

**Lecture 13: Diagnostic Checking**  
Bus 41910, Time Series Analysis, Mr. R. Tsay

We now discuss the last stage of Box-Jenkins iterative modeling procedure, model checking or diagnostic checking. The objective of this stage is two-fold. First, it checks for possible discrepancy of an entertained model. Secondly, it corrects any model discrepancy if necessary.

Model checking in time series analysis is very much the same as that in the traditional regression analysis. It emphasizes on residual analysis. Basically, one typically considers the following statistics:

- Residual plots
- Residual serial correlations
- Outlier detection.

For an entertained ARMA( $p, q$ ) model (or ARIMA( $p, d, q$ ) model), the residual series  $\hat{a}_t$  is defined by

$$\hat{a}_t = Z_t - \hat{\phi}_1 Z_{t-1} - \cdots - \hat{\phi}_p Z_{t-p} + \hat{\theta}_1 \hat{a}_{t-1} + \cdots + \hat{\theta}_q \hat{a}_{t-q}$$

where the mean of  $Z_t$  is assumed to be zero,  $\hat{\phi}$ 's and  $\hat{\theta}$ 's are MLE of  $\phi$ 's and  $\theta$ 's, respectively, and the starting residuals  $\hat{a}_t$  for  $t \leq 0$  are defined by the method of estimation used. For instance,  $\hat{a}_t = 0$  for  $t \leq 0$  under conditional MLE.

There are various ways to plot the residuals. For instance, the time plot is useful in spotting possible serial correlation, non-constant variance, and outliers and normal-score plot (or histogram) is used to check the normality assumption. Residual plots are integral parts of time series analysis.

To check for residual serial correlation, one can use residual ACF and PACF. For an adequate model, the residuals should behave as white noise. The residual ACF at lag- $\ell$  is defined by

$$\hat{r}_\ell = \frac{\sum_{t=\ell+1}^n (\hat{a}_t - \bar{a})(\hat{a}_{t-\ell} - \bar{a})}{\sum_{t=1}^n (\hat{a}_t - \bar{a})^2}$$

where  $\bar{a}$  is the average of the residuals  $\hat{a}_t$ 's. For a white noise series, the asymptotic variance of the sample ACF is  $\frac{1}{n}$ . Therefore, one can use  $\frac{1}{\sqrt{n}}$  as an approximate standard error to check the significance of individual  $\hat{r}_\ell$ . More efficiently, one may wish to check simultaneously that several residual ACFs are not significantly different from zero, that is to say, in hypothesis testing, to test the null hypothesis

$$H_0 : r_1 = r_2 = \cdots = r_m = 0$$

where  $r_i$  is the theoretical ACF of  $a_t$  at lag  $i$ . To this end, Ljung and Box (1978, BKA) propose a modified Box-Pierce statistic:

$$Q(m) = n(n+2) \sum_{v=1}^m \frac{\hat{r}_v^2}{n-v}$$

which, under the null hypothesis that the entertained ARMA( $p, q$ ) model is the “true” model, is asymptotically  $\chi_{m-p-q}^2$ .

**Remarks:** The Ljung-Box  $Q(m)$  statistic is asymptotically equivalent to testing for an ARMA( $p, q$ ) model against an ARMA( $p, q+m$ ) or ARMA( $p+m, q$ ) model. This  $Q$  statistic is commonly used. However, it is not a powerful statistic in detecting residual series correlation and its power depends on the choice of  $m$ . In the literature, simulation results by Newbold and his associates suggest that  $m = O(\ln n)$  is close to an optimal choice.

Next, we consider outlier detection. This problem has gained much attention in the 1980s and various methods are available. In time series analysis, outliers can cause biases in parameter estimation as well as model misspecification, resulting in misleading conclusion. For this reason, several outlier detection and robust estimation procedures have been proposed in the literature for time series analysis. Here we shall consider a simple regression approach which is closely related to the “Intervention Analysis” of Box and Tiao (1975, JASA).

Following Fox (1974), there are basically two types of outliers in a time series. The first type of outlier is called the “additive outlier (AO)” which represents that a disturbance is committed to a particular observation. Mathematically, the observed time series is

$$Y_t = Z_t + \omega_a I_t^{(d)}$$

where  $Z_t$  is an outlier-free time series,  $\omega_a$  denotes the magnitude of the disturbance and  $I_t^{(d)}$  is an indicator variable defined by

$$I_t^{(d)} = \begin{cases} 1 & \text{if } t = d \\ 0 & \text{if } t \neq d \end{cases}$$

In other words, for an AO model

$$Y_t = Z_t \quad \text{if } t \neq d \quad \text{and} \quad Y_d = Z_d + \omega_a.$$

A typical example of an AO is a typo or a recording error.

Another type of outlier is called an “innovational outlier (IO)”, which is a disturbance in the innovational series  $\{a_t\}$  and may affect every subsequent observation of the series. Mathematically, an IO model is

$$Y_t = \frac{\theta(B)}{\phi(B)}(a_t + \omega_v I_t^{(d)})$$

where  $I_t^{(d)}$  is defined as before and  $\omega_v$  denotes the magnitude of the disturbance. Rewriting the model as

$$Y_t = Z_t + \frac{\theta(B)}{\phi(B)} \omega_v I_t^{(d)}$$

we see that an IO affects the series through its own dynamic  $\frac{\theta(B)}{\phi(B)}$  and, in effect, becomes part of the system thereafter. In practice, an IO often indicates an onset of certain changes in the system. For instance, in a manufacturing process, changing an operator or a measurement instrument may result in an IO.

Of course, many other types of disturbance can happen to a time series. The AO and IO models only two of many possibilities. In Chen and Tiao (1990) and Tsay (1988), two types of disturbances were introduced. They are the level shift and temporary change in level. Mathematically, a level shift (LS) can be described by

$$Y_t = Z_t + \frac{\omega_s}{(1-B)} I_t^{(d)}$$

where  $\omega_s$  is the amount of shift in the level of  $Z_t$ . Writing

$$\frac{1}{(1-B)} = 1 + B + B^2 + \dots$$

we see that for the above model

$$Y_t = \begin{cases} Z_t & \text{for } t < d \\ Z_t + \omega_s & \text{for } t \geq d. \end{cases}$$

Thus, the fixed constant  $\omega_s$  is added to every observation one or after  $d$ . Such a level shift is permanent.

In some cases, the effect of a level shift is only temporary. A mathematical model which is capable of describing such a shift is

$$Y_t = Z_t + \frac{\omega_c}{(1-\delta B)} I_t^{(d)}, \quad 0 < \delta < 1.$$

Since

$$\frac{1}{1-\delta B} = 1 + \delta B + \delta^2 B^2 + \delta^3 B^3 + \dots$$

the magnitudes of level shift at times  $d, d+1, d+2, \dots$  are  $\omega_c, \delta\omega_c, \delta^2\omega_c, \dots$ . Thus, the initial shift is  $\omega_c$  and the subsequent shifts are discounted at the rate  $\delta$ . With  $0 < \delta < 1$ , the shift decays exponentially to zero. We refer to such a temporary level shift as a transient change (TC) model. In practice, the value of  $\delta$  is a prespecified constant. It may assume the value of 0.8 or 0.7.

Outlier Detection. In practice, outliers can occur at any time point in a series. Thus, to detect an outlier, we need to estimate the parameters  $\omega_a, \omega_v, \omega_s, \omega_c$  and check the significance of these estimates.

For simplicity, we assume that the time series parameters are known. In practice, the parameters need to be estimated and we employ an iterative procedure to detect outliers. The four outlier models discussed above can be put in the general form

$$Y_t = Z_t + \omega_0 \frac{\omega(B)}{\delta(B)} I_t^{(d)}$$

where

$$\omega_0 = \begin{cases} \omega_a & \text{AO case} \\ \omega_v & \text{IO case} \\ \omega_s & \text{LS case} \\ \omega_c & \text{TC case} \end{cases} \quad \frac{\omega(B)}{\delta(B)} = \begin{cases} 1 & \text{AO case} \\ \frac{\theta(B)}{\phi(B)} & \text{IO case} \\ \frac{1}{1-B} & \text{LS case} \\ \frac{1}{1-\delta B} & \text{TC case.} \end{cases}$$

Given  $\theta(B)$  and  $\phi(B)$ , define

$$y_t = \frac{\phi(B)}{\theta(B)} Y_t, x_t = \frac{\phi(B)}{\theta(B)} \frac{\omega(B)}{\delta(B)} I_t^{(d)}.$$

Then, we have

$$y_t = \omega_0 x_t + a_t$$

which is precisely a simple linear regression equation. Therefore,

$$\hat{\omega}_0 = \frac{\sum_{t=1}^n y_t x_t}{\sum_{t=1}^n x_t^2} \quad \text{and} \quad \text{Var}(\hat{\omega}_0) = \frac{\sigma_a^2}{\sum_{t=1}^n x_t^2},$$

where  $n$  is the sample size. Using this simple technique, we obtain

- IO case:  $\hat{\omega}_{v,d} = y_d$  and  $\text{Var}(\hat{\omega}_{v,d}) = \sigma_a^2$ .
- AO case:  $\hat{\omega}_{a,d} = \rho_{a,d}^2 (y_d - \sum_{i=1}^{n-d} \pi_i y_{d+i})$  and  $\text{Var}(\hat{\omega}_{a,d}) = \rho_{a,d}^2 \sigma_a^2$  where  $\pi$ 's are the  $\pi$ -weights of  $Z_t$  and  $\rho_{a,d}^2 = (1 + \pi_1^2 + \dots + \pi_{n-d}^2)^{-1}$ .
- LS case:  $\hat{\omega}_{s,d} = \rho_{s,d}^2 (y_d - \sum_{i=1}^{n-d} \eta_i y_{d+i})$  and  $\text{Var}(\hat{\omega}_{s,d}) = \rho_{s,d}^2 \sigma_a^2$  where  $\eta_i$ 's are the coefficient of  $B^i$  in the polynomial  $\eta(B) = \eta_0 - \eta_1 B - \eta_2 B^2 - \dots = \frac{\pi(B)}{1-B}$  and  $\rho_{s,d}^2 = (1 + \eta_1^2 + \dots + \eta_{n-d}^2)^{-1}$ .
- TC case:  $\hat{\omega}_{c,d} = \rho_{c,d}^2 (y_d - \sum_{i=1}^{n-d} \beta_i y_{d+i})$  and  $\text{Var}(\hat{\omega}_{c,d}) = \rho_{c,d}^2 \sigma_a^2$  where  $\beta_i$  is the coefficient of  $B^i$  in the polynomial  $\beta(B) = \beta_0 - \beta_1 B - \dots = \frac{\pi(B)}{1-\delta B}$  and  $\rho_{c,d}^2 = (1 + \beta_1^2 + \dots + \beta_{n-d}^2)^{-1}$ .

Based on the above results, we may employ the test statistics

- Existence of an IO at  $d$ :  $\lambda_{v,d} = \frac{\hat{\omega}_{v,d}}{\sigma_a}$

- Existence of an AO at  $d$ :  $\lambda_{a,d} = \frac{\hat{\omega}_{a,d}}{\rho_{a,d}\sigma_a}$
- Existence of a LS at  $d$ :  $\lambda_{s,d} = \frac{\hat{\omega}_{s,d}}{\rho_{s,d}\sigma_a}$
- Existence of an TC at  $d$ :  $\lambda_{c,d} = \frac{\hat{\omega}_{c,d}}{\rho_{c,d}\sigma_a}$ .

Under the null hypothesis of normality, no disturbance at  $d$  and knowing the time series parameters and  $d$ , all of the above four statistics are distributed as  $N(0, 1)$ . In practice, the parameters can be replaced by the MLEs. However, since  $d$  is unknown, we need to apply the tests to all possible values of  $d$ . Consequently, in other words, we need to consider the maximum of test statistics over  $d$ . The resulting statistics are no longer normal. However, one can obtain certain percentiles via simulation or using distributions of certain extreme-value statistics. Experience suggests that using a critical value of 3.0 or 3.5 works reasonably well in practice.

We now consider an iterative procedure for time series analysis in the presence of outliers, level-shifts and temporary changes. See Tsay (1988, packet) and Chang, et al. (1988, Technometrics). The procedure considered here is the very basic one. Some variants can be used to improve the efficacy.

1. Identify an ARMA model for  $Y_t$ , estimate the associated parameters. (Here we pretend that there are no outliers in  $Y_t$ .)
2. Based on the model of step 1, compute the four test statistics for each time point and identify

$$\lambda_{v,max} = \max_d \{|\lambda_{v,d}|\}, \lambda_{a,max} = \max_d \{|\lambda_{a,d}|\}, \lambda_{s,max} = \max_d \{|\lambda_{s,d}|\}, \lambda_{c,max} = \max_d \{|\lambda_{c,d}|\},$$

and denote the time points of these maximum by  $d_v, d_a, d_s, d_c$ , respectively.

3. Let  $\lambda = \max\{\lambda_{v,max}, \lambda_{a,max}, \lambda_{s,max}, \lambda_{c,max}\}$  and compare  $\lambda$  with the pre-specified critical value  $C$ . If  $\lambda < C$ , there is no outlier and stop. If  $\lambda \geq C$ , continue to the next step.
4. Compute a modified series  $Y_t^*$  by removing the effect of the identified outlier and go to step 1 with  $Y_t$  replaced by  $Y_t^*$ .

Here we use “identification-detection-removing” cycle to remove the effect of outlier one-by-one.

Illustration: Air-passenger-miles data: Cryer’s book. The data consists of the logarithm of monthly air-passenger-miles (in millions) within the U.S. from January 1960 to December 1977 for 216 observations.

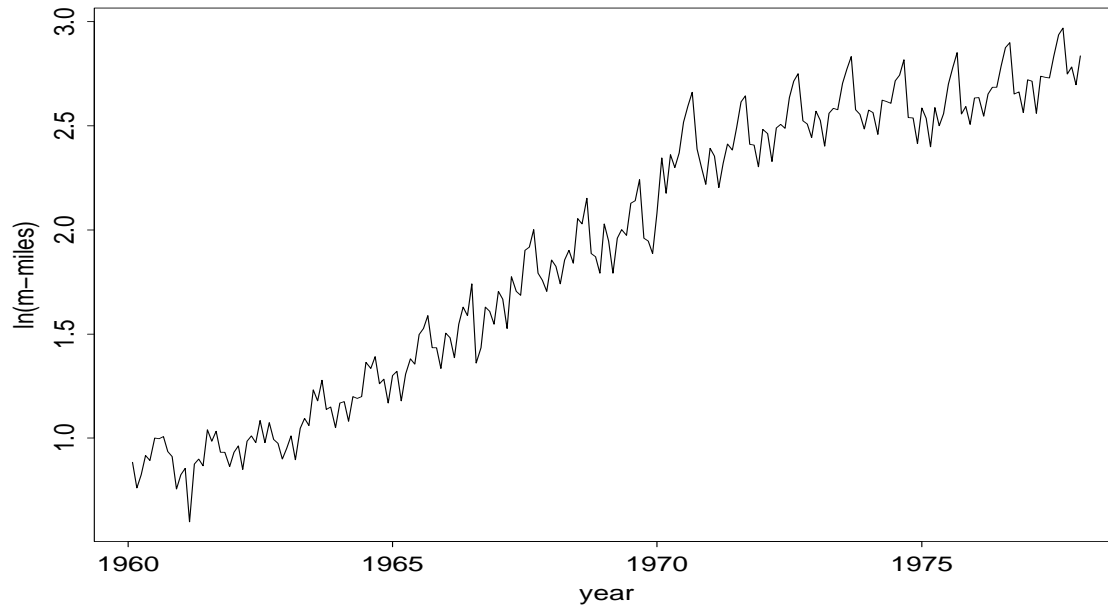


Figure 1: Time plot of the logarithm of monthly air-passenger-miles

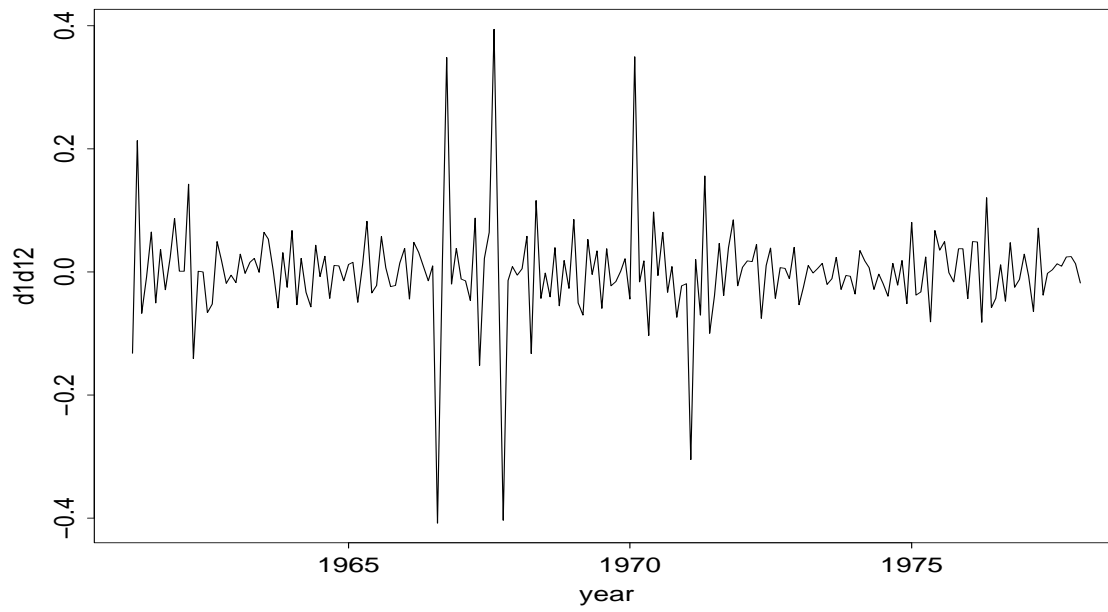


Figure 2: Time plot of the regular and seasonal differenced series

\*\*\* SCA demonstration of outlier detection \*\*\*

--

input x. file 'airpml.txt'

X , A 216 BY 1 VARIABLE, IS STORED IN THE WORKSPACE

--

acf x. dfor 1, 12.

	1	12
DIFFERENCE ORDERS. . . . .	(1-B )	(1-B )
NAME OF THE SERIES . . . . .		X
TIME PERIOD ANALYZED . . . . .	1 TO	216
MEAN OF THE (DIFFERENCED) SERIES . . .	0.0007	
STANDARD DEVIATION OF THE SERIES . . .	0.0805	
T-VALUE OF MEAN (AGAINST ZERO) . . . .	0.1257	

# AUTOCORRELATIONS

1- 12	-.18	-.16	-.07	-.01	-.00	.07	-.06	-.02	.06	.07	.07	-.51
ST.E.	.07	.07	.07	.07	.07	.07	.07	.07	.08	.08	.08	.08
Q	6.7	12.0	12.9	12.9	12.9	13.9	14.6	14.8	15.5	16.7	17.9	73.8
13- 24	.08	.12	-.01	-.03	.14	-.10	-.01	.02	.05	-.08	.06	.05
ST.E.	.09	.09	.09	.09	.09	.09	.09	.09	.09	.09	.09	.09
Q	75.2	78.2	78.2	78.4	83.1	85.4	85.5	85.6	86.1	87.7	88.6	89.2

-1.0 -0.8 -0.6 -0.4 -0.2 0.0 0.2 0.4 0.6 0.8 1.0  
+---+---+---+---+---+---+---+---+---+---+---+

		I	
1	-0.18	XX+XXI	+
2	-0.16	XXXXI	+
3	-0.07	+ XXI	+
4	-0.01	+ I	+
5	0.00	+ I	+
6	0.07	+ IXX	+
7	-0.06	+ XI	+
8	-0.02	+ XI	+
9	0.06	+ IXX	+
10	0.07	+ IXX	+
11	0.07	+ IXX	+
12	-0.51	XXXXXXXXX+XXXI	+
13	0.08	+ IXX	+
14	0.12	+ IXXX	+
15	-0.01	+ I	+

```

16  -0.03          +   XI   +
17   0.14          +  IXXXX+
18  -0.10          + XXXI   +
19  -0.01          +    I   +
20   0.02          +   IX   +
21   0.05          +   IX   +
22  -0.08          +  XXI   +
23   0.06          +   IXX  +
24   0.05          +   IX   +

```

--

```
tsm m1. model x(1,12)=(1,2)(12)noise.
```

--

```
estim m1. hold resi(r1).
```

THE FOLLOWING ANALYSIS IS BASED ON TIME SPAN 1 THRU 216

NONLINEAR ESTIMATION TERMINATED DUE TO:

RELATIVE CHANGE IN (OBJECTIVE FUNCTION)\*\*0.5 LESS THAN 0.1000D-02

SUMMARY FOR UNIVARIATE TIME SERIES MODEL -- M1

```

-----
VARIABLE   TYPE OF   ORIGINAL   DIFFERENCING
          VARIABLE OR CENTERED

```

```

          X      RANDOM   ORIGINAL          1      12
          (1-B ) (1-B )
-----

```

PARAMETER LABEL	VARIABLE NAME	NUM./ DENOM.	FACTOR	ORDER	CONS- TRAI NT	VALUE	STD ERROR	T VALUE
1	X	MA	1	1	NONE	.2863	.0689	4.16
2	X	MA	1	2	NONE	.2105	.0691	3.05
3	X	MA	2	12	NONE	.7119	.0499	14.26

EFFECTIVE NUMBER OF OBSERVATIONS . . . . . 203

R-SQUARE . . . . . 0.992

RESIDUAL STANDARD ERROR. . . . . 0.608897E-01

--

```
acf r1.
```

NAME OF THE SERIES . . . . . R1

TIME PERIOD ANALYZED . . . . . 14 TO 216

MEAN OF THE (DIFFERENCED) SERIES . . . . . 0.0038



STANDARD DEVIATION OF THE SERIES . . .	0.0608
T-VALUE OF MEAN (AGAINST ZERO) . . . .	0.8833

# AUTOCORRELATIONS

1- 12	.00	.01	-.04	-.04	.04	.08	-.02	-.04	.04	-.06	-.00	-.05
ST.E.	.07	.07	.07	.07	.07	.07	.07	.07	.07	.07	.07	.07
Q	.0	.0	.3	.7	1.0	2.3	2.4	2.8	3.2	3.9	3.9	4.6
13- 24	-.03	.03	.01	-.03	.13	-.02	.04	.01	.06	-.06	.08	.04
ST.E.	.07	.07	.07	.07	.07	.07	.07	.07	.07	.07	.07	.07
Q	4.8	5.1	5.1	5.3	9.1	9.2	9.5	9.5	10.3	11.3	12.7	13.0

-1.0	-0.8	-0.6	-0.4	-0.2	0.0	0.2	0.4	0.6	0.8	1.0
+	+	+	+	+	+	+	+	+	+	+

			I
1	0.00		+ I +
2	0.01		+ I +
3	-0.04		+ XI +
4	-0.04		+ XI +
5	0.04		+ IX +
6	0.08		+ IXX+
7	-0.02		+ I +
8	-0.04		+ XI +
9	0.04		+ IX +
10	-0.06		+ XI +
11	0.00		+ I +
12	-0.05		+ XI +
13	-0.03		+ XI +
14	0.03		+ IX +
15	0.01		+ I +
16	-0.03		+ XI +
17	0.13		+ IXXX+
18	-0.02		+ I +
19	0.04		+ IX +
20	0.01		+ I +
21	0.06		+ IXX +
22	-0.06		+ XXI +
23	0.08		+ IXX +
24	0.04		+ IX +

--

outlier m1. type ao,io,ls,tc.

INITIAL RESIDUAL STANDARD ERROR = 0.58670E-01

TIME	ESTIMATE	T-VALUE	TYPE
79	-0.37	-7.87	TC
121	0.30	7.85	LS
81	0.25	6.97	TC
80	-0.12	-3.52	TC
124	-0.10	-3.42	AO
87	0.10	3.61	AO
184	-0.09	-3.49	AO
31	-0.10	-3.19	TC
130	-0.09	-3.15	LS

ADJUSTED RESIDUAL STANDARD ERROR = 0.32881E-01

--

oestim m1. method exact. new-series v1,v2. hold resi(r1).

THE FOLLOWING ANALYSIS IS BASED ON TIME SPAN 1 THRU 216

SUMMARY FOR UNIVARIATE TIME SERIES MODEL -- M1

-----								
VARIABLE	TYPE OF	ORIGINAL		DIFFERENCING				
	VARIABLE	OR CENTERED						
				1	12			
X	RANDOM	ORIGINAL		(1-B )	(1-B )			
-----								
PARAMETER	VARIABLE	NUM./	FACTOR	ORDER	CONS-	VALUE	STD	T
LABEL	NAME	DENOM.			TRAINT		ERROR	VALUE
1	X	MA	1	1	NONE	.5622	.0735	7.65
2	X	MA	1	2	NONE	-.0876	.0732	-1.20
3	X	MA	2	12	NONE	.4810	.0606	7.94

SUMMARY OF OUTLIER DETECTION AND ADJUSTMENT

TIME	ESTIMATE	T-VALUE	TYPE
-----			
31	-0.097	-3.83	TC
79	-0.392	-14.46	TC
80	-0.119	-4.14	AO
81	0.182	6.68	TC

87	0.091	3.64	AO
121	0.310	12.68	LS
124	-0.099	-3.99	AO
130	-0.091	-3.71	LS
184	-0.086	-3.45	AO

```

-----
TOTAL NUMBER OF OBSERVATIONS. . . . . 216
EFFECTIVE NUMBER OF OBSERVATIONS. . . . . 203
RESIDUAL STANDARD ERROR (WITHOUT OUTLIER ADJUSTMENT). . 0.687643E-01
RESIDUAL STANDARD ERROR (WITH OUTLIER ADJUSTMENT) . . . 0.332751E-01

```

--  
acf v1.

```

NAME OF THE SERIES . . . . . V1
TIME PERIOD ANALYZED . . . . . 14 TO 216
MEAN OF THE (DIFFERENCED) SERIES . . . 0.0016
STANDARD DEVIATION OF THE SERIES . . . 0.0313
T-VALUE OF MEAN (AGAINST ZERO) . . . . 0.7473

```

#### AUTOCORRELATIONS

1- 12	-.00	-.06	.05	.01	-.01	.07	.08	.10	-.06	-.08	.08	.00
ST.E.	.07	.07	.07	.07	.07	.07	.07	.07	.07	.07	.07	.07
Q	.0	.7	1.2	1.2	1.3	2.2	3.5	5.8	6.7	8.0	9.4	9.4

13- 24	-.05	-.02	-.06	.14	.09	.01	.01	-.01	.00	.02	.12	.08
ST.E.	.07	.07	.07	.07	.07	.08	.08	.08	.08	.08	.08	.08
Q	9.9	10.0	10.8	15.0	16.9	16.9	16.9	16.9	16.9	17.0	20.3	21.7

--  
acf r1.

```

NAME OF THE SERIES . . . . . R1
TIME PERIOD ANALYZED . . . . . 14 TO 216
MEAN OF THE (DIFFERENCED) SERIES . . . 0.0017
STANDARD DEVIATION OF THE SERIES . . . 0.0663
T-VALUE OF MEAN (AGAINST ZERO) . . . . 0.3551

```

#### AUTOCORRELATIONS

1- 12	.25	-.15	-.18	-.10	.02	.08	-.03	-.03	.05	-.00	-.08	-.26
ST.E.	.07	.07	.08	.08	.08	.08	.08	.08	.08	.08	.08	.08

Q	13.1	17.8	24.7	26.6	26.7	28.0	28.2	28.5	28.9	28.9	30.3	45.1
13- 24	-.08	.04	.01	.01	.11	-.00	-.01	.02	.04	-.02	.05	-.03
ST.E.	.08	.08	.08	.08	.08	.09	.09	.09	.09	.09	.09	.09
Q	46.6	47.0	47.0	47.1	49.9	49.9	49.9	50.0	50.4	50.5	51.0	51.2

-1.0 -0.8 -0.6 -0.4 -0.2 0.0 0.2 0.4 0.6 0.8 1.0  
+-----+-----+-----+-----+-----+-----+-----+-----+-----+

		I	
1	0.25	+ IXX+XXX	
2	-0.15	XXXXI	+
3	-0.18	X+XXXI	+
4	-0.10	+ XXI	+
5	0.02	+ IX	+
6	0.08	+ IXX	+
7	-0.03	+ XI	+
8	-0.03	+ XI	+
9	0.05	+ IX	+
10	0.00	+ I	+
11	-0.08	+ XXI	+
12	-0.26	XXX+XXXI	+
13	-0.08	+ XXI	+
14	0.04	+ IX	+
15	0.01	+ I	+
16	0.01	+ I	+
17	0.11	+ IXXX+	
18	0.00	+ I	+
19	-0.01	+ I	+
20	0.02	+ IX	+
21	0.04	+ IX	+
22	-0.02	+ I	+
23	0.05	+ IX	+
24	-0.03	+ XI	+

--