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.
Además, si los errores están correlacionados, entonces la regresión de los mínimos cuadrados puede subestimar el error estándar de los coeficientes. Errores estándar subestimados pueden hacer que sus predictores parezcan significativos, cuando en realidad no lo son.

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 tres, la prueba de rachas, la prueba de Durbin-Watson y la prueba de Ljung-Box, 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:

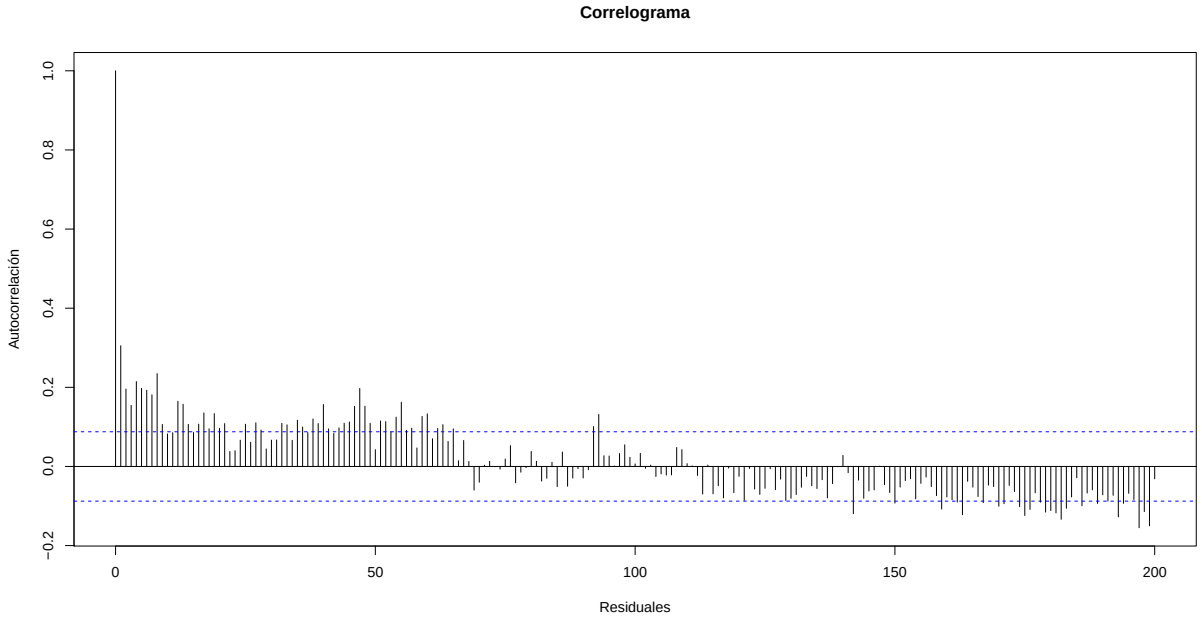

Figura 1: Correlograma de los residuales

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 :La muestra es aleatoria (Los residuales son independientes)
  H_1 :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 Zeileis & Hothorn (2002). 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:

  $H_0 : \rho = 0$ (La autocorrelación en los residuales es cero)
  $H_1 : \rho \neq 0$ (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.

- Prueba de Ljung-Box: Esta prueba también fue realizada en R, y plantea las siguientes dos hipotesis:

  H_0 :Los datos se distribuyen de forma independiente (es decir, las correlaciones en la población de la que se toma la muestra son 0, de modo que cualquier correlación observada en los datos es el resultado de la aleatoriedad del proceso de muestreo).
  H_1 :Los datos no se distribuyen de forma independiente.

Aplicando esta prueba en R, nos arroja el estadistico de prueba chi-cuadrado que fue $X^2 = 46.826$ y el valor p= $7.759 * 10^{-12}$ o 0.000000000007759, el cuál es un valor p demasiado bajo, muchisimo menor que cualquier nivel de significancia dado, esto nos indica que en la muestra(los residuales) hay demasiada evidencia en contra de la hipotesis nula, es decir, los datos dan evidencia de que los residuales no se distribuyen de forma independiente, por lo cuál se rechaza H_0.

5. Punto 3:Descripción de las pruebas:

- Prueba gráfica (correlograma): Esta gráfica nos muestra unas barras que corresponden a las correlaciones entre el residual e_i y el residual e_{i-1}, básicamente la decisión consiste en generar un intervalo para dichas correlaciones y observar cuantos de ellos se salen de ese intervalo, si varias barras se

salen de dicho intervalo, hay evidencia de que las correlaciones entre los errores son muy fuertes, por lo cual, no se puede decir que ellos son independientes. Está prueba es bastante subjetiva, ya que depende mucho de la experiencia y de la capacidad de la persona que la aplica para tomar su decisión sobre la validación del supuesto. Por lo cuál es preferible aplicar pruebas estadísticas como tal, como las siguientes tres que se van a describir.

- Prueba de Rachas:Para entender esta prueba primero se debe definir que es una racha. Dada una sucesión de n observaciones de una variable que solo puede tomar dos valores, una racha es una sucesión de uno o más datos con el mismo valor que están seguidos y precedidos por datos con el otro valor, o por ningún dato si se encuentran en el inicio o el final de la sucesión.
Ejemplo: En la sucesión A BBB AA BB A B se presentan 6 rachas.
Como se planteó, el problema que se desea resolver es decidir si el proceso mediante el cual se genera la sucesión pude considerarse aleatorio, tomando a la muestra como una sucesión de observaciones; se quiere conocer si ésta es aleatoria. Las hipótesis a considerar son:

$$H_0 \text{ :La muestra es aleatoria}$$
$$H_1 \text{ :La muestra no es aleatoria}$$

Muestra aleatoria es aquella que proviene de una misma distribución y las observaciones son independientes.

Estadistico de prueba: El estadistico de prueba es el número total de rachas, R, en la sucesión. Si llamamos R_1 y R_2 a las variables aleatorias número de rachas del tipo 1 (evento A) y del tipo 2 (evento B) respectivamente, entonces la el estadistico de prueba número total de rachas R será:

$$R = R_1 + R_2$$

Distribución del estadistico de prueba R: Su distribución bajo la hipótesis nula H_0 esta dada por:

$$P\{R=r\} = \frac{2\binom{n_1-1}{\frac{r}{2}-1}\binom{n_2-1}{\frac{r}{2}-1}}{\binom{n_1+n_2}{n_1}} \text{ si r es par}$$

$$P\{R=r\} = \frac{\binom{n_1-1}{\frac{r-1}{2}-1}\binom{n_2-1}{\frac{r-1}{2}}+\binom{n_1-1}{\frac{r-1}{2}}\binom{n_2-1}{\frac{r-1}{2}-1}}{\binom{n_1+n_2}{n_1}} \text{ si r es impar}$$

Para r=2,3,4,...,$n_1 + n_2$

Aproximación a la normal: La aproximación a la normal esta dada por:

$$Z = \frac{R - \left(\frac{2n_1n_2}{n_1+n_2}+1\right)}{\sqrt{\frac{2n_1n_2(2n_1n_2-n_1-n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}}$$

Región de rechazo: Se duda de la aleatoriedad cuando hay muchas o muy pocas rachas. Por tanto, la región de rechazo es de la forma:

$$R > R_{n_1,n_2;(1-\frac{\alpha}{2})} \text{ o } R < R_{n_1,n_2;\frac{\alpha}{2}}$$

Donde la parte derecha de cada inecuación son los percentiles $(1-\frac{\alpha}{2})$ y $\frac{\alpha}{2}$ de la distribución del número de rachas R, cuando hay n_1 observaciones del tipo A y n_2 del tipo B.

- Prueba de Durbin-Watson: El Test de Durbin-Watson permite evaluar si existe autocorrelación en una Regresión lineal, sea simple o múltiple. Con ello se pretende ver si los valores presentan algún tipo de dependencia en cuanto al orden de obtención. Las hipotesis que se plantean son las siguientes:

$$H_0 : \rho = 0$$
$$H_1 : \rho \neq 0$$

Estadístico de prueba: El estadístico de Durbin-Watson (D) está condicionado según el orden de las observaciones. El estadístico de Durbin-Watson determina si la correlación entre los términos de error adyacentes es o no es igual a cero. este estadistico esta dado por : $D = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2}$ donde e_t es el residuo t-esimo o a tiempo t.

D varía entre 0 y 4. Si D esta cerca de 0 los errores están correlacionados positivamente. Si D está cerca de 4 entonces la correlación es negativa. La distribución de D es simétrica con respecto a 2. Así que un valor de D cercano a 2 indica que no hay correlación de los errores.

Región de rechazo: Para rechazar se utiliza una tabla con lo valores críticos del estadístico, las regiones de rechazo son:

- Si $d < d_{L,\alpha}$ existe evidencia estadística de que los términos de error estén autocorrelacionados positivamente.
- Si $d > d_{U,\alpha}$ no hay evidencia estadística de que los términos de error estén autocorrelacionados positivamente.
- Si $d_{L,\alpha} < d < d_{U,\alpha}$ la prueba no es concluyente.

- Prueba de Ljung-Box:La prueba de Ljung-Box (llamada así por Greta M. Ljung y George E. P. Box) es un tipo de prueba estadística de si un grupo cualquiera de autocorrelaciones de una serie de tiempo son diferentes de cero. En lugar de probar la aleatoriedad en cada retardo distinto, esta prueba la aleatoriedad .$^{\text{en}}$ general"basado en un número de retardos.

En la prueba de Ljung-Box se puede definir o plantear las hipótesis de la siguiente manera:

H_0 :Los datos se distribuyen de forma independiente (es decir, las correlaciones en la población de la que se toma la muestra son 0, de modo que cualquier correlación observada en los datos es el resultado de la aleatoriedad del proceso de muestreo).
H_1 : Los datos no se distribuyen de forma independiente.

Estadistico de prueba: El estadistico de prueba es: $Q = n(n+2)\sum_{k=1}^{h}\frac{\hat{\rho_k^2}}{n-k}$
donde n es el tamaño de la muestra, $\hat{\rho}_k$ es la autocorrelación de la muestra en el retraso k y h es el número de retardos que se están probando.

Región de rechazo: Para un nivel de significancia α la región critica para rechazar la hipótesis de aleatoriedad es: $Q > \chi^2_{1-\alpha,h}$

donde $\chi^2_{1-\alpha,h}$ es el α-cuantil de la distribución chi-cuadrado con m grados de libertad.

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

Observando las 4 pruebas realizadas, podemos ver que en la primera(la prueba gráfica), hay muchas barras o correlaciones que se nos salen de los intervalos de confianza, por lo cuál se rechazo H_0 y se concluyo que con respecto a esa prueba no hay independencia en los errores, y con respecto a las pruebas estadísticas como tal (Rachas, Durbin-watson y Ljung-Box), vimos que los p valores de las tres pruebas son demasiado bajos, mucho mas bajos que cualquier nivel de significancia propuesto, por lo cuál en esas tres pruebas también se rechazo H_0, lo cuál confirma la existencia de correlación en los errores del modelo.

Teniendo en cuenta lo anterior, lo más común y razonable es concluir que nuestro modelo ajustado no cumple con este supuesto de independencia en los errores, por lo cuál diríamos que lo recomendable sería no usar este modelo. Sin embargo, los errores podrían ser independientes aún cuando los residuos no lo sean, y el supuesto se hace para los errores mas no para los residuos, pero eso no hay forma de probarlo a ciencia cierta, por lo cuál es mejor prevenir y usar un modelo el cuál por lo menos me muestre dicha independencia en sus residuales.

Referencias

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

Faraway, J. J. (2005), *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*, Vol. 1 edn., Chapman y Hall / CRC.

Gómez, N. G. (2006), *Notas de clase: Series de tiempo con R*, Universidad Nacional de Colombia.

Gutiérrez, R. B. (2003), 'Validación de supuestos en el modelo de regresión', *Serie Monografías* **1 edn.**

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

Zeileis, A. & Hothorn, T. (2002), 'Diagnostic checking in regression relationships', *R News* **2**(3), 7–10.
*https://CRAN.R-project.org/doc/Rnews/