International Journal of Advanced Scientific Technologies ,Engineering and Management Sciences (IJASTEMS-ISSN: 2454-356X) Volume.3,Special Issue.1,March.2017

OUT LIEARS AND THEIR DETECTION

B.SAROJAMMA[1*], B.HARI MALLIKARJUNA REDDY[2],B.VENKATESU [3], A.SRINIVASULU[4], S.C.THASLEEMA[5]

1,2,3,4,5: Department of statistics, s.v.university, Tirupati-517502.

****author for correspondence: saroja14397@gmail.com***

ABSTRACT: In time series and analysis it is assumed that the data or observations consist of a systematic pattern and stochastic component time series analysis widely used in atmospheric sciences, metrological sciences, business sciences etc. outlines play vital role in fitting and forecasting for data .in this paper outlines are detected using ACF and PACF, Kolmogrov smirnov testis used for goodness of fit.

KEYWORDS: ACF, PACF, Kolmogrov smirnov, ARIMA

I.INTRODUCTION

In the time series analysis the data or observations consist of a systematic pattern and stochastic component. Time series data are widely used in commercial, industrial, metrological, and sociological processes. In time series analysis, the major methods can be represented by a Univariate Box-Jenkins (1976) and ARIMA structure ("B-J model"). The basic univariate ARMA structure is used for model fitting with an intervention function to model outliers. A straight forward method for outlier detection in time series is based on forecasting.

An outlier can be defined as observation that deviates much from other observations. The identification and treatment of outliers constitute an important component of the time series analysis before modeling. Outlier detection is also known as anomaly detection and its important applications in a wide range of area in business, insurance, health care, security, engineering and many more.

An outlier in a set of data to be an observation which appears to be inconsistent with the remainder of that set of data. Outliers can again arise and cause difficulties.The outliers are values which seem either too large or too small as compared to rest of the observations (Gumbel, 1960).In another way, an outlying observation, or outlier is one that appears to deviate markedly from other members of the sample in which it occurs (Grubbs, 1969) outlier means an observation lies extreme to data. It may lie extreme upward or downward to the original data. A single outlier in a regression model can be detected by the effect of its detection on the residual sum of squares.Outliers could easily mislead the conventional time series analysis procedure resulting in erroneous conclusions.

Fox (1972) was first to consider outliers within time series, assuming an AR structure with Gaussian noise. Two broad categories of outlier are defined, one is "Additive Outliers (AO)" where a single point is affected and two "Innovative Outliers (IO)" where an innovation to the process affects both an observation and the subsequent series. He used likelihood ratio criteria for comparing the estimated error with the estimated standard error of that discrepancy.

Tsay (1988) was extended and generalized in detection of outliers, level shifts and variance changes in time series and he describes a method as a "Batch Procedure". he describes the different types of outliers as follows.

I. Additive Outlier (AO)

II. Innovation Outlier (IO)
III. Level Shift (LS)
IV. Permanent Level change (TC)
V. Transient Level Change (TC)
VI. Variance Change (VC)

The structure changes are allowed for the level ift (LS) and variance change (VC). Level shift is further classified as permanent level change (LC) and Transient Level change (TC).Parameter estimation is described for Additive Outliers (AO), Innovative Outliers (IO), Level Shift (LS) and Transient Level Shift (TC) using techniques of simple linear regression. Parameter estimation is slightly different for the VC model and the test statistic, follows an F-distribution is most powerful if the time point is known. i.Additive outliers (AO)

An outlier that affects a single observation. For example, data coding error might be identified as an additive outlier.

ii. Innovative outliers (IO)

An outlier that acts as an addition to the noise term at a particular series point. For stationary series, an innovative outlier affects several observations. For non-stationary series, it may affect every observation starting at a particular series point.

iii. Level Shift (LS)

An outlier that shifts all observations by a constant, starting at a particular series point. A level shift could result from a change in policy. An outlier whose impact decays exponentially to '0' is called Transient outliers (TO).

Outliers, level shifts, and variance changes are commonly used in applied time series analysis. The graphical representation of outlier detection method is Box-plot

II.DETECTION OF OUTLIERS

2.1. Detection of outliers for intraday data:

Outlier may occur due to mistyping of values, due to misunderstanding of observations, etc. Generally intraday contains numerous data introduced for analyses and forecasting model. We introduced a modified HWT (Holt Winter Taylor) model by adapting seasonal variations.

$$Y_t = S_T + W_T + R_T + \varepsilon_T$$

$$S_T = S_{T-1} + \varepsilon_{T-1}$$

$$W_T = W_{T-1} + \varepsilon_{T-2}$$

$$R_T = R_{T-1} + \varepsilon_{T-3}$$

where S_T is summer season

W_T is winter season

R_T is rainy season

$\varepsilon_1, \varepsilon_2, \varepsilon_3$ are error terms of summer, winter and rainy season respectively.

ACF (Auto correlation function) and PACF (Partial auto correlation functions)are used for detection of best model among different fitted $\varepsilon_1, \varepsilon_2, \varepsilon_3$ constants of modified HWT model.A traditional route of identification, estimation and diagnostic checking was followed with the ACF, PACF and Box and pierce Q-statistic being used to test for residual autocorrelations.

To determine a proper model for a given time series data, it is necessary to carry out the ACF and PACF analysis. These statistical measures reflect how the observations in a time series are related to each other. For modeling and forecasting purpose, it is often useful to plot the ACF and PACF against consecutive time lags. These plots help in determining the order of AR and MA terms i.e., *p* and *q*.

2.2. Auto Correlation Function (ACF):

The correlation between the error terms of current and past values is called *autocorrelation or serial correlation or lagged correlation*. The pattern of autocorrelation for lags 1,2, …., is known as the autocorrelation function or ACF.

The ACF plays a very important role in time series forecasting. The error term u_t at time period *t* is correlated with error terms u_{t+1}, u_{t+2}, ...and u_{t-1}, u_{t-2}, ... and so on.

The correlation between u_t and u_{t-k} is called an autocorrelation of order k.The *Autocorrelation Coefficient* at lag *k* is defined as:

The auto covariance at lag zero i.e. γ_0 is the variance of the time series. From the definition it is clear that the autocorrelation coefficient ρ_k is dimensionless and so $\rho_k = \frac{\gamma_k}{\gamma_0}$ is independent of the scale of measurement. Also, clearly $-1 \leq \rho_k \leq 1$

The correlation between u_t and u_{t-1} is the first-order autocorrelation and is usually denoted by ρ_1. The correlation between u_t and u_{t-2} is called the second-order autocorrelation and is denoted by ρ_2 and so on. There are *(n-1)* such autocorrelations if we have n observations. We denote the autocorrelation at lag k by r_k. The time series plot, the lagged scatterplot, and the autocorrelation function are three tools for assessing the autocorrelation of a time series.

A plot of the ACF against the lag is known as a *correlogram* and the ACF plot is a standard tool in time series before forecasting. It provides the useful check for seasonality, cycles, and other time series pattern.

2.3. Partial Auto Correlation Function (PACF):

The partial autocorrelation is used to identify the extent of relationship between current values of a variable with earlier values of that same variable with various time lags while holding the effects of all other time lags constant. Thus, it is completely analogous to partial correlation but refers to a single variable.

Partial autocorrelations are used to measure the degree of association between U_t and U_{t+k}, after the mutual linear dependency on the intervening variables U_{t+1}, U_{t+2},…., and U_{t+k-1} has removed.

The conditional correlation is as follows,

$$\text{Corr}\ (Ut,\ U_{t+k}\ |U_{t+1}, U_{t+2}, \ldots.., U_{t+k-1})$$

And is referred to as the partial autocorrelation in time series analysis. The partial autocorrelation coefficient of order *k* is denoted by α_k and can be calculated by regressing Y_t against Y_{t-1}, Y_{t-2}, ….,Y_{t-k}. The first partial autocorrelation is always equal to the first autocorrelation.with the ACF, the

partial autocorrelations should all be close to zero for noise series. The critical values of ± 1.96 / $\sqrt{n}$ can be used with a PACF to assess if the data are white noise. The ACF and PACF provide some guidance on how to select pure AR or pure MA models.

If the data becomes stationary, the next step is to determine the orders of the autoregressive (AR) and moving average (MA) terms using the Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF). The table below shows how p and q orders of ARMA models are identified.

Table 6.1: Identification of p and q orders in ARIMA Model

PROCESS	ACF	PACF
AR(P)	Tails off	Cut off after the order 'q' of the process
MA(q)	Cut off after the order 'q' of the process	Tails off
ARMA(p,q)	Tails off	Tails off

Outlier detection: Calculate error by using the formula difference of original value and forecasted value and denote it as ε_i .If difference is very high then it may be outlier and remove that outlier and proceed for forecasting equation.

2.4. Kolmogrov- Smirnov test (K-S Test):

Kolmogrov-Smirnov test is a non-parametric test (distribution-free test) used to test the goodness of fit and sometimes it used for testing the normality of the distribution for new exponential smoothing model and simple exponential smoothing model. Kolmogrov-Smirnov two sample test is generally used for comparing two samples.

The mathematical procedure for Kolmogrov-Smirnov test:

I. The data consist of a random sample $Y_1, Y_2, \ldots Y_n$ of size n associated with some unknown distribution function, denoted by F(y).

II. The sample is a random sample.

III. Let S(y) be the empirical distribution function based on the random sample $Y_1, Y_2, \ldots, Y_n$. Let $F^*(y)$ be a completely specified hypothesized distribution function.

IV. Let the test statistic T be greatest (denoted by "sup" for supremum) vertical distance between S(y) and $F^*(y)$. In symbols, we say

$$T = \sup F^*(y) - S(y)$$

V. For testing $H_0 : F(y) = F^*(y)$ for all y from $-\infty$ *to* ∞

i. $H_1 : F(y) \neq F^*(y)$ for at least one value of y

VI. f T exceeds the 1-α quantile then we reject H_0 at the level of significance α.

i. The approximate p-value can be found by interpolation.

Kolmogrov - Smirnov test procedure for detection of outliers in time series models:

I. Fit Modified HWT model
II. For different constants fit ACF and PACF.
III. Choose best model for data.
IV. Compute error using

$$\varepsilon_t = Y_t - u_t$$

Y_t = original value

u_t = forecast value

vi. An outlier has high error, remove that error.

vii. Fit Kolmogrov-Smirnov test for testing goodness of fit for errors by removing outliers.

vii. Fit it is good fit for errors, then fit model without outliers.

III. SUMMARY AND CONCLUSIONS

In time series data, there are different types of outliers. Some of them are Additive Outlier (AO), Innovation Outlier (IO) and Level Shift (LS). The structure changes allowed for are level shift (LS) and variance change (VC). Level shift is further classified as permanent level change (LC) and transient level change (TC). To detect the outliers in time series models, we have used detection of outliers for intraday data, auto correlation function, partial auto correlation function, auto regressive models and moving average models. We also used Kolmogrov-Smirnov test for

testing goodness of fit for detection of outliers in time series models.

BIBLIOGRAPHY

[1] Fong-Lin Chu (2006). A fractionally integrated auto regressive moving average approach to forecasting tourism demand.Tourism Management, 29(1), 79-88.

[2] Fox, A. J. (1972). Outliers in Time Series, Journal of the Royal Statistical Society, Ser. B. 43, 350-363.

[3] Frank E.Grubbs (1969). Procedures for Detecting Outlying Observations in Samples. Technometrics, 11(1), 1-21.

[4] James W. Taylor (2004). Smooth transition exponential smoothing. Journal of Forecasting, 23, 385-394.

[5] James W. Taylor (2004). Volatility forecasting with smooth transition exponential smoothing. International journal of Forecasting, 20, 273-286.

[6] James W. Taylor (2007). Forecasting daily supermarket sales using exponentially weighted quantile regression. European Journal of Operational Research, 178, 154-167.

[7] James W. Taylor (2012). Density forecasting of intraday call center arrivals using models based on exponential smoothing. Management science, 58, 534-549.

[8] James W. Taylor and Ralph, D. Snyder (2012). Forecasting intraday time series with multiple seasonal cycles using parsimonious seasonal exponential smoothing. Omega, 40, 748-75

[9] Makridakis, S., Wheelwright, S., & Hyndman, R.J. (1998). Forecasting: Methods and applications. 3rd Ed., New York, John Wiley and Sons.

[10] McGee, Victor E and Carleton, Willard T (1970). Piecewise Regression. Journal of American Statistical Association, 65, 1109-24.

[11] Synder, R.D., Ord, J.K., & Koehler, A.B (2001). Prediction intervals for ARIMA models. Journal of Business and Economic Statistics, 19(2), 217-225