Universidad del Valle
Facultad de Ingeniería
Escuela de Estadística
Programa Académico de Estadística
Estadística Aplicada II
Cali, marzo de 2018

Tarea 3:Supuestos en el modelo de regresión lineal

Kevin Steven García[a], Alejandro Vargas[b]

1. Introducción

En este trabajo evaluaremos el supuesto de independencia en los errores del modelo lineal ajustado para los 500 datos seleccionados de la base de datos 'cadata'. Se dará una presentación formal del supuesto de independencia de los errores y se mostrara la consecuencia que tiene el incumplimiento de este, además, se aplicaran pruebas tanto gráficas como estadísticas para verificar el incumplimiento o cumplimiento de este supuesto en nuestro modelo lineal ajustado y finalmente, se concluirá acerca de nuestro modelo teniendo en cuenta los resultados de las pruebas aplicadas.

2. Modelo ajustado

El modelo ajustado sobre el cuál se va a validar el supuesto de independencia en los errores, es el modelo que se ajusto en la tarea 1, que pretendía explicar la variable Y ='Valor mediano de las viviendas' con las variables explicativas X_1 ='Ingreso mediano',X_2 ='Edad mediana de la vivienda',X_3 ='Total de habitaciones',X_4 ='Total de dormitorios',X_5 ='Población' y X_6 ='Hogares'. El modelo ajustado para estas variables, sin hacer transformación ni selección de estas es:

$$Y = 57720.52 + 24261.20X_1 + 3443.94X_2 + 19.09X_3 - 67.72X_4 - 121.66X_5 + 315.92X_6$$

3. Punto 1:Supuesto de independencia (no correlación) en los errores

El supuesto de independencia o de no correlación de los errores establece que $COV(\varepsilon_i, \varepsilon_j) = 0 \; \forall i \neq j$
Esto en palabras quiere decir que todos los valores en posiciones diferentes a la diagonal de la matriz de varianzas y covarianzas deben ser igual a cero.

$$V[E] = \begin{pmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{pmatrix}$$

[a]Código: 1533173. E-mail: kevin.chica@correounivalle.edu.co
[b]Código: 1525953. E-mail: jose.alejandro.vargas@correounivalle.edu.co

El efecto que produce la correlación entre errores en el modelo de regresión, es que tiende a cambiar la matriz de varianzas y covarianzas $\sigma^2 I$ por $\sigma^2 V$, y esto tiene como consecuencia la perdida de la optimalidad del estimador de mínimos cuadrados ordinarios (MCO), es decir, los estimadores obtenidos por este método de estimación ya no son estimadores óptimos. La solución a este problema es utilizar mínimos cuadrados generalizados para tratar de recuperar la optimalidad de los estimadores, pero, este método nos exige conocer la matriz V, la cuál nunca conocemos, y estimarla resulta muy complicado en la practica.
El efecto de la dependencia puede ser un un problema muy grave y difícil de corregir, por ello es importante prevenirlo cuando los datos se están recopilando. El método más eficaz para prevenir la dependencia es realizar un procedimiento apropiado de aleatorización.

Para evaluar el cumplimiento o incumplimiento de este supuesto nos vamos a basar en una prueba gráfica y en dos pruebas estadísticas. La prueba gráfica consiste básicamente en realizar una gráfica de autocorrelación de los residuales, llamado correlograma y observar si estas correlaciones se salen del intervalo de confianza. Las pruebas estadísticas que vamos a aplicar son dos, la prueba de rachas y la prueba de Durbin-Watson, la explicación teórica de estas pruebas se dará en el punto 3 (literal 5).

4. Punto 2:Aplicación de las pruebas para evaluar el supuesto

- Prueba gráfica:

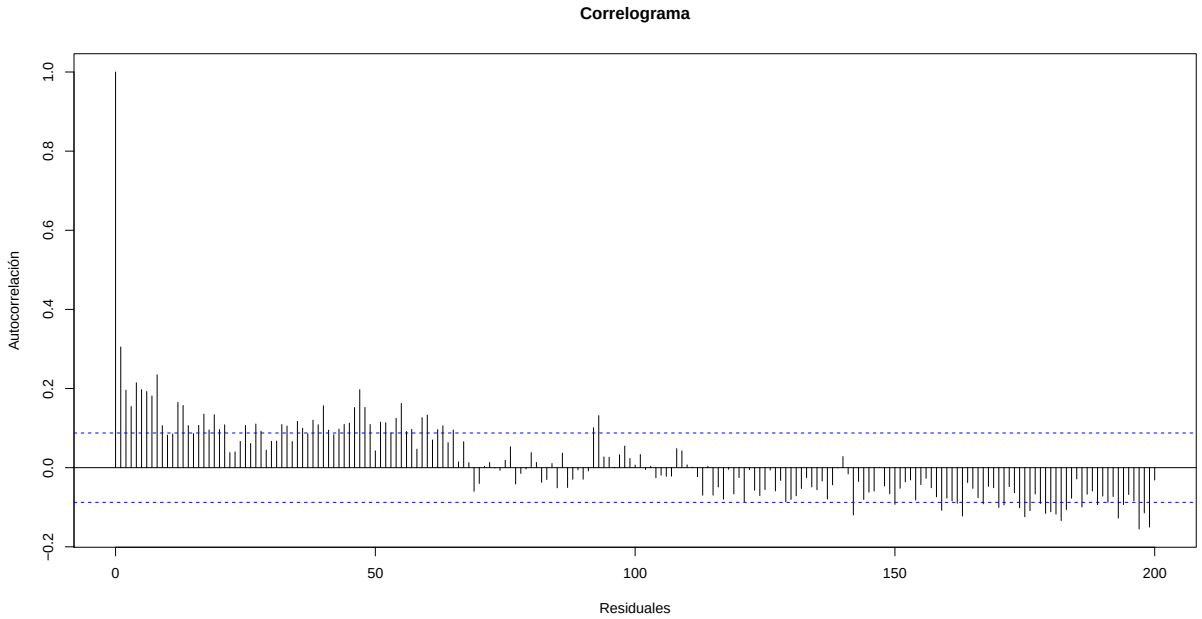

Figura 1: Correlograma de los residuales

En esta gráfica se puede observar que la autocorrelación de los errores, es distinta de cero, ya que hay varías barras que se salen del intervalo (lineas azules) que fue construido con un 95 % de confianza, entonces, se puede rechazar H_0 y concluir que si existe algún tipo de correlación en los errores.

- Prueba de Rachas: Esta prueba se realizo con la ayuda del software R Core Team (2017) con el paquete Trapletti & Hornik (2018). Se podría realizar manualmente, pero era algo tedioso por la cantidad de datos sobre los cuales se debe aplicar la prueba, recordemos que lo que se busca es que los residuales sean completamente aleatorios, en otras palabras, que entre ellos no exista algún tipo

de correlación.
La prueba realizada fue de dos colas, por lo tanto, las hipótesis planteadas son del siguiente tipo:

$$H_0 \text{ :La muestra es aleatoria (Los residuales son independientes)}$$
$$H_1 \text{ :La muestra no es aleatoria (Los residuales no son independientes)}$$

El resultado de R nos arroja el cuantil de la aproximación del estadístico a la normal estándar que fue Z=-7.5892 y el p-valor, el cual fue $3.219 * 10^{-14}$ o lo que es lo mismo 0.00000000000003219, lo cuál es un valor p demasiado pequeño, que nos lleva a rechazar H_0 y concluimos que los residuales no son independientes, o que tienen una correlación entre ellos.

- Prueba de Durbin - Watson: Esta prueba también fue realizada en R Core Team (2017) con el paquete ?. Al igual que la anterior, esta prueba fue realizada de dos colas o bilateral, ya que nos interesa si los residuales tienen correlación o no, es decir, por ahora no nos interesa saber si dicha correlación es menor o mayor que cero. Por lo tanto, las hipótesis planteadas fueron las siguientes:

$$H_0 : \rho = 0 \text{ (La autocorrelación en los residuales es cero)}$$
$$H_1 : \rho \neq 0 \text{ (La autocorrelación en los residuales es diferente de cero)}$$

El resultado de R nos arroja el estadístico de prueba que fue DW=1.3869 y el p valor el cual fue, p-value= $2.614 * 10^{-12}$ o lo que es lo mismo 0.000000000002614, el cuál es demasiado pequeño, esto nos dice que en la muestra hay mucha evidencia en contra de la hipotesis nula, por lo cuál esta se rechaza y concluimos que si hay autocorrelación en los errores.

5. Punto 3:Descripción de las pruebas:

- Prueba gráfica (correlograma):
- Prueba de rachas:
- Prueba de Durbin-Watson:

6. Punto 4:Conclusiones del modelo ajustado con respecto al supuesto

Referencias

Douglas C. Montgomery, Elizabeth A. Peck, G. G. V. (2002), *Introducción al análisis de regresión lineal*, tercera edn, Compañía Editorial Continental.

R Core Team (2017), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
*https://www.R-project.org/

Trapletti, A. & Hornik, K. (2018), *tseries: Time Series Analysis and Computational Finance*. R package version 0.10-43.
*https://CRAN.R-project.org/package=tseries

Uriel, E. (2013), 'Análisis de regresión múltiple con información cualitativa'.