Supuestos en el modelo de regresión

Kevin García - Alejandro Vargas

12 de abril de 2018

Introducción

En esta presentación 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.

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:

$$Y = 57720{,}52 + 24261{,}20X_1 + 3443{,}94X_2 + 19{,}09X_3 - 67{,}72X_4 - 121{,}66X_5$$

$$+ 315{,}92X_6$$

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.

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

Posibles causas de autocorrelación en los errores

1. Autocorrelación natural por el modelo y las variables.
2. Omisión de variables relevantes en el modelo especificado.
3. Mala especificación del modelo.
4. Transformaciones incorrectas de los datos.

Consecuencias de autocorrelación en los errores

La presencia de autocorrelación en los errores tiene diversos efectos sobre el procedimiento ordinario de regresión por mínimos cuadrados. Estos efectos se resumen a continuación:

1. 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.
2. Los coeficientes de regresión ordinaria por mínimos cuadrados siguen siendo insesgados, pero ya no son estimados con varianza mínima. Se dice entonces que esos estimados son ineficientes.

Consecuencias de autocorrelación en los errores

3. Cuando los errores están autocorrelacionados positivamente, el cuadrado medio de los residuales, MS_{Res} puede subestimar mucho a σ^2, en consecuencia, los errores estándar de los coeficientes de regresión pueden ser muy pequeños. Así, los intervalos de confianza son más cortos de lo que deberían ser, y las pruebas de hipótesis acerca de los coeficientes individuales de regresión pueden indicar que uno o más de los regresores contribuyen en forma significativa al modelo, cuando en realidad no contribuyen.
4. Los intervalos de confianza y las pruebas de hipótesis basados en las distribuciones T y F ya no son adecuados.

Posibles soluciones

1. 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.
2. Si se logra estimar la matriz V, se puede trabajar con los mínimos cuadrados generalizados, cabe decir que el proceso de estimación de dicha matriz es bastante complejo.

Pruebas para evaluar la independencia en los errores

Las pruebas que aplicamos para evaluar este supuesto fueron las 4 siguientes:

- Prueba gráfica (correlograma)
- Prueba de Rachas
- Prueba de Durbin-Watson
- Prueba de Ljung-Box

Descripción y resultados de las pruebas aplicadas

- 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.

Descripción y resultados de las pruebas aplicadas

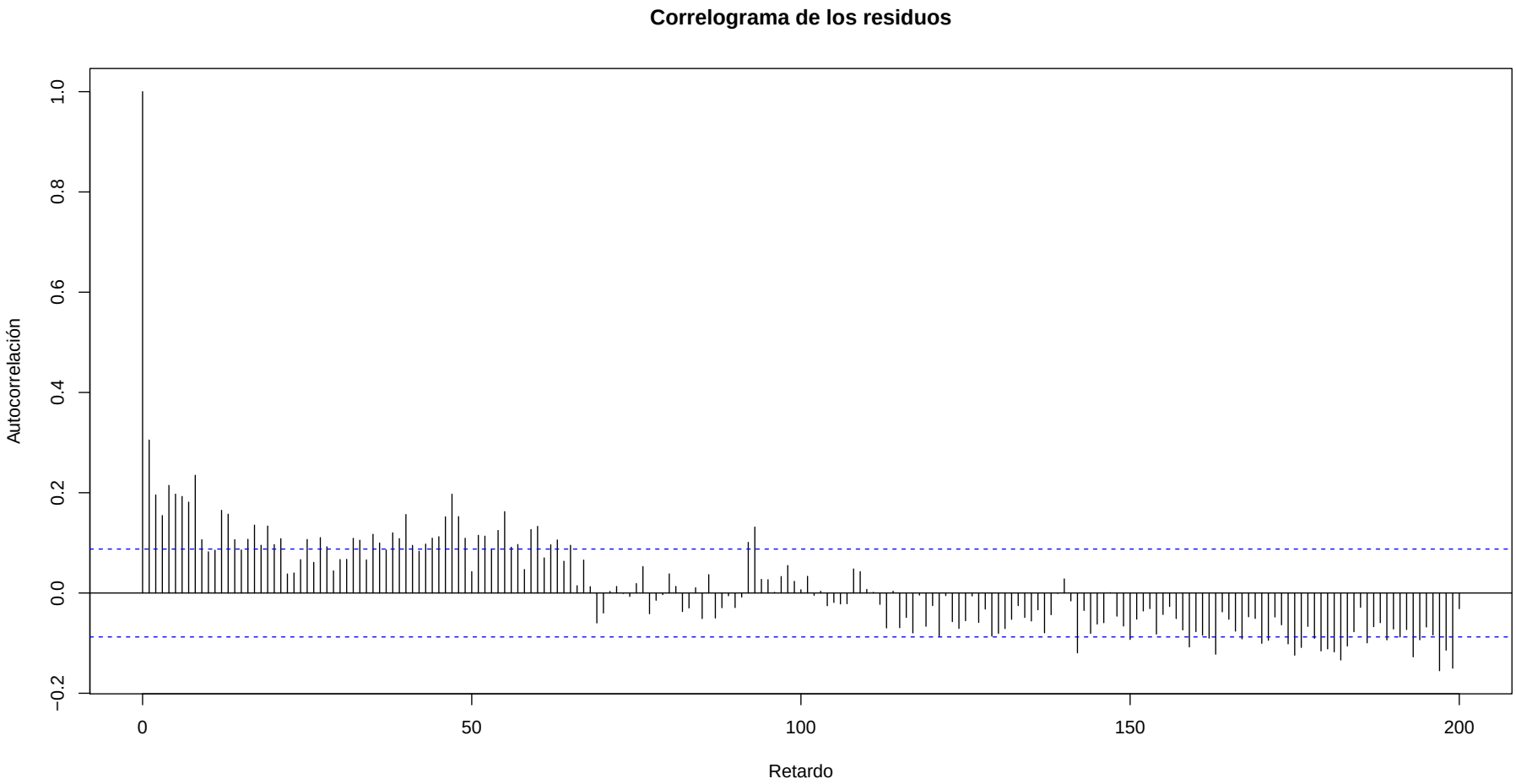

Figura: Correlograma de los residuos

Descripción y resultados de las pruebas aplicadas

- Prueba de Rachas: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.
  Hipótesis: Las hipótesis que se plantean en esta prueba son:
  H_0 :La muestra es aleatoria
  H_1 :La muestra no es aleatoria
  Estadistico de prueba: El estadistico de prueba es el número total de rachas, R, en la sucesión.

Descripción y resultados de las pruebas aplicadas

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:

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

Descripción y resultados de las pruebas aplicadas

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: $R > R_{n_1,n_2;(1-\frac{\alpha}{2})}$ 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.
Resultado:

```
        Runs Test

data:  factor(residualesfactor)
Standard Normal = -7.5892, p-value = 3.219e-14
alternative hypothesis: two.sided
```

Figura: Resultado prueba de rachas

Descripción y resultados de las pruebas aplicadas

- 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.

Descripción y resultados de las pruebas aplicadas

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.

Descripción y resultados de las pruebas aplicadas

Resultado:

```
        Durbin-Watson test

data:  Regresion
DW = 1.3869, p-value = 2.614e-12
alternative hypothesis: true autocorrelation is not 0
```

Figura: Resultado de la prueba de Durbin-Watson

Descripción y resultados de las pruebas aplicadas

- Prueba de Ljung-Box:La prueba de Ljung-Box es un tipo de prueba estadística que evalúa 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 ."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.

Descripción y resultados de las pruebas aplicadas

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.

Descripción y resultados de las pruebas aplicadas

Resultado:

```
        Box-Ljung test

data:  residuales
X-squared = 46.826, df = 1, p-value = 7.759e-12
```

Figura: Resultado de la prueba de Ljung-Box

Conclusiones sobre nuestro modelo ajustado

Teniendo en cuenta las pruebas aplicadas, donde vimos que en la prueba gráfica, muchas barras se nos salen de los intervalos y en las pruebas estadísticas los p valores de las tres pruebas son demasiado bajos llevándonos a rechazar siempre H_0, concluyendo que los residuos no son independientes, 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.