Geographically Weighted Regression

Gustavo A. García

ggarci24@eafit.edu.co

Econometría avanzada II

PhD/Maestría en Economía

Universidad EAFIT

Link slides en formato html

Link slides en formato PDF

En este tema

- Motivación
- Especificación de una GWR
- Función de densidad kernel y ancho de banda h
- Multicolinealidad
- GWR restringido
- GWR multi-escala (MGWR)
- Ejercicio aplicado en R

Lecturas

- Brunsdon,C., Fotheringham, A., Charlton, M. (1996). "Geographically weighted regression: a method for exploring spatial nonstationarity". *Geographical Analysis*, 28(4):281-298
- Fotheringham, A., Brunsdon,C., Charlton, M. (2002). *Geographically weighted regression: the analysis of spatially varying relationships*. John Wiley & Sons
- Wheeler, D.C. (2007). "Diagnostic tools and a remedial method for collinearity in geographically weighted regression." *Environment and Planning A*, 39(10):2464–2481
- Wheeler, D.C., Páez, A. (2010). Geographically Weighted Regression. In: Fischer, M., Getis, A. (eds) Handbook of Applied Spatial Analysis. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-03647-7_22
- Wheeler, D.C. (2014). Geographically Weighted Regression. In: Fischer, M., Nijkamp, P. (eds) Handbook of Regional Science. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-23430-9_77
- Algunas webs: crd230, quarcs-lab, MGWR
- Paquetes en R: `spgwr`, `gwrr`, `GWmodel`, `GWPR.light` y `mgwrsar`

Motivación

Si estamos interesados en la influencia o efecto de algunas variables x sobre alguna variable y, y asumiendo una relación lineal, se puede estimar el siguiente modelo de regresión lineal múltiple (RLM):

$$y_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} + u_i$$

Aquí i representa una unidad de observación espacial (bloques/manzanas, barrios, comunas, ciudades, regiones...)

Como sabemos, uno de los supuestos del modelo de RLM es que la relación entre las x y y es estacionaria

Efectos $(\beta's)$ estacionarios

- los coeficientes β no cambian a través del tiempo ni el espacio
- en estadística espacial, estacionariedad equivale a la homogeneidad de un efecto o, lo que es lo mismo, a que un proceso funciona igual independientemente de dónde se observe el proceso

El supuesto de estacionariedad de los β puede ser débil, y es posible preguntarnos si las x afectan a y en forma diferente dependiendo de la localización geográfica analizada

En este tema, se va a analizar la desigual distribución espacial en la relación entre dos o más variables x y y. El método que se va a cubrir intenta modelar la heterogeneidad espacial, esto es la Regresión Geográficamente Ponderada o Geographically Weighed Regression (GWR)

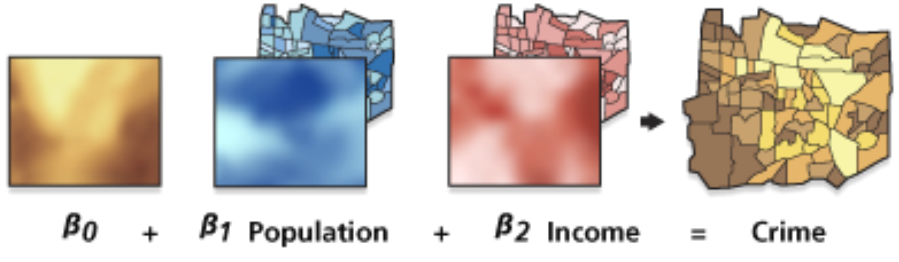
Especificación de una GWR

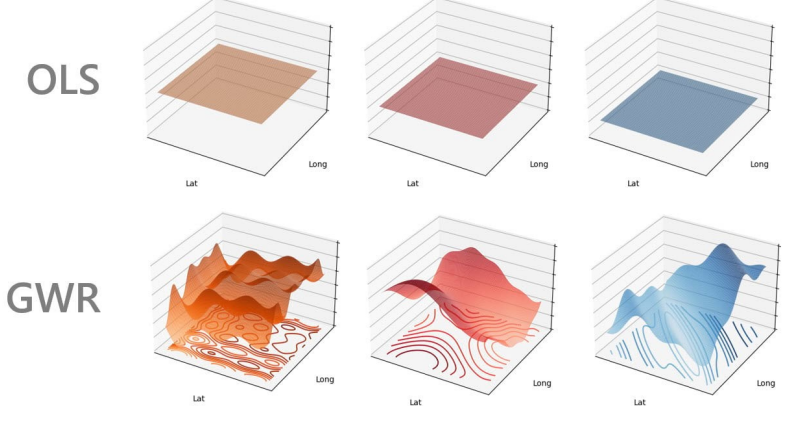
La GWR fue propuesta por Brunsdon et al. (1996) y tiene como objetivo estimar los β en cada localización i, usando los centróides de los poligonos de los datos utilizados. El modelo tiene la siguiente estructura:

$$y_i = \beta_{i0} + \sum_{k=1}^{p} \beta_{ik} x_{ik} + u_i$$

donde β_{ip} es la realización local de β_p en la localización i

Visualmente sería:





La GWR es una evolución de la regresión por mínimos cuadrados ordinarios (MCO) y añade un nivel de sofisticación al modelo al permitir que las relaciones entre las variables independientes y dependientes varíen según la localización

Especificación de una GWR

Los coeficientes de regresión son estimados para cada localización independientemente por mínimos cuadrados ponderados. La matriz de ponderación es una matriz diagonal en la que cada elemento diagonal w_{ij} es una función de la localización de la observación. La matriz de coeficientes estimados tiene la forma:

$$\hat{\boldsymbol{\beta}}(i) = [\mathbf{X}^T\mathbf{W}(i)\mathbf{X}]^{-1}\mathbf{X}^T\mathbf{W}(i)\mathbf{y}$$

donde $\mathbf{W}(i) = diag[w_1(i), \ldots, w_n(i)]$ es la matriz diagonal de pesos que varia en cada localización i. La matriz de ponderación da más valor a las observaciones que están cerca de i ya que se supone que las observaciones cercanas se influyen mutuamente más que las lejanas (ley de Tobler)

El modelo básico de regresión MCO es sólo un caso especial del modelo GWR, en el que los coeficientes son constantes en el espacio

Al estimar una GWR hay que tener en cuenta tres principales aspectos:

- la función de densidad kernel que asigna las ponderaciones w_{ij}
- el ancho de banda (*bandwidth*) h de la función, que determina el grado de decaimiento de la distancia
- a quién considerar como vecinos

Función de densidad kernel y ancho de banda h

La función de densidad kernel determina el peso asignado a las unidades vecinas

Existen varias funciones de densidad que se pueden utilizar, las más comúnes son:

- la función ponderada Gaussiana:

$$w_{ij} = exp\left(-\frac{d_{ij}^2}{h^2}\right)$$

donde d_{ij} es la distancia entre la localización i y j, y h es el ancho de banda

- la función bicuadrada

$$w_{ij} = 1 - \left(\frac{d_{ij}^2}{h^2}\right)^2$$

- la función tricúbica:

$$w_{ij} = 1 - \left(\frac{d_{ij}^3}{h^3}\right)^3$$

Función de densidad kernel y ancho de banda h

Escoger la función de ponderación también implica escoger el ancho de banda h. Existen diferentes formas de hacer esto, pero resaltan dos métodos comúnmente utilizados: el método *cross-validation* (CV) y la minimización del crítero de información de Akaike (AIC)

- CV
  En este método se intenta encontar la h que minimice la CV. La idea es minimizar la suma de los errores al cuadrado en todas las localizaciones i, y se llega a un ancho de banda óptimo. El CV toma la forma:

$$CV = \sum_i [y_i - \hat{y}_{\neq i}(\beta)]^2$$

donde $\hat{y}_{\neq i}(\beta)$ es el valor estimado de y_i con la observación diferente al punto i

- AIC
  Minimización del AIC

El procedimiento de estimación del modelo GWR implica:

1. el ancho de banda kernel es estimado por CV o AIC
2. Los ponderadores son calculados utilizando alguna de las funciones de densidad
3. los coefcientes de regresión son estimados en cada localización i

Multicolinealidad

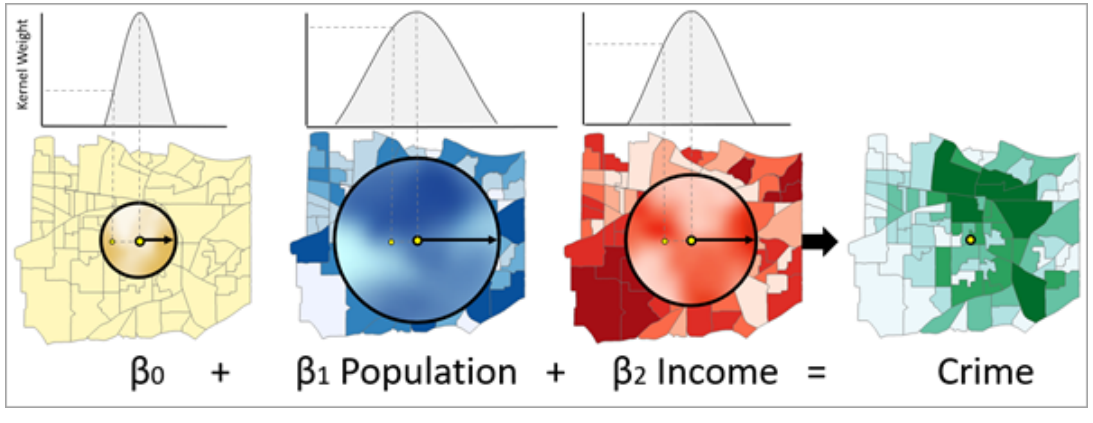
- Un problema con el modelo GWR es la correlación con los coeficientes estimados, parcialmente debido a la colinealidad en las variables explicatorias de cada modelo local
- El problema de multicolinealidad surge ya que se están usando valores de las variables explicativas en cada modelo local que son muy similares ya que son cercanas en el espacio, y al utilizar un ponderador similar para las observaciones cercanas, se está intensificando la similaridad entre las variables explicativas
- La multicolinealidad de las variables explicatorias localmente ponderadas puede llevar a potencial fuerte dependencia en los coeficientes locales estimados. Esta fuerte dependencia en los coeficientes estimados hace que la interpretación de los coeficientes individuales sea, en el mejor de los casos, tenue, y en el peor, engañosa
- Otro aspecto de la multicolinealidad es que en modelos lineales las varianzas de los coeficientes se inflan. Varianzas infladas en los coeficientes de regresión asociada a la colinealidad local en el modelo GWR puede llevar a sobreestimaciones de las magnitudes del efecto de las covariables y a la inversión del signo de los coeficientes, lo que puede dar lugar a interpretaciones incorrectas de las relaciones en el modelo de regresión
- Otro problema del modelo GWR son los errores estándar asociados a las estimaciones de los coeficientes de regresión. Los cálculos del error estándar en el modelo GWR son sólo aproximados debido a la reutilización de datos para la estimación de parámetros en múltiples ubicaciones y debido al uso de los datos para estimar tanto el ancho de banda del kernel con validación cruzada como los coeficientes de regresión
- Además, como ya se ha indicado, la colinealidad local puede aumentar las varianzas de los coeficientes de regresión estimados. Este problema con los errores estándar indica que los intervalos de confianza de los coeficientes GWR estimados son sólo aproximados y no son exactamente fiables para indicar los efectos estadísticamente significativos de las covariables y la selección de modelos

GWR restringido

- Los problemas derivados de la colinealidad pueden resolverse limitando la cantidad de variación de los coeficientes de regresión
- En el caso del modelo GWR, se han propuesto dos versiones de métodos que logran este objetivo: *geographically weighted ridge regression* (GWRR) y el *geographically weighted lasso regression* (GWL)
- Estas técnicas son basadas en las regresiones *ridge* y *lasso*, y los métodos funcionan penalizando la regresión para limitar la variación de los coeficientes. En ambos casos, se introduce una restricción en el tamaño de los coeficientes de regresión
- Los coeficientes de la regresión *ridge* minimizan la suma de una penalización sobre el tamaño de los coeficientes al cuadrado y la suma de los residuales al cuadrados
- Los coeficientes del *lasso* minimizan la suma del valor absoluto de los coeficientes y la suma de los residuales al cuadrados
- Tanto en la regresión *ridge* como en el *lasso*, es práctica común centrar la variable de respuesta, y centrar y escalar las variables explicativas para que tengan varianzas unitarias (estandarizar las variables), porque los métodos dependen de la escala

GWR multi-escala (MGWR)

- En un modelo GWR estándar un sólo ancho de banda o *bandwidth* es determinado y aplicado a cada variable explicativa.
- Sin embargo, en la realidad puede suceder que las relaciones en algunos procesos espacialmente heterogéneos operen sobre escalas más grandes que en otros
- En este caso, la escala de no estacionariedad de la relación determinada por una GWR estándar puede subestimar o sobrestimar la escala de las relaciones individuales entre la variable dependiente y las explicativas
- Para abordar esta limitación del GWR estándar, un GWR multi-escala (MGWR) puede ser usado (Yang 2014; Fotheringham, Yang, and Kang 2017; Oshan et al. 2019) $\Longrightarrow$ Este modelo determina el *bandwidth* para cada una de las variables explicativas, permitiendo así que varíen las relaciones individuales entre Y y cada X



- Trabajos recientes han sugerido que el MGWR debería ser el GWR por defecto (Comber et al. 2022), utilizándose un GWR estándar sólo en circunstancias específicas

Ejercicio aplicado en R

En este ejercicio se van a utilizar los datos de Columbus, una ciudad en el estado de Ohio en Estados Unidos. La idea es analizar los efectos del ingreso y el valor de la vivienda sobre el nivel de crimen a nivel de barrio.

Archivos a descargar:

- Descripción de los datos
- Código