UNIVERSIDAD NACIONAL DE COLOMBIA

Departamento de Estadística
Estadística Espacial
Análisis de área
10 de marzo - 2025

Jhon Alejandro Ramírez Daza jramirezda@unal.edu.co

---

1. Descripción de los datos:

Se usaron los datos Sudden Infant Death Syndrome presente en Anselin 2003 en geodacenter.

Estos datos corresponden a los casos de Síndrome de muestre infantil para lo condados de carolina del norte en dos periodos de tiempo (1974-1978 y 1979-1984) junto con las siguientes 17 variables siendo 100 observaciones, es decir 100 áreas territoriales de los condados de carolina del norte:

Variable	**Descripción**
AREA	Área del condado (computada por ArcView)
PERIMETER	Perímetro del condado (computado por ArcView)
CNTY-ID	ID interno del condado
NOMBRE	Nombre del condado
FIPS	Código de FIPS del condado, como carácter (código de Estado y código del condado)
FIPSNO	Código de FIPS del condado, numérico, utilizado en la Guía y tutoriales de GeoDa
CRESS-ID	ID del condado utilizado por Cressie (1993)
BIR74	Nacimientos en vida, 1974-78
SID74	Muertes por Síndrome de Muerte Súbita del Lactante (SIDS), 1974-78
NWBIR74	Nacimientos no blancos, 1974-78
BIR79	Nacimientos en vida, 1979-84
SID79	Muertes por Síndrome de Muerte Súbita del Lactante (SIDS), 1979-84
NWBIR79	Nacimientos no blancos, 1979-84
SIDR74	Tasa de mortalidad por SIDS, por 1.000 nacimientos (1974-78)
SIDR79	Tasa de mortalidad por SIDS, por 1.000 nacimientos (1979-84)
NWR74	Tasa de natalidad no blanca (no blancos por 1.000 nacimientos), 1974-78
NWR79	Tasa de natalidad no blanca (no blancos por 1.000 nacimientos), 1979-84

Cuadro 1: Descripción de las variables del conjunto de datos SIDS.

Estos datos provienen de Cressie 1993 como se dice en la pagina.

Con esta información la variable que sera nuestro principal interés resulta: SID79 Muertes por Síndrome de Muerte Súbita del Lactante (SIDS), 1979-1984, pero primero veamos el comportamiento espacial de estas variables en los siguientes graficos:

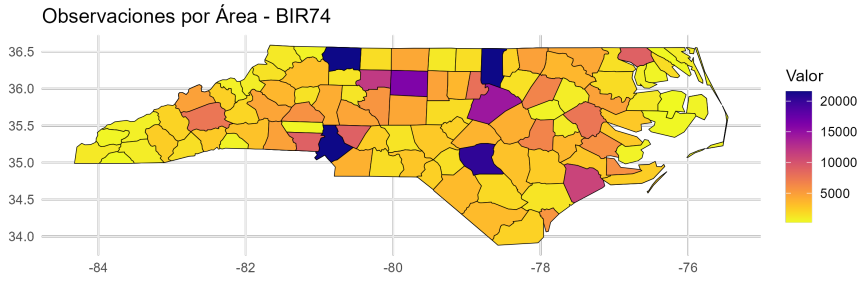

Figura 1: Mapa de valores observados de la variable BIR74 (Nacimientos en vida, 1974-78).

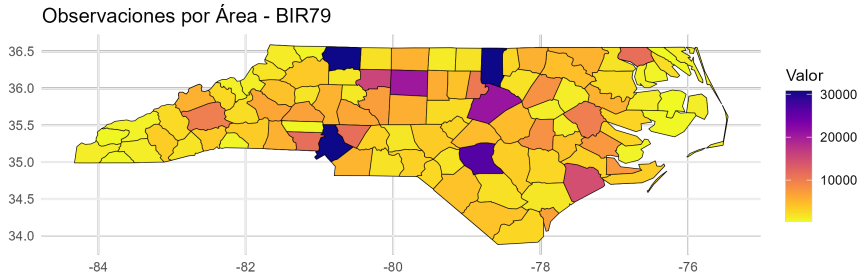

Figura 2: Mapa de valores observados de la variable BIR79 (Nacimientos en vida, 1979-84).

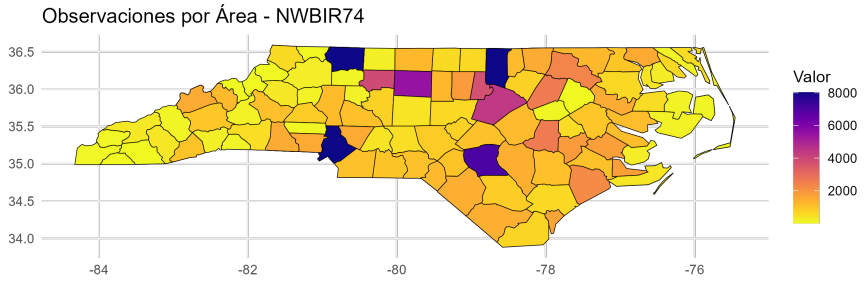

Figura 3: Mapa de valores observados de la variable NWBIR74 (Nacimientos no blancos, 1974-78).

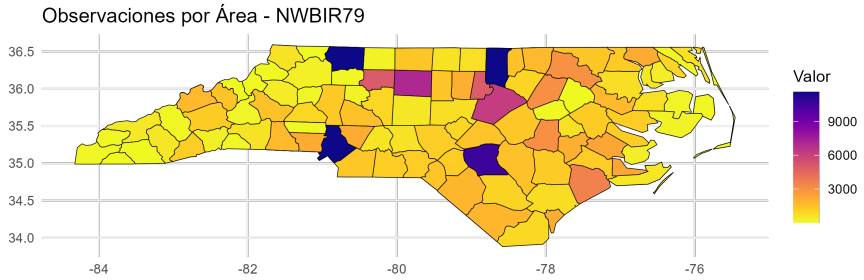

Figura 4: Mapa de valores observados de la variable NWBIR79 (Nacimientos no blancos, 1979-84).

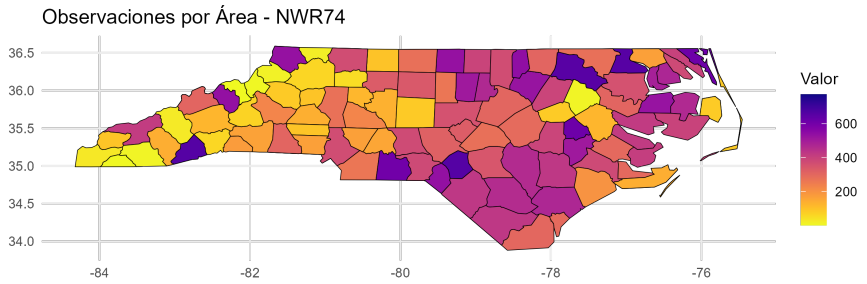

Figura 5: Mapa de valores observados de la variable NWR74 (Tasa de natalidad no blanca, 1974-78).

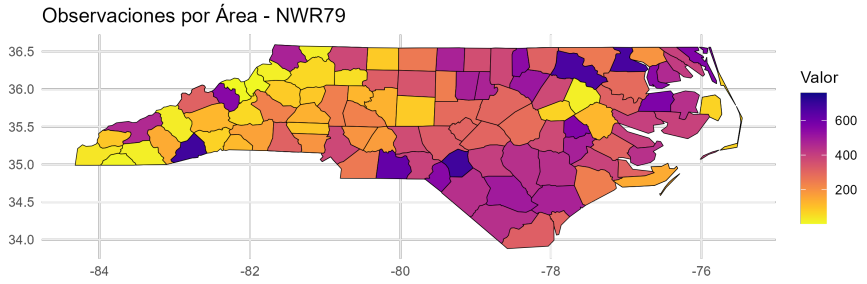

Figura 6: Mapa de valores observados de la variable NWR79 (Tasa de natalidad no blanca, 1979-84).

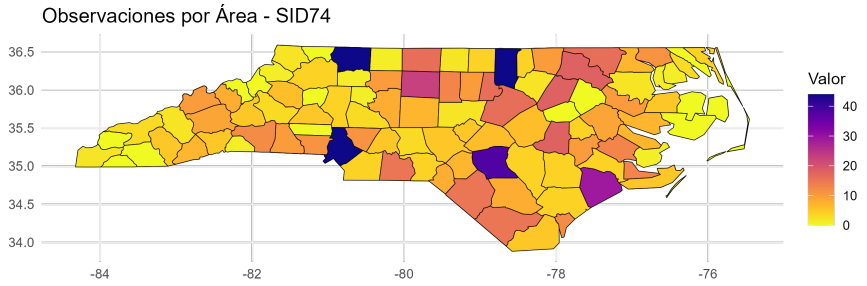

Figura 7: Mapa de valores observados de la variable SID74 (Muertes por SIDS, 1974-78).

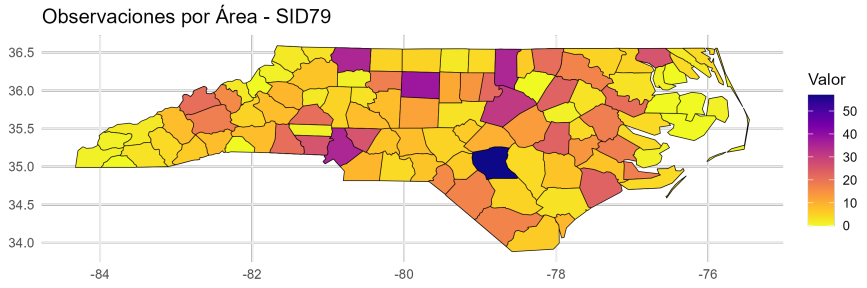

Figura 8: Mapa de valores observados de la variable SID79 (Muertes por SIDS, 1979-84).

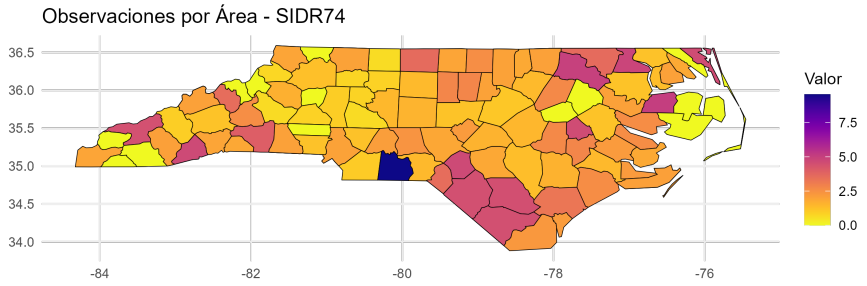

Figura 9: Mapa de valores observados de la variable SIDR74 (Tasa de mortalidad por SIDS, por 1.000, 1974-78).

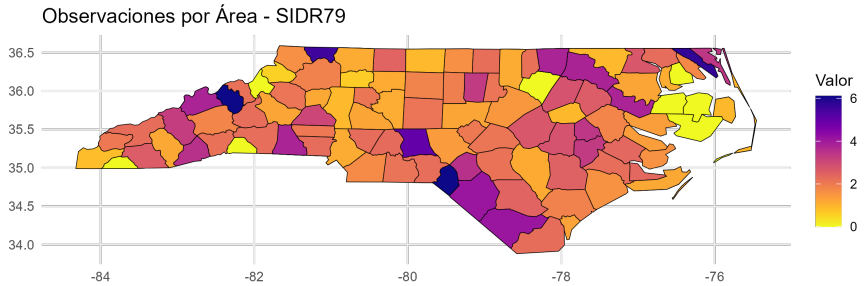

Figura 10: Mapa de valores observados de la variable SIDR79 (Tasa de mortalidad por SIDS, por 1.000, 1979-84).

En el contexto de norteamericano en los años 70´s y 80´s del siglo pasado, y marcado fuertemente por su historia el color de piel podría resultar resultar como un explicativo de condiciones de calidad de vida en ya que en estos años e históricamente las personas afroamericanas no contaban con las mejores ni igualitarias

condiciones de vida por lo que para el estudio correccional del síndrome de muerte súbita podría resultar importantes las variables presentadas donde la etnia juega un rol crucial.

por otro lado veamos un resumen numero de estas variables sin tener en cuenta su localización geográfica, esto a través del siguiente código en R:

```
summary(sids2 %>%
          st_drop_geometry() %>%
          select(BIR74, SID74, NWBIR74, BIR79, SID79, NWBIR79, SIDR74, SIDR79,
              NWR74, NWR79))

BIR74            SID74           NWBIR74            BIR79
 Min.   :  248   Min.   : 0.00   Min.   :   1.0   Min.   :  319
 1st Qu.: 1077   1st Qu.: 2.00   1st Qu.: 190.0   1st Qu.: 1336
 Median : 2180   Median : 4.00   Median : 697.5   Median : 2636
 Mean   : 3300   Mean   : 6.67   Mean   :1050.8   Mean   : 4224
 3rd Qu.: 3936   3rd Qu.: 8.25   3rd Qu.:1168.5   3rd Qu.: 4889
 Max.   :21588   Max.   :44.00   Max.   :8027.0   Max.   :30757
     SID79           NWBIR79            SIDR74          SIDR79
 Min.   : 0.00   Min.   :    3.0   Min.   :0.000   Min.   :0.000
 1st Qu.: 2.00   1st Qu.:  250.5   1st Qu.:1.084   1st Qu.:1.249
 Median : 5.00   Median :  874.5   Median :1.855   Median :2.075
 Mean   : 8.36   Mean   : 1352.8   Mean   :2.046   Mean   :2.039
 3rd Qu.:10.25   3rd Qu.: 1406.8   3rd Qu.:2.604   3rd Qu.:2.539
 Max.   :57.00   Max.   :11631.0   Max.   :9.554   Max.   :6.114
     NWR74            NWR79
 Min.   :  1.49   Min.   :  3.24
 1st Qu.:123.47   1st Qu.:120.20
 Median :304.99   Median :307.62
 Mean   :312.50   Mean   :310.45
 3rd Qu.:469.27   3rd Qu.:463.09
 Max.   :772.73   Max.   :759.22
```

2. Regresión espacial.

Con el entendimiento actual de los datos vamos a tratar de explicar nuestra principal variable de interés, por medio de la el resto de variables contemporáneas a esta y de sus propios resagos espaciales.

Para esto una de las partes más esenciales del trabajo es la matriz de pesos espaciales con las que se desarrollaran buena parte de los cálculos y estimaciones de modelos posteriores.

3. Diferentes Matrices de Ponderación

En esta sección, se presentan diferentes formas de ponderar la estructura espacial de los datos. Se comparan los estilos de ponderación `W`, `B`, `C`, `U` y `WW`.

Para llegar a estas matrices de ponderaciones espaciales se uso el siguiente código partiendo de la obtención de los centroides de cada uno de los condados estudiados:

```
#centroides
Centros=st_centroid(sids2)
#distancias
m_distancias=st_distance(Centros)

#matriz de vecindades queen
vecindades=poly2nb(sids2, queen = TRUE)

```

```r
#diferentes matrices de ponderación
# Alternativamente, usa diferentes estilos de ponderación
sids2.lw <- nb2listw(vecindades, zero.policy = TRUE )
sids2.lwb <- nb2listw(vecindades, style = "B",zero.policy = TRUE)
sids2.lwc <- nb2listw(vecindades, style = "C",zero.policy = TRUE)
sids2.lwu <- nb2listw(vecindades, style = "U",zero.policy = TRUE)
sids2.lww <- nb2listw(vecindades, style = "W",zero.policy = TRUE)
```

Con esto llegamos a estas 4 matrices que son:

Matriz de Pesos	**Descripción**
`sids2.lw`	Es la matriz de pesos por defecto (estilo `W`). Los pesos se **row-standardizan**, es decir, cada peso se divide por la suma total de la fila, de modo que la suma de los pesos de cada área es 1.
`sids2.lwb`	Utiliza el estilo `B` (binario). A cada par de áreas vecinas se le asigna un peso de 1 y 0 en caso contrario, sin aplicar estandarizació por filas.
`sids2.lwc`	Emplea el estilo `C` (globalmente estandarizado). Los pesos se ajustan en función del total de conexiones en el conjunto de datos, lo que permite una comparación a nivel global.
`sids2.lwu`	Aplica el estilo `U` (no estandarizado). Se mantienen los pesos originales (generalmente binarios o basados en alguna medida de cercanía), sin normalización; la suma de los pesos de cada área puede variar.
`sids2.lww`	Vuelve a calcularse la matriz con el estilo `W` (row-standardized) de forma explícita. Es funcionalmente equivalente a `sids2.lw` y se utiliza para enfatizar el efecto de la estandarización por filas.

Cuadro 2: Descripción de los diferentes estilos de ponderación utilizados.

4. Evaluación de Autocorrelación Espacial

Se ajusta un modelo OLS y se analiza la autocorrelación espacial en los residuos mediante el test de Moran.

Basado en el test de Moran tambien se escoja cual es la matriz de ponderación que mejor recoge las interacciones entre las diferentes zonas por vecindades en el estudio.

Con base en esto se llega por le test de Moran con un valor p de 0,01312067 que la matriz que mejor recoge esta interacción midiéndola con el valor P es la matriz llamada '`sids2.lw`'.

`sids2.lw` es la matriz de pesos por defecto utilizando el estilo `W`. En esta matriz, los pesos asignados a las áreas vecinas se **row-standardizan**. Esto significa que, para cada área, se calcula la suma total de los pesos (por ejemplo, 1 para cada vecino) y cada peso se divide por esa suma, de modo que la suma de los pesos de cada fila (o área) sea igual a 1.

Esta normalización tiene varias ventajas:

- **Comparabilidad:** Permite comparar áreas con distinto número de vecinos, ya que la contribución total de los vecinos se mantiene constante (1) en todas las áreas, evitando que las áreas con muchos vecinos tengan un peso acumulado mayor.

- **Interpretabilidad:** Facilita la interpretación de los coeficientes en modelos espaciales. El efecto espacial se expresa como una media ponderada de las variables de las áreas vecinas, haciendo que el coeficiente del término de retardo espacial se interprete en términos de un efecto promedio.
- **Homogeneidad:** Asegura que la influencia de la vecindad esté en una misma escala para todas las observaciones, lo que es especialmente útil en análisis donde la cantidad de vecinos varía significativamente entre áreas.

Por lo que para el ajuste de los sucesivos modelos se utilizara esta matriz.

5. Ajuste de Modelos Espaciales

Se presentan los modelos espaciales considerados:

- **Modelo SAR (Spatial Autoregressive Model)**
- **Modelo SDEM (Spatial Durbin Error Model)**
- **Modelo Manski (SAC - Spatial Auto-Covariance)**

Se comparan estos modelos utilizando el criterio de información de Akaike (AIC).

En este caso de usaron estos modelos ya que todos incorporan los resagos espaciales que resulta de gran interés ya que gracias al test de moran anteriormente realizado puede resultar muy significativa.

Por otro lado contamos con algunas variables que podrían ser explicativas aunque no son las mejores, por lo que el uso de los modelos SDEM y Manski trataran de capturan tanto el error espacial como estas variables explicativas y en el caso del modelo Manski los resagos de estas mismas, generando más parámetros como contra.

para la realización de estos modelos se uso el siguiente código:

```
# Ajustar el modelo SAR (Spatial Autoregressive Model)
modelo_sar <- lagsarlm(SIDR79 ~ NWBIR79 + NWR79 + AREA + PERIMETER + BIR79,
                       data = sids2, listw = mejor_matriz, zero.policy = TRUE)
summary(modelo_sar)

# Ajustar el modelo SDEM (Spatial Durbin Error Model)
modelo_sdem <- errorsarlm(SIDR79 ~ NWBIR79 + NWR79 + AREA + PERIMETER + BIR79,
                          data = sids2, listw = mejor_matriz, etype = "emixed",
                              zero.policy = TRUE)
summary(modelo_sdem)

# Ajustar el modelo Manski (SAC - Spatial Auto-Covariance)
modelo_manski <- sacsarlm(SIDR79 ~ NWBIR79 + NWR79 + AREA + PERIMETER + BIR79,
                          data = sids2, listw = mejor_matriz, type = "sacmixed",
                              zero.policy = TRUE)
summary(modelo_manski)

# Comparar los cuatro modelos con AIC
cat("\n--- Comparación de AIC entre modelos ---\n")
AIC_values <- AIC(modelo_ols, modelo_sar, modelo_sdem, modelo_manski)
print(AIC_values)
```

6. Selección del Mejor Modelo

Se elige el modelo SAR como el mejor modelo y se exploran diferentes configuraciones para seleccionar las variables explicativas adecuadas. Esto usando la comparacion dada por el AIC dando los siguientes resultados:

```
> print(AIC_values)
               df       AIC
modelo_ols      7 331.6628
modelo_sar      8 329.8675
modelo_sdem    13 337.9466
modelo_manski  14 339.7109
```

Ahora bien en este caso el modelo usa todas las variables como explicativas pero en este casi el modelo puede estar saturado por lo que empezamos a sacar variables del modelo según el valor de su significancia y usamos el AIC para ver que composición de modelo resulta mejor:

```
# Ajustar el modelo SAR (Spatial Autoregressive Model)
modelo_sar_1 <- lagsarlm(SIDR79 ~ NWR79 + AREA + PERIMETER + BIR79,
                         data = sids2, listw = mejor_matriz, zero.policy = TRUE)
summary(modelo_sar_1)


modelo_sar_2 <- lagsarlm(SIDR79 ~ NWR79  + PERIMETER + BIR79,
                          data = sids2, listw = mejor_matriz, zero.policy = TRUE)
summary(modelo_sar_2)

modelo_sar_3 <- lagsarlm(SIDR79 ~ NWR79   + BIR79,
                          data = sids2, listw = mejor_matriz, zero.policy = TRUE)
summary(modelo_sar_3)

AIC( modelo_sar,modelo_sar_1, modelo_sar_2, modelo_sar_3)
```

Dando como resultado lo siguiente:

```
> AIC( modelo_sar,modelo_sar_1, modelo_sar_2, modelo_sar_3)
              df       AIC
modelo_sar     8 329.8675
modelo_sar_1   7 327.8802
modelo_sar_2   6 325.9858
modelo_sar_3   5 324.0924
```

Por lo que resulta mejor el modelo 3 que en este caso contiene a las variables NWR79 y BIR79 aunque ninguno de estas dos es significativa, no obstante el modelo resultaba mejor incluyéndolas en todos los casos.

De acá podemos concluir que este modelo se ve beneficiado ampliamente de incluir el resago espacial.

Bajo este modelo construimos el siguiente gráfico con las predicciones:

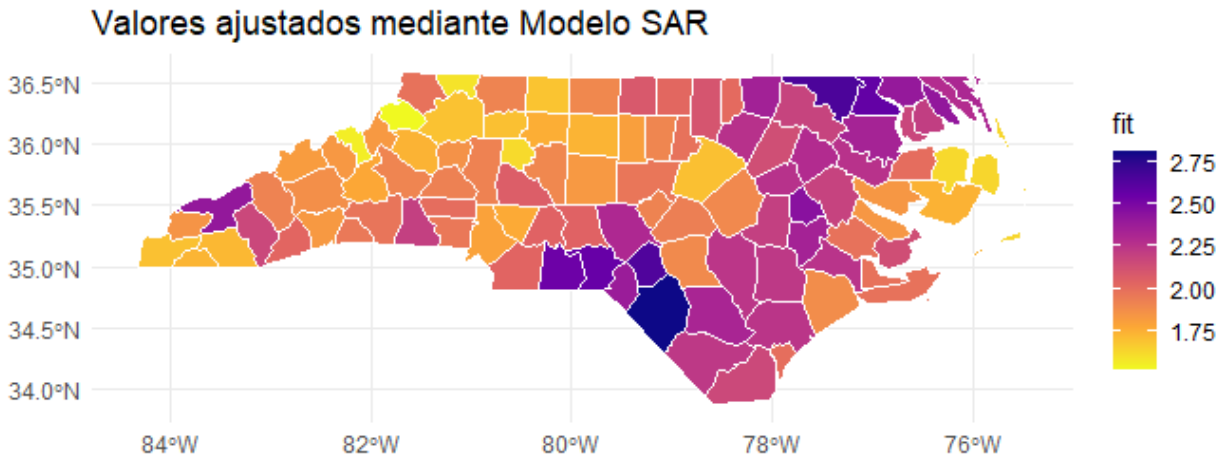

Figura 11: SIDR79 Predicho modelo SAR

Resultando razonablemente similar al gráfico descriptivo del principio, aunque menos concentrados los datos.

7. Verificación de Supuestos

Se realiza un análisis de los supuestos del modelo:

7.1. Normalidad de los residuos

Se presentan un histograma, un QQ-Plot y pruebas de normalidad como Shapiro-Wilk y Kolmogorov-Smirnov.

Figura 12: Histograma residuales

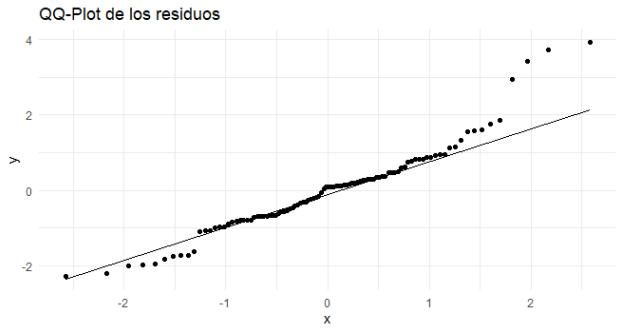

Figura 13: QQ plot de los residuales

Prueba de normalidad Shapiro-Wilk: p-valor = 0.0002587267 Prueba de Kolmogorov-Smirnov: p-valor = 0.3234714

La prueba de Kolmogorov es la única que no rechaza la hipótesis de normalidad en este caso, mientras que tanto por los metodos graficos como por el test de Shapiro se rechaza rotundamente.

7.2. Homocedasticidad

Se analizan los residuos con un gráfico de dispersión y la prueba de Breusch-Pagan.

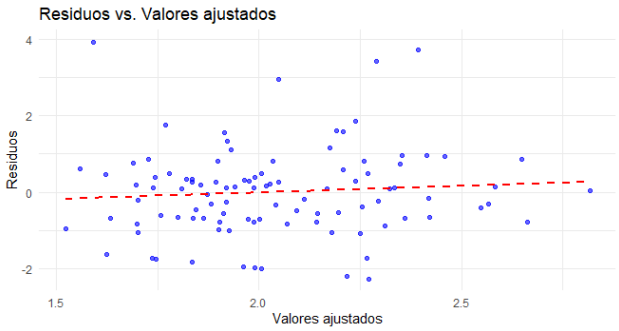

Figura 14: Dispersión de residuos vs ajustados

Prueba de Breusch-Pagan: p-valor = 0.9783141.

En este caso aunque se encuentra una ligera tendencia con el test de Breusch-Pagan llegamos a la conclusión de que no existen grandes problemas con la homocedasticidad.

7.3. Autocorrelación Espacial de los Residuos

Se utiliza la prueba de Moran's I para evaluar la autocorrelación espacial en los residuos del modelo SAR. Prueba de Moran's I: p-valor = 0.364924 En este caso vemos que no se rechaza la hipótesis de independencia espacial, con lo que el modelo cumplió filtrando buena parte de la autocorrelación espacial presente en los datos.

8. Conclusiones

En resumen las variables independientes que se presentan en los datos no resultan de gran significancia, posiblemente estas variables se consideraban útiles en el contexto estadounidense del siglo pasado pero no

resultan significativas para un estudio correccional del síndrome de muerte espontanea en lactantes no se vio en el articulo,

Es muy significativa la presencia de autocorrelación espacial por lo que posiblemente factores ambientales o sociales presentes en los condados puedan estar más ampliamente correlacionados con los hallazgos en cuanto a el síndrome de muerte espontanea en lactantes.

Referencias

Anselin, Luc (2003). *SIDS Data Set (North Carolina, 1974-84)*. Last updated June 16, 2003. Data provided .as is,"no warranties. URL: `https://geodacenter.github.io/data-and-lab/sids2/`.

Cressie, Noel (1993). *Statistics for Spatial Data*. New York: Wiley, págs. 386-389.