Chap 5: ARIMA Model

Ani Shabri

Department of Mathematical Sciences,
Faculty of Science, Universiti Teknologi Malaysia,
81310 UTM Johor Bahru, Malaysia
ani@utm.my

Jun 8, 2014

Chap 5: ARIMA models

Outline:

- Introduction to Box-Jenkins methodology
- Box-Jenkins methodology procedure
- Stationarity
- Transformations to achieve stationary
- Models for stationary time series
- Identification ARIMA models
- Parameter estimation technique
- Diagnostics Checking
- Forecasting

Introduction to Box-Jenkins methodology

- Box-Jenkins (BJ) methodology or Autoregressive Integrated Moving Average (ARIMA) models are a class of linear models that is capable of representing stationary as well as non-stationary time series.
- The BJ methodology refers to a set of procedures for identifying, fitting, estimating and checking ARIMA models with time series data. Forecast follow directly from the form of fitted model.
- The BJ methodology aims to obtain a model that is parsimony. Parsimony referred a model has the smallest number of parameters needed to adequately fit the patterns in the data observed.

Box-Jenkins methodology procedure

- **Stationary:** Stationary is a fundamental property underlying for ARIMA model. In this step, non-stationary to achieve stationary series usually by taking first and second difference of the data.
- **Identification**: When the data are confirmed stationary, one may proceed to tentative identification of models through visual inspection of both the autocorrelation function (ACF) and partial autocorrelation function (PACF).
- **Estimation**: Determine coefficients and estimate of the ARIMA model using various techniques such as the least squares, moment and maximum likelihood methods.
- **Diagnostics**: Having estimated the coefficients, the model is then tested for its adequacy. Test statistics, ACFs and PACFs of residuals were used to verify whether the model is valid. If valid then use the decided model, otherwise repeat the steps of Identification, Estimation and Diagnostics.
- **Forecast**: Once the model's fitness has been confirmed, the model then ready to be used to generate the forecasts for future value.

Stationarity

- A stationary process has the property that the mean, variance and auto-covariance structure do not change over time.
- A time series y_t is said to be stationary if it satisfies the following conditions:

  i. $E(Y_1) = E(Y_2) = \ldots = E(Y_n) = \mu$

  ii. $V(Y_1) = V(Y_2) = \ldots = V(Y_n) = \sigma^2$

  iii. $Cov(Y_t, Y_{t-k}) = Cov(Y_{t+h}, Y_{t+h-k}) = \gamma_k$

- Visually, it is a flat looking series, without trend, fluctuates around a constant mean and the autocorrelation function (ACF) tails off toward zero quickly.
- Transformation will be used when time series is not stationary in variances.

Transformations to achieve stationary

- Differencing is often used to made series stationary in mean.
- Number of times differencing is needed to achieve stationary is called "order of integration". In most cases, first and second order is sufficient.

  1st differencing : $\Delta y_t = y_t - y_{t-1} = y_t - By_t = (1-B)y_t$

  2nd differencing : $\Delta^2 y_t = \Delta y_t - \Delta y_{t-1} = (1-B)^2 y_t$

- For the series shows increasing in variability over time, normally we use
- logarithm : $x_t = \log y_t$
- or square root : $x_t = \sqrt{y_t}$

Models for stationary time series

Mixed Autoregressive & Moving Average model, ARMA(p, q)

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}$$

Autoregressive model, AR(p) or ARMA(p, 0)

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \ldots + \phi_p y_{t-p} + \varepsilon_t$$

Moving Average model, MA(q) or ARMA(0, q)

$$y_t = \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t$$

Determining a tentative ARIMA models

- Behavior of ACF and PACF were used to determine the appropriate ARIMA model. *ACF* measures the linear relationship between time series observations separated by a lag of *k* time units. The sample ACF is computed by

$$r_k = \frac{\sum_{t=1}^{n-k}(y_t - \bar{y})(y_{t+k} - \bar{y})}{\sum_{t=1}^{n}(y_t - \bar{y})^2}$$

The t_{r_k} statistic is $t_{r_k} = \frac{r_k}{s_{r_k}}$ where $s_{r_k} = \sqrt{\frac{1 + 2\sum_{j=1}^{k-1} r_j^2}{n}}$.

- The ACF is called cut off at 95% confidence interval if value of r_k lie in the range $[-2s_{r_k}, 2s_{r_k}]$

Interpretation of behavior of sample ACF

The sample ACF is said to die down if this function does not cut off but rather decreases in a 'steady fashion'. The sample ACF can die down in

(i) a damped exponential fashion

(ii) a damped sine-wave fashion

(iii) a fashion dominated by either one of or a combination of both (i) and (ii).

The SAC can die down fairly quickly or extremely slowly.

Note: Behavior of ACF and PACF usually drawn with 95% confidence interval.

Interpretation of behavior of sample PACF

- PACF is used to measure the degree of association between Y_t and Y_{t-k}, when the effects of other time lags (1, 2, 3, …, k – 1) are removed. The sample PACF is given by

$$r_{kk} = \frac{r_k - \sum_{j=1}^{k-1} (r_{k-1,j})(r_{k-j})}{1 - \sum_{j=1}^{k-1} (r_{k-1,j})(r_j)}$$

- The $t_{r_{kk}}$ statistic is $t_{r_{kk}} = \frac{r_{kk}}{s_{r_{kk}}}$ where $s_{r_{kk}} = \sqrt{\frac{1}{n}}$.

- The PACF is called cut off at 95% confidence interval if value of r_k lie in the range $[-2/\sqrt{n}, 2\sqrt{n}]$

- Behavior of sample PACF similar to its of the sample ACF.

Identification ARIMA models

Summary Of The Behaviour Of Autocorrelation And Partial Autocorrelation Functions

	ACF	PACF
AR(p)	Exponential decay/tails off towards zero/damped sine wave	Cut off after the order p
MA(q)	Cut off after the order q	Exponential decay/tails off towards zero/damped sine wave
ARMA(p,q)	Exponential decay/tails off towards zero/damped sine wave	Exponential decay/tails off towards zero/damped sine wave

* Note: Sometimes order of p and q cannot be determined from ACF and PACF. May use trail and error starting with simplest models AR(1), MA(1) and ARMA(1,1).

Parameter estimation technique

- Once a "tentative" model has been identified, the parameters for the models need be estimated.
- Many computer softwares have programs/algorithms will automatically find appropriate initial estimates of the parameters ARIMA model and then successively refine them until the optimum values of the parameters are found. Usually they use

  - maximum likelihood - for ARIMA process

  - non-linear least squares - for AR process

  - method of moments - for AR process

Estimating the parameters ARIMA model

- Once a tentative model has been identified, the estimates for constant and the coefficients of the parameter ARIMA models must be obtained.
- The model should be parsimonious (simplest form)
- All parameters and constant estimated should be significantly different from zero. Significance of parameters is tested using standard t-test

$$t_{stat} = \frac{\text{point estimate of parameter}}{\text{standard error of estimate}}$$

- The parameters model are significances if $|t_{stat}| > 2 \text{ for } \alpha = 0.05.$

Diagnostics Checking

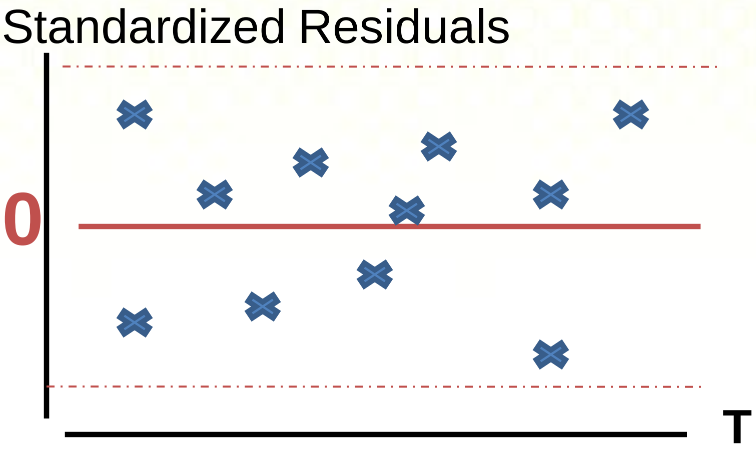
In the model-building process, if an ARIMA(p, d, q) model is chosen (based on the ACFs and PACFs), some checks on the model adequacy are required. A residual analysis is usually based on the fact that the residuals of an adequate model should be approximately white noise. Basically, a model is adequate if the residuals nearly the properties white noise process, i.e. the errors

- constant on variances
- Independent
- normally distributed with zero means and variance σ^2

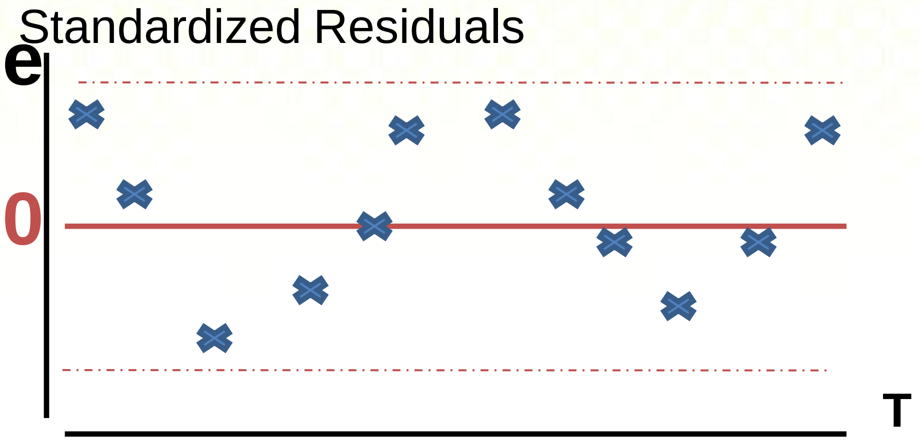
Constant on variances

- Variance of residuals are constant can be checked by plot the residuals or standardized residuals. Absence of any trends or pattern may also for suggestion of dependence residuals.
- The variance of errors is constant if standardized residuals are within ± 2 or almost all of them should be within ± 3 and should exhibit the random pattern

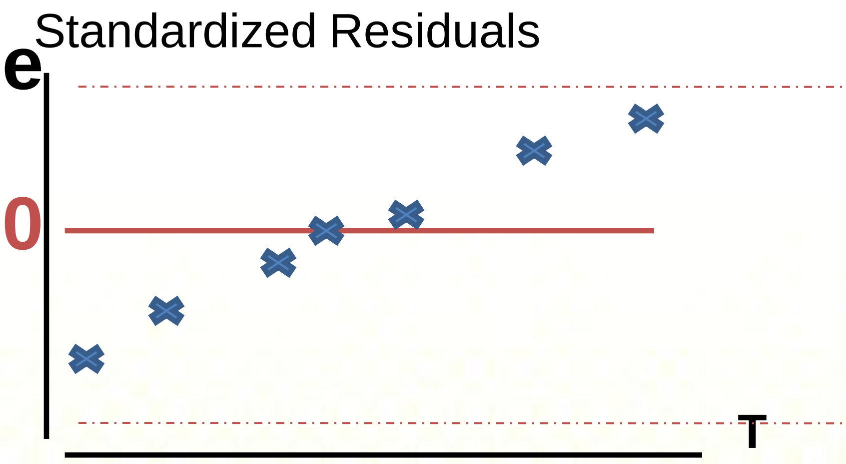
Constant on variances



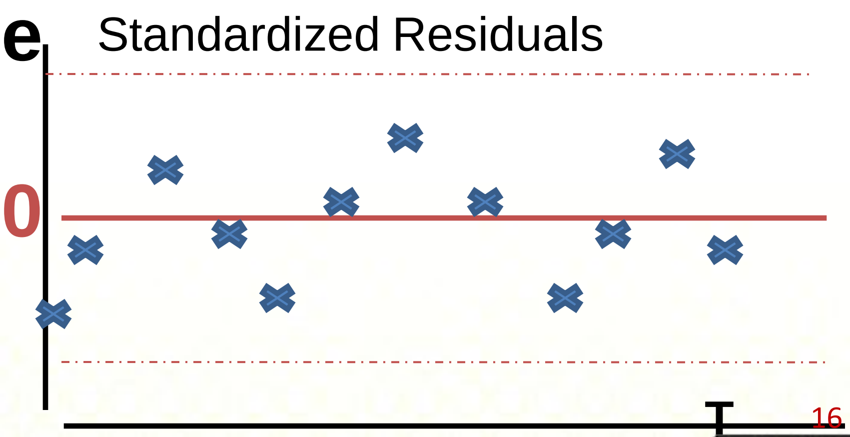
Random errors



Cyclical effects not accounted for



Trend not full accounted for



Seasonal effects not accounted for

Independent test

If an ARMA(p,q) model is an adequate representation of the data generating process, then the residuals should be independent. 2 Tests were considered

i. ACF of residuals mostly falls inside Barlett confidence interval.

ii. Portmanteau test statistic uses sample ACF of the residuals as a group to examine the following hypothesis:

$$H_0 : \rho_1 = \rho_2 = ... = \rho_k = 0$$

Portmanteau test statistic:

$$Q^*(k) = (n-d)(n-d+2)\sum_{l=1}^{k} \frac{r_l^2(e)}{n-d-k} \sim \chi^2_{(k-p-q)}$$

Hypothesis nol H_0 is rejected when $Q^*(k) > \chi^2_{(k-p-q)}$

If H_0 rejected, say up to 3, 6 and 12 lags, suggest to look for another better model.

Normality test

In statistics, the Jarque–Bera (JB) test is one procedure for determining whether sample data (residuals) are normal distribution. The test is named after Carlos Jarque and Anil K. Bera. The test statistic *JB* is defined as

$$JB = \frac{n}{6}\left[S^2 + \tfrac{1}{4}(K-3)^2\right]$$

where *n* is the number of observations (or degrees of freedom in general); *S* is the sample skewness, and *K* is the sample kurtosis:

$$S = \frac{\hat{\mu}_3}{\hat{\sigma}^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{3/2}} \qquad K = \frac{\hat{\mu}_4}{\hat{\sigma}^4} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2}$$

The data does not follows normal distribution if $\text{JB} > \chi^2_{\alpha,2}$

Model selection criteria

In many practical situation, many possible ARIMA models adequate to fit the data. AIC and SBC criteria can be used to choose the best model among all possible models.

- Akaike Information Criterion (AIC)

$$AIC = \ln \hat{\sigma}^2 + \frac{2}{n} r$$

- Bayesian Information Criterion (BIC)

$$BIC = \ln \hat{\sigma}^2 + \frac{\ln n}{n} r \qquad \hat{\sigma}^2 = \frac{SSE}{n}$$

r = number of parameters to be estimated,
n = number of observations.
SSE= sum of square error

- Ideally, the AIC and SBC should be as small as possible

Forecasting

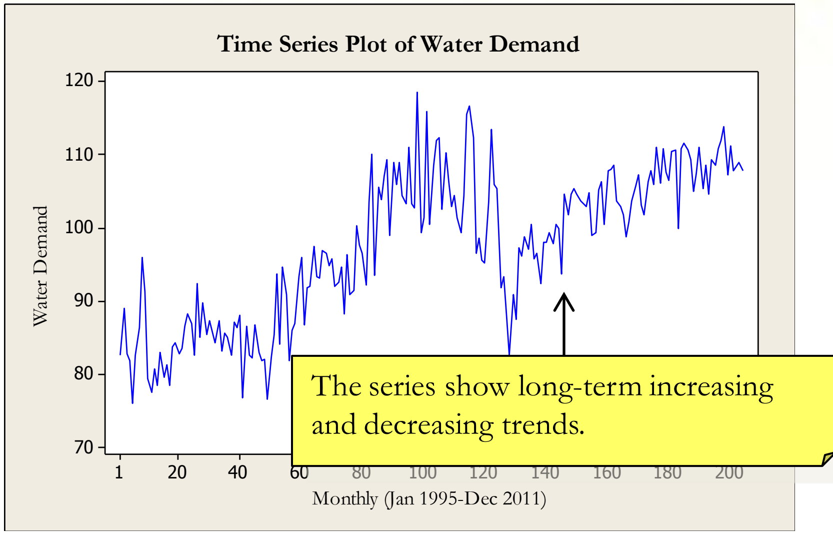
- Once the fitted model has been selected, it can be used to generate forecasts for future time periods.
- The forecast values of h-period ahead for ARMA(p,q) model is given by

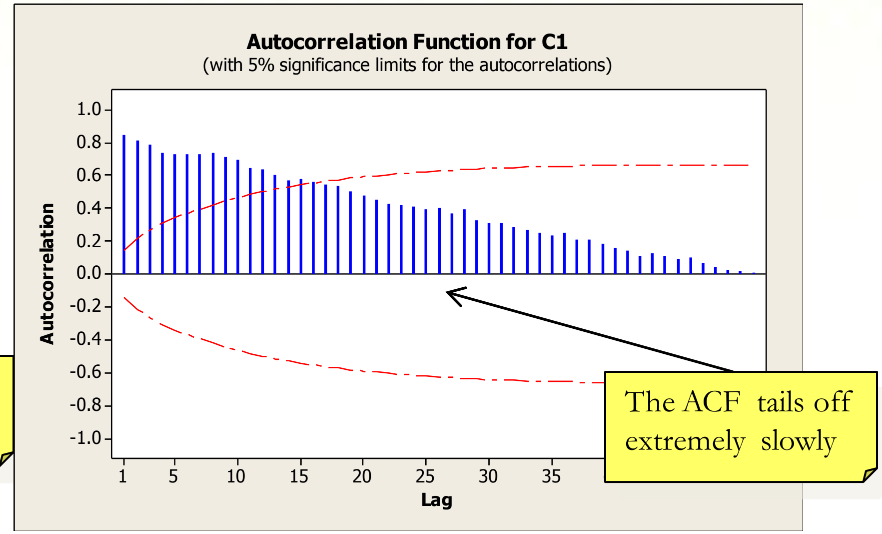
$$\hat{X}_{t+h} = \hat{\phi}_1 X_{t+h-1} + \ldots.\hat{\phi}_p X_{t+h-p} + e_{t+h} - \hat{\theta}_1 e_{t+h-1} - \ldots.\hat{\theta}_q e_{t+h-q}$$

where the forecast values of the ARIMA model may be found by replaced by their estimates when the actual values are not available.

Example

Monthly data of water demand in Kluang Johor in Malaysia from January 1995 to December 2011.

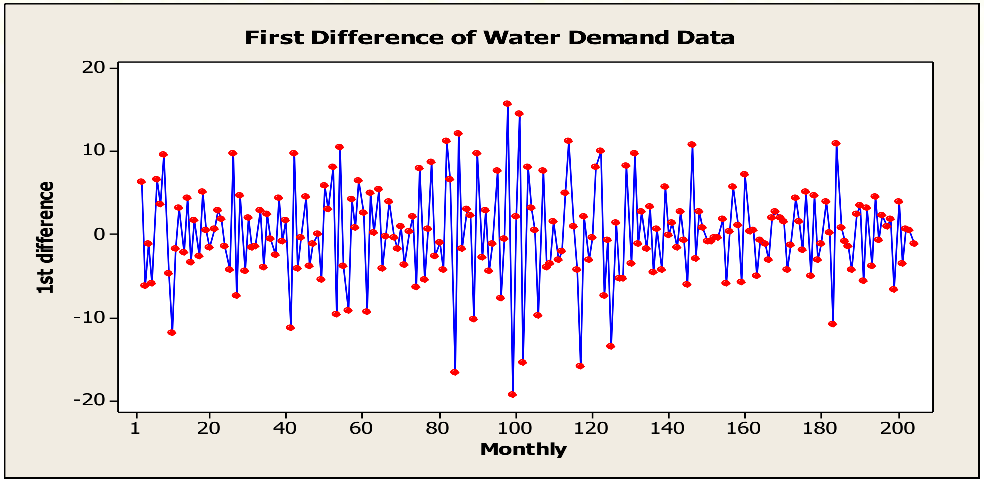


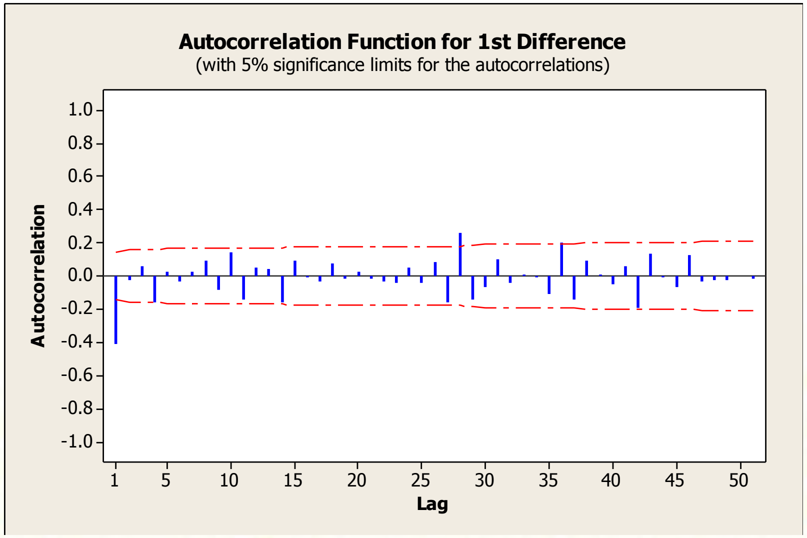


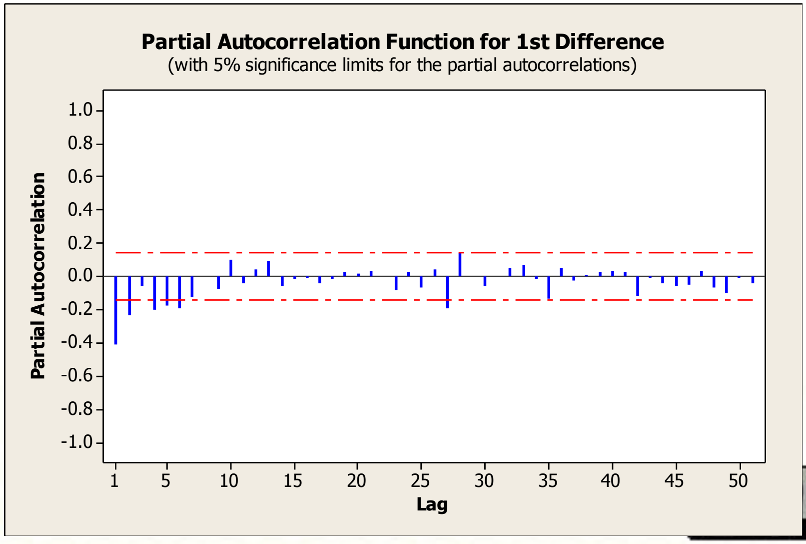
- The time series plot shows that it is non-stationary in the mean.
- The ACF also shows a pattern typical for a non-stationary series:
  - i. Large significant ACF for the first 16 time lag
  - ii. Slow decrease in the size of the autocorrelations.

We take the first differences of the data and reanalyze.

First difference of water demand data







Example

- The plot and ACF (cuts off quickly) of the 1st difference of water demand suggests the series is stationary.
- Based on ACF and PACF, 3 tentative models are identified

i. ARIMA(0,1,1)-ACF cuts off after lag 1 and PACF shows a exponential decay

ii. ARIMA (2,1,1)-ACF follows a damped cycle and PACF cuts off after lag 2.

iii. ARIMA(1,1,1)-ACF and PACF decay exponentially.

ARIMA with MINITAB

ARIMA(2,1,0)

Final Estimates of Parameters

Type		Coef	SE Coef	T	P
AR	1	-0.5031	0.0686	-7.33	0.000
AR	2	-0.2305	0.0686	-3.36	0.001

Differencing: 1 regular difference
Number of observations: Original series 204, after differencing 203
Residuals: SS = 5143.83 (backforecasts excluded)
MS = 25.59 DF = 201

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag	12	24	36	48
Chi-Square	28.8	40.2	64.5	83.0
DF	10	22	34	46
P-Value	0.001	0.010	0.001	0.001

ARIMA(0,1,1)

Final Estimates of Parameters

Type		Coef	SE Coef	T	P
MA	1	0.6793	0.0516	13.16	0.000

Differencing: 1 regular difference
Number of observations: Original series 204, after differencing 203
Residuals: SS = 4874.03 (backforecasts excluded)
MS = 24.13 DF = 202

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

Lag	12	24	36	48
Chi-Square	25.3	37.9	58.4	77.2
DF	11	23	35	47
P-Value	0.008	0.026	0.008	0.004

ARIMA(1,1,1)

Final Estimates of Parameters

Type		Coef	SE Coef	T	P
AR	1	0.1849	0.0997	1.85	0.065
MA	1	0.7765	0.0634	12.24	0.000

The t statistics are significant at α = 10%

Differencing: 1 regular difference
Number of observations: Original series 204, after differencing 203
Residuals: SS = 4789.01 (backforecasts excluded)
MS = 23.83 DF = 201

Modified Box-Pierce (Ljung-Box) Chi-Square statistic

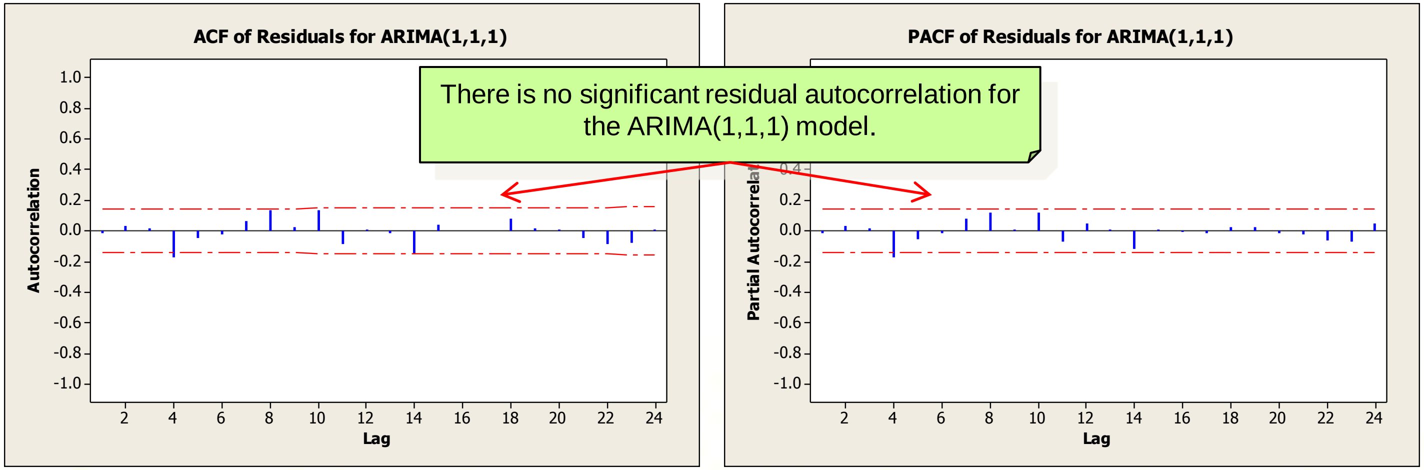
Lag	12	24	36	48
Chi-Square	17.1	27.6	51.3	70.8
DF	10	22	34	46
P-Value	0.072	0.190	0.029	0.011

The LBQ statistics are significant as indicated by the small p-values for either model.

The LBQ statistics are not significant at α = 10%

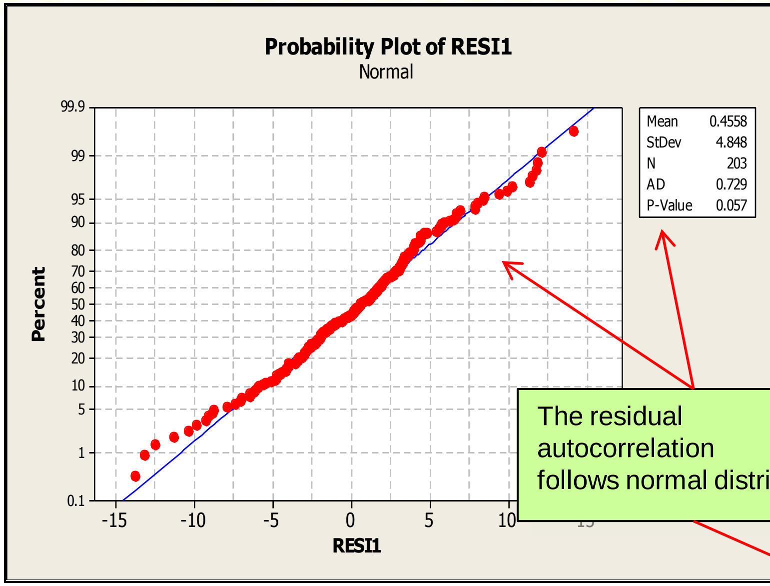
Example

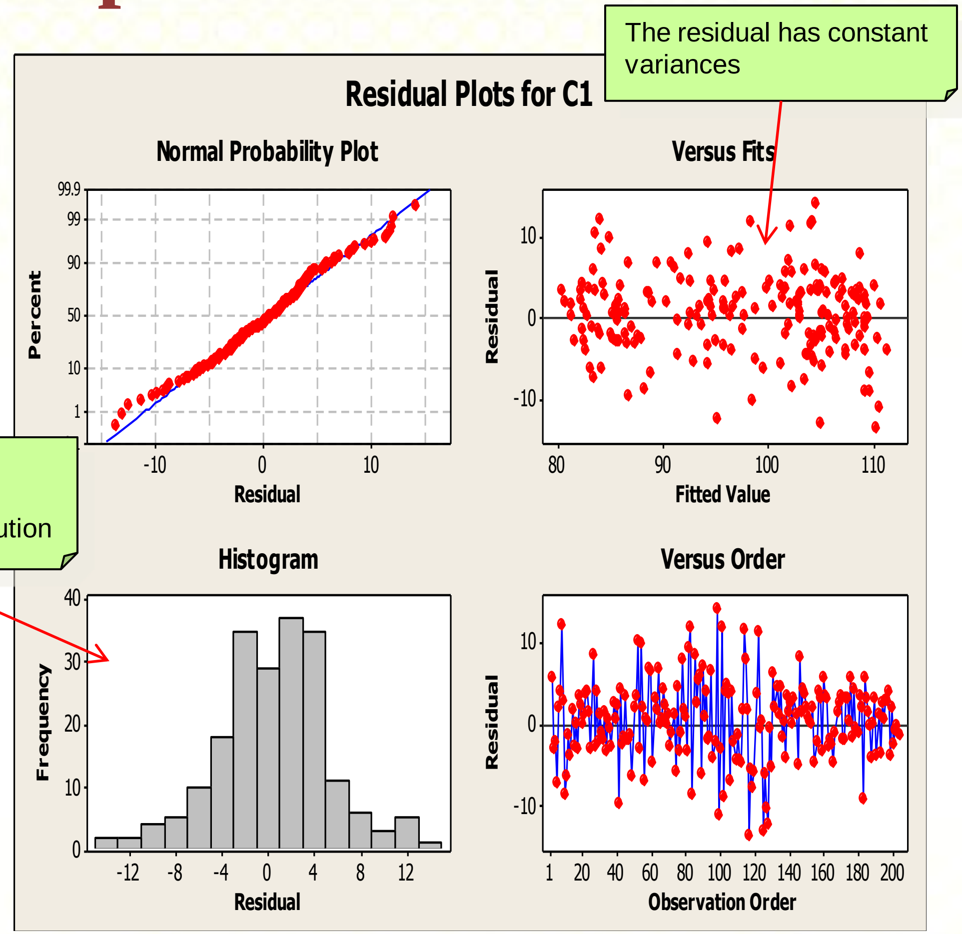
- The results indicate that ARIMA(1,1,1) residuals are uncorrelated at least up to lag 48, while ARIMA(2,1,0) and ARIMA(0,1,1) residuals are correlated.
- The ACF and PACF of residuals of ARIMA(1,1,1) are well within their two standard error limits indicating residuals are white noise.



ACF and PACF of residual of ARIMA(1,1,1)

Example





The Figure shows that the residuals follow normal distribution and have constant variances

The results shows that ARIMA(1,1,1) model is an adequate model.

Model selection criteria

	t-test	Q-test	AIC	BIC
ARIMA(2,1,1)	√	x	3.247	3.280
ARIMA(0,1,1)	√	x	3.183	3.200
ARMA(1,1,1)	√	√	3.176	3.208

Judging these results, it appears that the estimated ARIMA(1,1,1) model best fits the data.

Comparison of actual and forecasted value using ARIMA(1,1,1) model