Paper 327-2011

Testing Adequacy of ARMA Models using a Weighted Portmanteau Test on the Residual Autocorrelations

Thomas J. Fisher, Department of Mathematics & Statistics, University of Missouri-Kansas City, Kansas City, MO, 64110, USA

ABSTRACT

In examining the adequacy of a statistical model, an analysis of the residuals is often performed. This includes anything from performing a residual analysis in a simple linear regression to utilizing one of the portmanteau tests in time-series analysis. When modeling an autoregressive-moving average time series we typically use the Ljung-Box statistic on the residuals to see if our fitted model is adequate. In this paper we introduce two new statistics that are *weighted* variations of the common Ljung-Box and, the less-common, Monti statistics. A brief simulation study demonstrates that the new statistics are more powerful than the commonly used Ljung-Box statistic. The new statistics are easy to implement in SAS® and source code is provided.

INTRODUCTION

A plethora of situations are known to occur in which the errors in a regression model are not independent. This violates the underlying assumptions of regression and can lead to a multitude of problems. Modeling the errors via a time series analysis does not address all of the issues but allows us to have better predictive models. However, much like checking the adequacy of the regression through an F-test, checking the adequacy of the fitted time-series model is of the utmost importance. Let $\{X_t\}$ be a time series for $t = 1, \ldots, n$ where n is the number of observations. Suppose $\{X_t\}$ is generated by a stationary and invertible ARMA(p, q) process of the form

$$X_t = \sum_{i=1}^{p} \varphi_i X_{t-i} - \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t$$

where $\varepsilon \sim N(0, \sigma^2)$ are white-noise residuals. A model of this form is typically fitted with the autoregressive and moving average parameters, φ_i and θ_j, respectively, estimated by their maximum likelihood or conditional least squares counterparts, $\hat{\varphi}_i$ and $\hat{\theta}_j$. After we have fit the model for a given p and q, testing for the adequacy of the fitted model follows. Most diagnostic goodness-of-fit tests are based on the residual autocorrelation coefficients provided by

$$\hat{r}_k = \frac{\sum_{t=k+1}^{n} \hat{\varepsilon}_t \hat{\varepsilon}_{t-k}}{\sum_{t=1}^{n} \hat{\varepsilon}_t^2}$$

where $\hat{\varepsilon}_t$ is the observed residual at time t. If we correctly identified (the null hypothesis), or possibly overestimated, the ARMA process, the value of each autocorrelation should be approximately zero. However, if we underestimate the ARMA model (the alternative hypothesis) the values of the autocorrelations will deviate from zero towards ±1. Thus, developing a test statistic based on the autocorrelations has theoretical justification.

PORTMANTEAU TEST

The first widely used testing method based on the autocorrelation coefficients is the Box-Pierce (1970) statistic, provided by

$$Q = n \sum_{k=1}^{m} \hat{r}_k^2.$$

In most modern applications, it has been replaced by the Ljung-Box (1978) statistic

$$\tilde{Q} = n(n+2) \sum_{k=1}^{m} \frac{\hat{r}_k^2}{n-k}$$

that includes the standardizing term $\frac{n+2}{n-k}$ on each squared autocorrelation coefficient. These statistics are used to test for significant correlation up to lag m. It is well known that for independent and identically distributed data, as $n \to \infty$ the autocorrelations behave as independent normally distributed random variables, and therefore under the null hypothesis (correctly fitted model) both Q and $\tilde{Q}$ are shown to be asymptotically distributed chi-squared random variables with $m - (p+q)$ degrees of freedom, where p and q are the order of autoregressive and moving average terms estimated in the fitted model, respectively. If Q or $\tilde{Q}$ are large, compared to the chi-squared critical value at significance level α, we have evidence to suggest the fitted ARMA process does not adequately model the correlation in the data.

The Ljung-Box statistic is provided in the SAS procedure `ARIMA` for an assortment of lags m. For large n, the Box-Pierce and Ljung-Box statistics are essentially equivalent. The Ljung-Box (1978) statistic is typically used since it better approximates a chi-squared random variable for smaller n.

A similar statistic to the Ljung-Box statistic was introduced by Monti (1994) and uses the standardized partial autocorrelation function up to lag m:

$$\tilde{M} = n(n+2)\sum_{k=1}^{m}\frac{\hat{\pi}_k^2}{n-k}$$

where $\hat{\pi}_k$ is the residual partial autocorrelation at lag k. Recently, Peña and Rodríguez (2002) proposed a statistic based on the determinant of the residual autocorrelation matrix:

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

Under the null hypothesis that we have fitted an adequate model for the ARMA process, each $\hat{r}_k \approx 0$. Hence the matrix $\hat{R}_m$ should be approximately the identity matrix. Testing for model adequacy is equivalent to testing if $\hat{R}_m$ is approximately the identity matrix. They show

$$D = n\left(1 - \left|\hat{R}_m\right|^{1/m}\right)$$

is asymptotically distributed as a linear combination of chi-squared random variables and is approximately a Gamma distributed random variable for large values of m. In practice, they recommend the matrix $\hat{R}_m$ be constructed using the standardized residuals as this improves the Gamma distribution approximation. In Peña and Rodríguez (2006) they show that the log of the determinant follows the same asymptotic distribution as D and can be better in small sample time series. The statistic D determines whether the matrix $\hat{R}_m$ is an identity matrix, or equivalent, if the fitted model is adequate.

It has been demonstrated that both $\tilde{M}$ and D improve over the Ljung-Box and Box-Pierce statistics; see Monti (1994) or Peña and Rodríguez (2002, 2006). However, neither appears to be frequently implemented in applications of time series. Particularly, the Peña and Rodríguez statistic may be difficult to implement since it involves calculating the determinant of a matrix. As pointed out in Lin and McLeod (2006), the statistic D constructed using the standardized residuals may be degenerate in practice since the matrix $\hat{R}_m$ could be ill-conditioned or singular.

WEIGHTED PORTMANTEAU TEST

In this article we propose two new statistics that are easy to implement and improve over the frequently used Ljung-Box and Box-Pierce statistics. Define

$$\tilde{Q}_W = n(n+2)\sum_{k=1}^{m}\frac{(m-k+1)}{m}\frac{\hat{r}_k^2}{n-k}$$

and

$$\tilde{M}_W = n(n+2)\sum_{k=1}^{m}\frac{(m-k+1)}{m}\frac{\hat{\pi}_k^2}{n-k}.$$

The two statistics look similar to the Ljung-Box and Monti statistics with the exception a weight, $\frac{m-k+1}{m}$, on each autocorrelation or partial autocorrelation. The weights are derived using multivariate analysis techniques on the matrix of autocorrelations or matrix of partial autocorrelations (similar to that in Peña and Rodríguez). Note that the sample autocorrelation at lag 1, $\hat{r}_1$, is given weight $\frac{m}{m} = 1$. The sample autocorrelation at lag 2, $\hat{r}_2$, is given weight $\frac{m-1}{m} < 1$. We can interpret the weights as putting more emphasis on the first autocorrelation, and the least emphasis on the autocorrelation at lag m (corresponding weight $\frac{1}{m}$). This matches our intuition about statistical estimators. The first autocorrelation $\hat{r}_1$ is calculated using information from all n observations. The second autocorrelation is based on $n-1$ observations, and the m^{th} autocorrelation is based on $n-m$ observations. Intuitively, it makes sense to put more emphasis on the first autocorrelation as it should be the most accurate. This idea also holds true for the partial autocorrelations.

The two statistics are asymptotically distributed as a linear combination of chi-squared random variables. This is the same asymptotic distribution as the statistics in Peña and Rodríguez (2002, 2006). The weighted Ljung-Box $\tilde{Q}_W$ and weighted Monti $\tilde{M}_W$ statistics are asymptotically equivalent to D but have the added benefit of easy calculation and computational stability. When a small number of parameters have been fit under the null hypothesis of an adequate model, the statistics $\tilde{Q}_W$ and $\tilde{M}_W$ are approximately distributed as Gamma random variables with shape parameter

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

and scale parameter

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

The Gamma approximation is constructed to have the same theoretical mean and variance as the true asymptotic distribution.

IMPLEMENTATION IN SAS

The two statistics, $\tilde{Q}_W$ and $\tilde{M}_W$, require no difficult matrix calculations and are easy to implement in SAS. Although their asymptotic distribution is difficult to express, an easy approximation is possible with the gamma distribution. The shape and scale parameters may look complicated, but when testing at lag m with p autoregressive and q moving average fitted parameters, the shape and scale parameters are just constants. The following code will calculate both, $\tilde{Q}_W$ and $\tilde{M}_W$, for a given m after $p + q$ parameters were fit. This example fits an ARMA(2,1) model on the variable `x` in the dataset `myseries`; we then test the models adequacy at lag $m = 30$, where $p + q = 3$ parameters were estimated in the model:

```
PROC ARIMA DATA=myseries;
    IDENTIFY VAR=x NLAG=30 NOPRINT;
    ESTIMATE p=2 q=1 METHOD=ML;
    FORECAST OUT=getresiduals;
RUN;

PROC ARIMA DATA=getresiduals;
    IDENTIFY VAR=RESIDUAL NLAG=30 OUTCOV=acfs;
RUN;

DATA acfs;
    SET acfs;
    IF lag = 0 THEN r = 0;
           ELSE r = corr*corr;
    IF lag = 0 THEN pr = 0;
           ELSE pr = partcorr*partcorr;
    rLB = r/N;
    prM = pr/N;
    sampSize = N+lag;
    m = 30; pq = 3;
    shape = (3/4)*(m^2 + m -2*(m-1)*(pq))^2/(2*m^3 + 3*m^2 + m -6*(m^2-2*m-1)*pq);
    scale = (2/3)*(2*m^3 + 3*m^2 + m - 6*(m^2-2*m-1)*pq)/(m*(m^2+m-2*(m-1)*pq));
    weights = (m - lag + 1)/m;
    WLBtmp + (weights*rLb);
    WLB = WLBtmp*sampSize*(sampSize+2);
    WMtmp + (weights*prM);
    WM = WMtmp*sampSize*(sampSize+2);
    pValWLB = SDF('gamma', WLB, shape, scale);
    pValWM = SDF('gamma', WM, shape, scale);
RUN;

PROC PRINT DATA=acfs(where=(lag=30));
    VAR m WLB pValWLB WM pValWM;
RUN;
```

SIMULATION STUDY

To demonstrate an improvement over the existing methods consider the following simulation study. A sample of size $n = 100$ is generated from one of the ARMA(2,2) models provided in Table 1 or Table 2. An AR(1) or MA(1) is fitted for each model. Four statistics are calculated and compared to their theoretical critical value at lag $m = 20$. The process is repeated 10,000 times and the simulated power is reported as the proportion of times the calculated statistic exceeds the critical value. This study is similar to those in Monti (1994) and Peña and Rodríguez (2002, 2006). The test statistic with the highest power for any particular model is in bold.

Fitted by AR(1) Model

φ_1	φ_2	θ_1	θ_2	$\tilde{Q}$	$\tilde{M}$	$\tilde{Q}_W$	$\tilde{M}_W$
---	---	-0.50	---	0.2261	0.2057	0.3202	**0.3516**
---	---	-0.80	---	0.6170	0.8753	0.8494	**0.9804**
---	---	-0.60	0.30	0.6377	0.9482	0.8618	**0.9945**
0.10	0.30	---	---	0.3681	0.2930	**0.5113**	0.4928
1.30	-0.35	---	---	0.6204	0.5363	**0.7927**	0.7792
0.70	---	-0.40	---	0.4574	0.4634	0.6524	**0.7140**
0.70	---	-0.90	---	0.9366	0.9998	0.9982	**1.0000**
0.40	---	-0.60	0.30	0.6951	0.9853	0.9186	**0.9994**
0.70	---	0.70	-0.15	0.1677	0.1272	**0.2163**	0.1898
0.70	0.20	0.50	---	0.6329	0.5970	0.7831	**0.7862**
0.70	0.20	-0.50	---	0.3193	0.3112	0.4768	**0.5594**
0.90	-0.40	1.20	-0.30	0.5682	0.8962	0.8012	**0.9815**

Table 1: Power of $\tilde{Q}$, $\tilde{M}$, $\tilde{Q}_W$ and $\tilde{M}_W$ when data are fitted by an AR(1) under various ARMA(2,2) models

Fitted by MA(1) Model

φ_1	φ_2	θ_1	θ_2	$\tilde{Q}$	$\tilde{M}$	$\tilde{Q}_W$	$\tilde{M}_W$
0.50	---	---	---	0.2404	0.1769	**0.3321**	0.3011
0.80	---	---	---	0.9609	0.9323	**0.9888**	0.9849
1.10	-0.35	---	---	0.9834	0.9822	**0.9987**	**0.9987**
---	---	0.80	-0.50	0.6974	0.8097	0.8951	**0.9537**
---	---	-0.60	0.30	0.3137	0.3151	0.4672	**0.5325**
0.50	---	-0.70	---	0.7772	0.7093	**0.9049**	0.8931
-0.50	---	0.70	---	0.8053	0.7299	**0.9253**	0.9109
0.30	---	0.80	-0.50	0.4975	0.5663	0.7116	**0.8044**
0.80	---	-0.50	0.30	0.9521	0.9002	**0.9811**	0.9701
1.20	-0.50	0.90	---	0.3787	0.5694	0.5099	**0.7249**
0.30	-0.20	-0.70	---	0.2093	0.1713	0.2979	**0.3036**
0.90	-0.40	1.20	-0.30	0.6278	0.8227	0.8493	**0.9515**

Table 2: Power of $\tilde{Q}$, $\tilde{M}$, $\tilde{Q}_W$ and $\tilde{M}_W$ when data are fitted by an MA(1) under various ARMA(2,2) models

In Monti (1994) it is shown that the statistic $\tilde{M}$ dominates $\tilde{Q}$ when the fitted model underestimates the order of the moving average component. This can be seen in many of the examples in Tables 1 & 2. When the fitted model underestimates the order of the autoregressive component, $\tilde{Q}$ tends to perform better than $\tilde{M}$. Moreover, the results in the table demonstrate that the newly proposed statistics appear to outperform both, the Ljung-Box and Monti statistics, in every case. This result happens due to the weight associated with each autocorrelation or partial autocorrelation. Consider the simple case of an AR(2) process where we underestimate the order of the process and fit an AR(1) model (the fourth and fifth model in Table 1 are examples). The residuals will be determined by an AR(1) process with an unknown autoregressive parameter φ and the autocorrelations of the residuals should follow the form $r_k = \varphi^k$. Since $|\varphi| < 1$, as k grows the autocorrelations will get small, approaching the value zero. Under the null hypothesis that we've fit the correct model, we expect the autocorrelations to take on the value zero. The statistic $\tilde{Q}$ weighs everything equally and is more susceptible to not detecting an inadequate model that underestimates the autoregressive order. The new test statistic puts more emphasis (weight) on the first few autocorrelations (those most likely to deviate from zero) and hence is more likely to detect that the fitted model has underestimated the autoregressive order. A similar argument can be made to explain why $\tilde{M}_W$ appears to outperform $\tilde{M}$ when we underestimate the order of the moving-average.

CONCLUSIONS

This article introduced two new statistics for checking the adequacy of a fitted stationary ARMA process. A brief simulation study demonstrates the new statistics improve over those typically used in the application of times series. Source code is provided to implement these test statistics in the SAS statistical language.

REFERENCES

Box, G.E.P. and Pierce, D.A. (1970) "Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models," *Journal of the American Statistical Association*, Vol. 65, No. 332, pp. 1509—1526.

Lin, J.W. and McLeod, A.I. (2006) "Improved Peña-Rodríguez portmanteau test," *Computational Statistics and Data Analysis*, Vol. 51, No. 3, pp. 1731—1738.

Ljung, G.M. and Box, G.E.P. (1978) "On a measure of lack of fit in time series models," *Biometrika*, Vol. 65, No. 2, pp. 297—303.

Monti, A.C. (1994) "A Proposal for a Residual Autocorrelation Test in Linear Models," *Biometrika*, Vol. 81, No. 4, pp. 776—790.

Peña, D. and Rodríguez, J. (2002) "A Powerful Portmanteau Test of Lack of Fit for Time Series," *Journal of the American Statistical Association*, Vol. 97, No. 458, pp. 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*, Vol. 136, No. 8, pp. 2706—2718.

RECOMMENDED READING

The theoretical results of this work will appear in a journal article currently in preparation by the author and collaborator Dr. Colin M. Gallagher of Clemson University. Please contact either for more information on its publication and where to obtain a copy.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Dr. Thomas J. Fisher
Department of Mathematics & Statistics
University of Missouri-Kansas City
5100 Rockhill Road
Kansas City, MO 64110
(816) 235-2853
fishertho@umkc.edu
http://f.web.umkc.edu/fishertho/