Chap 3: Model Diagnostic Checking

Ani Shabri

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

Jun 8, 2014

Chap 3: Model Diagnostic Checking

Outline:

- Introduction to diagnostics checking
- Model adequacy checking
- Testing for zero mean
- Testing for constant variances
- Testing independent of error
- Normality tests
- Evaluating the accuracy of the model

Introduction to diagnostics checking

- It is important to check the adequacy of the model before it used in forecasting and becomes part of the decision making process.
- It is important to study outlying observations to decide whether they should be retained or eliminated.
- If retained, whether their influence should be reduced in the fitting process or revise the regression function.
- Residual plots can be used to check the model assumptions.

Model adequacy checking

The Time Series model is adequate if the error of the model meet 4 assumptions (error = $e_i = \hat{\varepsilon}_i = y_i - \hat{y}_i$)

1. The error component will have zero mean ; $E[\varepsilon_i] = 0$
2. The variance of ε_i , denoted by σ^2, is <u>the same</u> for all values of the independent variable(s), i.e.,

   $$V[\varepsilon_i] = \sigma_\varepsilon^2 \quad i = 1, 2, \ldots, n$$
3. The errors are independent.
4. The errors has a normal distribution with mean zero and variance constants.

Testing for zero mean

Hypothesis $H_0 : \mu = 0$ VS $H_1 : \mu \neq 0$

Statistics Test

For n < 30

$$T = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Reject H_0 *if* $|T| > t_{\alpha/2,v}$

For n > 29

$$Z = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Reject H_0 *if* $|Z| > z_{\alpha/2}$

Testing for constant variances

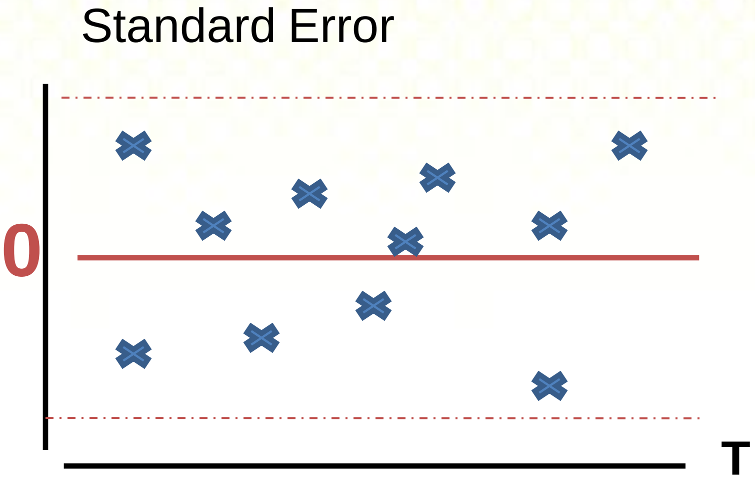
Residuals are often standardized so that they have mean zero and variance one. Since residuals are estimates of the errors, and since the observed errors have variance σ_ε^2, the **standardized residuals** are given by

$$d_i = \frac{e_i}{s} \quad \text{where} \quad s = \frac{\sum (e_i - \bar{e})^2}{n - p - 1}$$

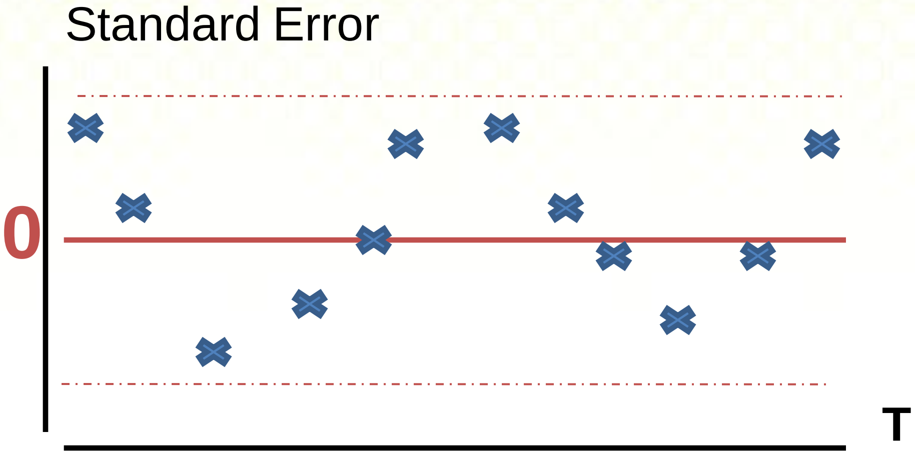
p = the number of parameters model.

The variance of errors is constant if d_i are within ± 2 or almost all of them should be within ± 3 and should exhibit the random pattern

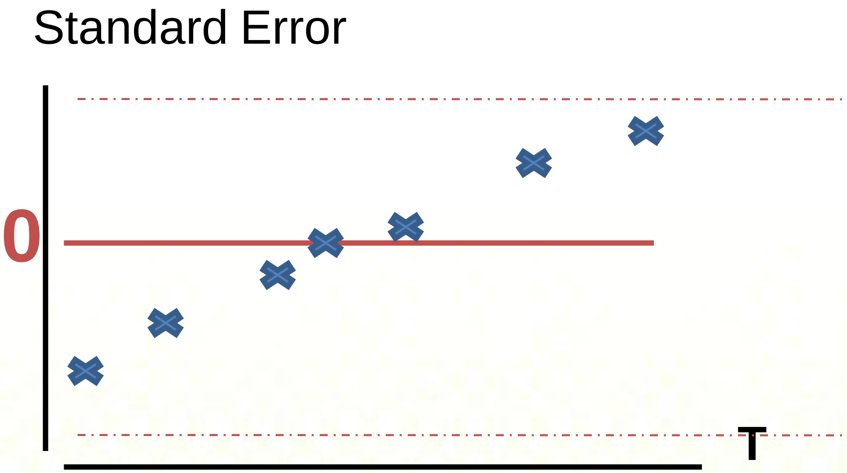
Pattern of forecast error



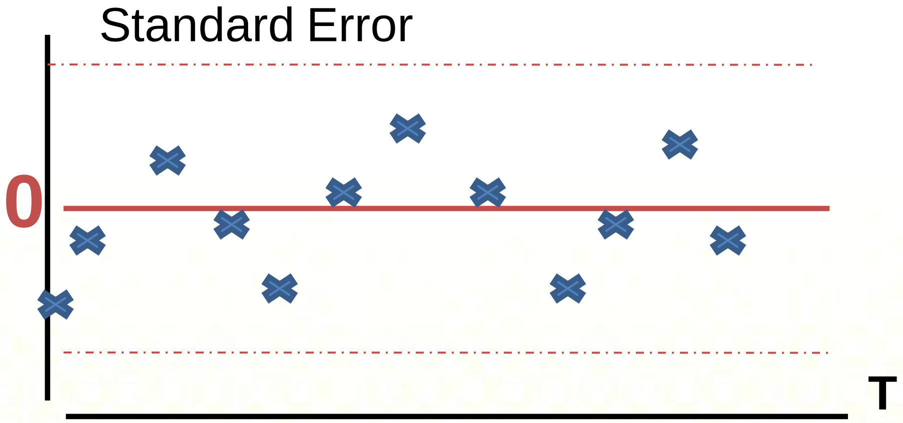
Random errors



Cyclical effects not accounted for



Trend not full accounted for



Seasonal effects not accounted for

Testing independent of error

Autocorrelation function (ACF) of errors is used to check independent of error. ACF of errors is given by

$$r_k = \frac{\sum_{t=k+1}^{n} (e_t - \bar{e})(e_{t+k} - \bar{e})}{\sum_{t=1}^{n} (e_t - \bar{e})^2}, \quad k = 1, 2, \ldots$$

The errors are independents if all (of most) of the ACF are within $\pm 2/\sqrt{n}$

Normality tests

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

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

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

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

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

Anderson–Darling test

In statistics, the Anderson–Darling (AD) test is usually used to test whether the data (errors) follows normal distribution. The AD test is given by

$$A^2 = -n - S \qquad S = \sum_{i=1}^{n} \frac{2k-1}{n} \{\ln[F(Y_i)] + \ln[1 - F(Y_{n+1-i})]\}$$

For normal distribution, the formula is

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^{n} (2i-1)\{\ln[\Phi(Y_i)] + \ln[1 - \Phi(Y_{n+1-i})]\}$$

$$Y_i = \frac{X_i - \hat{\mu}}{\hat{\sigma}} = \frac{X_i - \bar{x}}{s}$$

Φ is standard normal CDF

A modified statistic is calculated using

$$A^{*2} = A^2 \left(1 + \frac{0.75}{n} + \frac{2.25}{n^2}\right)$$

and normality is rejected if A^{*2} exceeds 0.631, 0.752, 0.873, 1.035, or 1.159 at 10%, 5%, 2.5%, 1%, and 0.5% significance levels, respectively; the procedure is valid for sample size at least n=8.

Evaluating the accuracy of the model

If the null hypothesis of the validating model assumptions are rejected, then we might look several common causes:

i. The wrong model was chosen (e.g. linear regression model used with non-linear data)

ii. There are other components in the time series (e.g. cycle or seasonal components that have not been modeled).

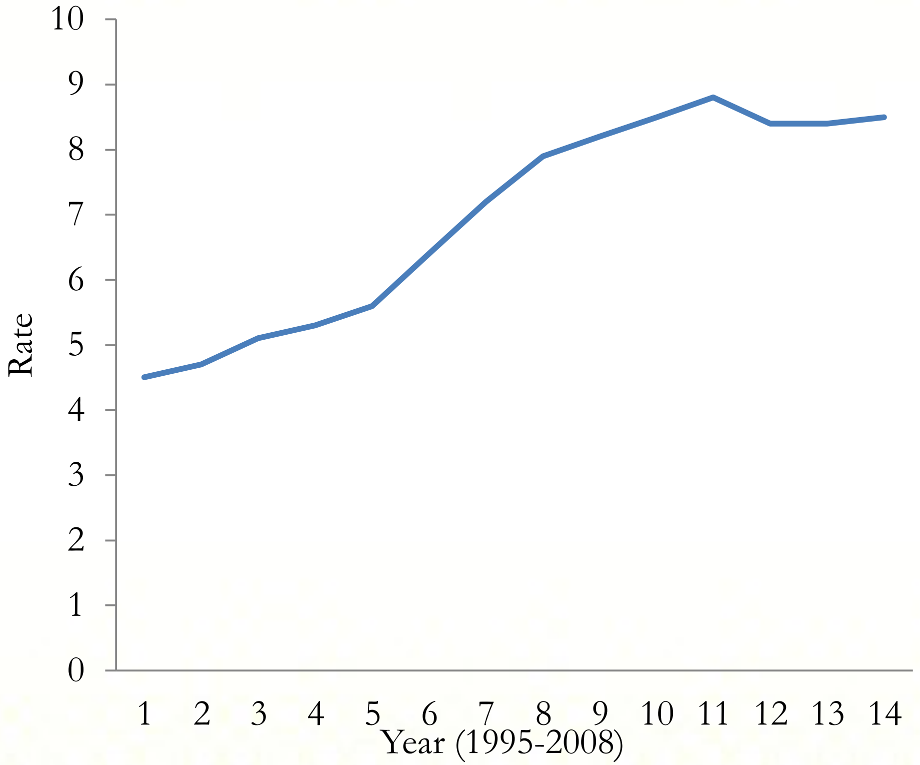
iii. The equation has not completely modeled the trend (i.e., there is trend left in the error)

iv. If error is large, either model being used is the wrong one, or parameters need adjusting.

Example

The following table gives the average cost (in cents per kilo watt hour) of electricity from 1995 to 2008.

Year	Rate
1995	4.5
1996	4.7
1997	5.1
1998	5.3
1999	5.6
2000	6.4
2001	7.2
2002	7.9
2003	8.2
2004	8.5
2005	8.8
2006	8.4
2007	8.4
2008	8.5



Example

- The scatter plot suggests that a linear regression model is appropriate.
- Least squares method was used to fit a regression line to the data

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	4.2033	0.3032	13.8642	9.53E-09
X Variable 1	0.3681	0.0356	10.3389	2.5E-07

- Linear regression model $Y = 4.2033 + 0.3681\ t$

Example

Testing the Mean of Error is Zero

Hypothesis $H_0 : \mu = 0$ VS $H_1 : \mu \neq 0$

Statistics Test

For $n < 30$

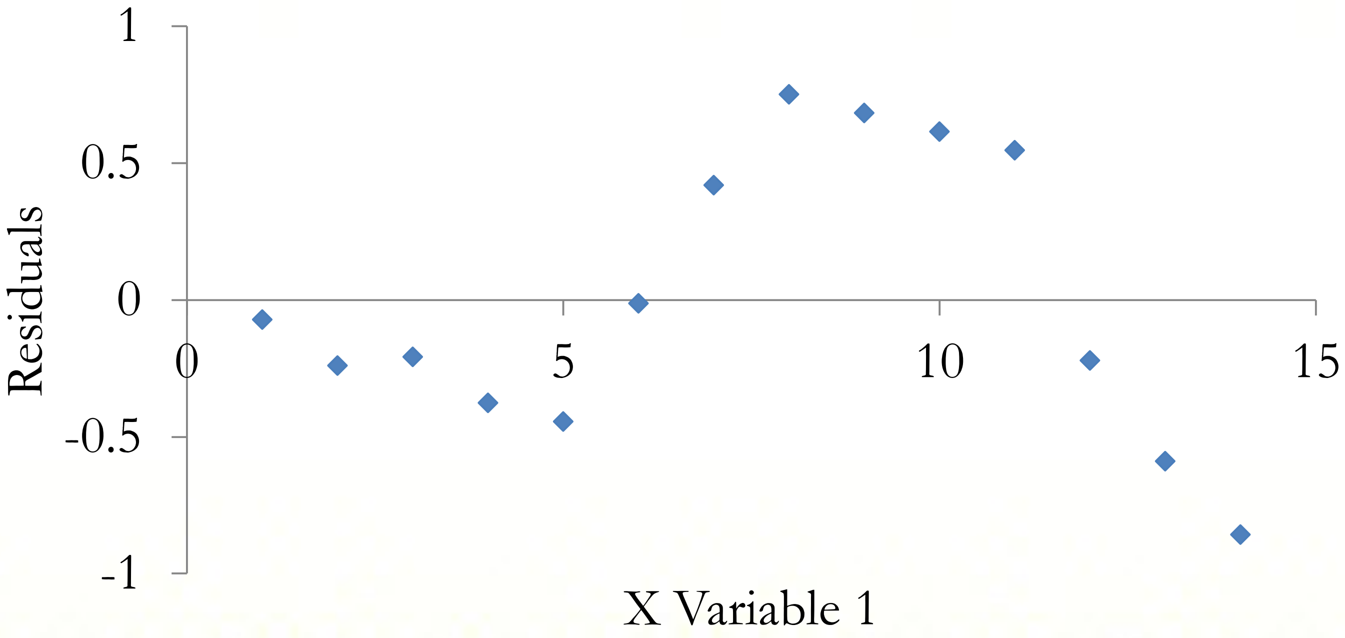
$$T = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{1.90E-16}{0.5160/\sqrt{14}} = 1.38E-15 \approx 0$$

Since $|T| < t_{\alpha/2,v}$ Accept H_0

Conclusion: The mean of Errors is Zero

Example

- The standard residuals were plotted against the fitted values.
- The plot shows that the residuals are not consistent and exists a seasonal effects.



Example

- To confirm this graphic diagnosis we will use the Autocorrelation function (ACF) of errors test for:

H_0 : Errors are independents

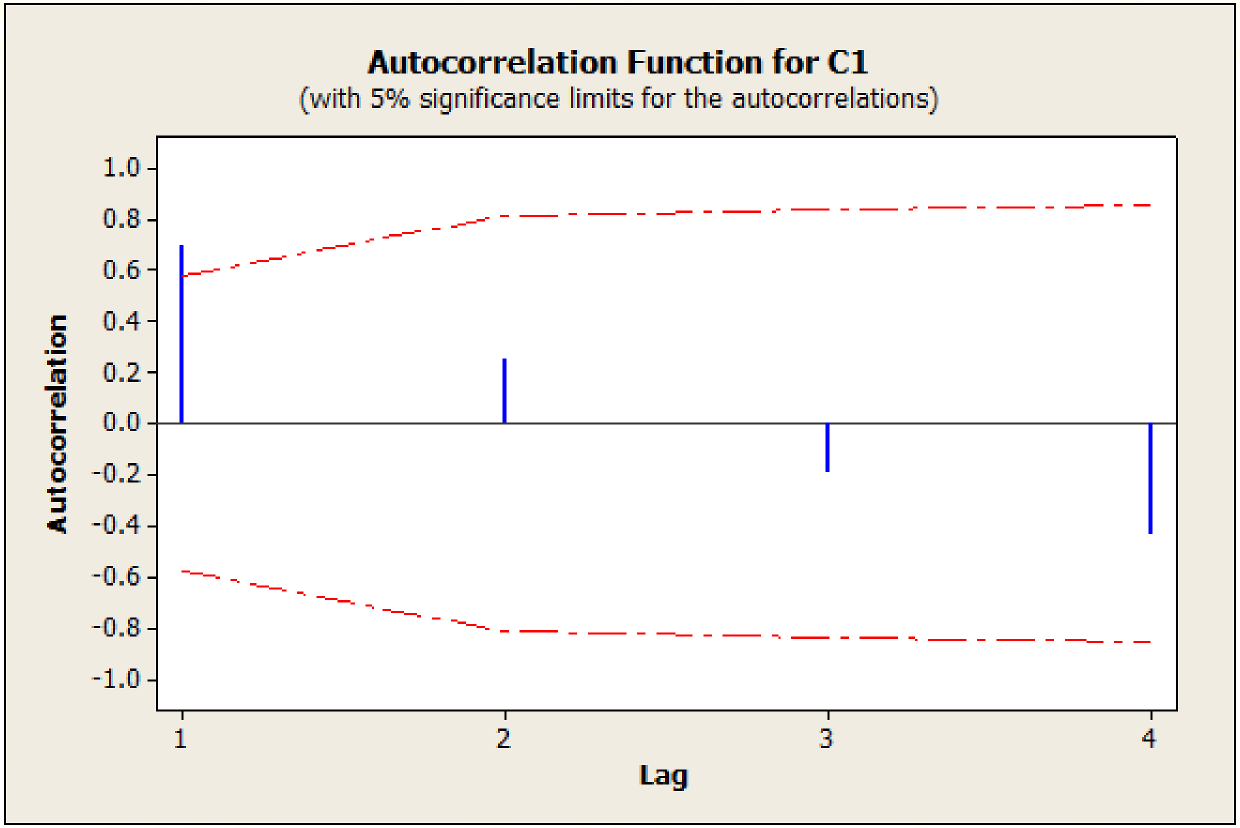
H_1 : Errors are dependents

- The test statistic is:

$$r_k = \frac{\sum_{t=k+1}^{n}(e_t - \bar{e})(e_{t+k} - \bar{e})}{\sum_{t=1}^{n}(e_t - \bar{e})^2}, \quad k = 1, 2, \ldots$$

The errors are independents if all (of most) of the ACF are within $\pm 2/\sqrt{n}$

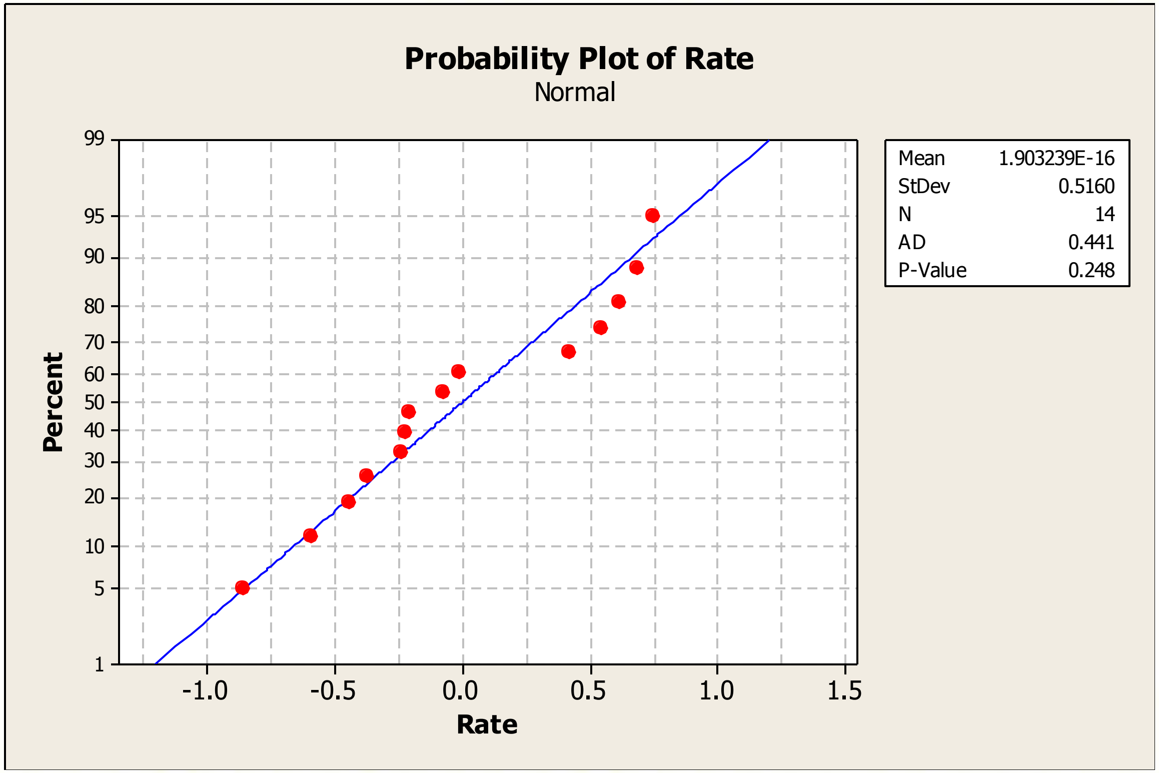
Example



The Figure shows that the errors are dependent due to one of the ACF (ACF at t=1) is outside $\pm 2/\sqrt{n}$

Example

The Figure and p-value of AD test shows that the errors are follow normal distribution.



Example

Conclusion

- The study shows that only 2 the criterion for assumption were satisfied (mean of errors is zero and the errors follows normal distribution).
- We can make conclusion that the linear regression might be inappropriate and it would be more appropriate to use non-linear regression model.