Model Selection, Prediction Intervals and Outlier Detection for Time Series

David J. Olive [*]

Southern Illinois University

January 11, 2014

Abstract

Model selection for ARIMA models, prediction intervals for a wide variety of time series models, and the use of response plots for detecting outliers is discussed. A robust method for time series models that only have AR parameters is also given.

KEY WORDS: ARIMA, prediction interval, outliers, variable selection.

[*] David J. Olive is Associate Professor, Department of Mathematics, Southern Illinois University, Carbondale, IL 62901-4408 (E-mail: *dolive@siu.edu*).

1. Introduction

Many time series models have the form

$$Y_t = \tau + \sum_i \psi_i Y_{t-ik_i} + \sum_j \nu_j e_{t-jk_j} + e_t \tag{1}$$

where $\{e_t\}$ are iid with 0 mean and variance σ_e^2 and $Y_1, ..., Y_n$ form the time series observed at times 1, ..., n while the errors e_t are unobserved random variables. For example, the Box, Jenkins, and Reinsel (1994) multiplicative seasonal ARIMA$(p, d, q) \times (P, D, Q)_s$ time series models have this form.

Next several important time series models will be given. We will use the *R* software notation and write a moving average parameter θ and seasonal moving average parameter Θ with a positive sign. Many references and software will write the model with a negative sign for the moving average parameters. The backshift operator or lag operator B satisfies $BW_t = W_{t-1}$ and $B^j W_t = W_{t-j}$.

A *moving average* MA(q) times series is

$$Y_t = \tau + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q} + e_t = \tau + (1 + \theta_1 B + \cdots + \theta_q B^q) e_t = \tau + \theta(B) e_t$$

where $\theta(B) = 1 + \theta_1 B + \theta_2 B^2 + \cdots + \theta_q B^q$ and $\theta_q \neq 0$.

An *autoregressive* AR(p) times series is

$$Y_t = \tau + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + e_t \text{ or } \ (1 - \phi_1 \mathrm{B} - \cdots - \phi_\mathrm{p} \mathrm{B}^\mathrm{p}) \mathrm{Y_t} = \tau + \mathrm{e_t},$$

or $\phi(B) Y_t = \tau + e_t$ where $\phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$ and $\phi_p \neq 0$.

An *autoregressive moving average* ARMA(p, q) times series is

$$Y_t = \tau + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q} + e_t,$$

or $\phi(B)Y_t = \tau + \theta(B)e_t$ where $\theta_q \neq 0$ and $\phi_p \neq 0$.

To describe ARIMA models, let the difference operator $\bigtriangledown = (1-B)$. Let $X_t = \bigtriangledown^d Y_t = (1-B)^d Y_t$ be the differenced time series. The first difference is $X_t = \bigtriangledown Y_t = (1-B)Y_t = Y_t - Y_{t-1}$. The second difference is $X_t = \bigtriangledown^2 Y_t = \bigtriangledown(\bigtriangledown Y_t) = Y_t - 2Y_{t-1} + Y_{t-2}$. If Y_t follows an ARIMA(p,d,q) model, want X_t to follow a stationary and invertible ARMA(p,q) = ARIMA$(p,0,q)$ model. Typically $d = 0$ or 1, but occasionally $d = 2$. Usually $\tau = 0$ if $d > 1$. The ARIMA$(p, d = 1, q)$ model is $Y_t = \tau + (1+\phi_1)Y_{t-1} + (\phi_2 - \phi_1)Y_{t-2} + \cdots + (\phi_p - \phi_{p-1})Y_{t-p} - \phi_p Y_{t-p-1} + \theta_1 e_{t-1} + \cdots + \theta_q e_{t-q} + e_t$. The ARIMA$(p,d,q)$ model can be written compactly as $\phi(B) \bigtriangledown^d Y_t = \tau + \theta(B)e_t$.

The multiplicative seasonal ARIMA models also have backshift and difference notation. Let $\Phi(B) = 1 - \Phi_1 B^s - \Phi_2 B^{2s} - \cdots - \Phi_P B^{Ps}$. Let $\Theta(B) = 1 + \Theta_1 B^s + \Theta_2 B^{2s} + \cdots + \Theta_Q B^{Qs}$. Let s be the seasonal period. Hence $s = 4$ for quarterly data and $s = 12$ for monthly date. Then the multiplicative ARMA(p,q)$\times(P,Q)_s$ model satisfies $\phi(B)\Phi(B)Y_t = \tau + \theta(B)\Theta(B)e_t$. This model is an ARMA$(p+Ps, q+Qs)$ model where the nonzero coefficients are determined only by $p+P+q+Q$ coefficients, the AR characteristic polynomial is $\phi(B)\Phi(B)$ and the MA characteristic polynomial is $\theta(B)\Theta(B)$.

Let $\bigtriangledown_s Y_t = (1-B^s)Y_t = Y_t - Y_{t-s}$ and $\bigtriangledown_s^D Y_t = (1-B^s)^D Y_t$ where usually $d \leq 1$ and $D \leq 1$, $d = 2$ is rare and $D = 2$ is very rare. The differenced time series $X_t = \bigtriangledown^d \bigtriangledown_s^D Y_t$. Then $Y_t \sim$ ARIMA$(p,d,q) \times (P,D,Q)_s$ if $X_t \sim$ ARMA$(p,q) \times (P,Q)_s$. Also, $\phi(B)\Phi(B) \bigtriangledown^d \bigtriangledown_s^D Y_t = \tau + \theta(B)\Theta(B)e_t$ where the default is $\tau = 0$ if $d > 0$ or $D > 0$.

2. Model Selection

Let I be a time series model. The $AIC(I)$ statistic is used to pick a model from several ARIMA models. The model I_{min} with the smallest AIC is always of interest but often overfits: has too many unnecessary parameters. Imagine fitting an ARIMA(p, d, q) model where $d = 0, 1$ or 2 is fixed and p and q run from 0 to j for small j. The number of parameters in the model for fixed d is $p + q + 2$ where $\sigma = \sqrt{V(X_t)}$, τ, $\phi_1, ..., \phi_p$, $\theta_1, ..., \theta_q$ are the parameters. $AIC(I)$ tends to be large when the model does not have enough terms, to drop as needed terms are added, and then to rise as unnecessary terms are added. If $\Delta(I) = AIC(I) - AIC(I_{min})$, then models with $\Delta(I) \leq 2$ are good, models with $4 \leq \Delta(I) \leq 7$ are borderline. See Brockwell and Davis (1987, p. 269), Duong (1984), and Burnham and Anderson (2004).

The initial model to look at is the model I_I with the smallest number of predictors such that $\Delta(I_I) \leq 2$, and also examine submodels I with fewer predictors than I_I with $\Delta(I) \leq 7$. Similar I_I rules are used Olive (2008, pp. 145, 456) and Olive and Hawkins (2005) for multiple linear regression and generalized linear models.

The aicmatrix computes $\Delta(I) = AIC(I) - AIC(I_{min})$ for ARIMA(p,d,q) models where d is fixed or for ARIMA$(p, d, q) \times (P, D, Q)_s$ models where d, P, D, Q and s are fixed, and p and q run from 0 to j for small $j = pmax$ such as $pmax = 5$. Here I_{min} is the ARIMA(p_m, d, q_m) model or the ARIMA$(p_m, d, q_m) \times (P, D, Q)_s$ model with the smallest AIC(I). This model will have a 0.00 in the aicmatrix. Look for model I_I with $p_I + q_I \leq p_m + q_m$ as small as possible such that the aicmatrix entry ≤ 2. It is possible that $I_I = I_{min}$. Also look at models I with $p + q \leq p_I + q_I$ with aicmatrix entries ≤ 7, especially models with entries ≤ 4. Check that the selected model I does not fail to reject Ho for Ho: $\phi_p = 0$ or Ho: $\theta_q = 0$. Make the usual model checks of plotting the time series, ACF, PACF, response and residual plots, the ACF and PACF of the residuals, and the plot of the Box–Ljung pvalues.

Another useful concept is that of a submodel. If d, P, D, and Q are fixed and model I_i has p_i and q_i for $i = 1, 2$, then I_1 is a submodel of I_2 if $p_1 \leq p_2$ and $q_1 \leq q_2$. If $\Delta(I_1) \leq \Delta(I) + 2$ where I_1 is a submodel of I, tentatively eliminate model I. Model I_1 will be a submodel of all models I with aicmatrix entries to the right and below the model I entry. Hence model I_1 is at the upper left corner of a block of models I such that I_1 is a submodel for each model I in the block.

These are rules of thumb: they do not always work but often lead to a good model. If I_I is the ARIMA(1,0,1) model, might take an AR(3) or MA(3) model even though these have 1 more parameter.

```
aicmat(WWWusage,dd=1,pmax=5,k=15)
$aics            q
  p      0     1    2     3    4    5  Find I_I by looking at models
  0 119.86 38.67 8.74  9.13 8.24 7.72  on and above the diagonal
  1  18.10  3.16 5.11  3.44 3.96 5.14  through (5,4) and (4,5) which have
  2  11.04  5.15 6.22  4.63 2.10 6.95  p+q <= 9. Interesting models are on
  3   0.85  2.80 4.48  3.27 3.62 5.29  or above the diagonal through (3,0),
  4   2.79  1.74 5.04  7.94 4.26 6.99  (2,1), (1,2) and (0,3) since they
  5   4.72  6.50 2.40 10.50 0.00 1.63  have p+q <= 3.
```

Example 1. Shown above is the aicmatrix of $\Delta(I) = AIC(I) - AIC(I_{\min})$ for the R `WWW usage` time series, which gives the number of users connected to the Internet through a server every minute where $n = 100$. First differences were used so $d = 1$.

From this output, I_{min} is the ARIMA(5,1,4) model and I_I is the ARIMA(3,1,0) model. Interesting models have $p + q \leq 3$ with entries ≤ 7. These are the ARIMA(2,1,1), ARIMA(1,1,2), and ARIMA(1,1,1) models. Since the ARIMA(1,1,1) model is a submodel of the ARIMA(2,1,1) and ARIMA(1,1,2) models, look at the ARIMA(3,1,0) model I_I first, and then at the ARIMA(1,1,1) model.

3. Prediction Intervals

For forecasting, predict $Y_{t+1}, ..., Y_{t+L}$ given the past $Y_1, ..., Y_t$. A large sample $100(1-\alpha)\%$ prediction interval (PI) for Y_{t+j} is (L_n, U_n) where the coverage $P(L_n \leq Y_{t+j} \leq U_n) = 1 - \delta_n \rightarrow 1 - \alpha$ as $n \rightarrow \infty$.

The shorth estimator will be defined below and used to create large sample PIs that do not require knowing the distribution of the errors e_t. If the data are $Y_1, ..., Y_n$, let $Y_{(1)} \leq \cdots \leq Y_{(n)}$ be the order statistics. Let $\lceil x \rceil$ denote the smallest integer greater than or equal to x (e.g., $\lceil 7.7 \rceil = 8$). Consider intervals that contain c cases $(Y_{(1)}, Y_{(c)}), (Y_{(2)}, Y_{(c+1)}), ..., (Y_{(n-c+1)}, Y_{(n)})$. Compute $Y_{(c)} - Y_{(1)}, Y_{(c+1)} - Y_{(2)}, ..., Y_{(n)} - Y_{(n-c+1)}$. Then the estimator shorth$(c) = (Y_{(d)}, Y_{(d+c-1)})$ is the interval with the shortest length.

Suppose the data $Y_1, ..., Y_n$ are iid and a large sample $100(1-\alpha)\%$ PI is desired for a future value Y_f such that $P(Y_f \in (L_n, U_n)) \rightarrow 1-\alpha$ as $n \rightarrow \infty$. The shorth(c) interval is a large sample $100(1-\alpha)\%$ PI if $c/n \rightarrow 1-\alpha$ as $n \rightarrow \infty$, that often has the asymptotically shortest length. If $c = \lceil n(1-\alpha) \rceil$, then for large n the coverage of the shorth(c) PI is about $1 - \alpha - 1.12\sqrt{\alpha/n}$. See Frey (2013).

Some more notation is needed before deriving PIs for time series. The l step ahead forecast for a future value Y_{t+l} is $\hat{Y}_t(l)$ and the l step ahead forecast residual is $\hat{e}_t(l) = Y_{t+l} - \hat{Y}_t(l)$. For example, a common choice for model (1) is

$$\hat{Y}_t(l) = \hat{\tau} + \sum_i \hat{\psi}_i Y^*_{t+l-ik_i} + \sum_j \hat{\nu}_j \hat{e}^*_{t+l-jk_j}$$

where $\hat{e}_t$ is the tth residual, $Y^*_{t+l-ik_i} = Y_{t+l-ik_i}$ if $l - ik_i \leq 0$, $Y^*_{t+l-ik_i} = \hat{Y}_t(l - ik_i)$ if $l - ik_i > 0$, $\hat{e}^*_{t+l-jk_j} = \hat{e}_{t+l-jk_j}$ if $l - jk_j \leq 0$, and $\hat{e}^*_{t+l-jk_j} = 0$ if $l - jk_j > 0$, and the forecasts $\hat{Y}_t(1), \hat{Y}_t(2), ..., \hat{Y}_t(L)$ are found recursively if there is data $Y_1, ..., Y_t$. Typically the residuals $\hat{e}_t = \hat{e}_{t-1}(1)$ are the 1 step ahead forecast residuals and the fitted or predicted values $\hat{Y}_t = \hat{Y}_{t-1}(1)$ are the 1 step ahead forecasts.

In the simulations, a moving average MA(2) = ARIMA(0,0,2)$\times(0,0,0)_1$ model, $Y_t = \tau + \theta_1 e_{t-1} + \theta_2 e_{t-2} + e_t$, will be used. Suppose data $Y_1, ..., Y_n$ from this model is available. The *R* software produces $\hat{e}_t$ and $\hat{Y}_t = Y_t - \hat{e}_t$ for $t = 1, ..., n$ where $\hat{Y}_t = \hat{Y}_{t-1}(1) = \hat{\tau} + \hat{\theta}_1 \hat{e}_{t-1} + \hat{\theta}_2 \hat{e}_{t-2}$ and $\hat{e}_t(1) = Y_{t+1} - \hat{Y}_t(1)$ for $t = 3, ..., n$. Also, $\hat{Y}_n(1) = \hat{\tau} + \hat{\theta}_1 \hat{e}_n + \hat{\theta}_2 \hat{e}_{n-1}$. Hence there are n 1 step ahead forecast residuals $\hat{e}_t = \hat{e}_{t-1}(1)$ available. Similarly, $\hat{Y}_t(2) = \hat{\tau} + \hat{\theta}_2 \hat{e}_t$ for $t = 1, ..., n$. Hence the 2 step ahead forecast residuals are available for $t = 3, ..., n-2$. Now $\hat{Y}_t(l) = \hat{\tau} \approx \overline{Y}$ for $l > 2$. Hence there are n l step ahead forecast residuals $Y_t - \overline{Y}$ for $l > 2$ and $t = 1, ..., n$.

Typically time series PIs assume normality and are similar to equation (2) below. There is a large literature on alternative PIs, especially for AR(p) models. See Clements and Kim (2007), Kabaila and He (2007), Panichkitkosolkul and Niwitpong (2012), Thombs and Schucany (1990), and Vidoni (2009) for references.

The following normal PI is often used, but typically does not work well unless the l step ahead forecast is normally distributed. For many time series models, a large sample normal $100(1-\alpha)\%$ PI for Y_{t+l} is

$$(L_n, U_n) = \hat{Y}_t(l) \pm t_{1-\alpha/2,n-p-q} SE(\hat{Y}_t(l)). \tag{2}$$

Suppose that as $n \to \infty$, $\hat{Y}_t(l) \to E(Y_{t+l}) = \mu_{t+l}$ and $SE(\hat{Y}_t(l)) \to SD(Y_{t+l}) = \sigma_{t+l}$. Then $P[Y_{t+l} \in (L_n, U_n)] \approx P[Y_{t+l} \in (\mu_{t+l} - z_{1-\alpha/2}\sigma_{t+l}, \mu_{t+l} + z_{1-\alpha/2}\sigma_{t+l})] =$ $P[|Y_{t+l} - \mu_{t+l}| < z_{1-\alpha/2}\sigma_{t+l}]$ " $\geq$ " $1 - \frac{1}{z^2_{1-\alpha/2}}$ assuming Chebyshev's inequality holds to a good approximation. Hence a 95% PI could have coverage as low as 75% and a 99.7% PI could have coverage as low as 89%.

The next PI ignores the time series structure of the data. Let $\overline{e}_t = Y_t - \overline{Y}$, and let shorth$(\lceil n(1-\alpha)\rceil) = (\tilde{L}_n, \tilde{U}_n)$ be computed from the $\overline{e}_t$. Then the large sample shorth(c_1) $100(1-\alpha)\%$ PI for Y_{t+l} is

$$(L_n, U_n) = (\overline{Y} + d_n \tilde{L}_n, \overline{Y} + d_n \tilde{U}_n) \tag{3}$$

where $d_n = (1 + \frac{15}{n})\sqrt{\frac{n-1}{n+1}}$. For stationary invertible ARMA(p,q) models, this PI is too long for l near 1, but should have short length for large l and if $l > q$ for an MA(q) model. This PI is the Olive (2013a) PI suggested for Y_f when $Y_1, ..., Y_t$ and Y_f are iid.

The following PI is new and takes into account the time series structure of the data. A similar idea in Masters (1995, p. 305) is to find the l step ahead forecast residuals and use percentiles to make PIs for Y_{t+l} for $l = 1, ..., L$. For ARIMA(p,d,q) models, let $c_2 = \lceil n(1-\alpha_n)\rceil$ and compute shorth$(c_2) = (\tilde{L}_n, \tilde{U}_n)$ of the l-step ahead forecast residuals $\hat{e}_t(l)$. Then a large sample $100(1-\alpha)\%$ PI for Y_{t+l} is

$$(L_n, U_n) = (\hat{Y}_n(l) + \tilde{L}_n, \hat{Y}_n(l) + \tilde{U}_n) \tag{4}$$

where $1-\alpha_n = \min(1-\alpha+0.05, 1-\alpha+(p+q)/n)$ for $\alpha > 0.1$ and $1-\alpha_n = \min(1-\alpha/2, 1-$

$\alpha + 10(p+q)\alpha/n)$ for $\alpha \leq 0.1$. Similar ideas to compensate for the undercoverage of estimated highest density regions, such as the shorth(c) interval, are used in Olive (2007, 2013b) to create prediction intervals for regression models and prediction regions for multivariate regression models.

Figure 1 shows a simulated MA(2) time series with $n = 100$, $L = 7$ and $U(-1,1)$ errors. The horizontal lines correspond to the 95% PI (3). Two of the one hundred time series points $Y_1, ...Y_{100}$ lie outside of the two lines. All seven of the future cases $Y_{101}, ..., Y_{107}$ lie within their large sample 95% PI.

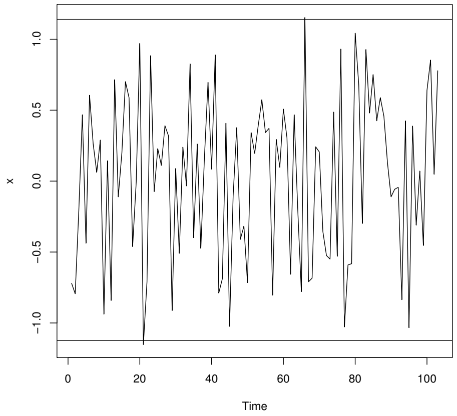

Figure 1: PIs for an MA(2) Time Series with Uniform$(-1,1)$ Errors

The simulations used the MA(2) model where the distribution of the white noise $\{e_t\}$ is N(0,1), t_5, $U(-1,1)$ or (EXP(1) - 1). All these distributions have mean 0, but the fourth distribution is not symmetric. The simulation generates 5000 time series of length $n + L$ and PIs are found for $Y_{n+1}, ..., Y_{n+L}$. The simulations used $L = 7$ and 95% and 50% nominal PIs. The two types of PI used were the normal PI (2), and the possibly asymptotically optimal PI used which is (3) for Y_{t+j} where $j > 2$ and (4) for $j = 1, 2$.

Table 1: Normal Errors

α	n	PI	j=1	j=2	j=3	j=4	j=5	j=6	j=7
0.05	100	N	0.9396	0.9432	0.9444	0.9436	0.9486	0.9498	0.9462
0.05	100		3.889	4.072	4.198	4.198	4.198	4.198	4.198
0.05	100	A	0.9482	0.9582	0.9550	0.9496	0.9556	0.9590	0.9532
0.05	100		4.143	4.509	4.461	4.461	4.461	4.461	4.461
0.05	1000	N	0.9520	0.9464	0.9476	0.9474	0.9496	0.9524	0.9474
0.05	1000		3.919	4.080	4.179	4.179	4.179	4.179	4.179
0.05	1000	A	0.9520	0.9488	0.9482	0.9446	0.9478	0.9500	0.9482
0.05	1000		3.913	4.086	4.170	4.170	4.170	4.170	4.170
0.5	100	N	0.4840	0.4896	0.5052	0.4980	0.4908	0.4984	0.4910
0.5	100		1.328	1.390	1.433	1.433	1.433	1.433	1.433
0.5	100	A	0.4912	0.4866	0.5052	0.4950	0.4956	0.4920	0.4974
0.5	100		1.391	1.456	1.497	1.497	1.497	1.497	1.497
0.5	1000	N	0.4936	0.4994	0.5000	0.4980	0.5114	0.5030	0.5072
0.5	1000		1.347	1.401	1.436	1.436	1.436	1.436	1.436
0.5	1000	A	0.4876	0.4920	0.4950	0.4940	0.5000	0.4964	0.4952
0.5	1000		1.338	1.392	1.427	1.427	1.427	1.427	1.427

Table 2: t_5 Errors

α	n	PI	j=1	j=2	j=3	j=4	j=5	j=6	j=7
0.05	100	N	0.9366	0.9456	0.9448	0.9422	0.9422	0.9428	0.9434
0.05	100		4.995	5.226	5.385	5.385	5.385	5.385	5.385
0.05	100	A	0.9444	0.9576	0.9504	0.9468	0.9460	0.9492	0.9480
0.05	100		4.995	5.226	5.730	5.730	5.730	5.730	5.730
0.05	1000	N	0.9466	0.9432	0.9484	0.9510	0.9480	0.9494	0.9464
0.05	1000		5.058	5.267	5.396	5.396	5.396	5.396	5.396
0.05	1000	A	0.9466	0.9436	0.9472	0.9510	0.9486	0.9472	0.9454
0.05	1000		5.100	5.336	5.429	5.429	5.429	5.429	5.429
0.5	100	N	0.5568	0.5414	0.5480	0.5436	0.5472	0.5580	0.5554
0.5	100		1.704	1.783	1.838	1.838	1.838	1.838	1.838
0.5	100	A	0.4928	0.4920	0.4910	0.5018	0.4990	0.4986	0.5044
0.5	100		1.518	1.621	1.678	1.678	1.678	1.678	1.678
0.5	1000	N	0.5690	0.5658	0.5514	0.5632	0.5586	0.5574	0.5572
0.5	1000		1.739	1.890	1.855	1.855	1.855	1.855	1.855
0.5	1000	A	0.4968	0.4968	0.4868	0.4940	0.4924	0.4950	0.4878
0.5	1000		1.443	1.533	1.591	1.591	1.591	1.591	1.591

Table 3: Uniform Errors

α	n	PI	j=1	j=2	j=3	j=4	j=5	j=6	j=7
0.05	100	N	0.9904	0.9796	0.9820	0.9794	0.9780	0.9818	0.9800
0.05	100		2.254	2.359	2.433	2.433	2.433	2.433	2.433
0.05	100	A	0.9816	0.9756	0.9756	0.9702	0.9730	0.9776	0.9754
0.05	100		2.132	2.342	2.388	2.388	2.388	2.388	2.388
0.05	1000	N	1.0000	0.9898	0.9826	0.9830	0.9834	0.9822	0.9844
0.05	1000		2.263	2.357	2.416	2.416	2.416	2.416	2.416
0.05	1000	A	0.9548	0.9486	0.9494	0.9512	0.9514	0.9506	0.9478
0.05	1000		1.913	2.094	2.182	2.182	2.182	2.182	2.182
0.5	100	N	0.3860	0.4096	0.4088	0.4136	0.4178	0.4078	0.4258
0.5	100		0.770	0.805	0.831	0.831	0.831	0.831	0.831
0.5	100	A	0.4644	0.4698	0.4780	0.4798	0.4738	0.4916	0.4886
0.5	100		0.954	0.984	1.004	1.004	1.004	1.004	1.004
0.5	1000	N	0.3996	0.4094	0.4138	0.4164	0.4094	0.4146	0.4260
0.5	1000		0.778	0.810	0.829	0.829	0.829	0.829	0.829
0.5	1000	A	0.4894	0.4904	0.4868	0.4940	0.4924	0.4950	0.4878
0.5	1000		0.963	0.974	0.982	0.982	0.982	0.982	0.982

Table 4: EXP(1) - 1 Errors

α	n	PI	j=1	j=2	j=3	j=4	j=5	j=6	j=7
0.05	100	N	0.9348	0.9470	0.9392	0.9448	0.9426	0.9432	0.9432
0.05	100		3.861	4.043	4.164	4.164	4.164	4.164	4.164
0.05	100	A	0.9430	0.9580	0.9472	0.9484	0.9508	0.9512	0.9458
0.05	100		3.485	4.075	4.041	4.041	4.041	4.041	4.041
0.05	1000	N	0.9466	0.9504	0.9464	0.9438	0.9480	0.9442	0.9432
0.05	1000		3.914	4.078	4.178	4.178	4.178	4.178	4.178
0.05	1000	A	0.9490	0.9544	0.9452	0.9462	0.9466	0.9444	0.9430
0.05	1000		3.092	3.554	3.773	3.773	3.773	3.773	3.773
0.5	100	N	0.5204	0.5390	0.5480	0.5494	0.5512	0.5412	0.5430
0.5	100		1.322	1.385	1.428	1.428	1.428	1.428	1.428
0.5	100	A	0.4924	0.5078	0.5064	0.4930	0.5080	0.5070	0.4944
0.5	100		0.850	0.963	1.031	1.031	1.031	1.031	1.031
0.5	1000	N	0.5238	0.5434	0.5512	0.5686	0.5586	0.5504	0.5522
0.5	1000		1.346	1.401	1.436	1.436	1.436	1.436	1.436
0.5	1000	A	0.4956	0.4876	0.4908	0.4966	0.5012	0.4950	0.4944
0.5	1000		0.725	0.882	0.964	0.964	0.964	0.964	0.964

These two types of PIs are denoted by N and A respectively in the tables. The simulated coverages and average lengths of the PI are shown.

With 5000 runs, coverages between 0.94 and 0.96 suggest that there is no reason to believe that the nominal coverage is not 0.95, while coverages between 0.48 and 0.52 suggest that there is no reason to believe that the nominal coverage is not 0.5.

From table 1 for normal errors, note that for $n = 1000$, the coverages and lengths of PIs (3) and (4) were very similar to the those of PI (2). PIs (3) and (4) were longer than the normal PI (2) for $n = 100$ and normal errors. From table 2 for t_5 errors, the 95% normal PI (2) worked well, but the nominal 50% normal PI (2) had coverage that was too high and the average lengths were too large. The alternative PIs had coverage near 50% with shorter average lengths. From table 3 for uniform errors, the normal PIs (2) were too long and the coverage was too high for 95% PIs. The alternative PIs (3) and (4) had coverage closer to the nominal level with good coverage for $n = 1000$. From table 4 with EXP(1) - 1 errors, for 95% PIs the normal PIs (2) were longer than the alternative PIs (3) and (4). For the 50% PIs, the normal PIs (2) were too long with coverage that was too high. The alternative PIs (3) and (4) were shorter with good coverage.

4. Outlier Detection

Oultiers are cases that lie far away from the pattern set by the bulk of the data, and can be detected from the plot of t versus Y_t and from the response plot of $\hat{Y}_t$ versus Y_t with the identity line that has zero intercept and unit slope added as a visual aid. In both plots Y_t is on the vertical axis, and the vertical deviations of Y_t from the identity line are the residuals $\hat{e}_t = Y_t - \hat{Y}_t$. The residual plot of $\hat{Y}_t$ versus $\hat{e}_t$ is also useful.

Suppose equations $Y_t = (1, \boldsymbol{x}_t^T)\boldsymbol{\beta} + e_t$ can be put in matrix form $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{X}$ is of full rank with more rows than columns $p+1$ and $\boldsymbol{\beta} = (\phi_0, \boldsymbol{\phi}^T)^T = (\phi_0, \phi_1, ..., \phi_p)^T$. Then the least squares estimator $\hat{\boldsymbol{\beta}}_{LS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$, $\hat{\phi}_{0,LS} = \overline{Y} - \hat{\boldsymbol{\phi}}_{LS}^T\overline{\boldsymbol{x}}$, and $\hat{\boldsymbol{\phi}}_{LS} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x},Y}$. The population parameters are $\phi_0 = E(Y) - \boldsymbol{\phi}_{LS}^T E(\boldsymbol{x})$ and $\boldsymbol{\phi} = \boldsymbol{\phi}_{LS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x},Y}$. The stationary AR(p) model can be put in this form, with $\boldsymbol{x}_t = (Y_{t-1}, ..., Y_{t-p})^T$ and $Y = Y_t$. Here $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x},Y}$ are the usual estimated covariance matrices used when $\boldsymbol{w}_i = (\boldsymbol{x}_i, Y_i)^T$ are iid from some population. Write the AR(p) equations $Y_t = \phi_0 + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + e_t$ in matrix form $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ or

$$\begin{bmatrix} Y_{p+1} \\ Y_{p+2} \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & Y_p & Y_{p-1} & \dots & Y_1 \\ 1 & Y_{p+1} & Y_p & \dots & Y_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & Y_{n-1} & Y_{n-2} & \dots & Y_{n-p} \end{bmatrix} \begin{bmatrix} \phi_0 \\ \phi_1 \\ \vdots \\ \phi_p \end{bmatrix} + \begin{bmatrix} e_{p+1} \\ e_{p+2} \\ \vdots \\ e_n \end{bmatrix}.$$

Under mild conditions on the white noise $\{e_t\}$ with zero mean and variance σ_e^2 (normality is not needed), $\sqrt{n}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma_e^2\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1})$, and $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_{p+1}(\boldsymbol{0}, \sigma_e^2\boldsymbol{C})$ where $\lim_{n\to\infty} \dfrac{\boldsymbol{X}^T\boldsymbol{X}}{n} \to \boldsymbol{C}^{-1}$. Hence $\hat{\boldsymbol{\beta}} \approx N_{p+1}(\boldsymbol{\beta}, \hat{\sigma}_e^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$. So tests from ordinary multiple linear regression can be applied to AR(p) time series, and $SE(\hat{\boldsymbol{\beta}}_i) = \sqrt{\hat{\sigma}_e^2(\boldsymbol{X}^T\boldsymbol{X})_{ii}^{-1}}$.

A robust estimator for ARIMA$(p, d, 0)$ data can be created by plugging in a robust estimator of multivariate location and dispersion. Let

$$\boldsymbol{w} = \begin{pmatrix} \boldsymbol{x} \\ Y \end{pmatrix}, \quad E(\boldsymbol{w}) = \boldsymbol{\mu}_{\boldsymbol{w}} = \begin{pmatrix} E(\boldsymbol{x}) \\ E(Y) \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_{\boldsymbol{x}} \\ \mu_Y \end{pmatrix}, \quad \text{and} \quad \text{Cov}(\boldsymbol{w}) = \boldsymbol{\Sigma}_{\boldsymbol{w}} =$$

$$\begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{x},\boldsymbol{x}} & \boldsymbol{\Sigma}_{\boldsymbol{x},Y} \\ \boldsymbol{\Sigma}_{Y,\boldsymbol{x}} & \boldsymbol{\Sigma}_{Y,Y} \end{pmatrix}.$$

Let $(T, \boldsymbol{C}) = (\tilde{\boldsymbol{\mu}}_{\boldsymbol{w}}, \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{w}})$ be a robust estimator of multivariate location and dispersion. Then the robust plug in estimator $\tilde{\phi}_0 = \tilde{\mu}_Y - \tilde{\boldsymbol{\phi}}^T \tilde{\boldsymbol{\mu}}_{\boldsymbol{x}}$ and $\tilde{\boldsymbol{\phi}} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x},Y}$. The robust estimator $(T, \boldsymbol{C})$ used will be the RMVN estimator of Olive and Hawkins (2010) and Zhang, Olive and Ye (2012) that has been used to make robust estimators of multiple linear regression and multivariate linear regression. See Olive (2013c). The robust AR(p) estimator is not yet backed by theory and should be used as an outlier diagnostic.

Example 2. Here we examine outliers for the AR(p) model and use the Cryer and Chan (2008) *R* package `TSA` data set `deere1` which gives 82 consecutive values for the amount of deviation from a specified target value in an industrial machining process at Deere & Co. If there is an outlier at Y_k where k is not too close to 1 or n, then fitted values will use the outlier for $t = k + 1, ..., k + p$. So the outlier appears $p + 1$ times in the equations for the AR(p) model.

An AR(2) model will be used for the Deere time series, and the plot of the time series in Figure 2 shows that there is one large outlier. Figure 3 shows the response and residual plots for the AR(2) model. Only one outlier, instead of two, appears in the fitted values since $\hat{\phi}_1 = 0.027$ is quite small. The plots for the robust fit are similar and are not shown.

The outlier Y_{27} is changed from 30 to a more reasonable value 8 to create "cleaned data." The robust AR(2) model was refit using the cleaned data resulting in "cleaned fitted values." In the original data, cases Y_7 and Y_{76} were changed to 25 and 26. The fitted values from the robust AR(2) models versus the cleaned fitted values showed some tilt. Next cases Y_7 and Y_{76} were changed to 250 and 260. Figure 4 shows fitted values from the robust AR(2) models versus the cleaned fitted values with the identity line added as a visual aid. The two sets of fitted values for the bulk of the data are similar since big

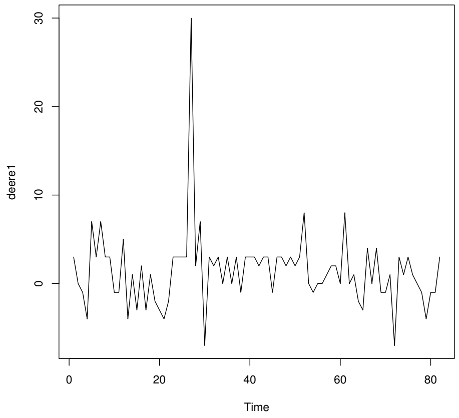

Figure 2: The Deere Time Series Has One Outlier

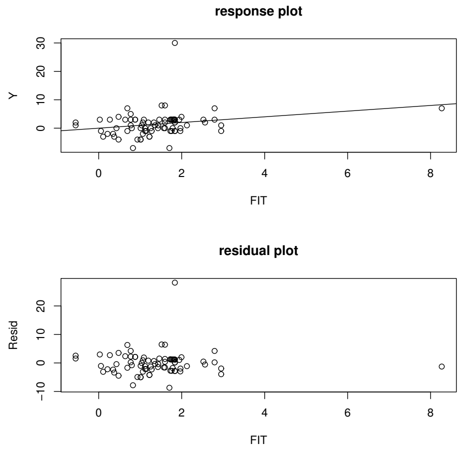

Figure 3: Response and Residual Plots for the AR(2) Model

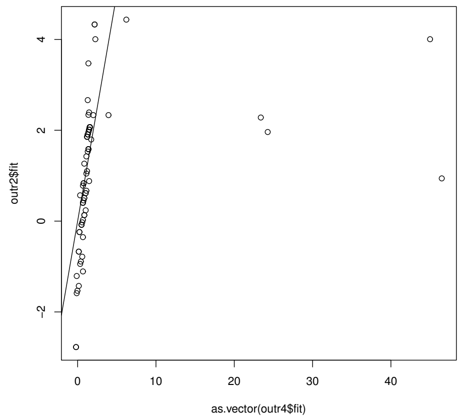

Figure 4: Fitted Values from the Cleaned Data Versus Robust Fitted Values from the Data with 3 Outliers

outliers are easier to detect.

5. Discussion

The aicmatrix is also somewhat useful for GARCH models. The aicmatrix is made in one of the R time series help files. Using I_I and submodels helps to quickly find a small number of good models to examine.

Plots and simulations were done in R. See R Development Core Team (2011). Programs are in the collection of functions *tspack.txt*. See (http://lagrange.math.siu.edu/Olive/tspack.txt). The function `aicmat` makes the aicmatrix for ARIMA(p, d, q) models with d fixed while the function `saics` makes the aicmatrix for ARIMA$(p, d, q) \times (P, D, Q)_s$ models with d, P, D, Q and s fixed. The function `pimasim` was used to simulate the prediction intervals. The function `robar` fits a robust AR(p) model.

REFERENCES

Box, G., Jenkins, G.M., and Reinsel, G. (1994), *Time Series Analysis: Forecasting and Control*, 3rd ed., Prentice Hall, Englewood Cliffs, NJ.

Brockwell, P.J., and Davis, R.A. (1987), *Time Series: Theory and Methods*, Springer, New York, NY.

Burnham, K.P., and Anderson, D.R. (2004), "Multimodel Inference Understanding AIC and BIC in Model Selection," *Sociological Methods & Research*, 33, 261-304.

Clements, M.P., and Kim, N. (2007), "Bootstrapping Prediction Intervals for Autoregressive Time Series," *Computational Statistics & Data Analysis*, 51, 3580-3594.

Cryer, J.D., and Chan, K.-S. (2008), *Time Series Analysis: with Applications in R*, 2nd ed., Springer, New York, NY.

Duong, Q.P. (1984), "On the Choice of the Order of Autoregressive Models: a Ranking and Selection Approach," *Journal of Time Series Analysis,* 5, 145-157.

Frey, J. (2013), "Data-Driven Nonparametric Prediction Intervals," *Journal of Statistical Planning and Inference*, 143, 1039-1048.

Kabaila, P., and He, Z. (2007), "Improved Prediction Limits for AR(p) and ARC(p) Processes," *Journal of Time Series Analysis*, 29, 213-223.

Masters, T. (1995), *Neural, Novel, & Hybrid Algorithms for Time Series Prediction*, Wiley, New York, NY.

Olive, D.J. (2007), "Prediction Intervals for Regression Models," *Computational Statistics & Data Analysis,* 51, 3115-3122.

Olive, D.J. (2008), *Applied Robust Statistics,* Unpublished Online Text, see (http://lagrange.math.siu.edu/Olive/ol-bookp.htm).

Olive, D.J. (2013a), "Asymptotically Optimal Regression Prediction Intervals and Prediction Regions for Multivariate Data," *International Journal of Statistics and Probability*, 2, 90-100.

Olive, D.J. (2013b), "Plots, Prediction and Testing in the Multivariate Linear Model," unpublished manuscript, (http://lagrange.math.siu.edu/Olive/ppmultreg.pdf).

Olive, D.J. (2013c), *Robust Multivariate Linear Regression*, unpublished manuscript, (http://lagrange.math.siu.edu/Olive/pprobmreg.pdf).

Olive, D. J., and Hawkins, D. M. (2005), "Variable Selection for 1D Regression Models," *Technometrics*, 47, 43-50.

Olive, D.J., and Hawkins, D.M. (2010), "Robust Multivariate Location and Dispersion," Preprint, see (http://lagrange.math.siu.edu/Olive/pphbmld.pdf).

Panichkitkosolkul, W., and Niwitpong, S.-A. (2012), "Prediction Intervals for the Gaussian Autoregressive Processes Following the Unit Root Tests," *Model Assisted Statistics and Applications*, 7, 1-15.

R Development Core Team (2011), "R: a Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna, Austria, (www.R-project.org).

Thombs, L.A. and Schucany, W.R. (1990), "Bootstrap Prediction Intervals for Autoregression," *Journal of the American Statistical Association*, 85, 486-492.

Vidoni, P. (2009), "A Simple Procedure for Computing Improved Prediction Intervals for Autoregressive Models," *Journal of Time Series Analysis*, 30, 577-590.

Zhang, J., Olive, D.J., and Ye, P. (2012), "Robust Covariance Matrix Estimation With Canonical Correlation Analysis," *International Journal of Statistics and Probability*, 1, 119-136.