ❖ Los errores estándar serán equivocados (P-valor equivocado)

Autocorrelación Espacial en DV
Modelo de retardo espacial

- Incorpora efectos espaciales mediante la incorporación de una variable dependiente retardada espacialmente como un predictor adicional

$$y = \rho Wy + x\beta + \varepsilon$$

- Donde,

Wy es el DV espacialmente retrasado para los pesos de la matriz W

x es una matriz de observaciones de las variables explicativas

ε *es un vector de errores*

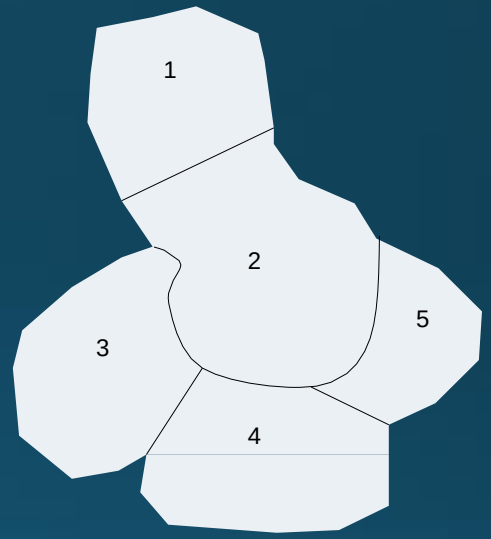
ρ *Es el coeficiente espacial t*

Si no hay dependencia espacial, y Y no depende de valores Y en la vecindad, $\rho=0$

Cómo calcular el retardo espacial?

- Y es el promedio de todos los vecinos

$$W = \begin{matrix} 0 & 1 & 0 & 0 & 0 \\ .25 & 0 & .25 & .25 & .25 \\ 0 & .5 & 0 & .5 & 0 \\ 0 & .33 & .33 & 0 & .33 \\ 0 & .5 & 0 & .5 & 0 \end{matrix}$$



Area	y	*Wy*
1	5	(1*7)=7
2	7	(.25*5)+(.25*9)+(.25*12)+(.25*11)=9.25
3	9	(.5*7)+(.5*12)=9.5
4	12	(.33*7)+(.33*9)+(.33*11)=8.91
5	11	(.5*7)+(.5*12)=9.5

¿Qué tipo de modelo SR usar?

- Si los residuales están espacialmente correlacionados (Moran's I), entonces usar el diagnóstico Multiplicador Lagrange para determinar el modelo apropiado
    - Residuales de la regresión (LM-Error)
        - No correspondencia del proceso y unidades espaciales errores sistemáticos, correlacionados a través de unidades espaciales →
    - Variable dependiente (LM-Retardo)
        - El proceso ha dado lugar a la distribución de variables agrupadas influencia de los valores vecinos en los valores unitarios →
    - Autocorrelación espacial en ambos

Regresión Espacial en R
Ejemplo: Precios de vivienda en Boston

CRIM	per capita crime rate by town
ZN	proportion of residential land zoned for lots over 25,000 ft^2
INDUS	proportion of non-retail business acres per town
CHAS	Charles River dummy variable (=1 if tract bounds river; 0 otherwise)
NOX	Nitrogen oxide concentration (parts per 10 million)
RM	average number of rooms per dwelling
AGE	proportion of owner-occupied units built prior to 1940
DIS	weighted distances to five Boston employment centres
RAD	index of accessibility to radial highways
TAX	full-value property-tax rate per $10,000
PTRATIO	pupil-teacher ratio by town
B	$1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town
LSTAT	% lower status of the population
MEDV	Median value of owner-occupied homes in $1000's

Regresión Espacial en R

- Cargar Boston.shp
- Definir vecindad (k más cercano w/datos punto)
- Crear matriz de pesos
- Test de Moran de DV, gráfica de Moran
- Correr regresión OLS
- Verificar residuales de dependencia espacial
- Determinar cual modelo de RE utilizar
- Correr modelo de regresión espacial

Moran I en el DV

- `moran.test(boston$LOGMEDV, listw= bost_kd1_w)`

- `Moran's I test under randomisation data: boston$LOGMEDV`
- `weights: bost_kd1_w`

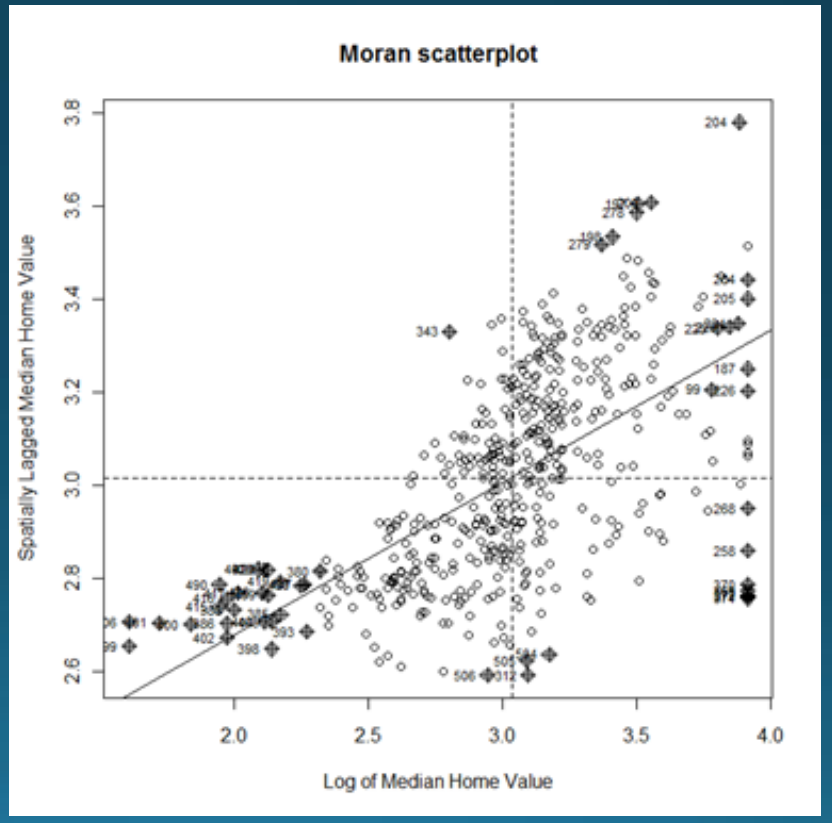
- `Moran I statistic standard deviate = 24.5658, p-value < 2.2e-16`
- `alternative hypothesis: greater sample estimates:`

```
Moran I statistic       Expectation          Variance

     0.3273430100      -0.0019801980     0.0001797138
```

Moran para el DV

```
> moran.plot(boston$LOGMEDV, bost_kd1_w,
labels=as.character(boston$ID))
```



Regresión OLS

```
bostlm<-lm(LOGMEDV~RM + LSTAT + CRIM + ZN + CHAS + DIS,
data=boston)

Residuals:
   Min   1Q Median 3Q   Max
-0.71552 -0.11248 -0.02159  0.10678  0.93024

Coefficients:
        Estimate Std. Error t value Pr(>|t|)

(Intercept)  2.8718878  0.1316376  21.817  < 2e-16 ***
RM           0.1153095  0.0172813   6.672 6.70e-11 ***
LSTAT       -0.0345160  0.0019665 -17.552  < 2e-16 ***
CRIM        -0.0115726  0.0012476  -9.276  < 2e-16 ***
ZN           0.0019330  0.0005512   3.507 0.000494 ***
CHAS         0.1342672  0.0370521   3.624 0.000320 ***
DIS         -0.0302262  0.0066230  -4.564 6.33e-06 ***
---

Residual standard error: 0.2081 on 499 degrees of freedom
Multiple R-squared: 0.7433,         Adjusted R-squared: 0.7402
F-statistic: 240.8 on 6 and 499 DF,  p-value: < 2.2e-16
```

Verificar residuales para la Autocorrelación espacial

```
> boston$lmresid<-residuals(bostlm)
> lm.morantest(bostlm,bost_kd1_w)

 Global Moran's I for regression residuals
 Moran I statistic standard deviate = 5.8542, p-value = 2.396e-09
alternative hypothesis: greater
Sample estimates

Observed Moran's I     Expectation            Variance
 0.0700808323          -0.0054856590          0.0001666168
```

Determinar el tipo de dependencia

```
> lm.LMtests(bostlm, bost_kd1_w, test="all")

   Lagrange multiplier diagnostics for spatial dependence

LMerr = 26.1243, df = 1, p-value = 3.201e-07

LMlag = 46.7233, df = 1, p-value = 8.175e-12

RLMerr = 5.0497, df = 1, p-value = 0.02463

RLMlag = 25.6486, df = 1, p-value = 4.096e-07

SARMA = 51.773, df = 2, p-value = 5.723e-12
```

- Tests robustos usados para encontrar una alternativa apropiada
- Solo use tests robustos cuando **AMBOS** LMErr y LMLag son significativos

Un diagnóstico adicional

```
> library(lmtest)

> bptest(bostlm)

        studentized Breusch-Pagan test

data:  bostlm
BP = 70.9173, df = 6, p-value = 2.651e-13
```

- Indica que los errores son HETEROSCEDÁSTICOS: no es de extrañar, puesto que tenemos dependencia espacial

Correr un modelo de retardo espacial

```
> bostlag<-lagsarlm(LOGMEDV~RM + LSTAT + CRIM + ZN + CHAS +
   DIS, data=boston, bost_kd1_w)
Type: lag
Coefficients: (asymptotic standard errors)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.94228260	0.19267675	10.0805	< 2.2e-16
RM	0.10158292	0.01655116	6.1375	8.382e-10
LSTAT	-0.03227679	0.00192717	-16.7483	< 2.2e-16
CRIM	-0.01033127	0.00120283	-8.5891	< 2.2e-16
ZN	0.00166558	0.00052968	3.1445	0.001664
CHAS	0.07238573	0.03608725	2.0059	0.044872
DIS	-0.04285133	0.00655158	-6.5406	6.127e-11

```
Rho: 0.34416, LR test value:37.426, p-value:9.4936e-10 Asymptotic
standard error: 0.051967
    z-value: 6.6226, p-value: 3.5291e-11 Wald statistic:
43.859, p-value: 3.5291e-11

Log likelihood: 98.51632 for lag model
ML residual variance (sigma squared): 0.03944, (sigma: 0.1986) AIC: -179.03, (AIC
for lm: -143.61)
```

Unos diagnósticos más

- Test LM de Autocorrelación residual

```
test value: 1.9852, p-value: 0.15884
> bptest.sarlm(bostlag)

        studentized Breusch-Pagan test
data:
BP = 60.0237, df = 6, p-value = 4.451e-11
```

- La prueba LM sugiere que no hay más correlación espacial en los datos
- La prueba BP indica que permanece la heteroscedasticidad en los residuales, probablemente debido a malas especificaciones

Correr un modelo de error espacial

```
> bosterr<-errorsarlm(LOGMEDV~RM + LSTAT + CRIM + ZN + CHAS +
   DIS, data=boston, listw=bost_kd1_w)
Type: error
Coefficients: (asymptotic standard errors)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	2.96330332	0.13381870	22.1442	< 2.2e-16
RM	0.09816980	0.01700824	5.7719	7.838e-09
LSTAT	-0.03413153	0.00194289	-17.5674	< 2.2e-16
CRIM	-0.01055839	0.00125282	-8.4277	< 2.2e-16
ZN	0.00200686	0.00062018	3.2359	0.001212
CHAS	0.06527760	0.03766168	1.7333	0.083049
DIS	-0.02780598	0.01064794	-2.6114	0.009017

```
Lambda: 0.59085, LR test value: 24.766, p-value:
6.4731e-07 Asymptotic standard error: 0.086787
   z-value: 6.8081, p-value: 9.8916e-12 Wald
statistic: 46.35, p-value: 9.8918e-12

Log likelihood: 92.18617 for error model
ML residual variance (sigma squared): 0.03989, (sigma:
0.19972) AIC: -166.37, (AIC for lm: -143.61)
```

Por qué no usar R^2

- R^2 no es una medida confiable del ajuste del modelo para la regresión espacial
- R^2 es calculado basado en la relación entre la variación explicada y no explicada (residual)
  - Requiere que los residuales sean independientes uno de otro
- La razón para usar la regresión espacial es que encontremos Autocorrelación espacial en los residuales
  - Ej. Variaciones explicadas y no explicadas no son independientes en este escenario