UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Diciembre 2008, Vol. 1, No. 1

Una nota sobre la prueba de Peña y Rodríguez para la bondad del ajuste en series de tiempo

A Note About the Peña-Rodríguez Test of the Goodness of Fit in Time Series

Hanwen Zhang[a]
predictive@telmex.net.co

Resumen

Este artículo tiene como fin divulgar a los lectores una prueba de bondad de ajuste para series de tiempo: la prueba de Peña y Rodríguez modificada (2002). Esta prueba es asintóticamente equivalente a la anterior, pero más potente. Se presentan dos aproximaciones de la estadística de prueba: por la distribución normal y la distribución Gamma. Mediante simulaciones de Monte Carlo, se muestra que la prueba de Peña y Rodríguez es más potente para la detección de series no lineales que la prueba de Ljung-Box y la prueba de Monti.

***Palabras clave*:** coeficiente de autocorrelación, autocorrelación parcial, prueba de no linealidad, modelos ARIMA.

Abstract

The aim of this paper is to divulge the modification of the Peña y Rodríguez test (2002) of goodness of fit. This test is asymptotically equivalent, but it is more powerful than the previous one. Two approaches are proposed using the gamma and the normal distributions. By an empirical example it is shown that the proposed test is more powerful than Ljung-Box test and Monti test of nonlinear models detection.

***Key words*:** autocorrelation coefficient, partial autocorrelation, nonlinearity test, ARIMA models.

1. Introducción

Se presenta la familia de pruebas de bondad de ajuste para el análisis de la independencia de los residuales.

[a]Investigador. Predictive Ltda.

Sea X_t un proceso de media cero generado por un modelo $ARMA(p,q)$, es decir $\phi(B)X_t = \theta(B)\varepsilon_t$, donde B es el operador de rezago, $\phi(v)$ es un polinomio de orden p, y $\theta(v)$ es un polinomio de orden q, y ε_t es ruido blanco. Y sean $\hat{\varepsilon}_1$, $\hat{\varepsilon}_2$,...,$\hat{\varepsilon}_T$ los residuales obtenidos después de estimar el modelo en una muestra de tamaño T, y sea

$$\hat{r}_j = \sum_{t=j+1}^{T} \hat{\varepsilon}_t \hat{\varepsilon}_{t-j} / \sum_{t=1}^{T} \hat{\varepsilon}_t^2 \qquad para\ j = 1, 2, ... \tag{1}$$

Los coeficientes de autocorrelación estimada de los residuales. Basados en estos coeficientes estimados, Box y Jenkins propusieron una familia de estadísticas para analizar la independencia entre los residuales en el año 1976, como sigue:

$$Q = T\left\{\delta \sum_{i=1}^{m} w_i g(\hat{r}_i^2) + (1-\delta)\sum_{i=1}^{m} \omega_i g(\hat{\pi}_i^2)\right\}, \tag{2}$$

Donde los $\hat{\pi}_i$ son los coeficientes de autocorrelación parcial estimados de los residuales, $0 \leq \delta \leq 1$, $m < T$, $w_i \geq 0$, $\omega_i \geq 0$, y g es una función de suavizamiento no decreciente con $g(0) = 0$. Algunos miembros conocidos de esta familia cuando $g(x) = x$ son:

- Box-Pierce (1970), cuando $\delta = 1$ y $w_i = 1$
- Ljung-Box (1978), cuando $\delta = 1$ y $w_i = (T+2)/(T-i)$
- Monti (1994), cuando $\delta = 0$ y $\omega_i = (T+2)/(T-i)$
- Hong (1996), cuando $\delta = 1$, $w_i = k^2(j/m)$, con $k(z) = sen(\pi z)/\pi z$

Peña y Rodriguez (2002) propusieron una prueba basada en el determinante de la matriz de autocorrelación, la estadística está dada por:

$$\hat{D}_m = T[1 - |\hat{R}_m|^{1/m}] \tag{3}$$

Donde $\hat{R}_m$ es:

$$\hat{R}_m = \begin{bmatrix} 1 & \hat{r}_1 & \cdots & \hat{r}_m \\ \hat{r}_1 & 1 & \cdots & \hat{r}_{m-1} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{r}_m & \hat{r}_{m-1} & \cdots & 1 \end{bmatrix} \tag{4}$$

Ellos demostraron que esta prueba pertenece a la familia de pruebas (1.2), además es más potente que las pruebas de Ljung-Box y Monti. A continuación, se presenta una modificación de esta prueba que tiene varias ventajas en comparación que (1.3).

2. La estadística de Peña-Rodríguez y su distribución asintótica

2.1. Prueba de pseudo-verosimilitud

Esta prueba, al igual que la anterior, se inspira en el estudio de las dependencias dentro del análisis multivariante. Los residuales estimados pueden ser considerados como una muestra de datos provenientes de una distribución normal: $(\hat{\varepsilon}_1, \cdots, \hat{\varepsilon}_T) \sim N_T(0, V_t)$ y se quiere probar si la matriz de covarianza V_T es diagonal o no. En el contexto del análisis multivariante, la prueba para chequear si un conjunto de p variables aleatorios tiene una matriz de covarianza diagonal está dada en términos de la matriz de correlación, R_p, mediante la estadística: $-2log\lambda = -Tlog|R_p|$, la cual tiene distribución χ^2 con $p(p+1)/2$ grados de libertad bajo H_0. Para evitar el problema de que cuando $p = T$, la estadística de razón de verosimilitud diverge, Peña y Rodriguez propone la siguiente estadística:

$$D_m^* = -\frac{T}{m+1} log|\hat{R}_m| \tag{5}$$

Donde se estandariza la matriz de correlación estimada por su dimensión $m + 1$.

Ramsey, en el año 1974, demostró que esta estadística pertenece a la familia (1.2) con $g(x) = log(1 - x)$, $\delta = 0$ y $\omega = (m + 1 - i)/(m + 1)$, dando más peso a los coeficientes de autocorrelación de orden bajo, y menos peso a los de orden superior.

2.2. Distribución asintótica y dos aproximaciones

En esta sección, presentamos la distribución asintótica de la estadística D_m^* y dos aproximaciones de esta distribución.

Teorema 2.1. *Si el modelo está correctamente identificado, $\hat{D}_m^*$ tiene distribución asintótica como $\sum_{i=1}^m \lambda_i \chi_{1,i}^2$, donde $\chi_{1,i}^2$ $(i = 1, \cdots, m)$ son variables aleatorios independientes con distribución χ_1^2 y λ_i son los valores propios de $(I_m - Q_m)W_m$, donde $Q_m = X_m V^{-1} X_m'$, V es la matriz de información de los parámetros ϕ y θ, X_m es una matriz de tamaño $m \times (p + q)$ con los elementos ϕ' y θ' definidos por $1/\phi(B) = \sum_{k=0}^{\infty} \phi_i' B^i$ y $1/\theta(B) = \sum_{k=0}^{\infty} \theta_i' B^i$ y $W_m = diag(m - i + 1)/(m + 1)$, con $i = 1, \cdots, m$*

Box y Pierce (1970) asumieron que $m = O(T^{1/2})$ cuando $T \to \infty$ y demostraron que la matriz Q_m puede aproximarse por la matriz de proyección:

$$Q_m = X_m(X_m'X_m)^{-1}X_m' \text{ cuando } m \text{ es relativamente grande.}$$

Basado en este resultado, presentamos dos aproximaciones de los percentiles de la distribución $\sum_{i=1}^{m} \lambda_i \chi^2_{1,i}$.

2.2.1. Aproximación por la distribución Gamma

Aproximamos la estadística D_m^* mediante la distribución Gamma, $G(\alpha, \beta)$, donde los parámetros se definen como:

$$\alpha = \frac{3(m+1)\{m - 2(p+q)^2\}}{2\{2m(2m+1) - 12(m+1)(p+q)\}} \tag{6}$$

y

$$\beta = \frac{3(m+1)\{m - 2(p+q)\}}{2m(2m+1) - 12(m+1)(p+q)} \tag{7}$$

y la distribución tiene media $\alpha/\beta = m/2 - (p+q)$ y varianza $\alpha/\beta^2 = m(2m+1)/(3m+3) - 2(p+q)$. Esta aproximación se denotará por GD_m^*.

2.2.2. Aproximación por la distribución Normal

La aproximación mediante la distribución normal se basa en el resultado de Chen y Deo(2004). Sea $Y_n = \sum_{j=1}^{n} a_{j,n} X_j$, donde X_j es una secuencia de variables aleatorias positivas independientes e idénticamente distribuidas con media conocida μ y varianza σ^2, y $a_{j,n}$ números reales positivos. Se puede demostrar que con normalización adecuada, Y_n es asintóticamente normal. Sin embargo, en muestras finitas, la distribución de Y_n es sesgada a la derecha y carece de distribución normal. Para solucionar este problema, Chen y Deo demostraron el siguiente resultado mediante la expansión Taylor.

Teorema 2.2.

$$\frac{Y_n^\beta - \mu_Y^\beta - \frac{1}{2}\beta(\beta-1)\mu_Y^{\beta-2}\sigma_Y^2}{\beta\mu_Y^{\beta-1}\sigma_Y} \to_d N(0,1)$$

donde

$$\beta = 1 - \frac{\mu E(X_1 - \mu)^3 (\sum_{j=1}^{n} a_{j,n}) \sum_{j=1}^{n} a_{j,n}^3}{3\sigma^4 \left(\sum_{j=1}^{n} a_{j,n}^2\right)^2}$$

El resultado de este teorema se aplica a la distribución Gamma de la sección anterior, teniendo en cuenta que en este caso el valor esperado y la varianza de Y están dados por: $\mu_Y = \alpha/\beta = m/2 - p - q$ y $\sigma_Y^2 = \alpha/\beta^2 = m(2m+1)/3(m+1) - 2p - 2q$, tenemos que la aproximación a Normal de la estadística D_m^* está dada por:

$$ND_m^* = (\alpha/\beta)^{-1/\lambda}(\lambda\sqrt{\alpha})\left((D_m^*)^{1/\lambda} - (\alpha/\beta)^{1/\lambda}\left(1 + \frac{\lambda-1}{2\alpha\lambda^2}\right)\right) \quad (8)$$

y

$$\lambda = \left\{1 - \frac{(m-2p-2q)(m^2/(4(m+1)) - p - q)}{3(m(2m+1)/(6m+6) - p - q)^2}\right\}^{-1} \quad (9)$$

donde para m moderadamente grande, $\lambda \simeq 4$, y los valores de α y β se obtiene mediante (2.2) y (2.3). La estadística ND_m^* tiene distribución normal estándar.

3. Simulación y resultados empíricos

En esta sección, presentamos un estudio sobre la robustez del valor m, el nivel de significación y la potencia de las dos aproximaciones de la estadística comparando con otras pruebas. Las simulaciones fueron hechas en el programa R.

3.1. Robustez frente al valor m

La estadística D_m^* tiene la propiedad deseada de ser robusta al valor m, corroboramos esta conclusión mediante las siguientes simulaciones.

Se compara el nivel de significación de las dos aproximaciones de D_m^* con la anterior estadística de Peña-Rodriguez(D_m) y la estadística de Ljung-Box(Q_{LB}) para varios modelos $AR(1)$ con $\alpha = 0.05$. Para cada modelo $AR(1)$, se generó 20000 series gaussianos de tamaño $T = 100$. Se considera tres valores de m: 10, 15 y 25. Algunos percentiles de la estadística D_m se encuentran en la Tabla 1.

Los resultados del estudio se encuentran en la Tabla 2, donde la estadística D_m^* tiene nivel de significación cercano al 0.05; aúnque las estadísticas D_m y Q_{LB} tienen mejor comportamiento pero claramente son más sensibles al valor m. La robustez frente al valor m de la nueva estadística es una ventaja comparado con estas otras dos estadísticas.

3.2. Modelos lineales

En esta sección la potencia de las pruebas se analiza ajustando modelos $AR(1)$ y $MA(1)$ a 20 modelos $ARMA(2,2)$. En cada caso se generaron 1000 series y la potencia de las pruebas se calculó para m=10. Los resultados se muestran en la Tabla 3. La estadística modificada de Peña-Rodriguez, D_m^* es la más potente en casi todos los modelos (excepto en los modelos 12, 14 y 20 donde Q_{MT} y D_m son más potentes). La diferencia entre las pruebas ND_m^*, GD_m^* y D_m es pequeña, sin

	$p+q$							
m	0	1	2	3	4	5	6	7
7	8.56	6.4	4.52	·	·	·	·	·
10	10.71	9.00	7.14	4.96	·	·	·	·
12	12.10	10.46	8.71	6.76	4.37	·	·	·
14	13.46	11.87	10.11	8.39	6.35	3.56	·	·
24	19.97	18.52	17.05	15.53	13.96	12.32	10.57	8.63
36	27.42	26.06	24.69	23.3	21.88	20.44	18.96	17.44

Tabla 1: Percentiles de la estadística D_m para algunos valores de m, p y q

ϕ	$m=10$			
ND_m^*	GD_m^*	D_m	Q_{LB}	
0.1	0.064	0.056	0.031	0.035
0.5	0.065	0.059	0.032	0.038
0.9	0.058	0.068	0.037	0.040
	$m=15$			
	ND_m^*	GD_m^*	D_m	Q_{LB}
0.1	0.063	0.056	0.031	0.042
0.5	0.066	0.058	0.034	0.043
0.9	0.062	0.052	0.037	0.044
	$m=25$			
	ND_m^*	GD_m^*	D_m	Q_{LB}
0.1	0.064	0.051	0.024	0.058
0.5	0.068	0.052	0.025	0.058
0.9	0.063	0.054	0.027	0.063

Tabla 2: Nivel de significación de ND_m^*, GD_m^*, D_m y Q_{LB} bajo $AR(1)$ con $\alpha = 0.05$ con tamaño muestral 100.

embargo las dos aproximaciones de la nueva prueba casi siempre son más potentes que D_m (excepto en el modelo 20).

3.3. Modelos no lineales

McLeod-Li(1983) hicieron la propuesta de detectar no linealidad en series de tiempo usando los coeficiente de autocorrelación de los residuales al cuadrado en la estadística Q_{LB} en vez de los coeficientes de autocorrelación corrientes. Esta propuesta se basa en la idea de que si los residuales $\hat{\varepsilon}_t$ son independientes, entonces $\hat{\varepsilon}_t^2$ también lo son; pero si el modelo no es lineal y los residuales $\hat{\varepsilon}_t$ no son independientes, este característica puede aparecer en la función de autocorrelación de $\hat{\varepsilon}_t^2$.

M	ϕ_1	ϕ_2	θ_1	θ_2	ND_m^*	GD_m^*	D_m	Q_{LB}	Q_{MT}
$AR(1)$									
1	–	–	-0.5	–	0.457	0.484	0.379	0.188	0.235
2	–	–	-0.8	–	0.987	0.993	0.986	0.657	0.950
3	–	–	-0.6	0.3	0.229	0.272	0.195	0.157	0.137
4	0.7	–	-0.9	–	0.324	0.320	0.234	0.098	0.213
5	0.7	–	-0.4	–	0.154	0.141	0.091	0.069	0.072
6	0.7	0.2	0.5	–	0.682	0.704	0.615	0.295	0.386
7	0.7	–	0.7	-0.15	0.999	1.000	0.999	0.879	0.998
8	0.4	–	-0.6	0.3	0.356	0.339	0.245	0.166	0.147
9	0.7	0.2	-0.5	–	0.814	0.857	0.795	0.691	0.697
10	0.9	-0.4	1.2	-0.3	1.000	1.000	1.000	0.999	1.000
$MA(1)$									
11	0.5	–	–	–	0.371	0.396	0.304	0.226	0.196
12	0.8	–	–	–	0.971	0.986	0.965	0.966	0.990
13	–	–	0.8	0.5	0.776	0.774	0.681	0.406	0.532
14	–	–	-0.6	0.3	0.582	0.611	0.523	0.322	0.603
15	-0.5	–	0.7	–	0.108	0.135	0.082	0.06	0.065
16	0.3	–	0.8	-0.5	0.125	0.138	0.084	0.071	0.059
17	0.8	–	-0.5	0.3	0.991	0.997	0.986	0.985	0.972
18	1.2	-0.5	–	0.9	0.732	0.735	0.714	0.563	0.702
19	0.3	-0.2	-0.7	–	0.719	0.726	0.661	0.406	0.449
20	0.9	-0.4	1.2	-0.3	0.996	1.000	0.998	0.976	0.99

Tabla 3: Potencias de las pruebas basadas en ND_m^*, GD_m^*, D_m y Q_{LB} cuando se ajusta $AR(1)$ y $MA(1)$ a datos de $ARMA(2,2)$ con $\alpha = 0.05$

Los coeficientes de autocorrelación de los residuales al cuadrado están dados por:

$$\tilde{r}_k(\hat{\varepsilon}_t^2) = \frac{T+2}{T-k} \frac{\sum_{t=k+1}^{n}(\hat{\varepsilon}_t^2 - \hat{\sigma})(\hat{\varepsilon}_{t-k}^2 - \hat{\sigma})}{\sum_{t=1}^{n}(\hat{\varepsilon}_t^2 - \hat{\sigma})^2} \tag{10}$$

donde $\hat{\sigma} = \sum \hat{\varepsilon}_t^2/T$. En este caso, la estadística D_m^* está dada por:

$$\hat{D}_m^*(\hat{\varepsilon}_t^2) = -\frac{T}{m+1} log|\tilde{R}_m(\hat{\varepsilon}_t^2)| \tag{11}$$

donde $\tilde{R}_m(\hat{\varepsilon}_t^2)$ es la matriz de autocorrelación dada por (1.4) basado en $\tilde{r}_k(\hat{\varepsilon}_t^2)$. Podemos aproximar la distribución de esta estadística mediante la distribución Gamma y la distribución normal análoga a las aproximaciones presentadas en la sección 2.2, donde ahora los grados de libertad de la distribución Gamma no depende del orden del modelo $ARMA$, es decir, los valores de p y q serán iguales a cero.

En el siguiente estudio, se compara la potencia de las estadísticas D_m^*, D_m, Q_{LB}

y Q_{MT} para detectar la no linealidad de los seis modelos no lineales de la Tabla 4. Los e_t son independientes con distribución normal estándar.

M1:	$y_t = e_t - 0.4e_{t-1} + 0.3e_{t-2} + 0.5e_te_{t-2}$
M2:	$y_t = e_t - 0.3e_{t-1} + 0.2e_{t-2} + 0.4e_{t-1}e_{t-2} - 0.25e_{t-2}^2$
M3:	$y_t = 0.4y_{t-1} - 0.3y_{t-2} + 0.5y_{t-1}e_{t-1} + e_t$
M4:	$y_t = 0.4y_{t-1} - 0.3y_{t-2} + 0.5y_{t-1}e_{t-1} + 0.8e_{t-1} + e_t$
M5:	$y_t = e_t\sigma_t, \quad \sigma_t^2 = 1.21 + 0.404y_{t-1}^2 + 0.153\sigma_{t-1}^2$
M6:	$y_t = 0.025e_t\sigma_t, \quad log\sigma_t^2 = 0.9log\sigma_{t-1}^2 + \eta_t, \quad \eta_t \sim N(0, 0.363)$

Tabla 4: Seis modelos no lineales para el estudio de potencia

Para cada modelo se simularon 1000 replicaciones con tamaño muestral de $T = 100$. Se ajusta un modelo $AR(p)$ a los datos, donde p fue escogido de acuerdo con el criterio AIC con $p \in \{1, 2, 3, 4\}$. Los resultados se presentan en la tabla 5. El estudio indica que la potencia de la nueva prueba de Daniel-Rodríguez D_m^* es más grande de la de D_m utilizando las dos aproximaciones ND_m^* y GD_m^* y la ventaja incrementa con el valor de m.

	$m = 10$				
	ND_m^*	GD_m^*	D_m	Q_{LB}	Q_{MT}
M1	0.154	0.146	0.145	0.121	0.071
M2	0.379	0.360	0.298	0.292	0.199
M3	0.843	0.838	0.789	0.717	0.600
M4	0.676	0.673	0.665	0.566	0.464
M5	0.557	0.522	0.412	0.469	0.376
M6	0.442	0.421	0.440	0.410	0.338
	$m = 25$				
	ND_m^*	GD_m^*	D_m	Q_{LB}	Q_{MT}
M1	0.153	0.142	0.139	0.123	0.051
M2	0.424	0.414	0.325	0.390	0.114
M3	0.810	0.802	0.732	0.663	0.543
M4	0.671	0.662	0.628	0.564	0.281
M5	0.550	0.552	0.421	0.527	0.241
M6	0.430	0.421	0.402	0.376	0.211

Tabla 5: Potencia de las pruebas basadas en ND_m^*, GD_m^*, D_m y Q_{LB} cuando se ajusta $AR(p)$ a series no lineales con $\alpha = 0.05$.

Agradecimientos

Gracias a Dios por introducirme al mundo de la estadística y a la Universidad Santo Tomás por invitarme a ser parte en esta edición de *Comunicaciones en Estadística*.

Referencias

Chen, W. & Deo, R. (2004), 'Power transformation to induce normality and their application.', *Journal of Royal Statistical Society* **Ser. B**(66), 117 – 130.

Ljung, G. & Box, G. (1978), 'On a measure of lack of fit in time series models', *Biometrika* **65**, 297 – 303.

McLeod, A. & Li, W. (1983), 'Diagnostic checking arima time series models using squared-residual autocorrelations', *Journal of Time Series Analysis* **4**, 269 – 273.

Peña, D. and Rodríguez,J. (2002), 'A powerful portmanteau test of lack of fit for time series', *Journal of the American Statistical Association* **96**, 601 – 610.

Peña, D. and Rodríguez,J. (2006), 'The log of the determinant of the autocorrelation matrix for testing goodness of fit in time series', *Journal of Statistical Planning and Inference* **136**(8), 2706 – 2718.