Entre Ciencia e Ingeniería, ISSN 1909-8367
Año 2. No. 4 – Segundo semestre de 2008, páginas 99 - 114

DISTRIBUCIÓN DE LA ESTADÍSTICA DE JARQUE Y BERA PARA LA PRUEBA DE NORMALIDAD EN UNA SERIE TEMPORAL ESTACIONARIA CON DATOS FALTANTES[1]

Joaquín González Borja
Maestría en Ciencias Estadísticas
Profesional en Matemáticas con énfasis en Estadística
Docente Universidad Católica Popular del Risaralda
Grupo de Investigación GEMA
jgb@ucpr.edu.co

Fabio Humberto Nieto Sánchez
Doctorado en Estadística
Maestría en Estadísticas
Licenciatura en Matemáticas
Docente Titular Universidad Nacional de Colombia, sede Bogotá
Grupo de investigación SERIES DE TIEMPO
fhnietos@unal.edu.co

Recibido Septiembre 20 de 2008 / Aceptado Noviembre 24 de 2008

Resumen

En este artículo se estudia el efecto que produce la estimación de datos faltantes en una serie temporal en la distribución de la estadística de prueba de normalidad (Jarque y Bera, 1980). Tal estudio, se realiza vía simulación y es de carácter exploratorio. Se recomienda en estas circunstancias el uso de la técnica de bootstraping para obtener la distribución empírica de la estadística en mención.

Palabras clave: Estimación de datos faltantes, la estadística de Jarque y Bera, distribución empírica, *bootstraping*.

[1] Producto derivado del proyecto "Distribución de la Estadística de Jarque y Bera Para la Prueba de Normalidad en una Serie Temporal Estacionaria con Datos Faltantes", que realizó el autor para optar al título de de Magíster en Estadística, otorgado por la Universidad Nacional de Colombia - Sede Bogotá; bajo la dirección del Doctor en Estadística Fabio Humberto Nieto Sánchez.

Abstract

In this paper, we study the effect that produces the estimate of missing values in a time series in the distribution of the (Jarque- Bera, 1980) test of normality. The study is carried out via simulation and its character is exploratory. In this situation, the use of the bootstraping technique is recommended in order to obtain the empiric distribution of the statistic under consideration.

Keywords: Missing values estimate, Jarque-Bera test, empiric distribution, bootstraping.

1. INTRODUCCIÓN

En el análisis de series temporales, vía modelos estadísticos, usualmente se realizan los siguientes pasos: identificación del modelo, estimación de los parámetros desconocidos y verificación de los supuestos del ruido del proceso (no autocorrelación, homocedasticidad y normalidad), en el último caso los supuestos se verifican mediante los residuales del modelo. La verificación del modelo ajustado es una parte esencial del análisis y debe ser realizado con mucho cuidado, ya que es un criterio de decisión, que para ello es necesario contar con pruebas de hipótesis potentes, para estar totalmente seguros de la decisión tomada.

En la práctica, es muy frecuente encontrar serie temporales que por diversas circunstancias, tienen datos faltantes. Sin embargo, el efecto que pueda acarrear la estimación de dichos datos faltantes en la distribución de la estadística de prueba, en

particular la estadística de (Jarque y Bera, 1980) *JB*, para validar el supuesto de normalidad del ruido, no ha sido considerado.

Paquetes especializados en series temporales, como por ejemplo TSW (Gómez y Maravall, 1996) ignoran el hecho de que la estimación de los datos faltantes producen seudoresiduales, en el sentido de que no todos los residuales son calculados con datos observados, y en este caso, la estadística *JB* es usada para validar el supuesto de normalidad del ruido con los residuales\seudoresiduales del modelo, comparando la estadística calculada *JB* con los cuantiles de su distribución asintótica Chi cuadrado con dos grados de libertad, que fue probada por (Jarque y Bera, 1987) para series completas.

En este artículo se estudia vía simulación de Monte Carlo el efecto que produce la estimación de datos faltantes en la distribución de *JB*, en procesos autoregresivos estacionarios de orden 1 y 2 como un primer paso en la dirección de obtener resultados más generales en el futuro. De los resultados del estudio de simulación, se recomienda el uso de la técnica de *bootstraping*, con el fin de hallar la distribución empírica de la estadística *JB*.

El artículo está organizado como sigue: En la Sección 2, se presenta algunas consideraciones teóricas básicas sobre la estimación de datos faltantes, en el contexto de modelos de estados y el modelo AR(p) estacionario. Se incluye la estadística de prueba de normalidad del ruido del proceso *JB*, al igual que la descripción de la técnica de *bootstraping*. En la Sección 3 se hace el estudio de simulación y se describen los resultados. En la Sección 4 se presenta una aplicación y finalmente, en la Sección 5 se dan las conclusiones.

2. CONCEPTOS BÁSICOS

Aquí, presentamos algunos conceptos teóricos tales como la estimación de datos faltantes en representación de modelos de estados de una serie temporal, procesos AR(p) estacionarios, la prueba de normalidad *JB* y la técnica de *bootstraping*.

2.1 Modelos de estados

Sea Y_t un proceso estocástico univariado que obedece la ecuación

$$Y_t = G_t X_t + W_t, \ t = 1,2,\ldots \quad (1)$$

Llamada de observación, donde $W_t \sim \mathrm{RB}(0, R_t)$ y $\{G_t\}$ es una sucesión de vectores determinísticos *v*-dimensionales y sea la ecuación

$$X_{t+1} = F_t X_t + V_t, \ t = 1,2,\ldots \quad (2)$$

Llamada de estado, donde $\{F_t\}$ es una sucesión de matrices determinísticas

$v * v, \ \{V_t\} \sim RB\ (0, Q_t)\ \ y\ E(W_s V_t) = 0, \ \forall s, t.$

Las ecuaciones (1) y (2), definen un modelo de estados para el proceso $\{Y_t\}$

Si se tiene un proceso estocástico irregularmente espaciado $Y_{i1,\ldots,} Y_{ir}$ se introduce un nuevo proceso $\{Y_t^*\}$ (Brockwell y Davis, 1991) y (Brockwell y Davis ,1996) relacionado al proceso X_t por la ecuación de observación modificada

$$Y_t^* = G_t^* X_t + W_t^*, \ t = 1,2,\ldots \quad (3)$$

Donde

$$G_t^* = \begin{cases} G_t \ si\ t \in \{i_{1,\ldots,} i_r\} \\ 0, \quad en\ otro\ Caso \end{cases}$$

$$W_t^* = \begin{cases} W_t \ si\ t \in \{i_{1,\ldots,} i_r\} \\ N_t, \quad en\ otro\ Caso \end{cases}$$

Con $N_t \sim N(0,1)$. Las ecuaciones (2) y (3) constituyen una representación en modelo de estados para el proceso $\{Y_t^*\}$.

2.2 Estimación de datos faltantes

Sea $X_{t/s}$ la predicción óptima de X_t con base en $y_1, y_2, \dots y_s$, $s = 1,2, \dots, n$, (conjunto de observaciones del proceso $\{Y_t\}$) y sea $\Omega_{t/s}$ su correspondiente matriz de error cuadrático medio. Entonces, para cada $t = n-1, \dots, 1$, se tiene:

$$X_{t/n} = X_{t/t} + \Omega_t^*(X_{(t+1)/n} - F_{t+1}X_{t/t})$$

Y

$$\Omega_{t/n} = \Omega_{t/t} + \Omega_t^*(\Omega_{(t+1)/n} - \Omega_{(t+1)/t})\Omega_t^*$$

Dónde $\Omega_t^* = \Omega_{t/t}\, F_{t+1}^{`}\Omega_{(t+1)/t}^{-1}$. Las cantidades $X_{t/t}$ y $\Omega_{t/t}$ son calculadas mediante el filtro de Kalman y $\Omega_{t+1/t}$ es la matriz de error cuadrático medio de la predicción un paso adelante de X_t (Harvey, 1989). La estimación del dato faltante en el tiempo *t* está dada por $\tilde{Y}_t = G_t X_{t/n}$, donde la predicción de X_t se hace en términos de $y_1^*, y_2^*, \dots, y_s^*$, $s = 1,2, \dots, n$.

2.3 Procesos AR(p) estacionarios

Consideremos un proceso autorregresivo AR(p), definido por

$$Y_{t+1} = \phi_1 Y_t + \phi_2 Y_{t-1} + \dots + \phi_p Y_{t-p+1} + Z_{t+1}, \quad t = 0,1,2 \dots$$

Dónde $\{Z_t\} \sim RB\,(0,\sigma^2)$. Además, se cumple $\phi(Z) \coloneqq 1 - \phi_1 z - \cdots - \phi_p z^p \neq 0$ para $|z| \leq 1$, condición que garantiza estacionaridad. Sea $P_n Y_{n+1}$ el mejor predictor lineal un paso adelante de Y_{n+1} con base en $Y_1, Y_2, \ldots, Y_n$. En un proceso AR(p) estacionario

$$P_n Y_{n+1} = \phi_1 Y_n + \phi_2 Y_{n-1} + \cdots + \phi_p Y_{n+1-p}$$

2.4 La prueba de normalidad

Para examinar el supuesto de normalidad del ruido del proceso, existen diversas pruebas tanto de tipo gráfico, como de tipo analítico.

Una prueba que tiene propiedades óptimas de potencia asintótica, es la referida por (Jarque y Bera, 1980) y (Jarque y Bera, 1987), que está dada por:

$$JB = n\left(\frac{SK^2}{6} + \frac{(KU-3)^2}{24}\right)$$

Con $SK = \frac{\hat{\mu}_3}{\hat{\mu}_2^{3/2}}$ y $KU = \frac{\hat{\mu}_4}{\hat{\mu}_2^2}$ donde $\hat{\mu}_j = \frac{\sum_{t=1}^{n}(\hat{W}_t - \overline{W})^j}{n}, j = 2,3,4$ y $\overline{W} = \frac{\sum_{t-1}^{n}\hat{W}_t}{n}$. Siendo SK y KU, los coeficientes muestrales de asimetría y apuntamiento, respectivamente.

Esta estadística bajo la hipótesis nula de normalidad del ruido blanco del proceso se distribuye asintóticamente como una Chi cuadrado con dos grados de libertad $X^2_{(2)}$ para el caso de series completas. Esta es una prueba que se encuentra implementada en los paquetes especializados en series temporales, tales como TSW (Gómez y Maravall, 1996) y RATS (Doan, 2000).

El conjunto de datos que se toma para realizar la prueba son los residuales del modelo, que están dados por la ecuación:

$$\widetilde{W}_t = Y_t - \hat{P}_{t-1}Y_t \qquad t = 1,2,3,\dots,n$$

,

cuando en la ecuación se involucran estimaciones de datos faltantes, esta recibe el nombre de seudoresidual y se nota $\widetilde{W}_t$.

2.5 La técnica del bootstraping

La técnica conocida como *bootstrap* fue propuesta por (Efron, 1979) con el propósito de hallar intervalos de confianza para parámetros desconocidos en circunstancias donde es imposible encontrar analíticamente la distribución muestral de la estadística de interés. Es una técnica de remuestreo de tipo computacional, que funciona de la siguiente forma:

1. Sea $Y_1, Y_2, \dots Y_n$ la muestra a nuestra disposición y $F_n(y)$ la función de distribución empírica.

2. Se utiliza un generador de números aleatorios para obtener *n* nuevos puntos $y_1^*, y_2^*, \dots, y_n^*$ independientemente y con reemplazo de $F_n(y)$. Este conjunto de nuevos valores se denomina muestra *bootstrap*.

3. Se calcula la estadística de interés para la muestra *bootstrap.*

4. Se repiten los pasos 1 y 2 un número grande de veces, digamos *N.* Denotemos la secuencia de estimadores *bootstrap* para la estadística de interés por Escriba aquí la ecuación. $\hat{\Theta}^1, \hat{\Theta}^2, \cdots, \hat{\Theta}^N$.

5. Con la secuencia de estimadores se hallan los cuantiles de la distribución empírica de la estadística de interés.

3. EL ESTUDIO DE SIMULACIÓN

Consideremos procesos AR(1) y AR(2) estacionarios, asumiendo los parámetros del modelo conocidos con $\sigma^2 = 1$.

Tabla No. 1: Modelos AR(p) estacionarios elegidos para las simulaciones

Valor de p	Parámetros	Raíces de $\phi(z)$
p=1	-0.95 -0.5 0.5 0.95	-1.0526 -2 2 1.0526
p=2	-0.3,0.5 0.7, -0.2 0.8, -0.16	1.7457,-1.1457 1.75 ± 1.3919i 2.5, 2.5

Fuente: Construcción de los autores

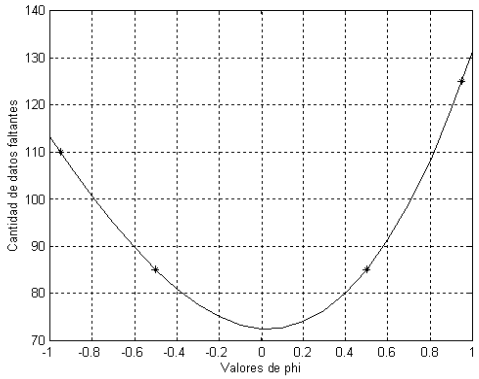
El diseño del experimento de simulación fue el siguiente: (1) se simularon series de longitud 50000, con parámetros dados en la Tabla 1, (2) se quitan datos de la serie en diferentes posiciones en forma aleatoria quedando una serie irregularmente espaciada, (3) se procede a estimar los datos faltantes con el suavizador de punto fijo definido en la sección precedente, (4) se "completa" la serie con dichas estimaciones, (5) se calcula el conjunto de ruidos\seudoruidos y con este se obtiene el valor de la estadística *JB*, que notaremos JB_s, (6) se replica los pasos anteriores, 10000 veces, lográndose 10000 valores de la estadística JB_s . Este conjunto se somete a una prueba de bondad de ajuste a una $X^2_{(2)}$, si no se ajusta a esta distribución, se procede a obtener los cuantiles empíricos de la estadística en mención. Con los ruidos simulados en (1), también se calcula la estadística *JB*, que se nota JB_r . En cada uno de los modelos escogidos para el estudio se

va aumentando la cantidad de datos faltantes en la serie, hasta lograr encontrar aproximadamente la cantidad de datos faltantes con que se pierde la distribución asintótica asumida para la estadística de prueba. En las Tablas 4 a 10 se presenta la cantidad de datos faltantes donde aproximadamente se pierde la distribución asintótica $\chi_{(2)}^{2}$ para JB_s, para cada una de las series estacionarias consideradas en el estudio, además con sus respectivos valores *p* de la prueba de bondad de ajuste Kolmogorov-Smirnov. Se puede ver que la cantidad de datos faltantes para que se preserve la distribución depende de los parámetros y del orden autorregresivo de la serie, hasta un umbral de 130 datos faltantes en series donde se pierde la distribución asumida para la estadística JB_s en todos los procesos en estudio. De las Tablas 4 a 7 se toma el valor de ϕ de un proceso AR(1) y sus respectiva cantidad de datos faltantes en la serie donde se pierde la distribución considerada, se obtiene un polinomio de grado tres,

$$12.099x^3 + 49.8084\,x^2 - 3.0248\,x + 72.5479$$

(Figura No. 1) con el cual el valor de ϕ de un proceso AR(1) estacionario se puede interpolar la cantidad de datos faltantes con que se pierde la distribución asumida para la estadística. Así por ejemplo, para ϕ=0.6 arroja que aproximadamente con 91 datos faltantes en una serie de longitud grande se pierde la distribución asumida para la estadística de prueba; También se puede ver en la figura, que cuando la autocorrelación es alta requiere de mayor cantidad de datos faltantes en la serie para que se pierda la distribución.

Figura No. 1: Pérdida de la Distribución de JB_s en un proceso AR(1)



Fuente: Construcción de los autores

En vista de que a partir de 130 datos faltantes en todos los modelos escogidos para el estudio de simulación se pierde la distribución asumida para JB_s, se calculan los cuantiles empíricos de la estadística de prueba para 200, 500 y 700 datos faltantes que se presentan en las Tablas 2 y 3. Es importante notar aquí, que la diferencia numérica es pequeña entre los valores de los cuantiles obtenidos para cada uno de los modelos en estudio y que los valores reportados en las Tablas 2 y 3 son las medianas de dichos valores. Por ejemplo, para 200 datos faltantes en un proceso AR(1), los cuantiles 0.01 obtenidos para los modelos AR(1) considerados (valor del parámetro del modelo en paréntesis) son 0.0233 (-0.95), 0.0238 (-0.5), 0.0194 (0.5) y 0.0202 (0.95). El valor de la mediana es 0.0218, que aparece en la Tabla 2. Lo anterior indica que no es relevante el valor de los parámetros del modelo en la pérdida de la distribución asumida para JB_s. Se puede observar también, que los cuantiles empíricos exceden a los cuantiles teóricos de la distribución $X^2_{(2)}$. Lo anterior se presenta con mayor notoriedad a medida que aumenta la cantidad de datos faltantes en la serie.

Tabla No. 2: Distribución empírica de $JB_{s,}$ en un modelo AR(1)

Orden del cuantil(\100)	Cantidad de datos faltantes		
	200	500	700
1.00	0.0218	0.0275	0.0401
2.50	0.0538	0.0756	0.0995
5.00	0.1092	0.1630	0.1954
10.0	0.2237	0.3369	0.3882
25.0	0.6037	0.9133	1.0940
50.0	1.4572	2.1420	2.5540
75.0	2.9571	4.1910	4.9220
90.0	4.8919	6.8398	8.0238
95.0	6.4423	8.7292	10.3204
97.5	8.0831	10.8448	12.4776
99.0	10.2083	13.7628	15.6741

Fuente: Construcción de los autores

Tabla No. 3: Distribución empírica de $JB_{s,}$ en un modelo AR(2)

Orden del cuantil(\100)	Cantidad de datos faltantes	
	200	500
1.00	0.0209	0.0242
2.50	0.0500	0.0685
5.00	0.1060	0.1322
10.0	0.2130	0.2702
25.0	0.6017	0.7549
50.0	1.4536	1.8016
75.0	2.9340	3.6300
90.0	4.8456	5.9891
95.0	6.3540	7.8507
97.5	7.8628	9.4773
99.0	10.0043	12.0699

Fuente: Construcción de los autores

En las Tablas 2 y 3, se puede ver el cambio sustancial en el valor de los cuantiles ante el aumento de datos faltantes en la serie. Por ejemplo, cuando se pasa de 200 a 700 datos faltantes en un proceso AR(1), el cuantil 0.90 cambia de 4.8919 a 8.0238 y de similar comportamiento el resto de cuantiles, indicando que la distribución empírica de la estadística JB_S, no son iguales en los casos mencionados. Por lo cual en términos de proporciones de datos faltantes en una serie no es adecuado crear rangos de números de datos faltantes, que nos sirva para elegir una distribución empírica, ya que esto carece de practicidad, debido a la longitud tan corta que presentaría dichos rangos. Por lo cual, por

las circunstancias enunciadas anteriormente se recomienda la técnica de *bootstraping* para la obtención de la distribución empírica de la estadística JB_s.

Tabla No. 4: AR(1) con ϕ=-0.95

Cantidad de datos faltantes			
100		**110**	
JB_r	JB_s	JB_r	JB_s
0.6636	0.1563	0.2787	0.0337

Tabla No. 5: AR(1) con ϕ =-0.50

Cantidad de datos faltantes			
85		**90**	
JB_r	JB_s	JB_r	JB_s
0.8632	0.3005	0.9394	0.0401

Tabla No. 6: AR(1) con ϕ =0.50

Cantidad de datos faltantes			
85		**90**	
JB_r	JB_s	JB_r	JB_s
0.9854	0.2247	0.2259	0.0029

Tabla No. 7: AR(1) con ϕ =0.95

Cantidad de datos faltantes			
120		**125**	
JB_r	JB_s	JB_r	JB_s
0.7214	0.0812	0.0603	0.0006

Tabla No. 8: AR(2) con $\phi_1 = -0.3 \ y \ \phi_2 = 0.5$

Cantidad de datos faltantes			
90		**95**	
JB_r	JB_s	JB_r	JB_s
0.5161	0.2617	0.1183	0.0138

Tabla No. 9: AR(2) con $\phi_1 = 0.7 \ y \ \phi_2 = -0.20$

Cantidad de datos faltantes			
125		**130**	
JB_r	JB_s	JB_r	JB_s
0.6999	0.0851	0.1855	0.0004

Tabla No. 10: AR(2) con $\phi_1 = 0.80 \ y \ \phi_2 = -0.16$

Cantidad de datos faltantes			
70		**75**	
JB_r	JB_s	JB_r	JB_s
0.5102	0.0951	0.1430	0.0202

4. UNA APLICACIÓN

A continuación se presenta la técnica de *bootstraping*, con el propósito de ilustrar la forma de obtener la distribución empírica de la estadística de prueba JB_s, en una serie irregularmente espaciada. La rutina de programación para el desarrollo de este procedimiento se elaboró en el paquete RATS versión 5, (Doan, 2000)). Consideremos el siguiente ejemplo:

1. Sea una serie AR(2) estacionaria con $\phi_1 = -0.3$, $\phi_2 = 0.5$, y $\sigma^2 = 1$ de longitud 200, con 20 datos faltantes en las posiciones 5, 7, 10, 16, 20, 22, 26, 31, 50, 73, 74, 100, 120, 125, 132, 158, 170, 187, 189, 198. Se estiman los datos faltantes y se calculan los ruidos\seudoruidos del modelo $w_1, w_2, \ldots, w_{200}$ que se consideran como una muestra.

2. Se crea un vector de enteros de longitud 200 con los cuales se hace la aleatorización para el remuestreo con reemplazamiento, obteniéndose $w_1^*, w_2^*, \ldots, w_{200}^*$.

3. Se calcula la estadística JB_s con la muestra *bootstrap* anterior.

4. Se repiten los pasos 1 y 2 un número grande de veces (5000 veces), y se obtiene la secuencia

$$JB_{s1}, JB_{s2}, \ldots, JB_{s5000}.$$

5. Con el anterior conjunto de datos se calculan los cuantiles de la distribución empírica de JB_s.

Llevando a cabo estos cinco pasos de la técnica *bootstrap*, se obtiene los cuantiles empíricos dados en la Tabla 11. Así, ante una serie real de longitud 200 con 20 datos faltantes, que se ajuste a un modelo AR(2) estacionario con parámetros descritos arriba, para examinar los supuestos probabilísticos de $\{z_t\}$ en particular el de normalidad, se calcula la estadística de prueba JB_s y a un nivel de significancia de $\alpha = 0.05$ se rechaza la normalidad de $\{z_t\}$, si el valor calculado de $JB_s > 6.6071$.

Tabla No. 11: Distribución empírica de JB_s vía bootstrap para 20 datos faltantes en una serie AR(2) estacionaria de longitud 200

Orden del cuantil(\100)	Cantidad de datos faltantes
	20
1.00	0.0503
2.50	0.1181
5.00	0.2280
10.0	0.4404
25.0	1.0233
50.0	2.0563
75.0	3.5419
90.0	5.3242
95.0	6.6071
97.5	7.9864
99.0	9.8114

Fuente: Construcción de los autores

5. CONCLUSIONES

Del estudio de simulación se concluye que no es conveniente chequear la hipótesis nula de normalidad del ruido del proceso comparando el valor calculado de la estadística JB_s con el valor teórico de los cuantiles de una $X^2_{(2)}$, ya que la obtención de los seudoresiduales a partir de las estimaciones de los datos faltantes en la serie si afecta la distribución de la estadística de prueba considerada.

En presencia de datos faltantes en una serie de cualquier longitud se recomienda el uso de la técnica de bootstraping con el propósito de hallar los cuantiles empíricos de la estadística y a un nivel de significancia escogido compararlo con el valor calculado de la estadística JB_s, encontrando así un criterio de decisión más confiable.

Bibliografía

- Brockwell, P.J. and Davis, R.A. (1991). *Time Series: Theory and Methods*, 2nd Edition, Springer- Verlag, New York.
- Brockwell, P.J. and Davis, R.A. (1996). *Introduction to Time Series and Forecasting*, Springer- Verlag, New York.
- Doan, T.A. (2000). *RATS Users Manual Version 5*, Evanston, USA: Estima.
- Efron, B. (1979). Computers and the theory of statistics: Thinking the unthinkable, *SIAM Review*, **21**(49), 460-480.
- Gómez, V. and Maravall, A. (1996). Programs TRAMO and SEATS, instructions for the user, beta version: september 1996, *Working Paper Number* **9628**, Bank of Spain.
- Harvey, A. (1989). Forecasting, Structural Time Series Models and Kalman Filter, Cambridge University Press, Cambridge.
- Jarque, C.M. and Bera, A.K. (1980). Efficient test for normality, homocedasticity and serial independence of regression residuals, *Economics Letters*, **6**, 255-259.
- Jarque, C.M. and Bera, A.K.(1987). A test for normality of observations and regression residual, *International Statistical Review*, **55**, 163-172.